Conception et évaluation d'un outil d'extraction, d'analyse et de classification de marqueurs linguistiques de la maladie d'Alzheimer

par

Camille Michèle GAGNÉ

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE
M. Sc. A.

MONTRÉAL, LE 3 SEPTEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

Mme. Sylvie Ratté, directrice de mémoire
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Pierre André Ménard, codirecteur
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Tony Wong, président du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Luc Duong, membre du jury
Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 26 AOÛT 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

To Nick Dudek, thank you for encouraging me to accomplish this project and for supporting me in every possible way. It would not have been possible without you.

À ma professeure et directrice de maîtrise, Sylvie Ratté, un énorme merci pour ta confiance, ton soutien et ton appui tout au long de cette recherche. Merci d'avoir partagé tes idées avec autant de générosité et de m'avoir offert la chance de contribuer à une grande variété de projets. Merci aussi de m'avoir ouvert les portes de ton propre projet et de m'avoir donné l'opportunité de travailler de nouveau en linguistique, un domaine qui m'avait manqué.

À mon codirecteur, Pierre André Ménard, merci pour ton encadrement rigoureux, tes relectures attentives et tes viennoiseries toujours appréciées. Merci d'avoir partagé ton expérience et d'avoir contribué à maintenir la rigueur et la régularité de ce projet. Je te suis également reconnaissante pour les sessions de *brainstorming* stimulantes qui m'ont aidée à approfondir ma réflexion et à faire de ce projet ce qu'il est aujourd'hui.

Merci à Félix Journet, du collège Sainte-Anne, pour ses annotations manuelles précises, sans lesquelles il m'aurait été impossible d'évaluer correctement notre outil.

Merci à Aurélie Gagné et Jérôme Ouellet-Ayotte pour leur lecture attentive et leurs diligentes corrections.

L'utilisation du *Pitt Corpus* a été réalisée avec le soutien financier des subventions NIA AG03705 et AG05133.

Conception et évaluation d'un outil d'extraction, d'analyse et de classification de marqueurs linguistiques de la maladie d'Alzheimer

Camille Michèle GAGNÉ

RÉSUMÉ

L'Alzheimer est une maladie neurodégénérative qui entraîne une détérioration progressive des fonctions cognitives et peut être observée par un déclin des fonctions langagières. Plusieurs chercheurs se sont tournés vers l'utilisation de processus informatiques automatisés permettant d'analyser des biomarqueurs provenant du langage et de détecter des signes de la maladie de manière non invasive. Au cours des cinq dernières années, le laboratoire d'ingénierie cognitive et sémantique (LiNCS) de l'école de technologie supérieure a contribué à cette recherche en mettant sur pied une application dédiée à l'extraction de caractéristiques linguistiques et à la classification de participants en fonction de la présence ou de l'absence de troubles cognitifs. Il demeure toutefois essentiel de pouvoir extraire et analyser des mesures pertinentes, les manipuler, les visualiser et s'appuyer sur un classificateur réutilisable et fiable.

Ce mémoire présente une solution qui s'appuie sur les travaux antérieurs du LiNCS et propose une application reconçue et améliorée. Notre approche met en place un système fiable, reposant sur des méthodes optimisées d'extraction de caractéristiques et mettant en place des outils pertinents de manipulation et de visualisation des données. Il explique l'observation de nouvelles caractéristiques significatives issues de l'analyse de trois catégories peu explorées dans la littérature : les chaînes de coréférence, les pauses et la complexité syntaxique. L'application comporte également un pipeline de classification optimisé utilisant un algorithme XGBoost comme base du modèle prédictif.

Les résultats obtenus à l'aide de notre système ont permis d'observer que les pauses courtes en début de phrase, les mesures de ratio de syntagmes ADVP, ADJP, VP, S et SBARQ, ainsi que les mesures de densité de syntagmes de types FRAG, VP, NP et ADJP, sont corrélées avec les étiquettes de diagnostic ($F > 4$, valeur $p < 0,05$) et contribuent à une meilleure classification. Nous avons également éliminé 17 caractéristiques générant du bruit et ayant été identifiées au moyen de notre outil. Le modèle de classification final a atteint un score F1 de 80,9% et un rappel de 80,5%, soit une amélioration respective de 5% et de 8,2% par rapport au modèle de base.

La conception de notre outil met en place un système performant permettant d'observer le déclin du langage et d'identifier des caractéristiques associées au diagnostic de la maladie d'Alzheimer. Il constitue une base solide pour la recherche et le développement futur et propose une application pouvant être utilisée par des professionnels de la santé pour suivre l'évolution des patients.

Mots-clés: Maladie d'Alzheimer, traitement automatique de la langue naturelle, algorithme de classification, chaînes de coréférence, pauses, complexité syntaxique

Design and Evaluation of a Tool to Extract, Analyze and Classify Linguistic Markers of Alzheimer's Disease

Camille Michèle GAGNÉ

ABSTRACT

Alzheimer's is a neurodegenerative disease that leads to a progressive deterioration in cognitive functions and can be observed by a decline in language functions. Many researchers have turned to the use of automated computer processes to analyze language biomarkers and detect signs of the disease in a non-invasive way.

Over the past five years, LiNCS has contributed to this research by developing an application dedicated to extracting linguistic features and classifying participants. However, it remains essential to be able to extract and analyze relevant measures, manipulate and visualize them, and rely on a reusable and reliable classifier.

This thesis presents a solution that builds on previous LiNCS work and proposes a redesigned and improved application. Our approach implements a reliable system, based on optimized feature extraction methods, with relevant data manipulation and visualization tools. It integrates to the classifier significant features derived from the analysis of three categories less explored in the literature : coreference chains, pauses, and syntactic complexity. The application also features an optimized classification pipeline using an XGBoost algorithm as the basis for the predictive model.

The results obtained using our system showed that short pauses at the beginning of sentences, ADVP, ADJP, VP, S and SBARQ phrase ratio measures, as well as FRAG, VP, NP and ADJP phrase density measures, correlated with diagnostic labels ($F > 4$, p-value < 0.05) and contributed to better classification. Our tool also enabled us to eliminate 17 noise-generating features. The final classification model achieved an F1 score of 80.9% and a recall of 80,5%, an improvement of 5% and 8,2% respectively.

The design of our tool provides a powerful system for observing language decline and identifying features associated with the diagnosis of Alzheimer's disease. It provides a solid basis for future research and development, and offers a tool that can be used by healthcare professionals to monitor patients' progress.

Keywords: Alzheimer's disease, natural language processing, classification algorithm, coreference chains, pauses, syntactic complexity

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADJP	Syntagme adjectival
ADP	Adposition
ADVP	Syntagme adverbial
ANN	*Artificial Neural Network*
ANOVA	*Analysis of Variance*
ASC	*Aire sous la courbe*
CCONJ	Conjonction de coordination
DAG	*Directed Acyclic Graph*
DRS	*Dementia Rating Scale*
DT	*Decision Tree*
FRAG	Fragment
KNN	*K-Nearest Neighbors*
LiNCS	Laboratoire d'ingénierie cognitive et sémantique
LR	*Linear Regression*
MCI	*Mild cognitive impairment*
MI	*Mutual Information*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
MLR	*Multilinear Logistic Regression*
NP	Syntagme nominal
NSE	*Noisy Speech Eval*
PCA	*Principal Component Analysis*
POS	*Part-of-speech tagging*
PP	Syntagme prépositionnel
RBF	*Radial Basis Function*

RF	*Random Forest*
Root	Root - Racine syntaxique
S	Proposition déclarative simple
SBAR	Proposition introduite par une conjonction de subordination
SBARQ	Question directe introduite par un mot ou groupe interrogatif
SINV	Phrase déclarative avec inversion sujet-verbe
SVM	*Support Vector Machine*
TALN	Traitement automatique du langage naturel
UFE	*UsAge Feature Extraction*
VP	Syntagme verbal
WHADVP	Syntagme adverbial interrogatif *(wh-adverb)*

LISTE DES SYMBOLES ET UNITÉS DE MESURE

γ	Gamma : hyperparamètre du SVM avec noyau RBF
C	Paramètre de régularisation

INTRODUCTION

L'Alzheimer est une maladie neurodégénérative qui entraîne une détérioration progressive des fonctions cognitives. Il s'agit de la forme de démence la plus répandue, touchant actuellement plus de 55 millions de personnes dans le monde, un chiffre qui devrait dépasser les 153 millions d'ici 2050 (Alzheimer's Research UK, 2022). Malgré sa forte incidence, la maladie n'est souvent diagnostiquée qu'à un stade avancé, alors que la personne en est atteinte depuis plusieurs années. Le diagnostic repose sur l'observation de changements comportementaux rapportés par les proches et implique parfois des méthodes invasives, telles que l'imagerie médicale ou les analyses biologiques (Alzheimer's Association, 2025a).

Un des principaux effets de la maladie est le déclin des fonctions langagières, observé par la difficulté d'une personne atteinte à maintenir une conversation fluide, à répéter un élément qui vient d'être mentionné ou à utiliser le bon vocabulaire pour nommer un item familier (Alzheimer's Association, 2025a). Pour ces raisons, les cliniciens utilisent souvent des tests élicitant la parole dans le but de détecter des signes d'Alzheimer ou d'autres maladies affectant le langage, tels que l'aphasie ou des troubles cognitifs légers.

Plusieurs chercheurs se sont tournés vers l'utilisation de processus informatiques automatisés pour analyser des caractéristiques linguistiques extraites de transcriptions ou d'enregistrements provenant d'échantillons de discours produits par les patients. Dans un premier temps, l'automatisation de ce processus permet l'ingénierie de caractéristiques significatives pour la détection précoce de la maladie, le traitement de volumes importants de données et l'utilisation de modèles d'apprentissage machine, contribuant à une meilleure compréhension des manifestations linguistiques permettant l'identification de la maladie. Dans un deuxième temps, l'utilisation d'outils automatisés permet d'identifier rapidement et de manière non invasive des biomarqueurs précoces, en détectant des altérations subtiles du langage souvent impossibles à repérer par des évaluations manuelles.

Le laboratoire d'ingénierie cognitive et sémantique (LiNCS) de l'École de technologie supérieure a mis au point un outil nommé UsAge, dont la finalité est d'automatiser le traitement de fichiers, d'analyser leurs caractéristiques linguistiques et d'appliquer un algorithme de classification en vue de regrouper les participants selon leur diagnostic. Néanmoins, en dépit des avancées réalisées, cet outil demeure limité par plusieurs contraintes techniques et méthodologiques qui le rendent actuellement difficilement exploitable.

0.1 Problématique

L'observation des troubles cognitifs et la détection de la maladie d'Alzheimer posent plusieurs difficultés. Les méthodes de détection actuellement utilisées sont souvent invasives (analyse du liquide céphalorachidien, imagerie structurelle par résonance magnétique), chronophages (Mini-Mental State Exam, Montreal Cognitive Assessment) (Alzheimer's Association, 2025b) et ne permettent pas toujours une comparaison directe dans le temps. Une solution possible réside dans le développement d'un outil flexible et automatisé, capable de suivre l'évolution de la maladie à partir de données linguistiques standardisées et exploitables par des algorithmes d'apprentissage automatique modernes. Or, la création d'un outil capable d'observer et de quantifier automatiquement le déclin cognitif à travers le langage représente un défi complexe. Un tel système doit d'abord reposer sur une architecture à la fois flexible et robuste, capable d'extraire de manière fiable et reproductible des caractéristiques linguistiques significatives, tout en assurant la manipulation des données et la classification de ces dernières.

0.2 Objectifs de recherche

Le but principal de cette recherche est de concevoir un tel outil automatisé permettant l'extraction, l'analyse et l'exploitation de marqueurs linguistiques de la maladie d'Alzheimer dans un cadre de classification. Ce travail s'appuie sur la base de code initiale de l'outil du LiNCS, qu'il

restructure, corrige et optimise en plus de l'enrichir de nouvelles fonctionnalités pour étendre sa couverture fonctionnelle et améliorer sa fiabilité.

À cette fin, trois objectifs spécifiques sont poursuivis :

0.2.1 Identification des problèmes et refonte du code du pipeline UsAge

Cet objectif vise à analyser le code afin d'identifier ses dysfonctionnements et à élaborer un plan d'action pour y remédier.

0.2.2 Observation et automatisation de nouvelles caractéristiques

Cet objectif vise à étudier trois nouveaux types de caractéristiques linguistiques ainsi qu'à évaluer leur impact sur la performance du classificateur afin de déterminer leur potentiel discriminant.

0.2.2.1 Chaînes de coréférence

Les chaînes de coréférence sont des suites d'expressions linguistiques qui renvoient à une même entité dans un texte ou un discours. L'entité principale, appelée « référent » , est désignée à plusieurs reprises par des « coréférents », souvent des pronoms ou des substitutions nominales, situés ailleurs dans la phrase ou dans le texte. Par exemple, dans les phrases « La jeune fille veut un biscuit. Elle en prend un dans la boîte. », le coréférent « elle » renvoie au référent « la jeune fille » alors que le pronom « en » réfère au biscuit.

Les chaînes de coréférence mobilisent des processus cognitifs complexes comme la mémoire de travail, laquelle est nécessaire pour maintenir et actualiser les référents au fil du discours et est souvent affectée par la maladie d'Alzheimer. Elles pourraient donc constituer un indicateur pertinent de troubles cognitifs. Cette hypothèse est en accord avec l'observation selon laquelle les patients atteints de troubles cognitifs ont tendance à privilégier l'usage des pronoms au

détriment des noms, ces derniers exigeant un effort cognitif plus important pour être retenus et conservés en mémoire. Plusieurs études ont en effet relevé un usage accru des pronoms chez ces patients ainsi qu'une corrélation négative entre l'utilisation de pronoms et celle de noms, indiquant ainsi un lien entre un nombre élevé de pronoms et un petit nombre de noms et vice versa ((Kavé & Levy, 2003) ; (Wilson *et al.*, 2010) ; (Fraser, Meltzer & Rudzicz, 2016) ; (Mueller *et al.*, 2018) ; (Cho *et al.*, 2020) ; (Cho *et al.*, 2022) ; (Williams, Theys & McAuliffe, 2023) ; (Gumus & Koo, 2023)).

Dans cette optique, Kavé & Levy (2003) ont analysé les erreurs relatives aux référents, c'est-à-dire l'utilisation incorrecte d'un pronom ou l'emploi d'un pronom n'ayant pas d'antécédent, en intégrant ces erreurs dans une catégorie plus large d'erreurs sémantiques. Les résultats montrent que les patients atteints de démence produisent une proportion significativement plus élevée d'erreurs sémantiques que les participants du groupe témoin. Les auteurs ne fournissent cependant pas de détails supplémentaires sur les modalités d'utilisation des référents.

Il apparaît donc pertinent d'examiner plus en profondeur les comportements liés aux chaînes de coréférence afin de déterminer si ces éléments permettent de faire une distinction entre les groupes de participants.

0.2.2.2 Pauses

L'utilisation de pauses dans un discours spontané marque une hésitation ou un moment de réflexion nécessaire pour compléter une idée. L'observation du nombre de pauses présentes dans un discours pourrait constituer un biomarqueur permettant de différencier des adultes atteints de troubles cognitifs de personnes en bonne santé ((Wilson *et al.*, 2010) ; (Ahmed, Haigh, de Jager & Garrard, 2013) ; (Gumus & Koo, 2023)). Les pauses permettraient également d'évaluer le débit de parole des participants, une caractéristique utile pour classifier automatiquement les deux groupes (Roark, Mitchell, Hosom, Hollingshead & Kaye, 2011). Les personnes atteintes

de troubles cognitifs auraient ainsi tendance à produire davantage de pauses ou des pauses d'une plus longue durée.

0.2.2.3 Complexité syntaxique

La complexité syntaxique est une étiquette dont la définition varie fortement selon les auteurs et dont les effets restent controversés (Fraser *et al.*, 2016). Plusieurs chercheurs se contentent de mesures basiques, comme la longueur des phrases, le nombre et le type de mots ou encore la structure syntaxique globale ((Wilson *et al.*, 2010) ; (Ahmed *et al.*, 2013) ; (Santander-Cruz, Salazar-Colores, Paredes-García, Guendulain-Arenas & Tovar-Arriaga, 2022)). D'autres s'intéressent au ratio de types de phrases et de syntagmes ainsi qu'à leur distribution ((Guinn & Habash, 2012) ; (Mueller *et al.*, 2018) ; (Gumus & Koo, 2023)).

Des mesures plus complexes, comme la longueur moyenne des dépendances ((Wilson *et al.*, 2010) ; (Roark *et al.*, 2011) ; (Fraser *et al.*, 2016) ; (Lindsay, Tröger & König, 2021)) ou des indices syntaxiques composites ((Onofre de Lira et al., 2011) ; (Mueller *et al.*, 2018)), sont également courantes dans la littérature. Ces approches semblent plus aptes à mettre en évidence des différences entre groupes.

Roark *et al.* (2011) et Fraser *et al.* (2016) ont aussi examiné la complexité syntaxique à travers des mesures telles que la profondeur et la largeur des arbres syntaxiques. Ces mesures offrent une évaluation plus raffinée que l'observation de ratios, mais restent limitées quant à leur observation des rôles syntaxiques ou encore à la représentation de la difficulté de traitement cognitif.

Pour observer la complexité syntaxique, nous avons mis en place des mesures de densité syntaxique, de ratio de syntagmes et de flux de dépendances.

La densité syntaxique est calculée à partir d'une structure générée à l'aide des pipelines de constituants et de dépendances de Stanza (Qi, Zhang, Zhang, Bolton & Manning, 2020). Cette

dernière est constituée des racines et des nœuds de l'arbre syntaxique et conserve leur relation pour permettre une manipulation plus efficace de la structure des phrases. Comme pour le module de coréférence, l'implémentation est modulaire, permettant de changer l'outil utilisé pour l'extraction de l'arbre de dépendances sans pour autant perdre les mesures intégrées.

Les mesures du ratio de syntagmes évaluent la diversité des types de syntagmes présents dans un texte et permettent d'apprécier la complexité syntaxique à un niveau global en analysant la densité des syntagmes selon leurs catégories. Leur calcul repose sur la structure récursive déjà implémentée pour l'analyse de la densité syntaxique. Cette dernière a été intégrée afin de compléter le système d'analyse existant, qui reposait exclusivement sur les étiquettes morphosyntaxiques et ne tenait pas compte des structures syntaxiques hiérarchiques.

Les mesures de flux de dépendance ajoutées sont issues de Kahane, Yan & Botalla (2017), qui définissent le flux de dépendance comme « un ensemble de dépendances dans [une] position donnée reliant un mot situé à gauche à un mot situé à droite » [traduction libre] (Kahane, Yan & Botalla, 2017, p.74). Cette approche vise à modéliser la charge cognitive impliquée dans la production linguistique en considérant le flux de dépendances syntaxiques à maintenir en mémoire pour produire un énoncé cohérent.

0.2.3 Évaluation et amélioration de l'algorithme de classification

Cet objectif consiste à analyser en profondeur l'algorithme de classification développé par le LiNCS et à proposer des améliorations visant à optimiser ses performances afin de déployer un modèle généralisable et performant sur des jeux de données ou des points de données inédits.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

L'analyse linguistique est devenue importante pour identifier le déclin cognitif associé à la maladie d'Alzheimer et à d'autres troubles neurodégénératifs. Des changements subtils dans la parole spontanée précèdent souvent l'apparition de symptômes cliniques évidents, faisant du langage une source de biomarqueurs prometteurs ((Ahmed *et al.*, 2013) ; (Gumus & Koo, 2023)). L'extraction automatisée de caractéristiques linguistiques à partir de transcriptions de la parole a gagné en popularité, car elle constitue un moyen non invasif, objectif et modulable pour détecter et suivre l'évolution de la maladie d'Alzheimer ((Fraser *et al.*, 2016) ; (Can & Kuruoğlu, 2019)). Elle permet également la mise en place de pipelines exploitant des modèles d'apprentissage automatique permettant de détecter précocement des patrons linguistiques associés à la maladie ou encore d'observer automatiquement sa présence.

Malgré les avancées, les pipelines automatisés actuels font face à des limitations importantes et reposent souvent sur des métriques simples basées sur la fréquence ou le ratio ou alors se concentrent uniquement sur certains domaines linguistiques, ce qui limite leur capacité à saisir la complexité des troubles du langage observés dans la maladie d'Alzheimer et les démences apparentées (Cho *et al.*, 2020). Kavé & Levy (2003) ont proposé des ratios d'observations verbales, visant à mesurer la densité des verbes dans un texte, une approche reprise et élargie par Ahmed *et al.* (2013) qui se sont concentrés sur les verbes avec inflexions. Onofre de Lira *et al.* (2011) ont développé un indice syntaxique basé sur le ratio de complexité des sous-types de phrases alors que d'autres études ont exploré les ratios de catégories grammaticales par nombre de mots ((Guinn & Habash, 2012) ; (Fraser *et al.*, 2016) ; (Lindsay *et al.*, 2021)). Cho *et al.* (2022) ont quant à eux mesuré le pourcentage de parole dans un texte et Santander-Cruz *et al.* (2022) ont introduit une métrique de diversité lexicale en comparant la longueur totale du texte à la taille du vocabulaire utilisé. Des ratios spécifiques, tels que le nombre de noms et de pronoms, ont été étudiés par Gumus & Koo (2023) et Williams *et al.* (2023) ont comparé le ratio des noms

et de copules par rapport aux verbes pour mieux capturer certaines structures syntaxiques et sémantiques. Enfin, Fromm *et al.* (2024) ont proposé une mesure du débit verbal en mots par minute.

Les pipelines sont également sensibles à la variabilité interlinguistique (Lindsay *et al.*, 2021) ainsi qu'à la taille réduite des échantillons, qui complique la généralisation des résultats (Guinn & Habash, 2012). Plusieurs études ont néanmoins identifié des tendances linguistiques robustes, telles qu'une réduction de la diversité lexicale ((Croisile *et al.*, 1996) ; (Kavé & Levy, 2003)), une utilisation accrue de pronoms et de mots vides (*thing, uhm, like*) (Fraser *et al.*, 2016) ; (Mueller *et al.*, 2018), une syntaxe simplifiée ((Onofre de Lira et al., 2011) ; (Cho *et al.*, 2020)) et une fréquence plus élevée de pauses (Williams *et al.*, 2023). Pourtant, peu d'outils intègrent ces caractéristiques dans des systèmes automatisés, interprétables et flexibles, adaptés à la recherche ou à l'usage clinique (Fromm *et al.*, 2024).

Cette revue examine les caractéristiques linguistiques principalement décrites dans la littérature et compare les processus automatisés ainsi que les pipelines de classification récents, en mettant en évidence les lacunes persistantes qui justifient le développement de systèmes plus complets.

1.2 Caractéristiques linguistiques observées

1.2.1 Caractéristiques lexicales

Les marqueurs lexicaux reflètent souvent la richesse du vocabulaire et les schémas d'utilisation des mots. De nombreuses études rapportent une réduction de la diversité lexicale et une préférence pour des mots ayant un usage fréquent et un âge d'acquisition plus bas chez les patients atteints de la maladie d'Alzheimer ((Kavé & Levy, 2003) ; (Cho *et al.*, 2022) ; (Williams *et al.*, 2023)). Par exemple, de Lira, Minett, Bertolucci & Ortiz (2014) et Kavé & Goral (2016) ont observé des difficultés d'accès au lexique dans les descriptions d'images produites par des patients. De même, Williams *et al.* (2023) ont mis en évidence des altérations dans l'utilisation des noms et des verbes dans la parole conversationnelle, comme une réduction globale du lexique employé dans

toutes les catégories de mots chez les patients atteints de la maladie d'Alzheimer et l'utilisation d'un plus grand nombre de verbes à usage fréquent (e.g : être). Par ailleurs, Cho *et al.* (2022) ont montré une corrélation entre la longueur des mots, leur fréquence, et les scores obtenus aux tests de *Mini Mental State Exam* et *Boston Naming Test*, utilisés pour déceler des troubles cognitifs. Fromm *et al.* (2024), quant à eux, ont conclu que les caractéristiques lexicales pouvaient être utilisées dans des approches de classification automatique.

Les mesures de fréquence lexicale, telles que le *type-token ratio*, l'indice de Brunet, la statistique d'Honoré ou encore la fréquence moyenne des mots, restent des indicateurs couramment utilisés, corrélés avec la sévérité du déclin cognitif (Hernández-Domínguez, 2019). Toutefois, elles tendent à simplifier à l'excès la complexité des troubles lexicaux, en donnant une image partielle, voire réductrice, de la production langagière du participant.

1.2.2 Caractéristiques morphosyntaxiques

Les marqueurs morphosyntaxiques sont utilisés pour quantifier la complexité syntaxique, morphologique et grammaticale des énoncés produits par les participants. Ils se caractérisent généralement par la longueur moyenne des phrases, la variété et le nombre de clauses syntaxiques, la proportion de types de mots utilisés — notamment les mots fonctionnels — , ainsi que par la présence d'erreurs grammaticales ((Kavé & Levy, 2003) ; (Onofre de Lira et al., 2011), (Ahmed *et al.*, 2013)). Selon Ahmed *et al.* (2013) et Fraser *et al.* (2016), ces marqueurs permettent de suivre la progression des troubles cognitifs en révélant un appauvrissement progressif de la structure syntaxique dans la parole des patients.

Fraser *et al.* (2016) ainsi que Mueller *et al.* (2018) ont réalisé une analyse factorielle exploratoire qui a mis en évidence la syntaxe comme l'un des facteurs centraux du langage, soulignant ainsi son rôle déterminant dans l'évaluation du déclin cognitif. Par ailleurs, de Lira *et al.* (2014) ont observé, chez des patients atteints de la maladie d'Alzheimer à un stade léger à modéré, une réduction de la longueur des énoncés accompagnée d'une raréfaction des constructions syntaxiques complexes. Lindsay *et al.* (2021) ont confirmé la constance de cette simplification

syntaxique à travers plusieurs langues, ce qui renforce son statut de marqueur robuste de la maladie.

Cependant, l'observation de la maladie par la syntaxe reste controversée : certains chercheurs suggèrent que les déficits syntaxiques observés pourraient résulter indirectement de troubles mnésiques ou sémantiques plutôt que d'un dysfonctionnement syntaxique isolé (Fraser *et al.*, 2016).

1.2.3 Caractéristiques discursives et sémantiques

Les marqueurs discursifs incluent les hésitations, les pauses, les mots vides et les répétitions, qui reflètent des difficultés de planification et de contrôle du discours. Ces phénomènes sont largement documentés dans la parole des patients atteints de troubles cognitifs, notamment dans la maladie d'Alzheimer ((Croisile *et al.*, 1996) ; (Wilson *et al.*, 2010) ; (Onofre de Lira et al., 2011) ; (Roark *et al.*, 2011) ; (Ahmed *et al.*, 2013) ; (Fraser *et al.*, 2016) ; (Orimaye, Wong, Golden, Wong & Soyiri, 2017) ; (Mueller *et al.*, 2018) ; (Gumus & Koo, 2023) ; (Hernández-Domínguez, 2019) ; (Lindsay *et al.*, 2021) ; (Cho *et al.*, 2022) ; (Fromm *et al.*, 2024)). Les pauses et l'usage accru de mots vides tendent à s'intensifier avec la progression de la maladie, constituant ainsi des indicateurs clés du déclin cognitif (Ahmed *et al.*, 2013).

La cohérence discursive et le contenu sémantique jouent également un rôle central dans une communication efficace, bien que leur quantification automatique demeure un défi pratique, en raison de l'absence de méthodes clairement définies pour les mesurer. Gumus & Koo (2023) ont mis en évidence un lien entre les altérations sémantiques et les symptômes cliniques, soulignant l'importance de ces marqueurs pour le suivi de la maladie. Des approches récentes ont recours à des modèles de plongements (*embeddings*) pour évaluer la similarité et la cohérence sémantique, offrant des indicateurs plus riches que de simples énumérations lexicales (Santander-Cruz *et al.*, 2022). Cependant, ces méthodes restent peu interprétables, ce qui limite leur compréhension ainsi que leur utilisation clinique directe.

Les Tableaux 1.1 et 1.2 résument les études récentes utilisant un pipeline automatisé ainsi que les principales constatations rapportées par les chercheurs.

Étude	Données	Langue	Méthodes	Catégorie	Découvertes
Fromm, D., Dalton G, Sarah et AL (2024)	Pitt Corpus	Anglais	CLAN : Analyse statistique	Lexicale	Patients produisent moins de mots lexicaux et ont une production plus lente.
Williams, E., Theys, C., & McAuliffe, M. (2023)	12 patients et 12 contrôles	Anglais	CLAN et WebCelex : Analyse statistique	• Production POS • Fréquence • Sémantique	Patients produisent des noms de fréquence plus basse, moins de noms, plus de pronoms, des noms avec un AoA plus élevée, une TTR plus basse et une diversité lexicale réduite.
Ahmed, S., & Haigh, A. (2023)	18 patients MCI, 15 contrôles	Anglais	Cookie Theft : Analyse statistique et longitudinale	• Parole • Complexité syntaxique • Contenu lexical • Erreurs de fluidité • Processus Sémantique	Déficience de la complexité syntaxique, du contenu sémantique, de la compréhension et de la pratique. Signification statistique de la complexité sémantique et syntaxique et du contenu lexical.
Gumus, A., & Koo, Y. (2023)	109 patients, 74 contrôles	Anglais	WinterLab processing pipeline, 'Brain-Code', SpaCy, Stanford NLP, Praat, Parselmouth, GloVe, FastText	• Lexicale • Temps • Acoustique • Sémantique • Syntactique • Discours • Cohérence • Sentiment	Catégorie lexicale et syntactique fortement corrélées. Scores faibles au MMSE et DRS associés à l'usage de mots plus courts et moins de phrases prépositionnelles. Patients avec atteinte sévère utilisent moins de noms et de verbes. Patients parlent moins longtemps et font plus de pauses.
Cho, S., Quilico Cousin, A. et Al. (2022)	93 patients, 28 contrôles	Anglais	Transcription manuelle du Cookie Theft, SpaCy, Natural Language Toolkit, Praat	• Lexicale • Acoustique	Corrélation entre fréquence des pauses, longueur et fréquence des mots et scores MMSE et BNT. Participants produisent moins de prépositions, de noms, d'adjectifs, des discours plus courts et plus lents, un temps de parole plus faible, plus d'adverbes et de répétitions.
Cho, S., Nevler N., et Al (2021)	138 patients, 37 contrôles	Anglais	POS tagging avec SpaCy, Analyse statistique, régression linéaire	• Mesures langagières • Mesures lexicales	Groupe contrôle produit plus de mots et de prépositions. Patients utilisent plus de noms et des noms avec un AoA plus précoce.
Can, S., & Kuruoglu, G. (2019)	39 patients, 39 contrôles	-	4 tests incluant Cookie Theft : Analyse statistique	Phrases coordonnées et composées	Patients produisent plus de phrases, plus de phrases courtes et moins complexes. Signification statistique pour les phrases composées nominales.

Tableau 1.1 Tableau récapitulatif des études utilisant un pipeline automatisé - 2024 à 2019

Étude	Données	Langue	Méthodes	Catégorie	Découvertes
Mueller, K., Koscik, R., et Al. (2018)	399 participants de WRAP	Anglais	Cookie Theft : CLAN et analyse statistique	• Sémantique • Lexicale	4 facteurs détectés pour mesurer les changements linguistiques : sémantique, syntaxe, fluidité, lexique.
Kave, G. & Goral, M. (2017)	20 patients, 20 contrôles	Hébreu	Cookie Theft : Analyse automatique et manuelle	Caractéristiques sémantiques	Signification statistique pour les tâches de dénomination et de fluidité. Différence significative pour les mots de contenu et le pourcentage de noms et de pronoms. Patients produisent plus de mots courts et plus fréquents et ont une richesse lexicale réduite.
Onofre de Lira, J., Soares Cianciarullo Minett, T. et Al (2014)	26 patients, 20 contrôles	Portugais	Cookie Theft : Analyse statistique	• Quantité de mots complets • Unité d'information	Diminution des performances en termes de quantité et de contenu chez les patients.
Onofre de Lira, A., & Ortiz, X. (2011)	60 patients, 61 contrôles	Portugais	Étude d'observation utilisant 7 images : Analyse statistique	• Erreurs lexicales • Syntaxe	Patients ont plus de difficultés à trouver les mots, plus de révisions et de répétitions et un indice syntaxique plus faible.
Wilson, S.M., Henri, Maya.L., et Al (2010)	60 patients PPA, 10 contrôles	Anglais	Analyse statistique	• Vitesse de parole • Erreurs de parole • Fluidité • Lexicale • Syntaxe	Patients ont un rythme lent, des distorsions, des erreurs syntaxiques et une complexité réduite. Ils produisent plus de mots de classe fermée, de pronoms et de verbes et de noms de fréquence plus élevée.
Kave, G. (2003)	14 patients, 48 contrôles	Hébreu	Cookie Theft : Analyse statistique	• Semantique conceptuelle • Syntaxe • Morphologie • Analyse d'erreur	Patients produisent moins d'unités d'information, plus de commentaires circonlocutoires, de pronoms, de mots, de phrases relatives et conjointes.
Croisile, B., Ska, B., et Al (1996)	22 patients, 24 contrôles	Français	Cookie Theft : Analyse statistique	• Lexicale • Syntaxique • Unité d'information	Patients produisent moins d'éléments de contenu, de mots, de clauses subordonnées et ont un contenu informatif médiocre. Ils produisent plus de phrases courtes, de répétitions, de paraphasies, de mots indéfinis, de déictiques, de phrases sémantiques, de répétitions de mots et d'informations non pertinentes.

Tableau 1.2 Tableau récapitulatif des études utilisant un pipeline automatisé - 2018 à 1996

1.3 Automatisation et classification

L'automatisation de l'extraction de caractéristiques linguistiques est cruciale pour permettre une analyse évolutive et objective des changements langagiers associés à la maladie d'Alzheimer. La majorité des pipelines automatisés suit une série d'étapes semblables : transcription, prétraitement, segmentation, étiquetage morphosyntaxique puis extraction des caractéristiques. Ces étapes sont réalisées à l'aide d'outils qui facilitent la saisie d'observations et permettent l'extraction de multiples catégories de caractéristiques. Le Tableau 1.3 présente des pipelines de classification automatisés entre participants témoins et participants atteints de démence ainsi que les meilleurs résultats obtenus. Nous avons exclu ceux contenant des données audio, puisque le présent projet n'intègre pas encore de caractéristiques audio ou ne permet pas une classification entre démence et d'autres types de troubles cognitifs. Nous avons également écarté les pipelines utilisant des modèles de *deep learning*, car la taille limitée de notre jeu de données ne permet pas d'utiliser ce type de modèles sans risquer un surapprentissage.

Tableau 1.3 Comparaison de modèles de classifications automatiques récentes

Étude	Participants	Langue	Nombre de caractéristiques	Algorithmes	Meilleur résultat
Santander-Cruz et al. (2022)	PittCorpus : 307 patients et 243 témoins	Anglais	17	KNN, RF SVM, ANN	Exactitude = 0,78
Lindsay, Tröger & König (2021)	76 patients et 78 témoins	Français, anglais	30	LR, SVM, MLP	ASC= 0,87
Abiven (2020)	Pitt Corpus : 300 patients et 217 témoins CRIUGM : 29 patients et 26 témoins	Anglais, français	37	SVM, DT, RF	ASC= 0,76
Hernández-Domínguez (2019)	Pitt corpus : 300 patients et 217 témoins	Anglais	37	SVM	ASC= 0,79
Hernández-Domínguez (2018)	Pitt corpus : 300 patients et 217 témoins BBVA corpus : 39 patients et 30 témoins	Anglais, espagnol	31+	RF, SVM	Exactitude = 0,98
Orimaye et al. (2017)	DementiaBank : 99 patients et 99 témoins	Anglais	1000	SVM	ASC= 0,93
Fraser, Meltzer & Rudzicz (2016)	DementiaBank : 167 patients et 97 témoins	Anglais	35	MLR	Exactitude = 0,82
Guinn & Habash (2012)	CCC : 31 patients et 57 témoins	Espagnol, anglais	14	KNN, DT, SVM	Exactitude = 0,79
Roark, Mitchell et al. (2011)	37 patients et 37 témoins	Anglais	21	SVM	ASC= 0,86

Les pipelines développés par (Fraser *et al.*, 2016), (Orimaye *et al.*, 2017) et (Lindsay *et al.*, 2021) intègrent des mesures complexes telles que le *parse tree score*, la fréquence normalisée SUBTL (Fromm *et al.*, 2024), les règles de production et la distance de dépendance (Orimaye *et al.*, 2017) ou encore la complexité syntaxique et la densité d'information (Lindsay *et al.*, 2021). Cependant, les autres pipelines, à savoir ceux de (Roark *et al.*, 2011), (Guinn & Habash, 2012), (Hernández-Domínguez, 2019), (Abiven, 2020) et (Santander-Cruz *et al.*, 2022) s'intéressent

plutôt à la conception d'outils automatisés de classification, se concentrant sur des métriques plus simples, telles que des dénombrements et des ratios, ou sur des dimensions limitées — par exemple la syntaxe ou la sémantique — négligeant ainsi la complexité réelle des troubles.

Un des principaux dilemmes est le compromis entre l'automatisation et l'interprétabilité. Les systèmes entièrement automatisés utilisent des métriques simples, faciles à extraire, mais peu représentatives de la richesse des troubles linguistiques liés à la maladie ((Roark *et al.*, 2011) ; (Guinn & Habash, 2012)). Ces caractéristiques sont interprétables, mais manquent de profondeur. Inversement, les pipelines qui intègrent des caractéristiques plus riches nécessitent souvent une intervention manuelle ou ne sont pas entièrement automatisés ((Santander-Cruz *et al.*, 2022) ; (Fromm *et al.*, 2024)). De plus, les systèmes existants séparent souvent l'extraction, la classification et la visualisation des caractéristiques, nécessitant de ce fait une expertise technique pour être utilisés efficacement (Williams *et al.*, 2023). Cette fragmentation limite leur adoption en milieu clinique et complique l'avancement de la recherche.

1.4 Conclusion

L'analyse automatisée du langage constitue un levier puissant pour la détection précoce de la maladie d'Alzheimer et le suivi de la progression des troubles cognitifs, en s'appuyant sur des indices subtils présents dans la parole spontanée. Cependant, les outils disponibles privilégient souvent l'automatisation au détriment de la richesse linguistique ou de l'interprétabilité. Ce compromis difficile entre automatisation, complexité des caractéristiques linguistiques et interprétabilité freine considérablement leur utilisation, surtout en milieu clinique.

Ces constats soulignent la nécessité de développer de nouveaux outils flexibles, capables non seulement de classifier, mais également d'expliquer et de visualiser les résultats, tout en permettant des analyses linguistiques plus poussées ainsi que l'exploration des données et la manipulation des caractéristiques.

CHAPITRE 2

MÉTHODOLOGIE

Ce chapitre décrit la méthodologie utilisée pour atteindre chacun de nos objectifs, soit l'identification des problèmes dans le code de base du pipeline UsAge, l'intégration de nouvelles caractéristiques discriminantes et l'amélioration de l'algorithme de classification. Nous présenterons également les données utilisées pour concevoir cet outil et entraîner les algorithmes de classification.

2.1 Objectif 1 : Identification des problèmes dans le code de base du pipeline UsAge

Afin d'identifier les problèmes dans le code source du pipeline UsAge, une analyse approfondie de celui-ci a d'abord été réalisée. Cette revue nous a permis de comprendre les requis initiaux du développement, tout en évaluant les bogues, la complexité, les duplications de code et la performance générale. Nous avons également testé individuellement chaque fonction chargée d'extraire des mesures linguistiques, afin de vérifier qu'elle produisait bien les résultats attendus. Ce processus nous a permis d'identifier des fonctions présentant des dysfonctionnements, des erreurs ou encore des fonctions mentionnées dans la documentation, mais absentes du code source.

Nous avons ensuite évalué la structure du code en détail et avons progressivement éliminé les redondances et mis en retrait les fonctions défaillantes ou dont le comportement s'écartait des spécifications initiales. Cette analyse a révélé que la base du code était trop fragile et inadaptée, ce qui nécessitait une réécriture complète avec une architecture repensée.

La phase de conception a été initiée en prenant en compte les contraintes inhérentes à la flexibilité, à la modularité et à la maintenabilité du système, avec une considération explicite des dépendances intermodules. Cette démarche architecturale vise à optimiser la réutilisabilité et la scalabilité du projet pour les futurs étudiants qui utiliseront l'application, tout en assurant une intégration cohérente et évolutive de nouvelles fonctionnalités.

2.2 Objectif 2 : Intégration de nouvelles caractéristiques discriminantes

Dans le cadre de cette recherche, nous avons évalué l'impact de trois catégories de caractéristiques linguistiques : les chaînes de coréférence, les pauses et la complexité syntaxique. Afin de quantifier leur influence respective, le modèle de classification de l'outil initial a été reconstruit, servant ainsi de base de référence pour les comparaisons ultérieures. Nous avons réalisé des tests statistiques ainsi que des évaluations de la contribution des caractéristiques. De plus, nous avons comparé les performances de deux outils automatiques en les confrontant aux résultats manuels pour en vérifier l'efficacité.

2.2.1 Modèle de classification de base

Afin de disposer d'un modèle de base permettant d'évaluer l'effet des nouvelles caractéristiques introduites, nous avons reconstitué le classificateur décrit dans la littérature d'UsAge, qui était absent de la base de code initiale. Tel que décrit par Cesari (2023), nous avons intégré un SVM linéaire utilisant *StandardScaler* pour normaliser les données et ayant été entraîné avec une validation croisée à 10 blocs qui retourne la moyenne de tous les résultats obtenus. Ce modèle a été entraîné sur un ensemble de 77 mesures, reproduisant aussi fidèlement que possible celles du pipeline de Cesari (2023) (voir le Tableau 3.1 pour la liste complète des mesures).

Le modèle de base affiche un score F1 de 0,759, une précision de 0,811, un rappel de 0,723 et une ASC de 0,848. Ces performances serviront de référence pour évaluer l'impact des nouvelles caractéristiques.

2.2.2 Tests statistiques

Afin d'évaluer la significativité statistique des nouvelles caractéristiques, nous avons calculé l'ANOVA en utilisant un seuil de $F > 4$ et un seuil de valeur p fixé à 0,05. Nous avons retenu l'ANOVA pour évaluer l'impact de variables continues, soit les mesures des caractéristiques, sur la cible binaire. Cette méthode, rapide et standard en sélection de caractéristiques, compare

directement la variance expliquée par les classes et permet de valider la pertinence de nouvelles variables dans un cadre de classification intergroupe.

Nous avons également calculé le coefficient bisérial de point afin d'évaluer la corrélation entre les nouvelles caractéristiques, ayant des valeurs continues, et les classes cibles, ayant des valeurs binaires.

2.2.3 Tests de contribution

Afin d'observer l'impact direct des nouvelles caractéristiques, nous avons eu recours à des tests d'attribution (Ribeiro, Singh & Guestrin, 2016) et à des tests d'ablation (Sheikholeslami, 2019).

Les tests d'attribution consistent à entraîner plusieurs modèles de classification en utilisant les caractéristiques individuellement afin d'évaluer leur contribution directe sur les performances.

Les tests d'ablation, quant à eux, consistent à retirer les caractéristiques une à une du modèle complet afin de mesurer la dégradation des performances et ainsi estimer l'importance relative de chaque caractéristique.

2.2.4 Comparaison des outils automatiques

Afin de calculer les mesures relatives aux chaînes de coréférence, nous avons intégré le pipeline de coréférence ((Dobrovolskii, 2021) ; (D'Oosterlinck *et al.*, 2023)) de Stanza ainsi que celui de SpaCy (Honnibal & Montani, 2017), afin de comparer les performances de ces deux outils de traitement automatique du langage naturel (TALN). Pour évaluer l'efficacité des deux modèles d'extraction automatique, 50% du corpus a été annoté manuellement afin d'obtenir des standards de référence pour les chaînes de coréférence. Les annotations ont été réalisées à l'aide de la plateforme INCEpTION (Klie, Agatonovic & Cimiano, 2018), un environnement collaboratif conçu pour la gestion, la visualisation et l'export automatisé d'annotations manuelles. Les annotations ont été effectuées par un stagiaire du laboratoire LiNCS, sous notre supervision, afin de garantir une cohérence suffisante dans l'interprétation des chaînes de coréférence.

Pour comparer les annotations manuelles et automatiques, nous avons calculé les mesures comparatives suivantes :

- **Ratio de clés communes** : Le ratio de clés communes correspond au nombre de coréférents de même longueur détectés par les différents outils par rapport au nombre total de coréférents détectés pour chaque référent. Il valide la similitude du contenu des chaînes de référence avec le standard manuel et est calculé à l'aide de l'indice de Jaccard (Jaccard, 1901) :

$$\text{ratio\_clés\_communes} = \frac{|\text{clé}(D_1) \cap \text{clé}(D_2)|}{|\text{clé}(D_1) \cup \text{clé}(D_2)|} \tag{2.1}$$

  où :

  Clé = entrée du dictionnaire des coréférents

  D_1 = Clés extraites manuellement

  D_2 = Clés extraites à l'aide d'un outils automatique

- **Ratio de correspondances exactes** : Le ratio de correspondances exactes correspond au nombre de paires présentant à la fois la même clé et la même valeur pour chaque référent — c'est-à-dire le même nombre de chaînes de coréférence pour une même longueur — rapporté au nombre total de clés communes entre les deux outils.

$$\text{ratio\_correspondance\_exactes} = \frac{\sum_{k \in K} \{D_1(k) = D_2(k)\}}{|K|} \tag{2.2}$$

  où :

  $K = \text{clés}(D_1) \cap \text{clés}(D_2)$

  Clé = entrée du dictionnaire des coréférents

  D_1 = Clés extraites manuellement

  D_2 = Clés extraites à l'aide d'un outils automatique

- **Score de proximité** : Le score de proximité (Manning & Schütze, 1999) quantifie la similarité entre les distributions des longueurs de chaînes de coréférence extraites par deux outils, en attribuant un crédit partiel aux différences faibles. Il est défini comme suit :

$$\text{score\_proximité} = \sum_{k \in K \setminus E} \frac{1}{1 + |D_1(k) - D_2(k)|} \tag{2.3}$$

où :

$$K = \text{clés}(D_1) \cap \text{clés}(D_2),$$
$$E = \{k \in K \mid D_1(k) = D_2(k)\}$$

et $|D_1(k) - D_2(k)|$ est la différence absolue entre les valeurs associées à la clé k.

2.3 Objectif 3 : Évaluation et amélioration du classificateur de base

Afin d'évaluer le modèle de base et de l'améliorer, nous avons utilisé des méthodes de sélection de caractéristiques ainsi que des techniques de comparaison des performances en classification.

Les méthodes de sélection de caractéristiques incluent des tests d'ablation et d'attribution, tels que décrits dans la section précédente. Nous utiliserons également le score d'information mutuelle, une mesure qui quantifie la dépendance entre deux variables en attribuant un score de 0 aux variables complètement indépendantes.

En ce qui concerne l'amélioration de la classification, nous avons optimisé les différents algorithmes implémentés dans notre outil en utilisant *GridSearch*, une méthode qui permet de tester systématiquement plusieurs combinaisons d'hyperparamètres pour chaque modèle et de retenir celle qui donne les meilleurs résultats. Les scores de performance observés sont le score F1, la précision, le rappel et l'aire sous la courbe (ASC).

2.4 Données

Les données utilisées pour la conception de l'outil et l'entraînement et l'évaluation des modèles de classification proviennent du Pitt Corpus (Becker et al., 1994), une base de données longitudinales créée par l'Université de Pittsburgh et disponible en anglais via la *DementiaBank*. Ce corpus contient des enregistrements et des transcriptions provenant du test *Cookie Theft* (Kaplan & Goodglass, 1983), une tâche de description issue du *Boston Diagnostic Aphasia Examination*, dans laquelle les participants sont appelés à décrire tout ce qu'ils perçoivent dans une image. L'image utilisée pour le test est montrée dans la Figure 2.2.

Demographics of DementiaBank data

	AD ($n = 240$)	Control ($n = 233$)
Age (years)	71.8 (8.5)	65.2 (7.8)
Education (years)	12.5 (2.9)	14.1 (2.4)
Gender (male/female)	82/158	82/151
Mini-Mental State Exam	18.5 (5.1)	29.1 (1.1)

Figure 2.1 Tableau descriptif des participants présents dans la base de données du Pitt Corpus, tiré de Fraser, Meltzer & Rudzicz (2016)

La base de données comprend 552 enregistrements issus de 233 participants faisant partie du groupe témoin et 240 participants ayant reçu un diagnostic de troubles cognitifs, tel qu'indiqué dans la Figure 2.1. Les participants sont âgés de plus de 44 ans et ont généralement un niveau d'éducation supérieur à sept années d'études. Les diagnostics recensés incluent la maladie d'Alzheimer probable, la maladie d'Alzheimer possible, le trouble cognitif léger, le trouble cognitif léger affectant principalement la mémoire ainsi que la démence vasculaire.

Les transcriptions et annotations manuelles des enregistrements ont été réalisées selon le protocole TalkBank CHAT (*Codes for the Human Analysis of Transcripts*) (MacWhinney, 2021), qui fait partie du système plus large CHILDES (*Child Language Data Exchange System*) utilisé pour l'analyse linguistique.

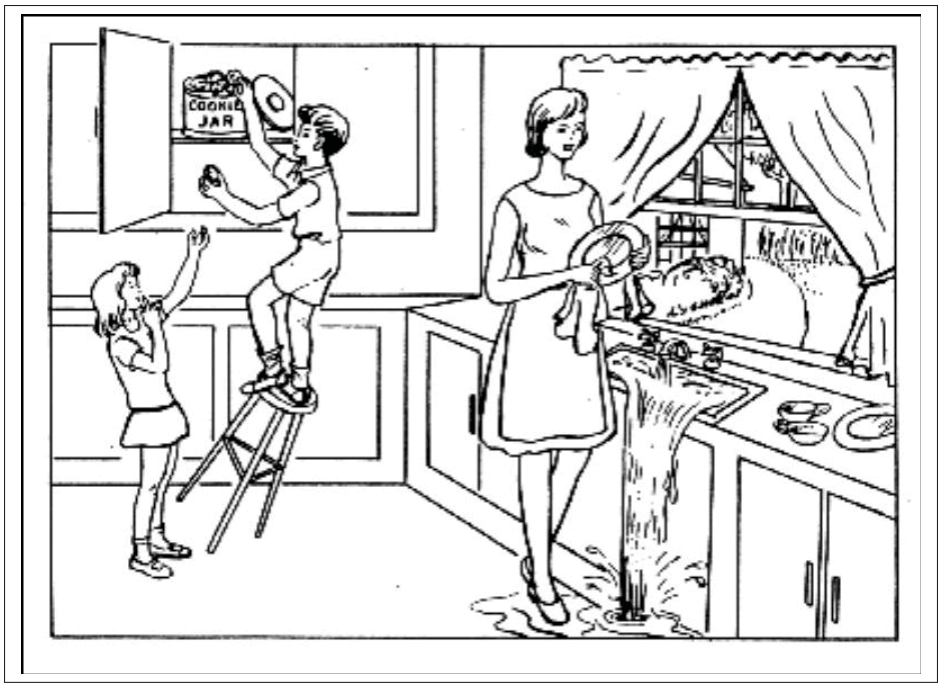

Figure 2.2 Image du test du Cookie Theft provenant du Boston Diagnostic Aphasia Examination (Kaplan & Goodglass, 1983)

CHAPITRE 3

OUTIL DE RECHERCHE DU LINCS

3.1 Introduction

UsAge est un outil de manipulation des données et de classification développé par le LiNCS depuis 2019. Il permet d'extraire des caractéristiques linguistiques initialement sélectionnées par Hernández-Domínguez (2019). Abiven (2020) a ensuite automatisé le processus d'extraction en développant un pipeline de traitement des transcriptions, puis Cesari (2023) a contribué à la création de graphiques dynamiques permettant de regrouper les données selon des ensembles de caractéristiques et de les visualiser de manière claire et interactive.

Cependant, l'application présentait plusieurs limitations majeures qui compromettaient son efficacité, en restreignaient l'usage et freinaient significativement le développement ultérieur, notamment l'ajout de nouvelles fonctionnalités et l'amélioration des modèles de classification.

Le code d'origine comportait plusieurs incohérences structurelles et la logique d'exécution des tâches affichait une redondance importante, résultant en un temps d'exécution élevé, une forte consommation de ressources et, surtout, de nombreuses erreurs d'exécution. Plusieurs mesures mentionnées dans la documentation ou dans la littérature associée n'étaient pas implémentées dans le code source ou bien ne retournaient pas de valeurs valides représentant fidèlement l'intention du calcul. C'est également le cas du processus de classification mentionné par Abiven (2020) et Cesari (2023), qui apparaît comme absent dans la base de code actuelle. Cette divergence entre la documentation et l'implémentation soulève des questions quant à la fiabilité et la complétude de l'outil tel que présenté.

Une nouvelle architecture a donc été mise en place afin de créer deux composants distincts :

1. *UsAge Feature Extraction* (UFE), une bibliothèque prenant en charge l'exécution des tâches d'extraction de caractéristiques de manière spécialisée et optimisée et permettant à l'utilisateur de passer une liste de mesures et d'obtenir un dictionnaire contenant les valeurs associées aux mesures demandées.

2. UsAge, un outil intégrant UFE pour récupérer les données extraites, qui se concentre sur la manipulation, l'observation, la visualisation et la classification des données.

Cette architecture, à la fois robuste et flexible, facilite l'intégration future de modules complémentaires et permet d'interchanger les outils intégrés — tels que les outils de TALN ou d'apprentissage machine (ML) — sans compromettre les mesures intégrées. Elle renforce également la stabilité du code en assurant une séparation claire des responsabilités : chaque composant est spécialisé et soumis à une batterie de tests rigoureux.

3.2 *UsAge Feature Extraction* : Bibliothèque d'extraction de caractéristiques

Nous avons conçu *UsAge Feature Extraction*, une bibliothèque développée en Python, destinée à l'extraction automatique de caractéristiques linguistiques visant à analyser et à quantifier les habitudes langagières. Elle repose sur un système de pipelines et de méthodes flexibles, offrant aux utilisateurs une grande liberté dans l'analyse de différentes caractéristiques. L'outil fonctionne en traitement multifil pour garantir une exécution efficace et met en place une liste de graphes acycliques dirigés (DAG) afin d'optimiser la gestion des dépendances entre les différentes mesures. La Figure 3.1 illustre le fonctionnement global de la bibliothèque. Les sous-sections suivantes détaillent deux composantes majeures de l'application et leur refonte : la gestion des dépendances, ainsi que les mesures ajoutées et l'extraction des caractéristiques.

3.2.1 Gestion des dépendances

Deux types de dépendances doivent être pris en compte : le type de fichier d'entrée et les transformations préalables nécessaires à l'extraction de certaines mesures.

La gestion des différents types de fichiers est assurée par des *readers* spécialisés capables d'identifier le format d'entrée et de déterminer les mesures exploitables. Par exemple, seuls les fichiers au format .cha seront soumis aux mesures nécessitant les annotations CHAT spécifiques à ce format. Les fichiers dans d'autres formats, comme .txt ou .srt, ne seront évalués qu'avec les mesures adaptées aux caractéristiques de ces fichiers.

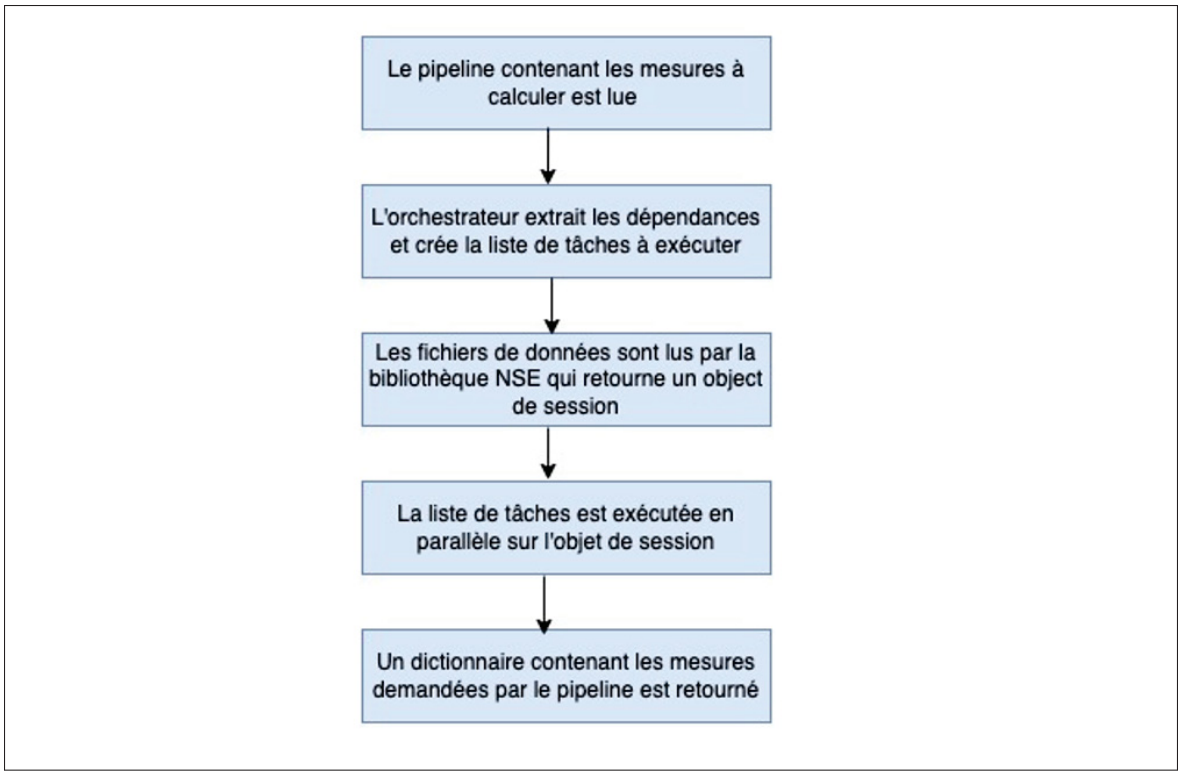

Figure 3.1 Fonctionnement logique de la bibliothèque *UsAge Feature Extraction*

Certaines mesures extraites par l'outil dépendent de traitements préalables, par exemple de l'extraction des étiquettes morphosyntaxiques ou de l'initialisation d'une classe spécifique permettant d'extraire une structure nécessaire à un calcul, tel un dictionnaire syntaxique. Ces dépendances doivent être résolues avant que les mesures ne puissent être correctement calculées.

Pour garantir une gestion efficace des tâches, un orchestrateur basé sur le principe de DAG et exploitant le traitement multifil simultané pour paralléliser l'exécution des opérations a été implémenté. Cette architecture permet de s'assurer qu'aucune tâche ne soit extraite de manière redondante. Le fonctionnement de l'orchestrateur est illustré dans la Figure 3.2. Il permet de construire dynamiquement une liste d'exécution des tâches en fonction des mesures demandées par l'utilisateur et des configurations requises. Pour chaque mesure, le système identifie les dépendances nécessaires, soit les transformations à appliquer aux données. Ces transformations sont ensuite consolidées afin d'éviter toute redondance : chaque transformation n'est effectuée qu'une seule fois, même si elle est requise par plusieurs mesures. Une même transformation

appliquée avec des paramètres différents est considérée comme deux transformations distinctes. Enfin, les mesures n'ayant aucune dépendance commune sont exécutées en parallèle, ce qui améliore significativement la performance globale du processus. La Figure 3.3 illustre le processus du DAG mis en place, où les dépendances sont représentées par des rectangles et les tâches nécessitant ces dépendances sont représentés par des cercles. Chaque colonne correspond à une liste de tâches pouvant être exécutées en parallèle. Les mesures nécessitant des transformations spécifiques sont, quant à elles, exécutées uniquement après que ces transformations aient été complétées, garantissant ainsi la cohérence et la validité des résultats.

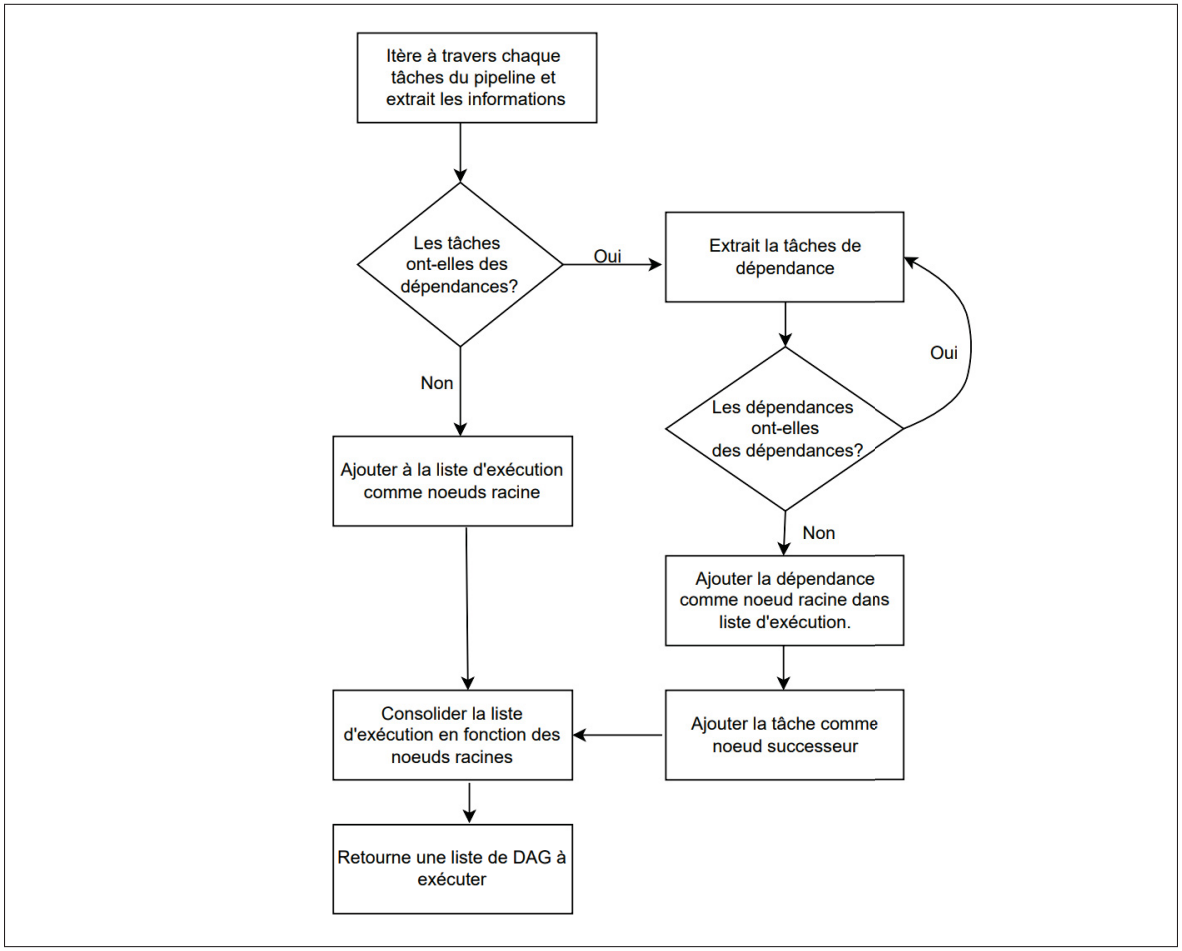

Figure 3.2 Fonctionnement logique de l'orchestrateur utilisé par UFE pour extraire les dépendances de chaque tâches et créer une liste de DAG d'exécution

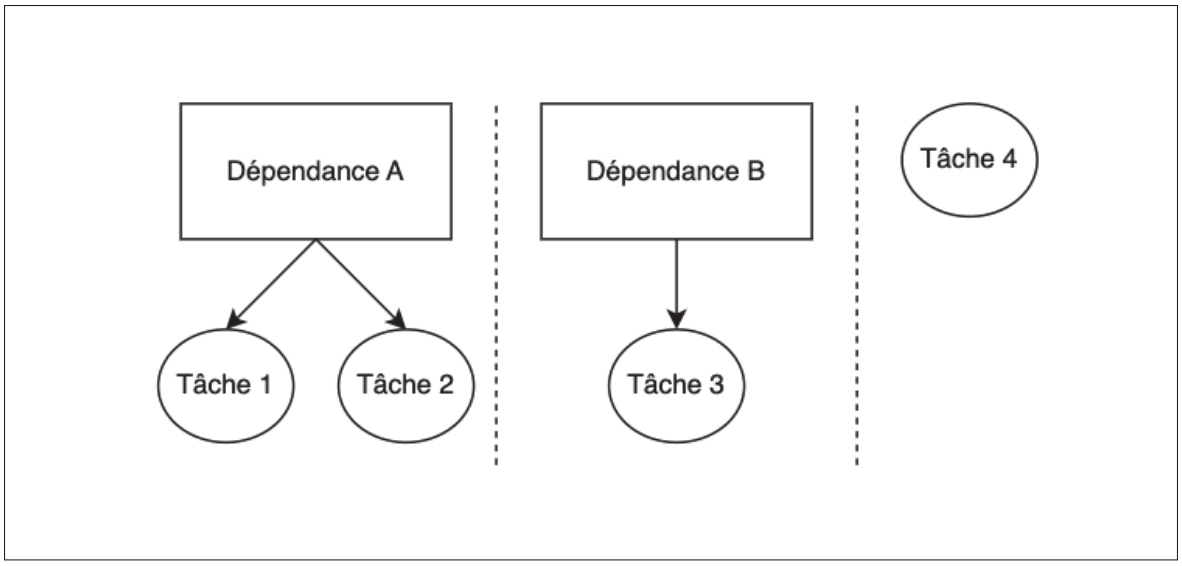

Figure 3.3 Représentation de la liste des DAG de tâches où chaque nœud racine est un élément de la liste et peut être exécuté en parallèle

3.2.2 Mesures et extraction des caractéristiques

Toutes les mesures implémentées dans la base de code d'UsAge ont été testées individuellement et manuellement afin de vérifier qu'elles mesuraient bien ce pour quoi elles avaient été conçues.

Lorsque des écarts ont été identifiés entre le comportement attendu et les résultats obtenus, les mesures concernées ont été réimplémentées afin d'assurer leur conformité aux objectifs initiaux. Une suite de tests automatisés a également été mise en place afin de garantir la pérennité et la fiabilité des mesures implémentées.

Des 120 mesures de base (Cesari, 2023), seules 57 ont été intégrées à la nouvelle structure. Les autres fonctions ont été écartées, car jugées non pertinentes : certaines ne calculaient qu'une partie des mesures attendues, d'autres étaient totalement inexploitables, soit parce qu'elles produisaient des résultats incorrects, soit parce qu'elles ne réalisaient pas du tout les opérations décrites. Sur ces 57 mesures intégrées, seules 10 ont été ajoutées telles quelles. Les 47 autres ont été réécrites pour corriger des erreurs de calcul ou de logique ou alors, elles ont été implémentées à l'aide de la bibliothèque *Noisy Speech Eval* (NSE).

NSE est une bibliothèque d'évaluation de transcription basée sur des fichiers au format CHAT, développée par le laboratoire LiNCS (Dupuis-Desroches, 2025). Cette bibliothèque a été intégrée à l'outil afin de réutiliser la logique de calcul de certaines mesures et de centraliser les implémentations développées au sein du laboratoire, favorisant ainsi leur maintenance et leur réutilisabilité. 24 mesures supplémentaires — indiquées dans le tableau 3.1 — ont été extraites à l'aide de cette bibliothèque. Nous avons intégré ces mesures afin d'assurer la complétude de l'outil et parce qu'elles sont largement représentées dans la littérature consacrée aux biomarqueurs linguistiques de la maladie d'Alzheimer.

Nous avons également intégré l'outil de TALN Stanza en remplacement de FreeLing 4.0 (Cesari, 2023), en raison de ses fonctionnalités plus étendues et de ses meilleures performances. Cette intégration a été conçue de façon flexible, permettant l'utilisation interchangeable d'autres bibliothèques de TALN sans compromettre la logique existante.

Nous avons ensuite introduit les mesures testées et améliorées dans la nouvelle base de code selon une classification linguistique des caractéristiques, favorisant une organisation cohérente et facilitant leur gestion. Cette classification améliore le regroupement des caractéristiques lors de la manipulation des résultats extraits. Une liste exhaustive des mesures extraites par la bibliothèque, regroupées selon leur classe linguistique, est présentée dans le tableau 3.1.

3.3 UsAge : Manipulation des données et classification

Nous avons ensuite développé UsAge, une application développée en Python qui permet d'explorer, de manipuler et de regrouper dynamiquement des données, ainsi que d'appliquer des algorithmes de classification à ces regroupements. Elle intègre également UFE afin d'extraire les caractéristiques linguistiques nécessaires aux observations.

Cet outil permet d'expérimenter différents regroupements des données selon plusieurs configurations d'étiquetage. Il est ainsi possible de sélectionner dynamiquement uniquement les participants témoins et ceux ayant reçu un diagnostic de MCI, ou encore de regrouper l'ensemble des

Tableau 3.1 Caractéristiques extraites par UFE

	Catégorie	**Caractéristique**	**Provenance**
Discours	Expressions	nb_expressions	Ancien UsAge
	Pauses	filled_pauses	NSE
		syllable_pauses	NSE
		unfilled_pauses	NSE
		short_pauses	NSE
		medium_pauses	NSE
		long_pauses	NSE
		pauses_in_duration	NSE
		total_time_pauses	NSE
		total_nb_pauses	NSE
		SHORT_pauses_POS	Nouvelle caractéristique
		MEDIUM_pauses_POS	Nouvelle caractéristique
		LONG_pauses_POS	Nouvelle caractéristique
		beg_of_sentence_POS	Nouvelle caractéristique
	Chaînes de références	missing_referent	Nouvelle caractéristique
		nb_anaphoras	Nouvelle caractéristique
		nb_cataphoras	Nouvelle caractéristique
		mean_coref_length	Nouvelle caractéristique
		coref_ratio	Nouvelle caractéristique
		coref_density	Nouvelle caractéristique
		coref_mean_distance	Nouvelle caractéristique
	Parole	duration	NSE
		empty_speech	NSE
		false_start	NSE
		jargon	NSE
		mean_length_utterance	NSE
		overlaps	NSE
		unintelligible_sequence	NSE
		words_per_minute	NSE
Lexicale	Disfluences	nb_repetitions	Ancien UsAge
		nb_retracing	Ancien UsAge
	Diversité	brunet_index	Ancien UsAge
		entropy	Ancien UsAge
		hapax_legomena	Ancien UsAge
		hapax_dislegomena	Ancien UsAge
		honore_r	Ancien UsAge
		sichel_s	Ancien UsAge
		ttr	Ancien UsAge
		yule_k	Ancien UsAge
	Erreurs	agramatic_utt	NSE
		disfluencies_error	NSE
		generic_errors	NSE
		neologism	NSE
		phonological_errors	NSE
		semantic_errors	NSE
		morphological_errors	NSE
Morphosyntaxique	Distribution	ratio	Ancien UsAge (Stanza)
		freq	Ancien UsAge (Stanza)
	Complexité syntaxique	mean_syntactic_density	Nouvelle caractéristique
		phrase_type_ratio	Nouvelle caractéristique
		flux_distribution	Nouvelle caractéristique
		mean_flux_weight	Nouvelle caractéristique
		max_flux_weight	Nouvelle caractéristique

participants diagnostiqués dans un même groupe afin de les comparer au groupe témoin. Il permet également l'isolement de participants spécifiques en vue d'analyser en détail leurs caractéristiques linguistiques individuelles. Cela favorise la recherche et le développement de nouvelles observations et expose des méthodes permettant l'analyse de données en plus de mettre en place plusieurs algorithmes de classification optimisés. La structure d'UsAge est illustrée dans la Figure 3.4.

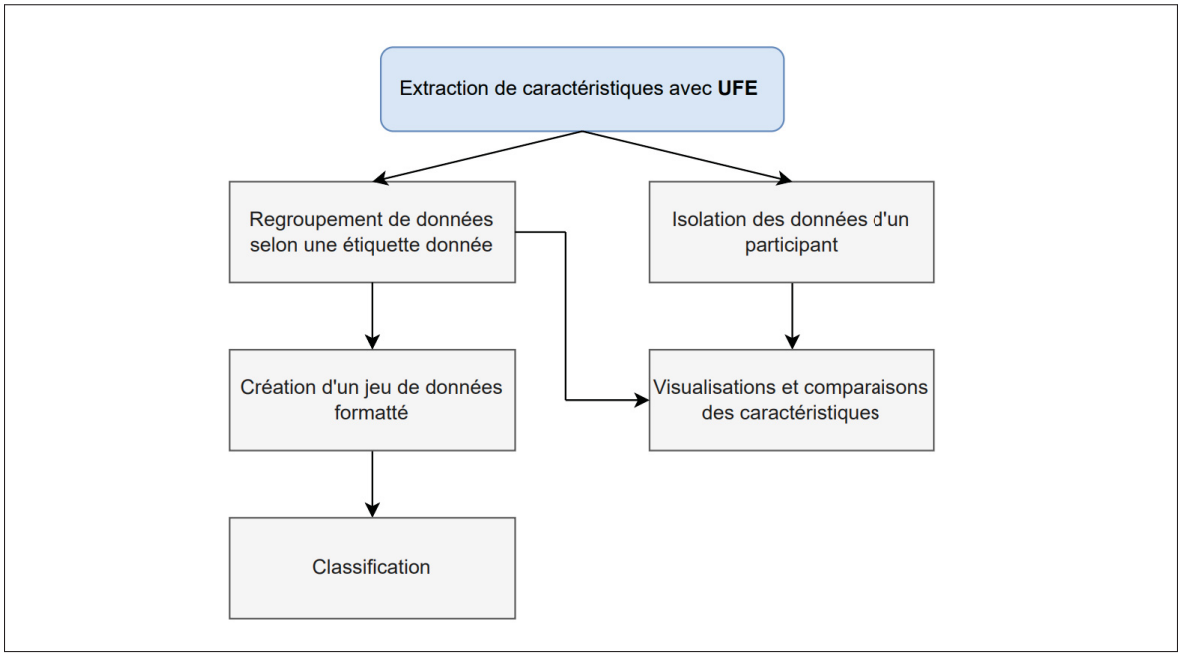

Figure 3.4 Structure d'UsAge

3.3.1 Création dynamique de jeux de données

UsAge implémente plusieurs classes permettant de structurer et de séparer les données, afin de dynamiquement former des groupes en fonction de critères définis (étiquette, âge, identification, etc.). Ces classes partagent un ensemble de méthodes communes permettant d'inspecter et de manipuler les données de manière uniforme, quel que soit le type de regroupement. Elles assurent également le formatage des groupes afin de les rendre directement exploitables par les méthodes de classification.

Les exemples suivants illustrent des graphiques générés par la méthode *compare*, qui permet de comparer les valeurs obtenues par plusieurs sujets ou profils pour une caractéristique donnée. La Figure 3.5 montre un résultat obtenu lors de l'affichage des caractéristiques de diversité linguistique pour le participant numéro 051. La Figure 3.6 permet de comparer les mesures de disfluence entre le groupe de participants témoins et le groupe de participants ayant reçu un diagnostic de trouble cognitif.

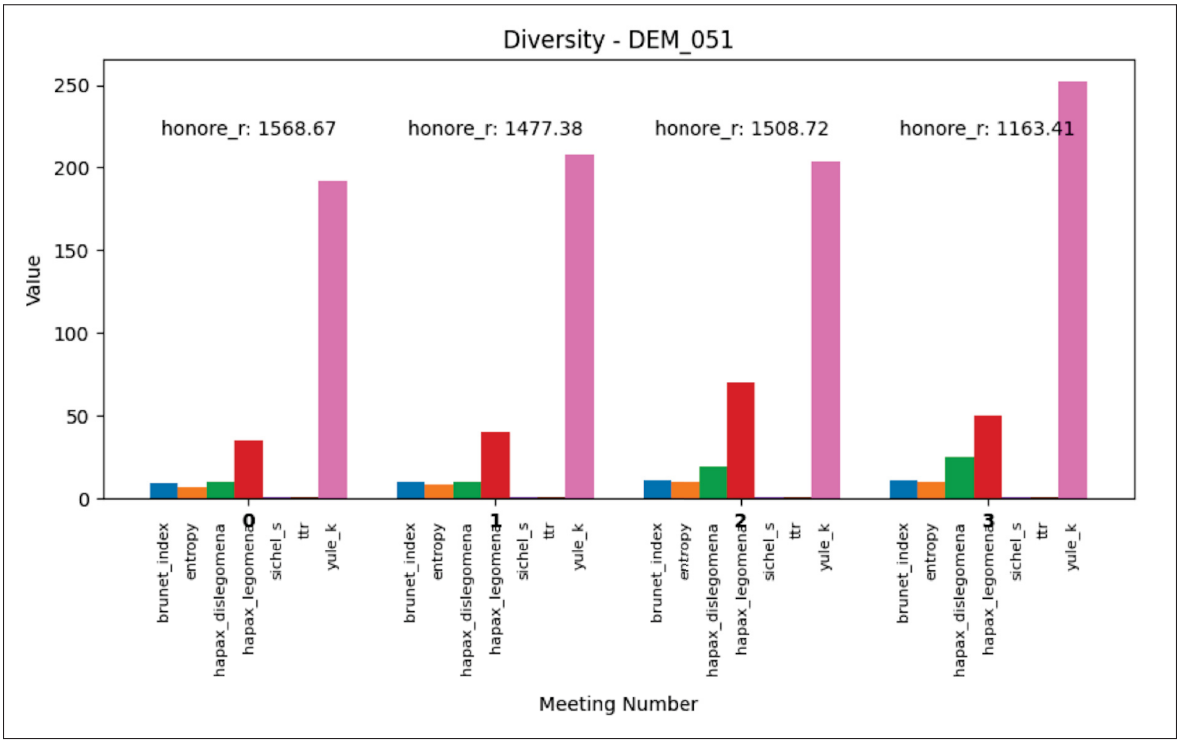

Figure 3.5 Affichage d'UsAge pour la comparaison des caractéristiques de diversité linguistique pour les participants 051 et 010

Ces méthodes de séparation des données facilitent la manipulation du jeu de données pour effectuer des observations basées sur un participant, une caractéristique, une rencontre ou un diagnostic.

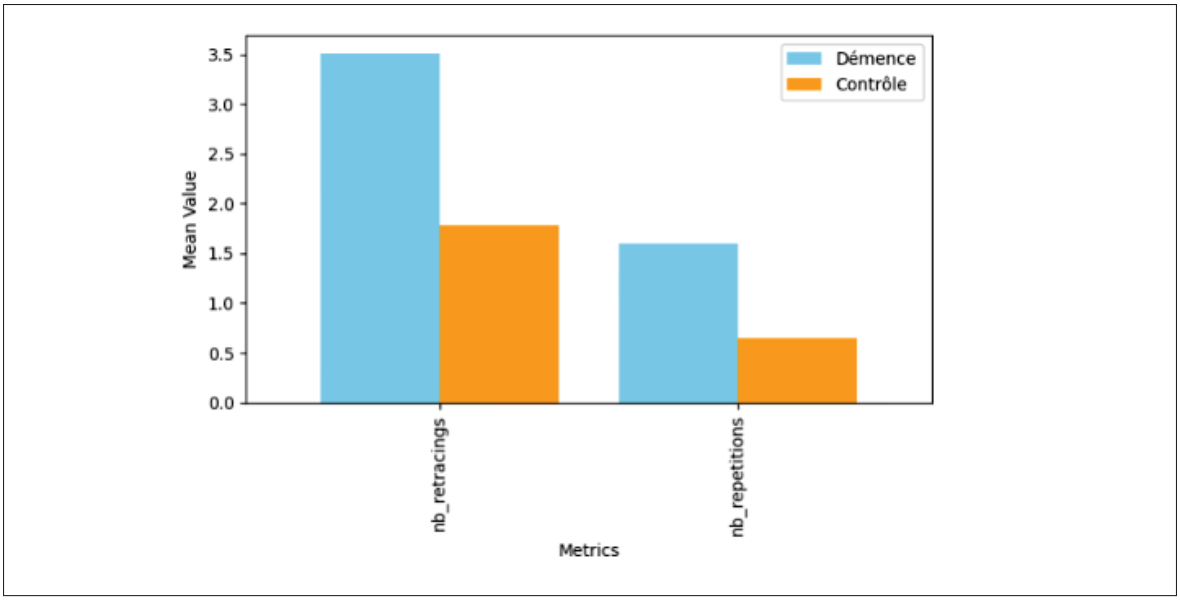

Figure 3.6 Comparaison des caractéristiques de disfluence linguistique pour les groupes démence et contrôle, toutes rencontres combinées

3.3.2 Analyse et sélection des caractéristiques

La compréhension des caractéristiques utilisées dans un modèle est impérative pour sa conception et son interprétation. L'analyse de la corrélation et de la significativité statistique des caractéristiques permet d'identifier des tendances cruciales pour l'interprétation et la découverte de nouvelles variables pertinentes. De plus, la variance et l'échelle des variables influencent fortement les résultats : des caractéristiques non normalisées peuvent dominer d'autres variables et biaiser l'apprentissage du modèle.

La sélection des caractéristiques est essentielle à la performance d'un modèle. Elle prévient à la fois le surapprentissage et le sous-apprentissage, tout en évitant l'utilisation de données bruitées susceptibles de tromper le modèle et de fausser son analyse.

Les sections suivantes décrivent les méthodes intégrées à UsAge qui permettent l'analyse et l'évaluation des caractéristiques.

3.3.2.1 *Show_correlation*

La corrélation des variables avec une étiquette permet d'observer la présence d'une relation linéaire indiquant un lien entre ces dernières. Bien qu'elle ne soit pas un indicateur direct de la performance d'un modèle, elle permet d'observer les interactions dans un jeu de données.

La valeur F et la valeur p issues de l'ANOVA permettent d'évaluer la présence d'une relation statistiquement significative entre les variables étudiées. Le coefficient de corrélation bisériale de point, quant à lui, permet d'indiquer une association entre une valeur continue et une valeur binaire. La méthode *show_correlation* calcule ces valeurs pour des caractéristiques données et produit un graphique combiné qui synthétise ces indicateurs en indiquant clairement les seuils statistiques, à savoir F > 4 et valeur p < 0,05.

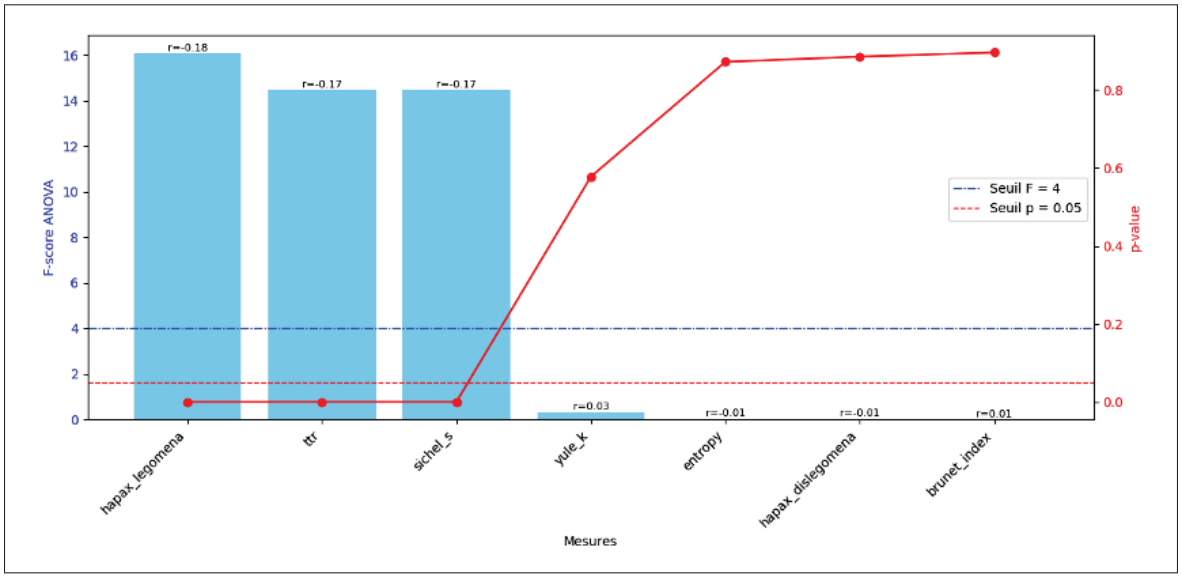

Figure 3.7 Graphique des mesures de corrélation montrant les scores F, la valeur p et le coefficient de corrélation bisériale de point

La Figure 3.7 présente les résultats des différentes mesures de diversité linguistique. On peut voir que le score F des indices *hapax legomena* (16), TTR (Type-Token Ratio) (14) et S de Sichel (14) dépasse le seuil critique de F>4, accompagné d'une valeur p (courbe continue rouge) inférieure au seuil de 0,05. Cela signifie qu'elles sont statistiquement significatives.

En revanche, les autres mesures — K de Yule, entropie, *hapax dislegomena* et index de Brunet — ont des scores F en dessous du seuil et des valeurs p supérieures au seuil, ce qui indique qu'elles ne sont pas statistiquement significatives.

3.3.2.2 *Show_MI*

La mesure d'information mutuelle (MI) permet d'évaluer à quel point une caractéristique apporte de l'information sur la classe cible. Plus le score est proche de zéro, plus la caractéristique est indépendante de l'étiquette et moins elle contribue à la prédiction de cette dernière.

La méthode *show_MI* retourne une liste des caractéristiques ayant un score nul, pouvant être directement utilisées comme paramètres dans un algorithme de classification pour exclure les variables non informatives. Elle affiche également les dix caractéristiques ayant le score d'information mutuelle le plus élevé, mettant en évidence les variables les plus pertinentes vis-à-vis l'étiquette cible. La mesure peut également être visualisée par le graphique de variance retourné par la méthode *show_variance* (voir Figure 3.8).

3.3.2.3 *Show_variance*

La variance des caractéristiques d'un modèle permet d'observer la dispersion des données et d'identifier d'éventuelles tendances ou redondances entre les variables. Il est donc essentiel de pouvoir visualiser ces valeurs de variance afin d'analyser les résultats en profondeur.

L'outil met en place plusieurs méthodes permettant d'extraire des observations concernant la variance du jeu de données. La première méthode, *show_variance*, affiche un tableau récapitulatif du score de variance et du score d'information mutuelle des caractéristiques utilisées, permettant d'identifier rapidement les valeurs aberrantes.

La Figure 3.8 permet d'observer que la caractéristique K de Yule (yule_k) possède une grande variance et un score d'information mutuelle peu élevé, susceptible d'introduire du bruit.

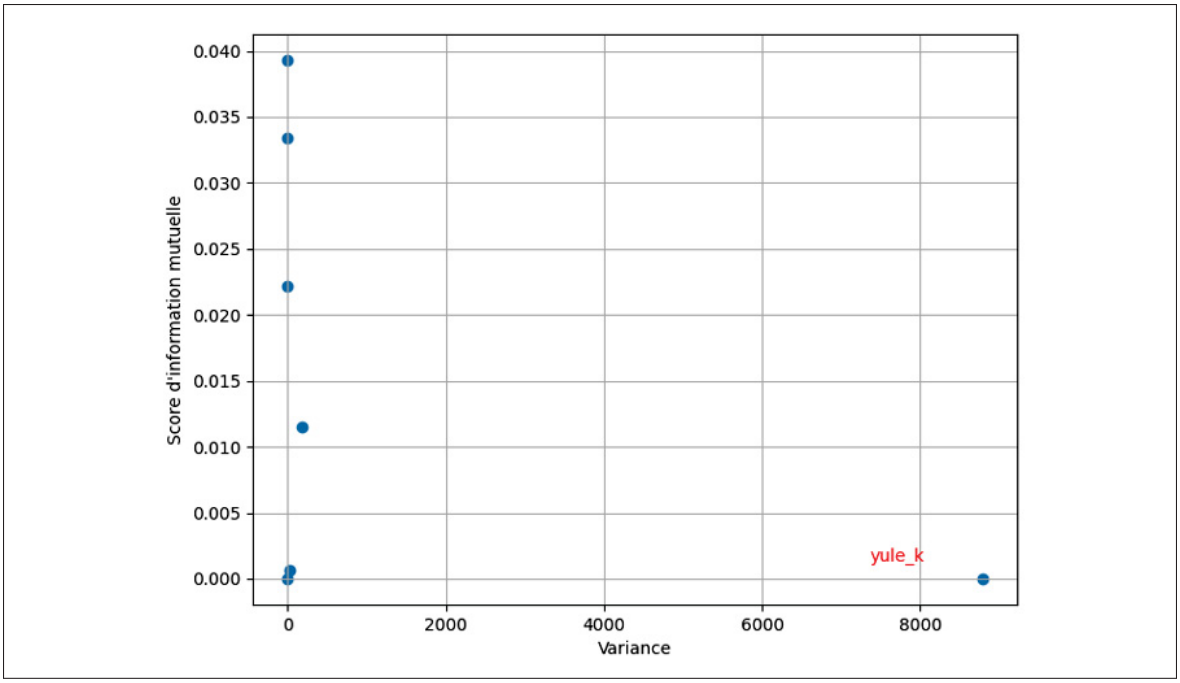

Figure 3.8 Graphique comparatif des scores de variance et d'information mutuelle pour les caractéristiques de diversité linguistique

3.3.2.4 *Show_PCA*

La méthode *show_PCA* permet également d'analyser la variance en étudiant les composantes principales à l'aide d'une analyse principale des composants (PCA), qui capture la plus grande part de variation dans les données. La section 2.2.1 montre un exemple d'utilisation de cette méthode exécutée sur le classificateur de base de l'application. Elle génère tout d'abord un graphique de la frontière de décision (Figure 3.9) permettant de visualiser la classification des données projetées dans un espace réduit à deux dimensions. Elle fournit également une explication de la variance capturée par chacune des deux premières composantes principales ainsi que les influences déterminantes (*loadings*), c'est-à-dire les contributions de chaque caractéristique de ces composantes, exprimées en pourcentage (Figure 3.10) et exposé dans le terminal.

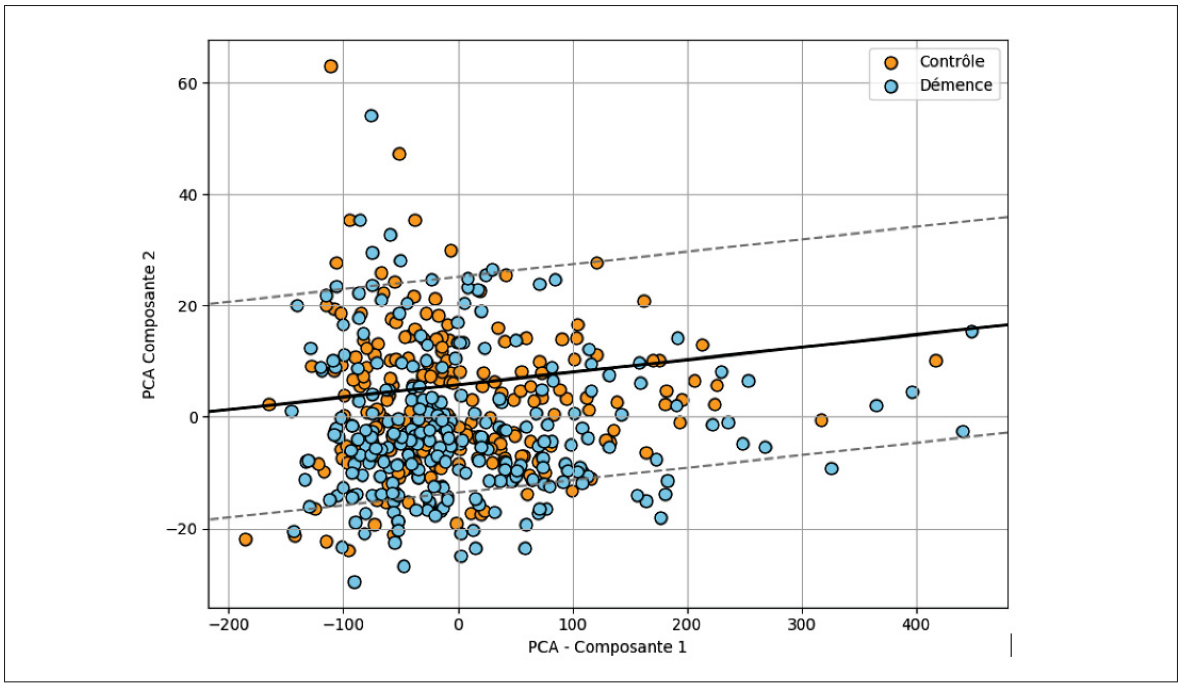

Figure 3.9 Frontière de décision de la méthode show_PCA, utilisée avec les caractéristiques de diversité linguistique

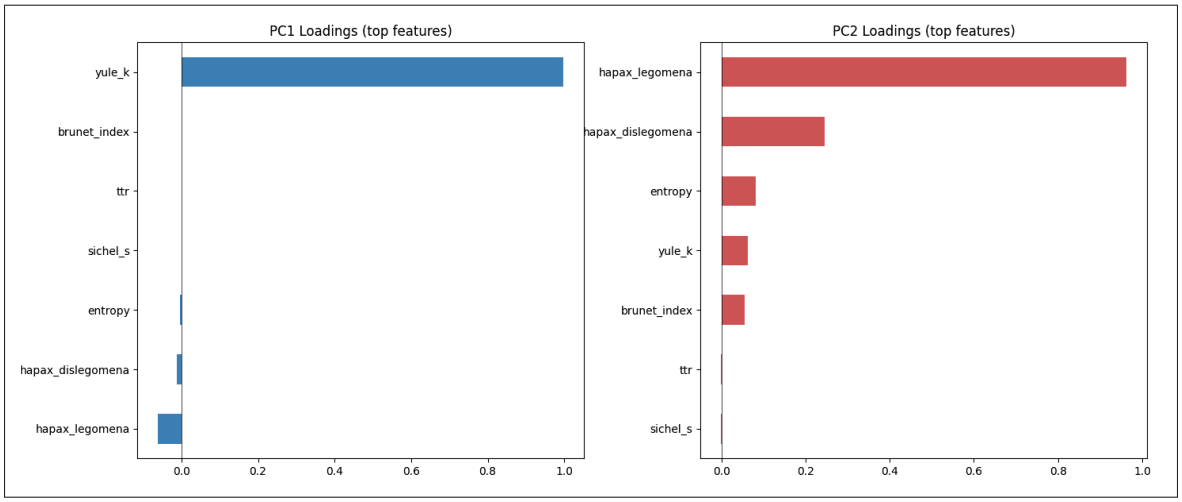

Figure 3.10 Influences déterminantes des deux composants principaux de la PCA

3.3.2.5 *Get_most_relevant*

La méthode *get_most_relevant* intègre *SelectKBest* de Scikit-Learn, un outil de sélection de caractéristiques qui choisit les meilleures variables en fonction d'un test statistique univarié, pour identifier les dix caractéristiques ayant le plus grand poids décisionnel dans un modèle. La méthode retourne un graphique permettant de rapidement visualiser les résultats, ainsi qu'une liste pouvant être utilisée dans les algorithmes de classification pour manipuler les caractéristiques.

Le graphique de la Figure 3.11 permet d'observer les résultats : un coefficient négatif, indiqué en orange, correspond aux caractéristiques favorisant la classification vers le groupe témoin — indiqué par l'étiquette 'contrôle' — tandis qu'un coefficient positif, indiqué en bleu, représente les caractéristiques favorisant la classification vers le groupe ayant reçu un diagnostic de démence.

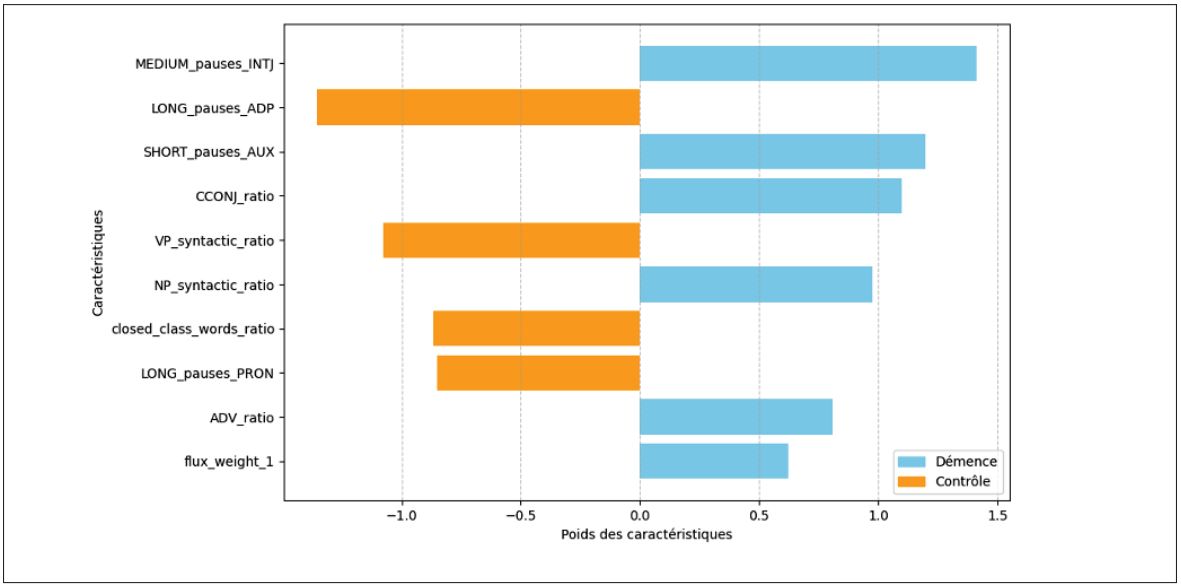

Figure 3.11 Affichage des 10 caractéristiques les plus importantes pour différencier deux groupes

3.3.3 Classification

La section suivante énumère les classificateurs que nous avons intégrés dans l'outil, lesquels peuvent être facilement utilisés avec les jeux de données créés par l'application.

Nous avons implémenté tous les classificateurs en intégrant *StandardScaler*, une méthode de standardisation qui transforme les données pour qu'elles aient une moyenne nulle et un écart-type égal à 1. Cela permet de réduire l'impact des valeurs aberrantes et des différences d'unités entre les variables. Cette standardisation peut être appliquée dynamiquement lors de l'appel au classificateur, assurant ainsi une mise à l'échelle des variables pour améliorer la performance et la stabilité du modèle.

Nous avons également intégré *GridSearch* dans les modèles, une méthode permettant d'identifier les hyperparamètres optimaux d'un algorithme par recherche exhaustive en testant plusieurs combinaisons de valeurs. Cette approche offre aux utilisateurs la possibilité de réoptimiser les classificateurs si nécessaire, en fonction des caractéristiques utilisées.

Tous les classificateurs sont implémentés par défaut avec une validation croisée à 10 blocs retournant la moyenne des résultats obtenus. Ils disposent d'une propriété *scores* qui retourne les valeurs moyennes des métriques de performance, soit le score F1, la précision, le rappel et l'ASC, calculées au cours des différentes itérations.

3.3.3.1 SVM

Cette méthode implémente un classificateur de type *Support Vector Machine* (SVM) utilisant par défaut un noyau linéaire. Le modèle a été optimisé sur les données d'entraînement avec un hyperparamètre C fixé à 0,01. L'outil offre également la possibilité de modifier le noyau du SVM pour utiliser un noyau *Radial Basis Function* (RBF). Les valeurs optimales des hyperparamètres pour le noyau RBF sont $C = 100$ et $\gamma = \text{scale}$.

L'architecture du SVM linéaire s'appuie sur les descriptions fournies par Cesari (2023) et Abiven (2020). Nous l'avons réimplémenté afin de constituer un point de départ pour notre propre

pipeline de classification, puisque le modèle initial n'était pas disponible dans le code source de base.

3.3.3.2 Régression Logistique

Nous avons ajouté un classificateur de régression logistique à notre ensemble d'algorithmes de classification afin d'explorer et de comparer les performances d'un autre modèle linéaire sur les jeux de données. Cette inclusion permet d'établir une référence supplémentaire pour évaluer l'efficacité des techniques de classification linéaire. Le modèle a été optimisé avec un paramètre de régularisation C=0,01.

3.3.3.3 XGBOOST

L'algorithme XGBoost a été intégré pour sa capacité à modéliser efficacement des relations complexes. Dans sa forme canonique, l'algorithme utilise des arbres de décision comme estimateurs de base et combine ces modèles de manière itérative pour améliorer ses performances. XGBoost peut également être configuré pour utiliser un amplificateur linéaire, où les arbres de décision sont remplacés par des régressions linéaires servant de modèles de base. L'algorithme a été optimisé dans sa forme canonique avec les hyperparamètres suivants : mesure d'évaluation qui utilise *logloss*, taux d'apprentissage fixé à 0,2, profondeur maximale de 3, nombre d'estimateurs de 200 et taux de sous-échantillonnage fixé à 1,0.

3.3.3.4 KNN

Cette méthode implémente un algorithme *K-Nearest Neighbors* (KNN) permettant d'identifier des regroupements au sein des jeux de données. Bien que ses performances ne soient pas toujours optimales pour la classification linéaire, cet algorithme peut s'avérer utile pour détecter d'éventuels sous-groupes dans les données. L'algorithme a également été optimisé avec les hyperparamètres suivants : métrique de distance euclidienne, nombre de voisins fixé à 5 et poids uniformes.

3.4 Conclusion

La création de la bibliothèque *UsAge Feature Extraction* permet d'encapsuler la complexité liée à l'extraction des caractéristiques tout en mettant en place une architecture modulaire capable d'activer uniquement les composants requis pour une tâche donnée. Cette approche allège considérablement le processus d'exécution, améliore les performances globales et garantit une gestion rigoureuse des dépendances nécessaires. De plus, l'intégration d'une suite de tests automatisés assure la fiabilité et la pérennité des résultats produits.

La nouvelle structure modulaire d'UsAge corrige les limitations structurelles de la version précédente et permet une utilisation fluide de l'outil. Elle recentre son usage sur des opérations ciblées et implémente des méthodes permettant la manipulation, l'observation et la visualisation des données, en plus de mettre à la disposition de l'utilisateur plusieurs algorithmes de classification.

Dans la suite de cet ouvrage, le terme UsAge fera référence à la combinaison des deux outils, dans la mesure où UsAge intègre UFE de manière transparente.

CHAPITRE 4

ÉVALUATION DE NOUVELLES CARACTÉRISTIQUES DISCRIMINANTES

4.1 Introduction

De nouvelles caractéristiques linguistiques ont été étudiées afin de tenter de combler certaines lacunes identifiées dans la littérature et d'améliorer les performances de détection de notre outil. Trois grands axes peu explorés ont été ciblés : l'analyse des chaînes de coréférence, la prise en compte de la disposition des pauses dans le discours et la complexité syntaxique. Ces dimensions linguistiques sont soit prises en compte dans des analyses statistiques et proviennent d'annotations manuelles, limitant la reproductibilité et l'extensibilité des analyses, soit analysées de manière superficielle, ne reflétant pas la complexité des informations susceptibles d'être extraites.

4.2 Chaînes de coréférence

Les chaînes de coréférence sont obtenues par l'entremise d'un outil de TALN qui extrait d'un échantillon de texte une structure contenant un référent, lié à tous ses coréférents. La structure est ensuite intégrée à UsAge afin d'extraire les mesures ajoutées. Cette architecture modulaire permet d'isoler le modèle d'extraction des coréférents de la logique des mesures, afin que ce dernier puisse être remplacé sans difficulté, que ce soit pour manipuler les données ou pour améliorer le pipeline. Les mesures extraites sont identifiées dans le Tableau 4.1.

4.2.1 Performances des outils automatiques

Nous avons comparé les chaînes de référence extraites par les outils Stanza et SpaCy aux chaînes de référence extraites manuellement. Les référents ayant une longueur de 1 ont été exclus de tous les calculs, car ils correspondent à des référents sans coréférence, ce qui fausserait l'évaluation de la similarité de la résolution.

Tableau 4.1 Mesures des chaînes de coréférence ajoutées

Mesure	Description
Nombre de référents manquants	Coréférents sans référent identifié
Nombre d'anaphores	Coréférents présents après le référent
Nombre de cataphores	Coréférents présents avant le référent
Longueur moyenne des chaînes	Moyenne des longueurs de chaînes par référent
Ratio de coréférents	Nombre moyen de chaînes par phrase
Densité des coréférents	Nombre de mots faisant partie d'une chaîne de référence divisé par le nombre total de mots
Distance moyenne	Moyenne des distances référent–coréférent calculée en mots
Distribution des longueurs	Répartition des chaînes par longueur

Tableau 4.2 Comparaison des résultats d'annotation automatique de coréférents

Métrique	Stanza	SpaCy
Ratios clés communes	0,424	0,492
Ratios correspondances exactes	0,318	0,388
Scores de similarité	0,243	0,248

Les résultats obtenus (Tableau 4.2) montrent que l'extraction automatique des coréférents ne parvient pas à approcher de manière satisfaisante les résultats obtenus manuellement. Parmi les deux modèles, celui de SpaCy présente une performance légèrement supérieure à celle proposée par Stanza. En effet, SpaCy est capable d'identifier près de la moitié des référents (0,492) présents dans le jeu de données de référence, et affiche un score de 0,388 pour les correspondances exactes de référent et de longueur de chaîne.

Les outils commettent plusieurs erreurs similaires, notamment en considérant comme coréférents deux mots identiques. Par exemple, si le mot « *cookie* » apparaît plusieurs fois dans un document, les outils considéreront parfois que ces occurrences sont coréférentes. De plus, ces derniers semblent confondre certains pronoms, en reliant « *you* » et « *I* » comme coréférents.

4.2.2 Pertinence des chaînes de coréférences

Nous avons utilisé les annotations manuelles afin d'examiner plus en profondeur les impacts des chaînes de référence sur la classification de base. Puisque les annotations ne permettaient d'observer que la longueur des chaînes et le nombre de leurs occurrences, seule la mesure de la longueur moyenne des coréférents sera étudiée en détail.

Puisque seulement 50% du corpus a été annoté manuellement, un nouveau standard a été constitué en ne conservant que les fichiers pour lesquels une annotation manuelle était disponible. Parmi ces fichiers, 132 proviennent de patients atteints de démence et 135 de participants témoins. Les données ont été soumises au modèle de classification de base (section 2.2.1) sans inclure la mesure de la longueur moyenne des coréférents, puis soumises de nouveau en l'incluant afin de mesurer son impact. Les résultats sont présentés dans le Tableau 4.3.

Tableau 4.3 Comparaison de la classification de base vs la classification avec mesures de coréférences

Modèle	F1	Précision	Rappel	ASC
Classification avec mesure de longueur moyenne de coréférents	0,854	0,837	0,871	0,912
Classification de base modifiée	0,796	0,787	0,809	0,813

Le modèle entraîné avec la mesure manuelle de longueur moyenne de coréférents atteint de meilleures performances que la classification faite avec le modèle de base modifié. Les résultats montrent une amélioration de 5,8% du score de F1, de 5% de la précision, de 6,2% du rappel et de 9,9% de l'ASC. La mesure présente également un score F de 4,67 au test ANOVA avec une valeur p de 0,03 et un coefficient de corrélation bisériale de point de -0,097, indiquant que la différence observée entre les deux groupes est statistiquement significative, bien que la corrélation soit faible.

Les résultats suggèrent que les patients atteints de démence tendent à produire des chaînes de coréférence plus courtes que les participants témoins et que cette mesure mérite d'être intégrée à une analyse plus large des variables linguistiques liées au déclin cognitif. Toutefois, les outils

automatiques de traitement du langage naturel utilisés dans cette étude ne parviennent pas à capter ces dépendances avec une précision suffisante pour permettre une utilisation fiable dans des systèmes automatisés.

4.3 Pauses

L'emplacement des pauses dans le discours pourrait constituer un indicateur pertinent de la complexité langagière en permettant d'identifier où et quand ont lieu des moments d'hésitation. Nous avons donc intégré à notre pipeline un module permettant d'analyser le type de mot précédé par chaque type de pause détecté, soit les pauses courtes, moyennes et longues, ainsi que les pauses ayant lieu en début de phrase, marquant ainsi une hésitation avant la prise de parole.

Ce module génère une structure contenant le dénombrement de chaque étiquette de partie du discours, qui constitue la base des mesures que nous allons analyser et manipuler dans le but d'améliorer la performance du modèle de classification. L'étiqueteur morphosyntaxique de l'outil Stanza, basé sur les étiquettes universelles (*Universal POS tags*), a été utilisé pour effectuer cette tâche.

4.3.1 Pertinence des pauses

Nous avons évalué la signification statistique des pauses en calculant le score F, la valeur p et le coefficient de corrélation bisériale de point pour chaque mesure observée, en les distinguant selon les catégories de pauses (courte, moyenne ou longue).

La Figure 4.1 met en évidence que le nombre de pauses courtes en début de phrase constitue une observation statistiquement significative. Le nombre de pauses précédant une conjonction de coordination (CCONJ) ne franchit pas le seuil de signification, mais sa proximité avec ce seuil justifie son inclusion dans les analyses ultérieures.

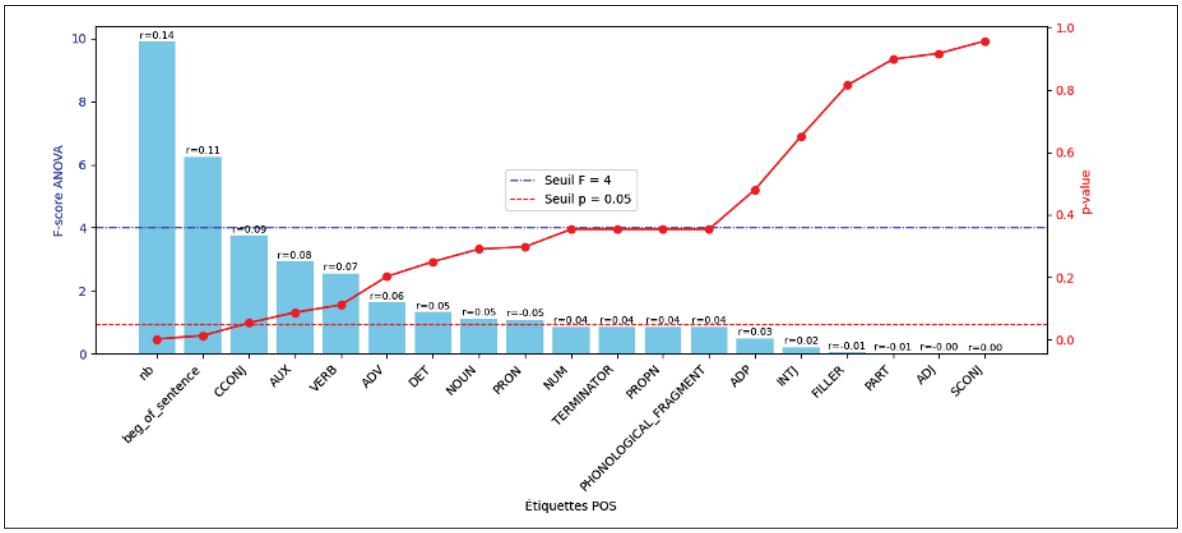

Figure 4.1 Visualisation des scores F, valeur p et de la corrélation bisériale pour les mesures de pauses courtes

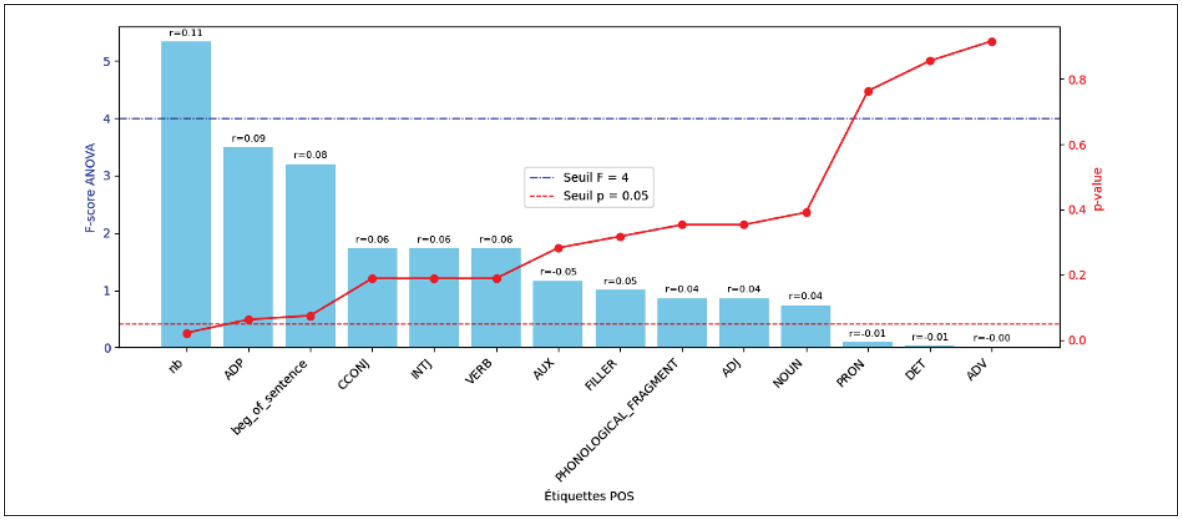

Figure 4.2 Visualisation des scores F, valeur p et de la corrélation bisériale pour les mesures de pauses moyennes

La Figure 4.2 montre que la mesure de pauses de longueur moyenne précédant une adposition (ADP) ne dépasse pas le seuil de signification, mais sa proximité avec celui-ci justifie un examen plus approfondi.

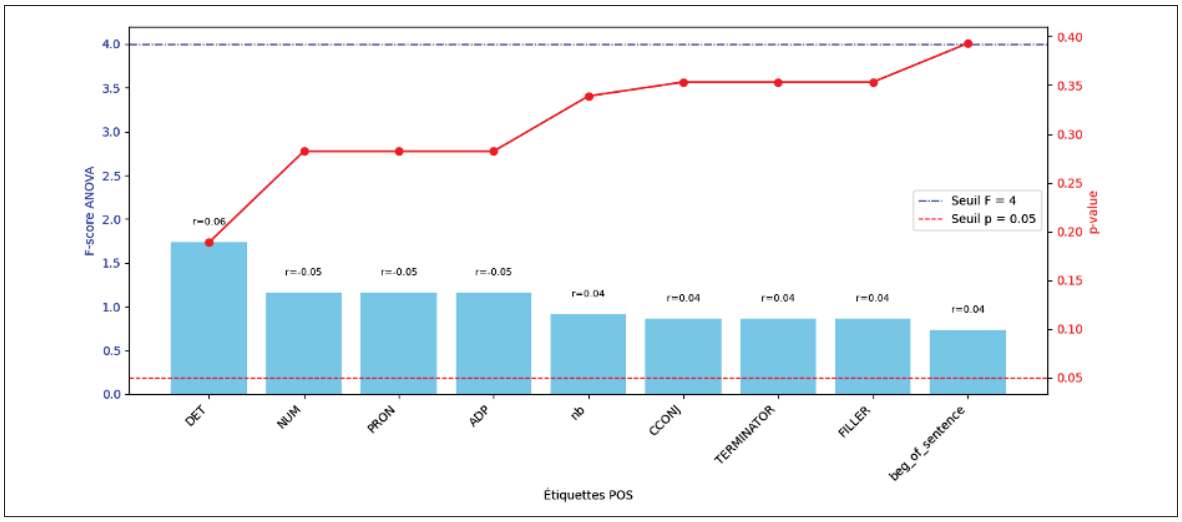

Figure 4.3 Visualisation des scores F, valeur p et de la corrélation bisériale pour les mesures de pauses longues

Le graphique des mesures de pauses longues (Figure 4.3) ne semble indiquer aucune importance particulière parmi les mesures observées. Il convient toutefois de souligner que les occurrences de pauses longues sont extrêmement faibles.

Nous avons ensuite procédé à une classification en utilisant l'ensemble des mesures de pauses observées (Tableau 4.5). Le modèle a affiché des performances inférieures sur l'ensemble des mesures. Plutôt que de réaliser des tests d'ablation, nous avons réalisé des tests d'attribution de caractéristiques additives. Le Tableau 4.4 présente les résultats obtenus. Les calculs du nombre de pauses de chaque catégorie n'ont pas été inclus, car ils sont déjà intégrés aux caractéristiques de l'algorithme de base.

Tableau 4.4 Résultats des tests d'attribution de caractéristiques

Modèle	**F1**	**Précision**	**Rappel**	**ASC**
Avec pauses courtes avant CCONJ	0,773	0,789	0,766	0,782
Avec pauses courtes en début de phrase	0,768	0,784	0,784	0,763
Classification de base	0,759	0,811	0,723	0,848
Avec pauses moyennes avant ADP	0,753	0,776	0,741	0,828

Le résultat des différentes classifications nous permet d'observer un effet positif des mesures de pauses courtes devant une conjonction (CCONJ) et de pauses courtes en début de phrase sur la performance du score F1 et du rappel de la classification.

Nous avons donc poursuivi l'expérimentation avec de nouvelles mesures combinées, soit le nombre moyen de pauses en début de phrase par catégorie, le nombre moyen de pauses en début de phrase toutes catégories confondues et le nombre moyen de pauses tous types confondus. Les résultats de ces classifications ainsi que de la classification faite avec toutes les mesures de pauses sont présents dans le Tableau 4.5.

Tableau 4.5 Comparaison des modèles de classification : différentes combinaisons de mesures de pauses

Modèle	F1	Précision	Rappel	ASC
Classification de base	0,759	0,811	0,723	0,848
Moyenne des pauses devant chaque POS, toutes catégories confondues	0,753	0,768	0,747	0,828
Moyenne des pauses en début de phrase pour chaque catégorie	0,753	0,780	0,738	0,827
Moyenne des pauses en début de phrase pour toutes catégories confondues	0.750	0,776	0,732	0,828
Moyenne des pauses toutes catégories confondues	0,746	0,769	0,732	0,828
Toutes les mesures de pauses	0,745	0,769	0,729	0,830

Dans l'ensemble, les modèles utilisant les mesures combinées présentent des performances inférieures au modèle de base pour le score F1 et la précision, mais augmentent légèrement le score de rappel. La distribution des pauses dans notre jeu de données présente un certain déséquilibre et résulte d'une quantification subjective des annotateurs (Tableau 4.6). Il serait pertinent de poursuivre ces observations sur un volume de données plus important et permettant l'observation de plus de pauses de chaque type.

Tableau 4.6 Distribution des types de pauses

Groupe	Pauses courtes	Pauses moyennes	Pauses longues
Démence	272	100	25
Contrôle	153	55	15

4.4 Complexité syntaxique

La complexité syntaxique est évaluée à l'aide de trois types de mesures : la densité syntaxique et le ratio de syntagmes, tous deux calculés à partir des étiquettes morphosyntaxiques extraites avec Stanza, ainsi que des mesures liées aux poids des dépendances transphrastiques.

4.4.1 Densité syntaxique

Cette mesure permet d'analyser récursivement le nombre de mots par syntagme dans le but d'identifier les schémas d'imbrication syntaxique, et donc de densité, permettant possiblement d'apercevoir une différence entre les groupes.

Les étiquetages syntaxiques employés reposent sur la typologie du *Penn Treebank* (Penn Treebank Project, 2003). 12 catégories distinctes de syntagmes ont été identifiées à l'aide de l'analyseur de constituants de Stanza et ont été soumises à une analyse statistique afin d'examiner leur corrélation avec les différents groupes étudiés.

La Figure 4.4 met en évidence que la densité des syntagmes de type fragment (FRAG) — phrases incomplètes (e.g. : « Peut-être. », « Oui, demain. ») — verbal (VP), nominal (NP) et adjectival (ADJP) est significativement liée à la distinction entre les groupes, leurs scores F étant supérieurs à 4 et leurs valeurs p inférieures ou égales à 0,05. Cette association est confirmée par le coefficient de corrélation bisériale de point, qui indique cependant une différence faible entre les groupes. Le coefficient de corrélation bisériale de point révèle également que le groupe de patients présente des valeurs plus élevées pour les mesures FRAG et VP, tandis que les mesures NP et ADJP sont plus élevées dans le groupe témoin.

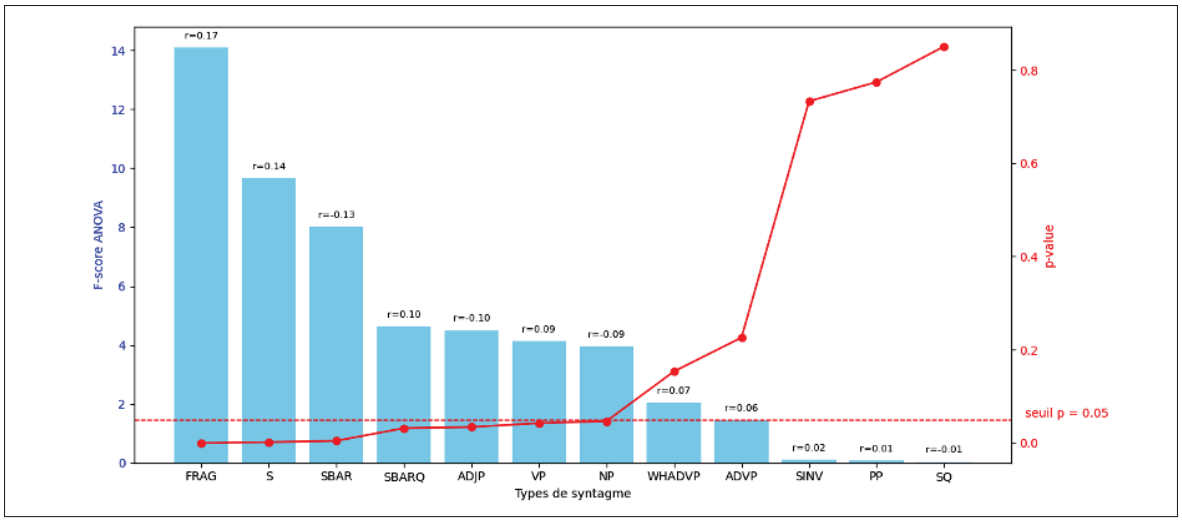

Figure 4.4 Visualisation des scores F, valeur p et de la corrélation bisériale pour les mesures de densité syntaxique

Par conséquent, nous avons ajouté les quatre mesures statistiquement significatives au modèle de classification afin de comparer ses performances avec celles du modèle de base ainsi que celles du modèle de base enrichi de toutes les mesures de densité. Les résultats sont présentés dans le Tableau 4.7. L'ajout de toutes les mesures de densité permet une amélioration du score de F1 et de rappel, mais les scores de précision et d'ASC sont plus bas. L'ajout des mesures significatives améliore le score de F1, la précision et le rappel, mais le score d'ASC est plus bas. Il serait intéressant de tester cette observation sur un autre jeu de données, ou sur un jeu de données plus volumineux, afin de déterminer si les quatre mesures significatives identifiées sont spécifiques à notre jeu de données — risquant ainsi un surapprentissage — ou si elles constituent une observation généralisable.

Tableau 4.7 Comparaison des classifications : Modèle de base vs observations de la densité syntaxique

Modèle	**F1**	**Précision**	**Rappel**	**ASC**
Mesure de densités significatives	0,783	0,813	0,766	0,831
Toutes les mesures de densités	0,766	0,783	0,759	0,824
Classification de base	0,759	0,811	0,723	0,848

4.4.2 Ratio de syntagme

Le Tableau 4.8 présente les résultats des analyses statistiques pour chaque mesure de ratio de syntagme et expose la liste des syntagmes observés.

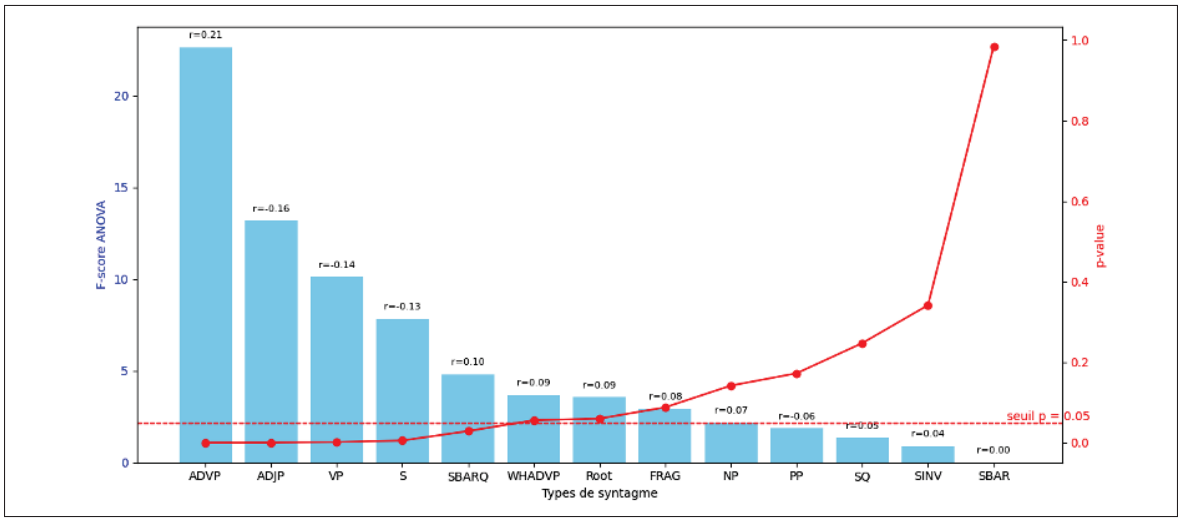

Figure 4.5 Visualisation des scores F et valeur p et de la corrélation bisériale pour les mesures de ratio syntaxique

Les scores obtenus indiquent que les ratios de syntagmes ADVP, ADJP, VP, S et SBARQ constituent des caractéristiques statistiquement significatives avec les étiquettes de groupes. La corrélation bisériale de point positive pour les syntagmes ADVP et SBARQ suggère que le groupe témoin tend à produire davantage de ces structures, tandis que les patients présentent une fréquence réduite de syntagmes ADJP, VP et S, ce qui pourrait refléter une simplification syntaxique du discours.

Les ratios de syntagmes WHADVP, Root et FRAG ne dépassent pas les seuils de signification statistique ($F > 4$, valeur $p < 0{,}05$), mais leur proximité avec ces derniers suggère qu'ils méritent d'être pris en considération dans le cadre de la classification. Leur potentiel informatif pourrait se révéler pertinent lorsque combiné à d'autres caractéristiques.

Nous avons par la suite effectué une comparaison de la performance du classificateur en utilisant toutes les mesures de ratio syntaxique, en utilisant seulement les quatre mesures montrant une

corrélation et en combinant les mesures ayant une corrélation et les mesures ayant des scores proches du seuil significatif. Les résultats sont présentés dans le Tableau 4.8.

Tableau 4.8 Comparaison des modèles de classification : Modèle de base vs modèles avec différentes observations du ratio des syntagmes

Modèle	F1	Précision	Rappel	ASC
Classification avec ADVP, ADJP, VP, S et SBARQ	0,779	0,801	0,771	0,829
Classification avec ADVP, ADJP, VP, S, SBARQ, NP, FRAG, Root, WHADVP	0,775	0,795	0,767	0,828
Classification avec toutes les mesures	0,766	0,783	0,759	0,824
Classification de base	0,759	0,811	0,723	0,848

Les trois modèles créés performent mieux que le modèle de base pour les mesures de F1 et de rappel. Les résultats obtenus pour la précision et l'ASC sont cependant inférieurs aux résultats du modèle de base. L'utilisation des ratios syntaxiques de ADVP, ADJP, VP, S et SBARQ permet d'atteindre la meilleure classification en termes de score F1.

4.4.3 Flux de dépendance

Nous avons intégré au pipeline un module dédié à l'extraction du poids des flux de dépendances, incluant le calcul du poids maximal détecté ainsi que la distribution complète de ces poids. Le poids d'un flux de dépendance mesure « les encastrements centraux et la construction imbriquée » [traduction libre] (Kahane, Yann & Botalla, 2017, p.73). Il se calcule en identifiant, pour un intervalle donné entre deux mots, le plus grand sous-ensemble de dépendances disjointes — c'est-à-dire le nombre maximal de dépendances qui ne partagent ni tête ni dépendant — qui traversent cet espace intermot.

La mesure de poids moyen correspond à la moyenne des cardinalités des plus grands sous-ensembles de dépendances disjointes pour chaque intervalle considéré, tandis que le poids maximal désigne la valeur la plus élevée de ces poids parmi tous les intervalles analysés. Dans la Figure 4.6, le poids moyen est de 1,15, tandis que le poids maximal atteint une valeur de 2. Concernant la distribution, l'exemple contient onze poids de longueur 1 et 2 poids de longueur 2.

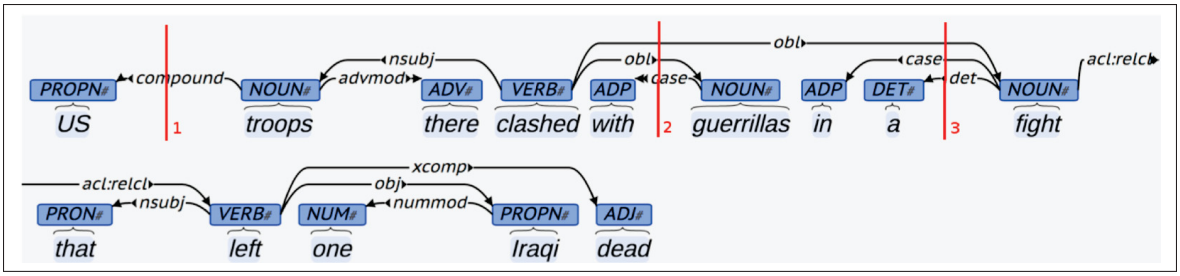

Figure 4.6 «A UD dependency tree with three inter-word positions marked», provenant de 'What are the limitations on the flux of syntactic dependencies ? Evidence from UD treebanks' (Kahane, Yann & Botalla, 2017)

Tableau 4.9 Résultats des analyses du score F, de la valeur p et du coefficient de corrélation bisériale du flux de dépendance

Mesure	Valeur F	Valeur p	Corrélation bisériale
Moyenne du poids	0,22	0,64	0,02
Maximum du poids	1,79	0,182	-0,05

Les mesures du flux de dépendance ne sont pas significatives pour distinguer les deux groupes, avec un score F inférieur à 4, une valeur p supérieure à 0,05 et un coefficient de corrélation bisériale de point très faible (voir Tableau 4.9). De plus, la distribution des valeurs de flux de dépendance dans notre jeu de données se limite essentiellement à deux modalités : un poids de 1 ou de 2.

Kahane note que le flux de dépendance « est inférieur à 3 dans 99,62% des positions intermots et est borné à 6, ce qui pourrait s'expliquer par des limitations de la mémoire à court terme » [traduction libre] (Kahane, Yann & Botalla, 2017, p.73). Dans notre cas, les résultats semblent bornés à 2. Cela pourrait s'expliquer par la nature de la tâche suscitant la parole chez les participants, qui limite les échanges relatifs aux éléments visuels et induit un discours principalement descriptif.

Puisqu'aucune des mesures de poids extraites ne présente de signification statistique et que leurs valeurs sont globalement similaires entre les échantillons, nous n'avons pas jugé pertinent de les intégrer au modèle de classification. En l'absence de variance discriminante ou de corrélation

significative avec les classes cibles, leur inclusion risquerait de dégrader les performances du classificateur en introduisant du bruit inutile.

Il serait pertinent de reproduire cet exercice avec des transcriptions susceptibles de générer une plus grande amplitude de poids de dépendances syntaxiques, telles que des discours libres. Cela permettrait d'observer plus finement les limites de la performance cognitive du locuteur à travers la complexité structurelle des phrases produites.

4.5 Conclusion

Dans cette section, nous avons étudié l'impact des mesures des chaînes de coréférence sur le modèle de classification et avons constaté que la longueur moyenne des chaînes de coréférence permet d'améliorer ses performances. Cependant, les outils testés pour l'extraction automatique, soit Stanza et SpaCy, se sont révélés insuffisants, ne détectant respectivement que 31,8% et 38,8% des chaînes identifiées manuellement. Il serait pertinent, dans le cadre de recherches futures, de calculer manuellement les mesures de coréférence extraites par le pipeline afin d'évaluer leur efficacité réelle.

Nous avons également analysé des mesures de distribution des pauses en fonction de leur durée — courte, moyenne ou longue —, ce qui a révélé un impact statistiquement significatif du nombre de pauses courtes en début de phrase. L'absence d'observations significatives pour les autres catégories de pauses — moyennes et longues — soulève des interrogations quant à la qualité des mesures de pauses présentes dans les données, d'autant plus que les pauses de courte durée y sont largement surreprésentées en comparaison.

Enfin, l'analyse des mesures de complexité syntaxique a révélé un effet positif des mesures de densité des syntagmes de types FRAG, VP, NP et ADJP sur toutes les mesures à l'exception de l'ASC. Les mesures significatives de ratio syntaxique, soit ADVP, ADJP, VP, S et SBARQ, ont également un effet positif sur la précision, le rappel et le score de F1. Les mesures de flux de dépendance n'ont quant à elles pas permis d'identifier de schémas significatifs. En effet, les données ne présentaient que deux valeurs distinctes, limitant toute conclusion. Cela incite

à s'interroger sur l'impact potentiel de cette mesure dans un corpus plus complexe, tel qu'un corpus de parole spontanée, qui, contrairement à nos transcriptions de descriptions du test du *Cookie Theft*, ne repose pas sur l'observation d'une image fixe.

CHAPITRE 5

SÉLECTION DES DONNÉES ET AMÉLIORATION DE LA CLASSIFICATION

5.1 Introduction

Dans ce chapitre, nous exploitons l'outil UsAge afin d'analyser les mesures introduites aux chapitres précédents. Nous observerons les tendances structurelles que ces mesures révèlent afin d'effectuer une sélection de caractéristiques et ne conserver que celles présentant une réelle pertinence pour la classification. Nous utiliserons par la suite les méthodes de classification implémentées dans l'outil afin d'optimiser notre modèle.

5.2 Tests d'ablation

Dans le chapitre 4, nous avons observé que l'ajout de certaines mesures liées aux pauses, à la densité syntaxique et aux ratios syntaxiques contribuait à améliorer les performances de classification du modèle. Afin de mesurer l'impact combiné de ces types de mesures, nous avons effectué des tests d'ablation, en retirant successivement chaque groupe de caractéristiques. Les résultats sont présentés dans le Tableau 5.1.

Les tests d'ablation ont révélé que l'ajout de l'ensemble des mesures n'était pas la solution optimale, car cela ne permet pas d'atteindre les meilleures performances de classification. Les mesures contribuant le plus efficacement au modèle sont les ratios syntaxiques associés à ADJP, SBARQ et ADVP. À l'inverse, certaines mesures — telles que les pauses devant CCONJ, le ratio syntaxique VP ainsi que les pauses en début de phrase — semblent introduire du bruit lorsqu'elles sont utilisées conjointement.

Puisque la mesure de pause courte devant CCONJ est la seule à ne présenter aucune corrélation significative et qu'elle s'avère nuisible à la performance du modèle, nous l'avons exclue des caractéristiques observées.

Tableau 5.1 Tests d'ablation des caractéristiques significatives par ordre de score F1

Modèle	F1	Précision	Rappel	ASC
Sans CCONJ	0,779	0,822	0,751	0,840
Sans ratio syntaxique VP	0,774	0,823	0,745	0,837
Sans pause en début de phrase	0,770	0,816	0,746	0,841
Avec toutes les mesures significatives	0,766	0,810	0,740	0,830
Sans densité moyenne FRAG	0,766	0,817	0,733	0,839
Sans ratio syntaxique S	0,765	0,810	0,736	0,842
Sans densité moyenne VP	0,763	0,812	0,733	0,839
Sans densité moyenne NP	0,762	0,811	0,732	0,836
Sans ratio syntaxique ADVP	0,761	0,810	0,729	0,839
Sans densité moyenne ADJP	0,761	0,813	0,729	0,840
Sans ratio syntaxique SBARQ	0,759	0,810	0,729	0,839
Classification de base	0,759	0,811	0,723	0,848
Sans ratio syntaxique ADJP	0,756	0,799	0,727	0,834

5.3 Sélection des caractéristiques

Nous avons utilisé la méthode *show_MI* pour observer les valeurs d'information mutuelle des caractéristiques, donc celles qui contiendraient de l'information sur l'étiquette, indépendamment des autres mesures. Dans un premier temps, nous pouvons constater que les valeurs sont généralement basses (voir Tableau 5.2). Cela suggère qu'aucune caractéristique ne se démarque clairement dans la distinction entre les deux groupes et que notre sélection actuelle de variables pourrait être insuffisante. Nous pouvons également remarquer que les mesures ayant les dix valeurs les plus élevées appartiennent aux catégories lexicale et syntaxique.

La méthode d'information mutuelle a également retourné une liste de caractéristiques ayant obtenu un score nul, indiquant une absence d'utilité informative pour la tâche de classification. La liste suivante recense l'ensemble de ces caractéristiques. Celles-ci ont été retirées, puis le modèle a été entraîné une seconde fois. Les résultats obtenus sont présentés dans le Tableau 5.3

Tableau 5.2 Les 10 caractéristiques ayant le plus haut score d'information mutuelle

Caractéristique	Score d'information mutuelle
ADV_ratio	0,0793
words_per_minute	0,0709
NOUN_ratio	0,0674
ADV_freq	0,0595
nb_incomplete_sentences	0,0591
total_pauses_nb	0,0552
nb_interjections	0,0539
agrammatic_utterances	0,0506
VP_syntactic_ratio	0,0503
mean_syntactic_density_FRAG	0,0498

Liste des caractéristiques retirées :

- ADJ_ratio
- DET_freq
- DET_ratio
- NUM_freq
- PART_freq
- PART_ratio
- MEDIUM_pauses_nb
- SHORT_pauses_nb
- disfluency_errors
- entropy
- false_start
- mean_length_utterance
- morphological_errors
- neologism
- syllable_pauses
- yule_k

Tableau 5.3 Impact des mesures ayant un score d'information mutuelle nul sur la classification

Modèle	F1	Précision	Rappel	ASC
Sans mesures ayant un score d'information mutuelle nul	0,778	0,817	0,756	0,848
Meilleur modèle à date	0,774	0,823	0,745	0,837

Ces résultats confirment que les mesures ayant un score nul n'avaient effectivement aucun impact sur le modèle. Une légère amélioration du score de F1, du rappel et de l'ASC suggère que ces mesures pouvaient introduire un certain bruit. Bien que modeste, cette observation souligne l'importance cruciale de la sélection des caractéristiques.

Nous avons ensuite utilisé la méthode *Get_Most_Relevant* pour analyser le poids des caractéristiques dans le modèle.

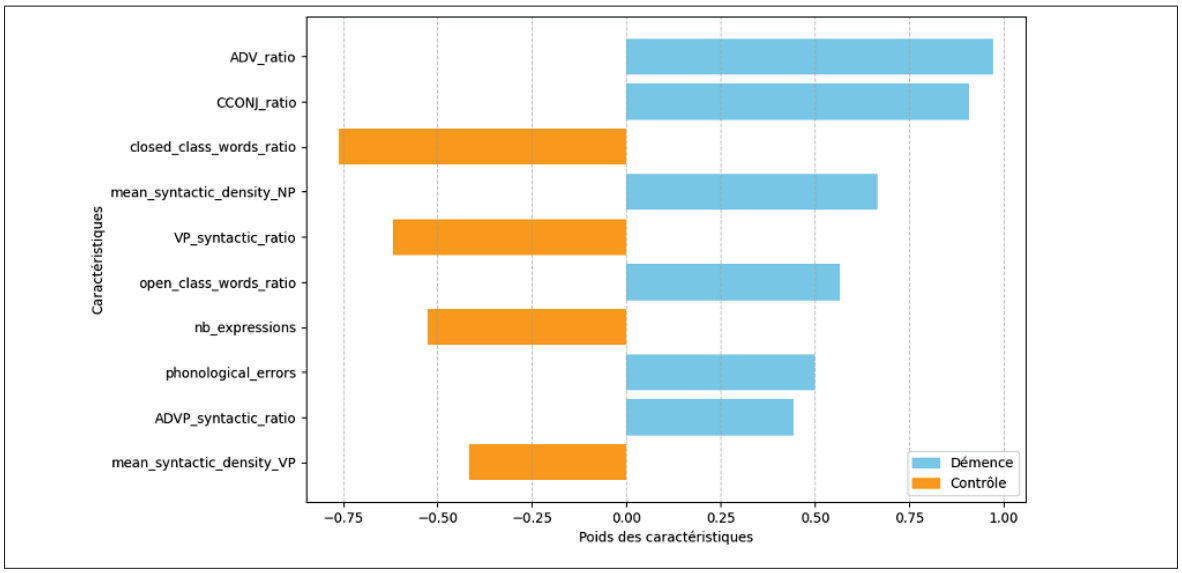

Figure 5.1 Dix caractéristiques les plus pertinentes dans la classification du modèle

La Figure 5.1 montre les mesures présentant le poids décisionnel le plus élevé pour nos étiquettes de classe. Parmi les caractéristiques ajoutées dans ce travail, nous pouvons observer que la densité syntactique moyenne des syntagmes NP (*mean_syntactic_density_NP*) et le ratio syntaxique des syntagmes ADVP (*ADVP_syntactic_ratio*) détiennent un poids important dans la discrimination des personnes atteintes de démence. Pour ce qui est des personnes faisant partie du groupe témoin, le ratio syntactique VP (*VP_syntactic_ratio*) ainsi que la densité syntaxique moyenne des VP (*mean_syntactic_density_VP*) détiennent un poids important. Le score élevé du ratio syntactique VP, associé à sa position parmi les caractéristiques ayant les meilleurs scores en information mutuelle, souligne son importance dans notre modèle.

La comparaison entre les scores des caractéristiques importantes et ceux issus de l'information mutuelle peut s'avérer utile pour détecter des caractéristiques redondantes ou corrélées entre elles. Dans de futurs travaux, l'intégration d'une mesure de colinéarité au sein de l'outil constituerait un apport pertinent pour une analyse plus en profondeur des interactions entre les variables.

La méthode *show_variance* nous a permis d'analyser les différentes caractéristiques de notre modèle. Nous avons pu observer que, de manière générale, la plupart des variables présentent une faible variance. Ce phénomène s'explique par la nature même des données, majoritairement constituées de ratios et intrinsèquement bornées et peu dispersées. Quatre caractéristiques se distinguent toutefois par une variance élevée : *honore_r, words_per_minute, word_count* et *duration*. Cette observation justifie l'utilisation de *StandardScaler*, une méthode de l'outil appliquée dans nos classifications pour normaliser les données.

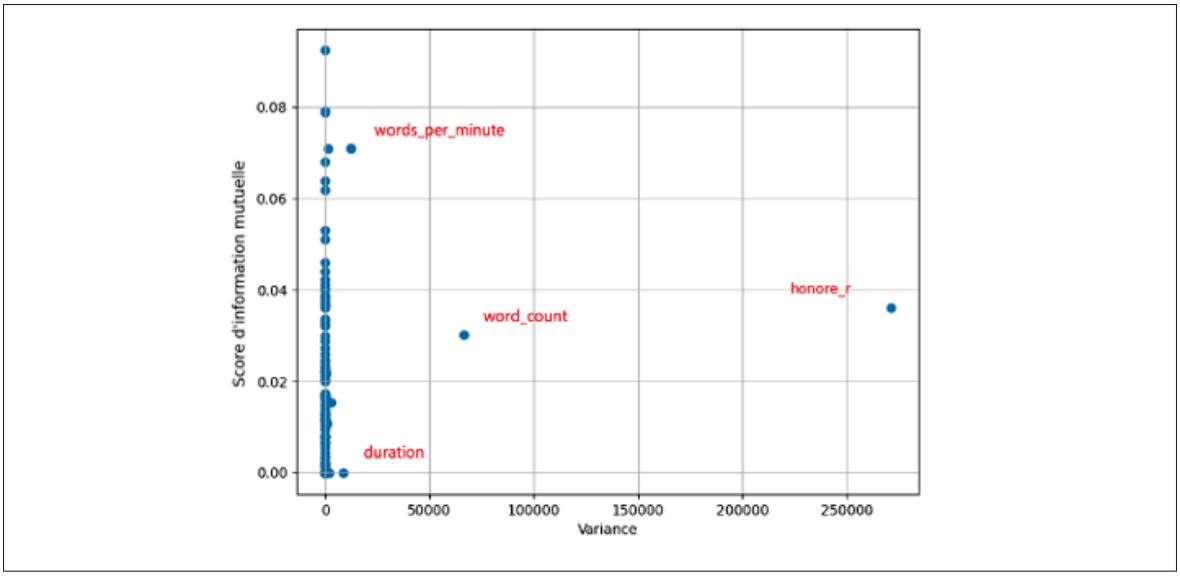

Figure 5.2 Graphique des scores de variance du modèle intégré

La méthode *show_variance* a également retourné des informations permettant d'observer que 98% de la variance de notre modèle est expliquée par la première composante (PCA1), qui est elle-même composée à 99,93% de la caractéristique *honore_r*. La deuxième composante (PCA2) est quant à elle responsable de 0,01% de la variance et est composée à 84,4% de *word_count*, 2,9% de *duration* et de 2,71% de *words_per_minute*.

La majorité de la variance est expliquée par *honore_r*, une caractéristique qui présente toutefois un faible score d'information mutuelle. Nous avons réalisé un test d'ablation pour mesurer son impact direct sur la classification et avons obtenu un résultat quasiment identique au modèle de base, soit sans amélioration notable. Les résultats sont présentés dans le Tableau 5.4. Puisque

honore_r génère une grande variance sans contribuer à la performance du modèle, nous avons décidé de retirer cette caractéristique de la classification.

Tableau 5.4 Impact de la mesure Honoré R sur la classification

Modèle	F1	Précision	Rappel	ASC	
Sans *honore_r*	0,780	0,776	0,824	0,752	0,841
Meilleur modèle à date	0,774	0,772	0,823	0,745	0,837

Après la suppression de la caractéristique *honore_r*, la variance se répartit à 67,29 % pour la première composante principale (PCA1) et à 26,28 % pour la deuxième (PCA2). La PCA1 est principalement composée de *word_count* (0,849), *duration* (0,318) et *words_per_minute* (0,240). Quant à la PCA2, elle est dominée par *words_per_minute* (0,829), *duration* (-0,554) et *hapax_legomena* (0,029).

5.4 Amélioration de la classification

Les caractéristiques retenues après la sélection des caractéristiques présentées dans la section précédente ont été utilisées pour entraîner les différents algorithmes de classification implémentés dans l'outil. Ces méthodes ont été optimisées en utilisant l'option *GridSearch* et testées avec divers hyperparamètres et configurations. Les meilleurs résultats obtenus pour chaque algorithme sont présentés dans le Tableau 5.5 qui présente les résultats en ordre de score de F1.

Tableau 5.5 Comparaison des différents modèles implémentés

Modèle	F1	Précision	Rappel	ASC
XGBOOST	0,809	0,814	0,805	0,847
XGBOOST linéaire	0,803	0,794	0,813	0,855
SVM - noyau linéaire	0,786	0,824	0,752	0,841
Régression linéaire	0,774	0,815	0,750	0,848
Classification de base	0,759	0,811	0,723	0,848
SVM - noyau RBF	0,748	0,749	0,756	0,787
KNN	0,694	0,772	0,512	0,611

Le modèle ayant obtenu la meilleure performance est le XGBoost, avec un score F1 de 0,809, une précision de 0,814, un rappel de 0,805 et une ASC de 0,847. Ce modèle utilise *StandardScaler* pour la normalisation des données ainsi que la validation croisée *Stratified K-Fold*, une technique adaptée aux jeux de données déséquilibrés. L'optimisation a été réalisée via *GridSearch*. Les paramètres optimaux sont les suivants : métrique d'évaluation *logloss*, échantillonnage par arbre de 1,0, γ de 0, taux d'apprentissage de 0,2, profondeur maximale de 3, nombre d'estimateurs de 200 et sous-échantillon de 1,0. Nous avons ajouté ces paramètres comme implémentation par défaut dans la méthode d'UsAge.

Le modèle de régression logistique a été principalement implémenté pour comparer le modèle de base, un SVM à noyau linéaire, à un autre modèle linéaire. Ses performances en classification sont légèrement inférieures à celles du SVM linéaire, mais restent très similaires. Le SVM avec noyau RBF obtient des performances inférieures à celles du SVM linéaire. Le meilleur résultat obtenu par le KNN est nettement inférieur à celui des autres classificateurs. Par ailleurs, la visualisation des résultats n'a pas révélé de sous-groupes significatifs au sein des données.

Nous avons comparé notre modèle le plus performant, le XGBoost entraîné avec 68 caractéristiques, en le ré-entraînant avec les caractéristiques retirées dans la section précédente afin de confirmer nos observations. Le Tableau 5.6 montre que les classifications faites avec ces caractéristiques sont moins performantes.

Tableau 5.6 Comparaison du meilleur modèle avec les caractéristiques retirées

Modèle	**F1**	**Précision**	**Rappel**	**ASC**
XGBOOST	0,809	0,814	0,805	0,847
Avec mesures MI score nul	0,788	0,793	0,785	0,842
Avec *Honore_r*	0,787	0,790	0,785	0,845
Avec pause CCONJ	0,780	0,797	0,765	0,850

5.5 Informations personnelles et données médicales

Les données du Pitt Corpus incluent des informations personnelles et médicales sur les participants, qui ne sont pas intégrées aux données linguistiques. Elles contiennent des scores médicaux ainsi que de l'information sur le niveau d'études et l'âge des participants. Les données non linguistiques présentes dans le corpus sont décrites dans le Tableau 5.7.

Tableau 5.7 Informations personnelles et médicales présentes dans le Pitt Corpus

Mesure	Description	Type de données
mms	Mini-examen de l'état mental	Médical
cds	Échelle d'évaluation clinique de la démence	Médical
Blessed	Échelle de démence de Blessed	Médical
Hamilton	Échelle de dépression de Hamilton	Médical
htotal	Échelle ischémique de Hachinski	Médical
Mattis	Échelle d'évaluation de la démence de Mattis	Médical
nyu	Échelle d'évaluation NYU (mesure de la maladie de Parkinson)	Médical
age	Âge du participant	Personnel
race	Race du participant	Personnel
sex	Sexe du participant	Personnel
educ	Années d'éducation	Personnel

Nous avons intégré ces données de manière singulière afin que chaque entretien avec un participant soit considéré comme une entrée distincte. Cette organisation nous permettra ainsi d'observer chaque rendez-vous comme un point de données et facilitera, à l'avenir, la réalisation de classification en classes multiples et l'exploration approfondie des corrélations entre données personnelles et données linguistiques. Elle permettra également d'examiner les données longitudinalement en sélectionnant une entrevue spécifique tout en conservant l'intégralité de l'information associée.

Nous avons entraîné l'algorithme de base avec l'ensemble des informations personnelles, puis uniquement avec les scores médicaux et, enfin, uniquement avec les informations démographiques, à savoir l'âge, le sexe, la race et le niveau de scolarité. Les résultats sont présentés dans le Tableau 5.8.

Tableau 5.8 Comparaison des classifications avec différent groupes de caractéristiques provenant des données personnelles et médicales

Modèle	F1	Précision	Rappel	ASC
Avec informations personnelles et médicales	0,902	0,913	0,894	0,974
Avec informations médicales	0,881	0,851	0,913	0,943
Classification XGBoost de base	0,809	0,814	0,805	0,847
Avec informations personnelles	0,768	0,749	0,789	0,794

Les résultats de la classification s'améliorent nettement avec l'ajout des données personnelles et médicales. L'apport des données médicales se distingue particulièrement, comme le montrent les tests d'ablation : l'utilisation des données personnelles seulement donne de moins bons résultats.

L'analyse des dix caractéristiques ayant le plus de poids ainsi que l'analyse des composantes principales montrent un mélange intéressant des caractéristiques linguistiques et des caractéristiques personnelles et médicales.

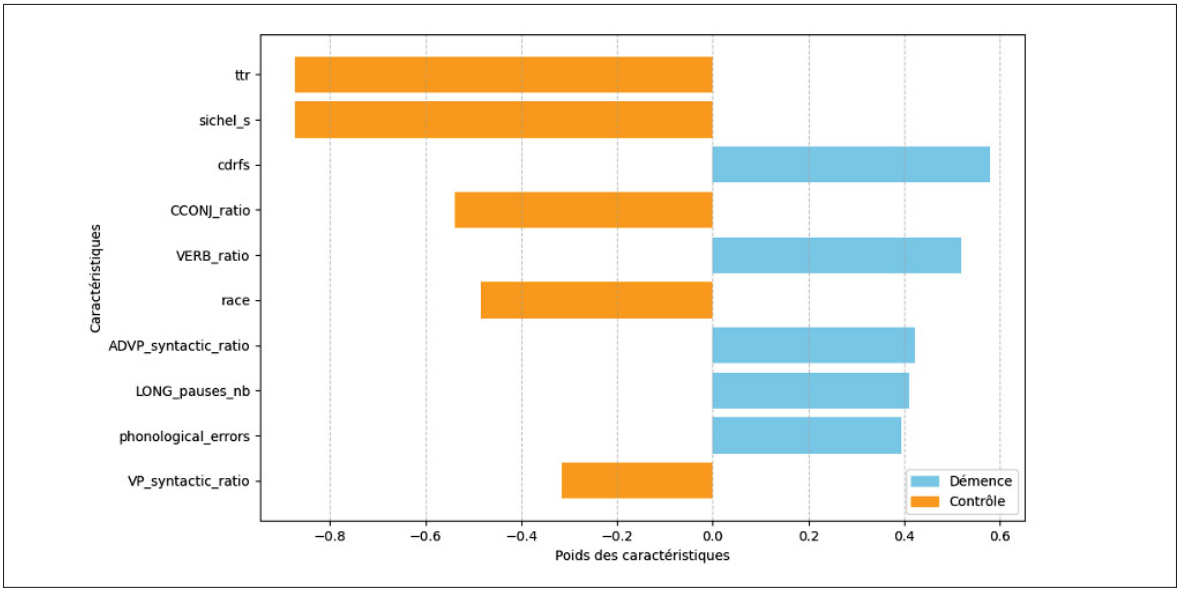

Figure 5.3 10 caractéristiques les plus pertinentes dans la classification du modèle en incluant les données personnelles et médicales

La première composante principale (PC1) explique 38,4 % de la variance totale du modèle. Elle est principalement dominée par la variable *mattis* (0,946), ce qui indique que cette dimension est largement influencée par ce facteur. D'autres variables comme *word_count* (-0,240), *age_apx* (-0,130) et *duration* (-0,121) y contribuent également, mais dans une moindre mesure et de manière négative, indiquant une importance pour la classification du groupe témoin. La deuxième composante principale (PC2) explique quant à elle 33,4 % de la variance. Elle est surtout structurée par *word_count* (0,814), suivie de *duration* (0,293), *mattis* (0,282) et *words_per_minute* (0,244). Ces résultats montrent que les dimensions principales du modèle captent des combinaisons différentes de variables, avec un poids très marqué de *mattis* sur PC1 et de *word_count* sur PC2.

Les informations médicales semblent significativement améliorer les performances du modèle. Toutefois, cet effet positif pourrait également traduire un risque de surapprentissage, notamment si ces variables captent des biais spécifiques à l'échantillon. Il serait donc pertinent d'approfondir les interactions entre ces données et les autres variables.

5.6 Conclusion

Nous avons débuté ce chapitre avec un algorithme de classification utilisant 85 caractéristiques et atteignant un score F1 de 0,779. Des tests d'ablation nous ont permis d'identifier et de supprimer une caractéristique qui semblait ajouter du bruit au modèle. Par la suite, en analysant les valeurs d'information mutuelle, nous avons retiré 16 caractéristiques présentant un score nul, indiquant qu'elles n'apportaient aucune contribution utile à la classification.

L'analyse de la variance a révélé que la caractéristique *honore_r* était majoritairement responsable de la variance des résultats. Un test d'ablation a confirmé que le retrait de cette variable permettait d'améliorer les performances du modèle. Par la suite, plusieurs classificateurs ont été testés et optimisés en fonction de ce nouvel ensemble de caractéristiques.

La meilleure performance a été obtenue avec un classificateur XGBoost, atteignant un score F1 de 0,809. Finalement, nous avons testé l'ajout des données personnelles des participants en

les intégrant individuellement. Cette approche a permis d'atteindre un score F1 de 0,902, des résultats particulièrement élevés qui suggèrent toutefois un risque important de surapprentissage.

La manipulation des caractéristiques nous a permis d'identifier et d'éliminer celles qui n'apportaient aucune valeur à notre algorithme. Les méthodes de classification mises en œuvre nous ont offert la possibilité d'explorer différentes approches et d'ajuster les paramètres afin de déterminer le classificateur le plus adapté à nos données.

CHAPITRE 6

DISCUSSION

L'objectif de cette recherche était de concevoir un outil permettant l'extraction et l'analyse de caractéristiques linguistiques pouvant être utilisées dans un modèle de classification optimisé, capable de distinguer des participants atteints de la maladie d'Alzheimer de participants témoins.

Dans ce contexte, nos contributions principales sont les suivantes :

- La conception d'*UsAge Feature Extraction*, une bibliothèque dédiée à l'extraction de caractéristiques linguistiques pertinentes ;
- La conception d'*UsAge*, un outil facilitant la manipulation des données et l'utilisation de modèles de classification ;
- L'ajout de nouvelles caractéristiques discriminantes pour la détection de la maladie d'Alzheimer ;
- La mise au point d'un algorithme de classification atteignant score F1 de 0,809 et un rappel de 0,805 ;

6.1 Outil de recherche

Nous avons présenté un outil composé de deux modules : *UsAge Feature Extraction* et UsAge. *UsAge Feature Extraction* est une bibliothèque d'extraction de caractéristiques optimisant l'exécution grâce à un orchestrateur basé sur un graphe orienté acyclique (DAG) et le traitement multifil. Conçue de manière modulaire, cette bibliothèque permet de remplacer facilement les outils externes utilisés, tels que Stanza, ce qui facilite la recherche et le développement futur de l'outil. La bibliothèque intègre également NSE, une bibliothèque développée par le LiNCS et spécifiquement conçue pour le traitement des fichiers CHAT.

UsAge est une application qui utilise *UsAge Feature Extraction* pour extraire des caractéristiques exploitables dans ses analyses. Elle propose plusieurs méthodes d'analyse des données permettant non seulement de mesurer l'impact d'une caractéristique sur un modèle, mais aussi d'observer des tendances. Elle met en place un système flexible qui permet de créer dynamiquement

des groupes afin de comparer les données entre ceux-ci. L'application intègre également des méthodes de classification pour tester plusieurs modèles sur ces groupes créés de manière dynamique. UsAge s'inscrit en continuité avec des études antérieures menées par le LiNCS et inclut des mesures de base ajoutées aux modèles pour leur capacité à distinguer les patients atteints de démence des participants témoins.

Malgré ses capacités d'observation, de classification et de soutien à la recherche, l'application et ses bibliothèques ne fonctionnent actuellement majoritairement qu'avec les fichiers au format CHAT, ce qui limite fortement le nombre de jeux de données utilisables ou oblige à formater d'éventuels nouveaux jeux de données dans ce format. De plus, les mesures observées sont optimisées pour la langue anglaise. De futurs développements devraient se concentrer sur l'expansion de pipelines applicables à différents formats de fichiers et à différentes langues, ce qui permettrait l'observation d'un plus grand nombre de jeux de données.

6.2 Nouvelles caractéristiques

Nous avons introduit trois nouvelles catégories de caractéristiques qui, selon nous, méritaient d'être explorées plus en profondeur. Tout d'abord, nous nous sommes penché sur les chaînes de coréférence en introduisant plusieurs mesures d'observation. Nous avons comparé les performances des outils Stanza et SpaCy à l'annotation manuelle en ce qui concerne l'extraction automatique des chaînes de coréférence. Malheureusement, les deux outils se sont révélés peu performants et nous avons choisi de ne pas inclure ces mesures dans notre modèle final. Cependant, l'ajout de mesures issues de l'annotation manuelle a permis une amélioration de 6% du score F1, ce qui met en évidence la nécessité de développer des méthodes d'automatisation précises pour ces mesures.

Nous nous sommes ensuite penchés sur les types de mots suivant une pause, en distinguant trois types de pauses : courtes, moyennes et longues. Nous avons constaté que les mesures liées aux pauses courtes en début de phrase ainsi que le nombre total de pauses présentaient une certaine signification statistique. Le nombre de pauses étant déjà utilisé, notre analyse vient confirmer

sa pertinence. En revanche, nous avons également observé une forte prédominance des pauses courtes par rapport aux pauses moyennes et longues. Il serait intéressant dans le futur de tester les mesures implémentées sur un jeu de données contenant plus d'exemples.

Nous avons également observé la complexité syntaxique en introduisant des mesures de densité syntaxique, de ratio de syntagme et de poids de dépendance.

La densité syntaxique est mesurée de manière récursive en calculant le nombre de mots par syntagme, afin de permettre une évaluation plus fine de la complexité structurelle des énoncés. Ces mesures ont permis d'identifier une corrélation entre la densité syntaxique des syntagmes incomplets (FRAG), verbaux (VP), nominaux (NP) et adjectivaux (ADJP) et les étiquettes qui contribuent à l'amélioration des performances de classification.

Les mesures du ratio de syntagmes, définies comme le nombre de syntagmes par document, permettent d'évaluer la complexité syntaxique à un niveau plus global, soit celui du document, de manière analogue à la densité syntaxique, qui opère à un niveau plus local. Ces mesures ont permis d'observer une corrélation entre les syntagmes ADVP, ADJP, VP, S et SBARQ, améliorant également la classification.

Des mesures de calcul du flux syntaxique, visant à quantifier les dépendances au sein d'une phrase, ont également été mises en œuvre. Toutefois, ces mesures ne contribuaient pas à la classification en raison d'une plage de valeurs très restreintes, comprise uniquement entre 1 et 2.

L'observation de mesures manquantes dans le pipeline original nous a permis d'améliorer le modèle final en y ajoutant les mesures significatives identifiées. Cependant, plusieurs mesures qui ne relevaient pas de l'étendue de cette recherche devraient être examinées dans de futurs travaux afin de déterminer si leur ajout est pertinent. L'intégration de mesures phonétiques constitue une perspective pertinente pour les travaux futurs, celles-ci n'ayant pas été incluses dans le pipeline actuel en raison de leur interprétabilité limitée, ce qui compromettait leur utilité analytique. L'ajout de mesures de couverture d'information, mentionnées par (Hernández-Domínguez, 2019)

et (Abiven, 2020), mais absentes de la base de code, constituerait également une piste pertinente pour enrichir les métriques utilisées.

6.3 Sélection des données et classification

Nous avons utilisé notre outil pour observer les caractéristiques extraites et effectuer une sélection des données en fonction de leur contribution à la classification. Cette étape nous a permis d'éliminer 18 caractéristiques qui introduisaient du bruit, ce qui a conduit à une amélioration des performances du score F1 et du rappel du modèle de base de 2%. Par la suite, plusieurs modèles de classification ont été testés. Nous avons finalement implémenté un modèle XGBoost, atteignant un score F1 de 80,9%, une précision de 81,4%, un rappel de 80,5% et une ASC de 84,7%, surpassant les résultats obtenus avec le pipeline de base. Nous avons ensuite intégré les informations cliniques des participants au modèle pour observer leur impact et avons atteint un score F1 de 90,2%, une précision de 91,3%, un rappel de 89,4% et une ASC de 97,4%.

Bien que l'ajout des mesures médicales et personnelles permette d'améliorer les performances du modèle, il est important de se questionner sur la pertinence d'ajouter ces mesures, qui sont elles-mêmes indicatrices de la maladie, dans un outil linguistique. Une analyse plus approfondie des données pourrait cependant permettre d'identifier des observations permettant d'extraire des caractéristiques linguistiques pertinentes.

6.4 Recherches futures

Une analyse des caractéristiques utilisées par notre modèle a révélé que plusieurs d'entre elles présentaient un faible score d'information mutuelle, ce qui suggère qu'il serait pertinent d'envisager une expansion ou un enrichissement des caractéristiques afin d'améliorer la capacité discriminante du modèle. Une analyse des composantes principales (PCA) nous a permis de constater que la variance de notre modèle s'explique principalement par deux grands axes : l'un regroupant des observations liées à la verbosité et au lexique, et l'autre incluant des observations relatives à la durée et à la vitesse du discours.

Dans de futures recherches, il serait pertinent d'enrichir l'outil par l'intégration de mesures visant à analyser de manière plus fine la temporalité des échantillons de discours pour tirer pleinement parti de la longitudinalité des données. Il serait également pertinent d'intégrer des méthodes avancées de sélection de caractéristiques visant à évaluer la colinéarité entre les différentes mesures utilisées. Cette approche permettrait de détecter les redondances et les chevauchements entre les variables, ce qui permettrait d'améliorer l'efficacité du modèle tout en facilitant son interprétabilité.

Enfin, il serait pertinent d'explorer des modèles de classification en classes multiples ainsi que des approches capables de traiter des séries temporelles, afin de permettre une analyse plus raffinée des groupes selon leurs diagnostics et d'exploiter pleinement les données longitudinales disponibles.

CONCLUSION ET RECOMMANDATIONS

L'objectif de cette recherche était de concevoir un outil automatisé pour l'extraction, l'analyse et la classification de biomarqueurs provenant du langage liés à la maladie d'Alzheimer. L'utilisation de cet outil permet d'analyser les phénomènes linguistiques et d'évaluer leur impact sur l'observation et la compréhension de la maladie. Cette application constitue une base robuste et flexible, permettant aux futurs chercheurs de poursuivre et d'élargir les travaux existants. Elle constitue également une fondation solide pour l'évolution de l'outil vers des usages cliniques, notamment pour le suivi de la maladie et sa détection précoce par les professionnels de la santé.

Nous avons présenté UsAge ainsi que sa bibliothèque intégrée, *UsAge Feature Extraction*, en détaillant les méthodes implémentées pour l'extraction et l'analyse de biomarqueurs linguistiques. Nous avons motivé les choix architecturaux, tels que l'utilisation d'un orchestrateur basé sur un DAG et le traitement multifil, et les décisions méthodologiques adoptées, notamment la conception modulaire de l'outil permettant l'interchangeabilité des composants d'analyse linguistique.

Nous avons ensuite utilisé l'outil pour évaluer l'impact de nouvelles caractéristiques sur le modèle, en procédant à une sélection de données et en justifiant nos choix d'inclusion et d'exclusion de certaines variables. Plusieurs modèles de classification ont été testés et nous avons affiné notre approche jusqu'à l'obtention d'un score F1 de 80,9% en utilisant uniquement des caractéristiques linguistiques et de 90,2% en y ajoutant des données personnelles et médicales.

Nous avons présenté une solution complète combinant extraction, analyse, classification et visualisation des données, une intégration encore peu représentée dans les solutions disponibles. Nous y avons intégré un classificateur performant, offrant une bonne capacité à distinguer les participants témoins des participants atteints de la maladie d'Alzheimer.

Ce travail ouvre la voie à des applications cliniques potentielles, notamment pour le suivi et la détection précoce de la maladie. Dans le futur, il conviendra d'enrichir l'ensemble des caractéristiques analysées et de mener une étude approfondie de leurs interactions afin de renforcer notre compréhension des effets des troubles cognitifs sur le langage. Ces avancées contribueront à améliorer les outils diagnostiques et à soutenir la recherche dans ce domaine crucial.

ANNEXE I

LISTE DES CARACTÉRISTIQUES EXTRAITES PAR L'ANCIEN PIPELINE USAGE

1. Listes des caractéristiques

Mesure	Equation
text_size	Nombre total de mots (N)
vocab_size	Nombre de différents lemmes (peut être un mot, un mot composé ou un groupe de mots (syntagme)) (V)
hapax_legomena	Nombre de lemmes utilisés une seule fois (V_1)
hapax_dislegomena	Nombre de lemmes utilisés deux fois (V_2)
brunet_index (Brunet, 1978)	$W = N^{V^{-c}}$ avec c=0.172 (Tweedie &Baayen, 1998) (Plus la valeur est basse, plus le vocabulaire est riche, la valeur est entre 10 et 20)
honore_r_statistics	$R = \frac{100 \log N}{1 - \frac{V_1}{V}}$
TTR	$TTR = \frac{V_1}{V}$
sichel_s	$S = \frac{V_2}{V \neq}$
yule_k (Yule's characteristic K (Miranda-Garcia & Calle-MartÍn, 2005))	$10^4 \frac{[\sum_{i=1}^{N} i^2 V(i,N)]}{N^2} - \frac{1}{N}$ (plus c'est grand moins le vocabulaire est riche)
entropy	$H(x) = -\sum_{x \in X} p(x)\ log_2\ p(x)$ où p(x) est la probabilité que le mot x soit dans le test X

Figure-A I-1 Marqueurs linguistiques extraits par Cesari (2023)

Mesures
Mean_(1 à 13)
Kurtosis_(1 à 13)
Skewness_(1 à 13)
Variance_(1 à 13)

Figure-A I-2 Marqueurs phonétiques extraits par Cesari (2023)

Marqueurs	Explication
nbIncPhrases	Nombre de phrases incomplètes
nbIncPhrasesRatio	Nombre de phrases incomplètes par rapport au nombre de mots total
nbRepetitions	Nombre de répétitions (phrases ou mots)
nbRepetitionsRatio	Nombre de répétitions (phrases ou mots) par rapport au nombre de mots total
nbPausesShort	Nombre de courtes pauses
nbPausesShortRatio	Nombre de courtes pauses par rapport au nombre de mots total
nbPausesMedium	Nombre de pauses moyennes
nbPausesMediumRatio	Nombre de pauses moyennes par rapport au nombre de mots total
nbPausesOther	Nombre de pauses (autres que petites, moyennes ou longues)
nbPausesOtherRatio	Nombre de pauses (autres que petites, moyennes ou longues) par rapport au nombre de mots total
nbIncWords	Nombre de mots incomplets
nbIncWordsRatio	Nombre de mots incomplets par rapport au nombre de mots total
nbRetracings	Nombre de reprises (par exemple répétition d'un mot ou d'une phrase)
nbRetracingsRatio	Nombre de reprises (par exemple répétition d'un mot ou d'une phrase) par rapport au nombre de mots total
nbPausesTotal	Nombre de pauses total
nbPausesTotalRatio	Nombre de pauses total par rapport au nombre de mots total
nbExpressions	Nombre d'expressions utilisées
nbExpressionsRatio	Nombre d'expressions utilisées par rapport au nombre de mots total
nbPausesLong	Nombre de pauses longues
nbPausesLongRatio	Nombre de pauses longues par rapport au nombre de mots total
nbInterjections	Nombre d'interjections utilisées
nbInterjectionsRatio	Nombre d'interjections utilisées par rapport au nombre de mots total
nbSynonyms	Nombre de synonymes utilisés
nbSynonymsRatio	Nombre de synonymes utilisés par rapport au nombre de mots total
nbErrors	Nombre d'erreurs (erreurs grammaticales, mot pas bien utilisé...)
nbErrorsRatio	Nombre d'erreurs (erreurs grammaticales, mot pas bien utilisé) par rapport au nombre de mots total
totalWordCount	Nombre de mots total

Figure-A I-3 Marqueurs discursifs extraits par Cesari (2023)

Marqueurs	Définition
adjFreq	Nombre d'adjectifs total
adjRatio	Nombre d'adjectifs par rapport au nombre total de mots
adpFreq	Nombre d'adpositions total
adpRatio	Nombre d'adpositions par rapport au nombre total de mots
advFreq	Nombre d'adverbes total
advRatio	Nombre d'adverbes par rapport au nombre total de mots
auxFreq	Nombre d'auxiliaires total
auxRatio	Nombre d'auxiliaires par rapport au nombre total de mots
conjFreq	Nombre de conjonctions total
conjRatio	Nombre de conjonctions par rapport au nombre total de mots
cconjFreq	Nombre de conjonctions de coordination total
cconjRatio	Nombre de conjonctions de coordination par rapport au nombre total de mots
detFreq	Nombre de déterminants total
detRatio	Nombre de déterminants par rapport au nombre total de mots
intjFreq	Nombre d'interjections total
intjRatio	Nombre d'interjections par rapport au nombre total de mots
nounFreq	Nombre de noms total
nounRatio	Nombre de noms par rapport au nombre total de mots
numFreq	Nombre de nombres total
numRatio	Nombre de nombres par rapport au nombre total de mots
partFreq	Nombre de particules total
partRatio	Nombre de particules par rapport au nombre total de mots
pronFreq	Nombre de pronoms total
pronRatio	Nombre de pronoms par rapport au nombre total de mots
punctFreq	Nombre de ponctuations total
punctRatio	Nombre de ponctuations par rapport au nombre total de mots
sconjFreq	Nombre de conjonctions de subordination total
sconjRatio	Nombre de conjonctions de subordination par rapport au nombre total de mots
verbFreq	Nombre de verbes total
verbRatio	Nombre de verbes par rapport au nombre total de mots
totalWordCount	Nombre de mots total

Figure-A I-4 Marqueurs du discours extraits par Cesari (2023)

BIBLIOGRAPHIE

Abiven, F. (2020). Automatisation multilingue du prétraitement de transcriptions dans la détection de la maladie d'Alzheimer [Dissertation]. Repéré à https://espace.etsmtl.ca/id/eprint/2659.

Ahmed, S., Haigh, A. M., de Jager, C. A. & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(Pt 12), 3727–3737. doi : 10.1093/brain/awt269.

Alzheimer's Association. [Consulté en avril 2025]. (2025a). 10 Early Signs and Symptoms of Alzheimer's. Repéré à https://www.alz.org/alzheimers-dementia/10_signs.

Alzheimer's Association. [Consulté en juin 2025]. (2025b). Medical Tests for Diagnosing Alzheimer's. Repéré à https://www.alz.org/alzheimers-dementia/diagnosis/medical_tests.

Alzheimer's Research UK. [Consulté en juin 2025]. (2022). Worldwide dementia cases to triple by 2050 to over 150 million. Repéré à https://www.alzheimersresearchuk.org/news/worldwide-dementia-cases-to-triple-by-2050-to-over-150-million/.

Becker, J. T., Boller, 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. doi : doi:10.21415/CQCW-1F92.

Can, E. & Kuruoğlu, G. (2019). Language Changes in Late-Onset Alzheimer's Disease. *Psycholinguistics*, 25(2), 50–68. doi : 10.31470/2309-1797-2019-25-2-50-68.

Cesari, M. (2023). Intégration d'outils et étude des améliorations possibles des algorithmes pour la détection et le suivi de maladies affectant la cognition. [Dissertation].

Cho, S., Nevler, N., Ash, S., Shellikeri, S., Irwin, D. J., Massimo, L., Rascovsky, K., Olm, C., Grossman, M. & Liberman, M. (2020). Automated analysis of lexical features in Frontotemporal Degeneration. *medRxiv [Preprint]*, 2020.09.10.20192054. doi : 10.1101/2020.09.10.20192054. Updated in : Cortex. 2021 Apr ;137 :215–231. doi : 10.1016/j.cortex.2021.01.012.

Cho, S., Quilico Cousins, K. A., Sanjana Shellikeri, Ash, S., Irwin, D. J., Liberman, M. Y., Grossman, M. & Nevler, N. (2022). Lexical and Acoustic Speech Features Relating to Alzheimer Disease Pathology. *Neurology Journals*, 99(4), 313–322. doi : https://doi.org/10.1212/WNL.0000000000200581.

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. doi : 10.1006/brln.1996.0033.

Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J. & Anderberg, P. (2017). Machine learning and microsimulation techniques on the prognosis of dementia : A systematic literature review. *PLoS One*, 12(6), e0179804. doi : 10.1371/journal.pone.0179804.

de la Fuente Garcia, S., Ritchie, C. W. & Luz, S. (2020). Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease : A Systematic Review. *Journal of Alzheimer's Disease*, 78(4), 1547–1574. doi : 10.3233/JAD-200888.

de Lira, J. O., Minett, T. S. C., Bertolucci, P. H. F. & Ortiz, K. Z. (2014). Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dementia & Neuropsychologia*, 8(3), 260–265. doi : 10.1590/S1980-57642014DN83000010.

Dobrovolskii, V. (2021). Word-Level Coreference Resolution. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dupuis-Desroches, J. (2025). Évaluation des systèmes de reconnaissance automatique de la parole pour les personnes atteintes d'aphasie : un cadre d'analyse de performance pertinent au milieu clinique [Dissertation].

D'Oosterlinck, K., Bitew, S. K., Papineau, B., Potts, C., Demeester, T. & Develder, C. (2023). CAW-coref : Conjunction-Aware Word-level Coreference Resolution. *Proceedings of the CRAC 2023 Workshop*.

Fraser, K. K., Meltzer, J. A. & Rudzicz, F. (2016). Linguistic Features Identify Alzheimer's Disease in Narrative Speech. *Journal of Alzheimer's Disease*, 49(2), 407–422. doi : 10.3233/JAD-150520.

Fromm, D., Dalton, S. G., Brick, A., Olaiya, G., Hill, S., Greenhouse, J. & MacWhinney, B. (2024). The Case of the Cookie Jar : Differences in Typical Language Use in Dementia. *Journal of Alzheimer's Disease*, 100(4), 1417–1434. doi : 10.3233/JAD-230844.

Guinn, C. & Habash, A. (2012). Language analysis of speakers with dementia of the Alzheimer's type. *AAAI Fall Symposium - Technical Report*, 8-13.

Gumus, A. & Koo, Y. (2023). Linguistic changes in neurodegenerative diseases relate to clinical symptoms. *Journal of Neurodegenerative Research*, 15(3), 123–145. doi : 10.1234/jnr.2023.56789.

Hernández-Domínguez, L. E. (2019). *Computer-based characterization of language alterations throughout the Alzheimer's disease continuum*. (PhD thesis, École de technologie supérieure). Repéré à Thèse de doctorat électronique.

Honnibal, M. & Montani, I. (2017). *spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547–579.

Jiskoot, L. C., Poos, J. M., van Boven, K., de Boer, L., Giannini, L. A. A., Satoer, D. D., Visch-Brink, E. G., van Hemmen, J., Franzen, S., Pijnenburg, Y. A. L., van den Berg, E. & Seelaar, H. (2023). The ScreeLing : Detecting Semantic, Phonological, and Syntactic Deficits in the Clinical Subtypes of Frontotemporal and Alzheimer's Dementia. *Assessment*, 30(8), 2545–2559. doi : 10.1177/10731911231154512. Epub 2023 Feb 17.

Kahane, S., Yan, C. & Botalla, M.-A. (2017, sep). What are the limitations on the flux of syntactic dependencies ? Evidence from UD treebanks. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pp. 73–82. Repéré à https://aclanthology.org/W17-6510/.

Kaplan, E. F., Goodglass, H. & Weintraub, S. (1983). *The Boston Naming Test* (éd. 2nd). Philadelphia : 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. doi : 10.1080/13803395.2016.1179266.

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. doi : 10.1044/1092-4388(2003/027).

Kim, H., Obermeyer, J. & Wiley, R. W. (2024). Written discourse in diagnosis for acquired neurogenic communication disorders : current evidence and future directions. *Frontiers in Human Neuroscience*, 17, 1264582. doi : 10.3389/fnhum.2023.1264582.

Klie, J.-C., Agatonovic, M. & Cimiano, P. (2018). INCEpTION : Semantic annotation platform with self-learning assistance. *Proceedings of the 27th International Conference on Computational Linguistics : System Demonstrations*, pp. 5–9.

Lindsay, H., Tröger, J. & König, A. (2021). Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13, 642033. doi : 10.3389/fnagi.2021.642033.

Lundholm Fors, K., Fraser, K. & Kokkinakis, D. (2018). Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. *Studies in Health Technology and Informatics*, 247, 705–709. doi : 10.3233/978-1-61499-896-9-705.

Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D. & MacWhinney, B. (2021). Editorial : Alzheimer's Dementia Recognition through Spontaneous Speech. *Frontiers in Computer Science*, 3, 780169. doi : 10.3389/fcomp.2021.780169.

MacWhinney, B. [Online ; accessed 2025-06-05]. (2021). Tools for Analyzing Talk. Part 1 : The CHAT Transcription Format. Repéré à https://doi.org/10.21415/3mhn-0z89.

Makara-Studzinska, M., Gustaw, K. & Krys, K. (2019). Difficulties in communication with a patient with Alzheimer's disease. *Psycholinguistics*, 25(2), 50-68. doi : 10.31470/2309-1797-2019-25-2-50-68.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA : MIT Press.

Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English : The Penn Treebank.

Mueller, K. D., Koscik, R. L., Clark, L. R., Hermann, B. P., Johnson, S. C. & Turkstra, L. S. (2018). The Latent Structure and Test-Retest Stability of Connected Language Measures in the Wisconsin Registry for Alzheimer's Prevention (WRAP). *Archives of Clinical Neuropsychology*, 33(8), 993–1005. doi : 10.1093/arclin/acx116.

Onofre de Lira, J., Zazo Ortiz, K., Carvalho Campanha, A., Ferreira Bertolucci, P. H. & Cianciarullo Minett, T. S. (2011). Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23(3), 404–412. doi : doi:10.1017/S1041610210001092.

Orimaye, S. O., Wong, J. S., Golden, K. J., Wong, C. P. & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18(1), 34. doi : 10.1186/s12859-016-1456-0.

Penn Treebank Project. [Accessed : 2025-07-01]. (2003). Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Repéré à https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*. Repéré à https://aclanthology.org/2020.acl-demos.14.

Rezaii, N., Mahowald, K., Ryskin, R., Dickerson, B. & Gibson, E. (2022). A syntax-lexicon trade-off in language production. *Proceedings of the National Academy of Sciences*, 119(25), e2120203119. doi : 10.1073/pnas.2120203119.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). "Why Should I Trust You ?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1135–1144. doi : 10.1145/2939672.2939778.

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. doi : 10.1109/TASL.2011.2112351.

Santander-Cruz, Y., Salazar-Colores, S., Paredes-García, W. J., Guendulain-Arenas, H. & Tovar-Arriaga, S. (2022). Semantic Feature Extraction Using SBERT for Dementia Detection. *Brain Sciences*, 12(2), 270. doi : 10.3390/brainsci12020270.

Sheikholeslami, S. (2019). *Ablation Programming for Machine Learning*.

Sitek, E. J., Barczak, A., Kluj-Kozłowska, K., Kozłowski, M., Barcikowska, M. & Sławek, J. (2015). Is descriptive writing useful in the differential diagnosis of logopenic variant of primary progressive aphasia, Alzheimer's disease and mild cognitive impairment ? *Neurologia i Neurochirurgia Polska*, 49(4), 239–244. doi : 10.1016/j.pjnns.2015.06.001.

Swinburn, K., Porter, G. & Howard, D. (2005). *Comprehensive Aphasia Test*. UK : Psychology Press.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. (2020). SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272. doi : 10.1038/s41592-019-0686-2.

Williams, E., McAuliffe, M. & Theys, C. (2021). Language changes in Alzheimer's disease : A systematic review of verb processing. *Brain and Language*, 223. doi : https://doi.org/10.1016/j.bandl.2021.105041.

Williams, E., Theys, C. & McAuliffe, M. (2023). Lexical-semantic properties of verbs and nouns used in conversation by people with Alzheimer's disease. *PLoS ONE*, 18(8), e0288556. doi : https://doi.org/10.1371/journal.pone.0288556.

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