Modèle d'amélioration des transcriptions automatiques des narrations de patients dans le contexte restreint des tâches de description d'images

par

Eric Ulises GARCÍA CANO CASTILLO

MÉMOIRE PAR ARTICLE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE EN GÉNIE DES TECHNOLOGIES DE L'INFORMATION
M. Sc. A.

MONTRÉAL, LE 26 JANVIER 2022

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. Patrick Cardinal, président du jury
Département de génie logiciel et des technologies de l'information 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 18 JANVIER 2022

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

En premier lieu, j'aimerais adresser ma profonde gratitude à ma directrice de recherche, Prof. Sylvie Ratté. Je vous remercie énormément pour votre soutien inestimable, l'aide financière et pour toute la confiance que vous m'avez manifestée dès le début. Merci aussi pour vos conseils, vos avis et votre expertise que vous avez partagés avec moi semaine après semaine. Je vous remercie également pour le dévouement inlassable dont vous faites preuve envers vos étudiants et pour m'avoir permis d'apporter ma contribution au projet Cecilia.

Je suis reconnaissant au Centre de recherche de l'Institut universitaire de gériatrie de Montréal (CRIUGM) pour sa précieuse implication dans la société et sans qui ce travail ne serait pas possible.

Je tiens à t'exprimer ma gratitude, Laura, pour être mon guide spirituel dans ma carrière professionnelle, pour les conseils et les idées qui ont largement contribué à façonner ce projet, mais surtout pour tout le soutien, l'attention et l'affection que tu m'as apportés sur le plan personnel. Merci *mija.*

Merci à mes collègues du laboratoire LiNCS, j'apprécie votre camaraderie et je chéris les moments que nous avons partagés. Bon courage à tous !

Un grand merci à toi Osvaldo, pour ton soutien, ta compagnie et ton amour sincère dans ce voyage que nous avons entrepris ensemble, et de m'avoir encouragé à saisir le volant de ma vie et à le posséder. *Eres mi mar, mi montaña, eres mi tempestad y mi calma, eres mi noche y el alba.*

Je remercie infiniment Edgar, mon frère, pour son soutien indéfectible depuis 1988. Tu es et seras toujours mon modèle. *Mijo,* je n'ai pas assez de mots pour t'exprimer ma gratitude et mon amour.

J'adresse également mes remerciements à mes meilleurs amis dans la vie, qui ne m'ont jamais abandonné dans les moments les plus difficiles, Ana María, Ana Rosa, Semo et Toño ; même si vous êtes loin, je vous emporte avec moi.

À Nicole Trocherie, pour tout ce que tu m'as appris, tes leçons m'accompagnent jour après jour dans ma vie. Pour ton dévouement et ton engagement envers tes étudiants, et pour nous avoir sans cesse encouragés et préparés à connaître de nouveaux horizons, merci à jamais.

Finalement, je dédie ce travail et mes remerciements les plus profonds à ma mère, pour son amour inconditionnel et pour m'avoir appris par l'exemple à surmonter toutes les adversités.

Modèle d'amélioration des transcriptions automatiques des narrations de patients dans le contexte restreint des tâches de description d'images

Eric Ulises GARCÍA CANO CASTILLO

RÉSUMÉ

La reconnaissance automatique de la parole (ASR, selon son acronyme anglais) est une technologie largement utilisée dans la vie quotidienne, mais qui n'est pas complètement résolue. Les systèmes ASR sont toujours sujets à des erreurs, surtout lorsqu'ils sont confrontés à des conditions non standard, différentes de celles utilisées pour les entraîner. Cette technologie est particulièrement mise à l'épreuve lorsqu'elle est utilisée avec la parole de non-anglophones et de personnes âgées. Dans le domaine de l'étude des maladies neurodégénératives, on sait que les troubles du langage apparaissent à des stades précoces de la maladie et que l'analyse du discours narratif des patients permet d'obtenir un diagnostic opportun. L'analyse manuelle, telle qu'elle est réalisée jusqu'à présent, est coûteuse en termes de temps et de ressources. La reconnaissance automatique de la parole pourrait donc rendre le processus plus efficace. Cependant, les taux d'erreur élevés de ces systèmes les empêchent d'être largement utilisés dans la science et la recherche.

Dans cet article, nous proposons une nouvelle méthode de postédition de détection et de correction des erreurs pour un système ASR qui génère des transcriptions automatiques de la parole d'adultes et de personnes âgées francophones décrivant une image.

Au moyen de techniques de traitement du langage naturel, nous extrayons le vocabulaire le plus courant des transcriptions manuelles correctes pour construire un dictionnaire de correction phonémisé ; ensuite, nous extrayons des phrases hors contexte des transcriptions automatiques, qui sont ensuite comparées par une recherche phonétique floue avec le dictionnaire de correction, pour trouver et appliquer les meilleures corrections. Les résultats expérimentaux montrent une précision de détection des erreurs de 80 % et notre meilleur modèle permet une amélioration moyenne du WER de 1,9 %, avec des valeurs allant de 0,6 % à 6,4 %.

Mots-clés : reconnaissance automatique de la parole ; détection d'erreurs ASR ; correction d'erreurs ASR ; voix vieillissante ; adulte âgé ; langue française

A model for improving automatic transcriptions of patient narratives in the restricted context of image description tasks

Eric Ulises GARCÍA CANO CASTILLO

ABSTRACT

Automatic speech recognition (ASR) is a technology widely used in daily life, but not completely solved. ASR systems are still prone to errors, especially when confronted with non-standard conditions, different from those used to train them. This technology is especially challenged when used with speech from non-English speakers and aged voices. In certain domains, such as the study of neurodegenerative diseases, it is known that language impairments appear in early stages of the disease, and that the analysis of patients' narrative discourse helps to obtain a timely diagnosis. Manual analysis, as it is done so far, is costly in terms of time and resources, so automatic speech recognition could make the process more efficient. However, the high error rates in these systems prevent them from being widely used in science and research.

In this paper, we propose a new post-editing method of error detection and correction for an ASR system that generates automatic transcriptions of the speech of French-speaking adults and older adults describing an image.

By means of natural language processing techniques, we extract the most common vocabulary from correct manual transcriptions to build a phonemicized correction dictionary; then we extract out-of-context sentences from the automatic transcriptions, which are then compared through a fuzzy phonetic search with the correction dictionary to find and apply the best corrections. Experimental results show an error detection accuracy of 80% and our best system yields an average WER improvement of 1.9%, with values ranging from 0.6% to 6.4%.

Keywords: automatic speech recognition; ASR error detection; ASR error correction; aging voice; older adult; french language

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AD	Alzheimer's disease
AM	Apprentissage Machine
ASR	Automatic Speech Recognition
BDAE	Boston Diagnostic Aphasia Examination
CHAT	Codes for the Human Analysis of Transcripts
CSV	Comma-Separated Values
CRIM	Centre de Recherche Informatique de Montréal
CRIUGM	Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal
IDF	Term Frequency
IPA	International Phonetic Alphabet
MA	Maladie d'Alzheimer
TALN	Traitement Automatique du Langage Naturel
TDI	Tâche de description d'image
TF	Term Frequency
TF-IDF	Term Frequency Inverse Document Frequency
WER	Word Error Rate

INTRODUCTION

La démence n'est pas une maladie unique, mais le nom d'un groupe de symptômes qui affectent gravement la capacité d'une personne à effectuer ses activités habituelles. Il existe différents types de démence, mais la maladie d'Alzheimer (MA) est la plus répandue chez les personnes âgées, représentant environ 70 % de tous les cas de démence, étant également la plus étudiée (Greenblat, 2021).

Il n'existe actuellement aucun test spécifique permettant d'établir un diagnostic de déficience neurocognitive due à la maladie d'Alzheimer. Si la maladie est suspectée, plusieurs tests physiques et cognitifs seront très probablement réalisés. En combinant les résultats et la description des antécédents médicaux, les médecins disposeront des données nécessaires pour établir un diagnostic (Alzheimer's Association, 2021). La maladie est diagnostiquée lorsqu'elle a atteint un stade où les symptômes cognitifs et neuropsychiatriques interfèrent avec le fonctionnement social ou les activités de la vie quotidienne.

Les traitements actuels peuvent ralentir temporairement le développement des symptômes si la maladie est diagnostiquée à un stade précoce, c'est pourquoi un diagnostic précoce peut donner aux patients des résultats thérapeutiques comparativement meilleurs (Alzheimer's Association, 2021). Les méthodes de traitement automatique de la parole ont un grand potentiel pour détecter automatiquement les indicateurs prototypiques en temps réel et présenter les analyses et les résultats d'une manière que les spécialistes médicaux peuvent inclure comme une source d'information supplémentaire lors du diagnostic des déficits cognitifs (Weiner, Angrick, Umesh, & Schultz, 2018).

Un échantillon de discours narratif peut contenir des informations précieuses sur la capacité du locuteur à choisir un contenu approprié et des mots fonctionnels, à construire des phrases et à communiquer du sens. La parole et le langage spontanés d'un patient peuvent être obtenus par des transcriptions manuelles et des enregistrements de la parole. La description d'images est une méthode bien acceptée pour capturer le discours narratif et analyser les compétences

conversationnelles d'une personne. Les descriptions fournissent un ensemble de données dans un contexte restreint particulièrement riche qui nous permet d'étudier l'impact des maladies neurodégénératives par rapport aux compétences cognitives linguistiques (Cummings, 2019).

Cependant, les tâches de description d'images (TDI) peuvent produire une quantité considérable de données qui sont actuellement analysés en majeure partie à la main. Ces tâches manuelles, qui sont coûteuses et demandent beaucoup de temps, pourraient être automatisées à l'aide d'un système de reconnaissance automatique de la parole (ASR, selon son acronyme anglais) (Cummings, 2019; Fraser, Rudzicz, Graham, & Rochon, 2013; Fraser, Rudzicz, & Rochon, 2013).

Un système de reconnaissance automatique de la parole transforme un signal vocal humain en un texte qui peut être utilisé et analysé sur un ordinateur. Néanmoins, même avec le développement technologique d'aujourd'hui, les systèmes de reconnaissance automatique de la parole font toujours des erreurs pendant le traitement de la parole, surtout lorsqu'ils sont confrontés à diverses conditions acoustiques comme le bruit, les locuteurs multiples, les mots hors vocabulaire et les variations de prononciation (Errattahi, El Hannani, & Ouahmane, 2018).

La dégradation des performances des systèmes ASR, principalement lorsqu'ils sont utilisés avec des adultes âgés (qui présentent le plus souvent des difficultés perceptives, cognitives et physiques), a été constatée dans de multiples recherches (Aman, Vacher, Rossato, & Portet, 2012; Le Grand, Aman, Vacher, Rossato, & Portet, 2012; Tóth et al., 2018; Vipperla, Renals, & Frankel, 2010; Weiner et al., 2018; Young & Mihailidis, 2010). Ces mauvaises performances sont fondamentalement dues à la différence entre la parole d'une personne âgée et la parole standard utilisée pour entraîner les systèmes ASR, ce qui alourdit la reconnaissance et entraîne de mauvaises performances, notamment pour la parole spontanée.

Les systèmes ASR sont également affectés par des biais de conception, l'un des principaux est la langue ; l'anglais étant la langue la plus étudiée et la plus utilisée pour le développement d'outils d’apprentissage machine (AM), le niveau de maturité des modèles en anglais est

meilleur que dans les autres langues. De plus, comme le mentionnent Feng, Kudina, Halpern, et Scharenborg (2021), les systèmes ASR ont montré des déficiences avec les variations de la parole dues à des facteurs tels que le sexe, l'âge, les troubles de la parole, la race et les dialectes. Ce dernier point est dû au fait que les systèmes ASR sont généralement entraînés avec le discours de locuteurs d'une variante considérée comme « standard » dans la langue, ce qui signifie un biais pour les locuteurs d'une variante linguistique d'une région différente, comme le français européen et le français canadien.

En résumé, avec les progrès des modèles d'AM et du traitement automatique du langage naturel (TALN), un intérêt est apparu pour la création de méthodes de surveillance automatique des maladies cognitives à partir du langage. Par exemple, certaines méthodes utilisent les audios de TDI et des systèmes ASR pour examiner le déclin des capacités linguistiques.

Cependant, la mise en œuvre de ces technologies dans la pratique clinique présente des difficultés : 1) les modèles d'AM sont souvent entraînés sur des données transcrites manuellement par des humains, ce qui prend beaucoup de temps ; 2) il y a un manque d'ensembles de données pour plusieurs langues. Les modèles les plus performants ont été conçus pour l'anglais ; 3) la précision de l'ASR dans les voix vieillissantes diminue avec l'âge, car ils sont entraînés sur des voix jeunes à la diction claire et précise. Le discours des patients atteints de démence contient davantage d'incohérences, ce qui rend le discours moins intelligible.

Les applications qui se fondent sur la reconnaissance automatique de la parole, telles que la détection de maladies, peuvent être sérieusement affectées par toutes les lacunes de ces technologies aujourd'hui. Ce travail propose un modèle d'amélioration des transcriptions automatiques dans le contexte restreint de la description de l'image *Cookie Theft* de *Boston Diagnostic Aphasia Examination* (BDAE) dans le but de rendre plus fiable et moins coûteux le processus de génération d'ensembles de données à partir de tâches cognitives descriptives, notamment pour la langue française canadienne.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

Cette revue de la littérature a pour but d'aider le lecteur à bien comprendre les concepts développés dans cette étude. Ce chapitre est principalement divisé en deux sections, la première décrivant ce qu'est un système de reconnaissance de la parole et les défis qu'il doit relever avec des voix vieillissantes, la langue française et les personnes atteintes d'une forme de démence. La deuxième section se concentre sur la présentation des tâches de description d'images et du contexte restreint impliqué.

1.2 Systèmes de reconnaissance automatique de la parole

La reconnaissance automatique de la parole désigne la technologie qui reconnaît le langage parlé, sous la forme d'un signal vocal, et le convertit automatiquement en texte écrit (Li, Deng, Haeb-Umbach, & Gong, 2015).

Les systèmes ASR sont encore relativement imprécis et ne sont pas parfaitement résolus. Ils sont encore sujets à des erreurs qui se manifestent couramment par des fautes d'orthographe et des remplacements de mots, visibles dans le texte reconnu par le système (Bassil & Semaan, 2012).

1.2.1 Mesure de la performance

La performance d'un système ASR est généralement évaluée en calculant des mesures quantitatives telles que la précision, le score F1 ou le taux d'erreur de mots (WER, dans son acronyme anglais), ce dernier étant le plus courant (Errattahi et al., 2018).

Le WER correspond au nombre de mots incorrects divisé par le nombre total de mots corrects dans une transcription. En d'autres termes, il s'agit de la somme des substitutions (S), insertions (I) et suppressions (D) qui se produisent dans une séquence de mots reconnue, divisée par le nombre total de mots initialement prononcés (N1). Il est défini comme suit :

$$WER = \frac{S + D + I}{N_1} = \frac{S + D + I}{H + S + D} \tag{1.1}$$

D'après les ingénieurs de Microsoft Azure et Amazon Web Services, deux des principaux fournisseurs de services d'ASR au monde, un WER de 5 à 10 % est considéré comme excellent, un WER de 20 % est acceptable et un WER de 30 % ou plus est un signal de mauvaise qualité (Farley et al.; Seyfarth & Zhao, 2020).

Néanmoins, la précision des systèmes ASR varie au regard des caractéristiques de la personne qui les utilise. Par exemple, Hakkani-Tür, Vergyri, et Tur (2010) ont trouvé un WER de 34,1 % pour des locuteurs en bonne santé âgés de plus de 70 ans (effectuant un test cognitif fondé sur la parole) et un WER de 26,3 % pour ceux âgés de moins de 70 ans, compte tenu de la tâche et du sexe du locuteur. Inversement, le WER se situait entre 21,1 % et 28,2 % pour les locuteurs plus jeunes. Aman et al. (2012) ont observé une dégradation de leur système ASR lorsqu'il est utilisé avec des voix plus âgées avec une différence absolue de WER de 20,23 %, tandis que Vipperla et al. (2010) ont constaté que le WER sur des voix plus âgées est 10 % plus élevé en valeur absolue en comparaison à ceux des voix adultes. Peintner et al. (2008) ont obtenu un WER de 37 % chez les patients atteints de démence sémantique et de 61 % chez les patients atteints d'aphasie progressive non fluente, avec un groupe de contrôle présentant un WER de 20 % en moyenne. De son côté, Lehr, Prud'hommeaux, Shafran, et Roark (2012) ont signalé un taux d'erreurs de mots important, de l'ordre de 30 à 50 %, en expérimentant avec un discours de patients souffrant de troubles cognitifs légers.

1.2.2 Défis de la reconnaissance de la parole dans les voix vieillissantes

La reconnaissance automatique de la parole est un problème informatique ouvert dont le but est de reconnaître la parole humaine avec précision et de la convertir en texte qui peut être utilisé à des fins multiples (Bassil & Semaan, 2012). Cependant, la reconnaissance de la parole spontanée des voix vieillissantes est relativement délicate et représente l'un des principaux défis pour ce type de système (Vipperla et al., 2010).

Avec l'âge, plusieurs changements affectent le mécanisme de production de la parole humaine, limitant l'efficacité des systèmes ASR dans la reconnaissance des mots et des phrases prononcés par les personnes âgées. Bien que ces changements se produisent principalement dans le système respiratoire et la cavité orale, ces changements naturels entraînent une dégradation de certains aspects de la voix (Vipperla et al., 2010).

Premièrement, des tremblements involontaires et une fonte musculaire générale peuvent apparaître, entraînant une diminution de la puissance et du contrôle de certains gestes. Ensuite, l'élocution a tendance à devenir plus lente, les personnes structurant leurs phrases et cherchant leurs mots. Certaines pauses et hésitations durent plus longtemps que chez les personnes plus jeunes (Le Grand et al., 2012).

Aussi, Le Grand et al. (2012) mentionnent que le vieillissement entraîne une dégradation inévitable des fonctions cognitives chez les personnes âgées en bonne santé, surtout dans celles essentielles à la structuration du langage, à la prise de décision, à la mémoire de travail et à la mémoire à court terme. Cette dégradation fait que le discours devient plus ambigu et moins cohérent. Enfin, Young et Mihailidis (2010) ont résumé les caractéristiques de la voix d'une personne âgée en bonne santé, indiquant que les voix des personnes âgées présentent généralement une augmentation de la respiration et de la nervosité et une diminution de la vitesse d'élocution, de l'articulation et de l'intensité vocale.

La dégradation du discours est causée par l'âge et d'autres facteurs, tels qu'une condition médicale. Les résultats de l'ASR peuvent être davantage remis en question par la présence de troubles de la parole tels que ceux qui se produisent dans la MA. De multiples études (Fraser, Rudzicz, Graham, et al., 2013; Fraser, Rudzicz, & Rochon, 2013; König et al., 2015; Rudzicz, Wang, Begum, & Mihailidis, 2015; Shinkawa & Yamada, 2018; Tóth et al., 2018; Vipperla et al., 2010; You, Ahmed, Barr, Ballard, & Valenzuela, 2019; Young & Mihailidis, 2010) ont mentionné d'autres troubles du langage causés par une forme de démence, tels que des pauses fréquentes pour trouver les mots, l'utilisation de mots incomplets, l'invention de mots, le manque d'initiative, la lenteur, le manque d'expressivité, la diminution de la proportion d'adjectifs, la répétition de mots, de phrases et de sujets.

1.2.3 Défis de la reconnaissance de parole en français

Plusieurs chercheurs ont souligné les difficultés du français à être transformé en phonèmes et transcrit par les systèmes ASR (Adda-Decker & Snoeren, 2011; Boyer & Rouas, 2019; Brousseau et al., 1995; Carbajal, Bouchon, Dupoux, & Peperkamp, 2018; Dufour & Estève, 2008). Les principaux défis sont le phénomène de réduction de la parole, de nombreux homophones et lettres muettes existants (Tableau 1.1), l'élision, la liaison (qui peut subir une dénasalisation), la suppression de liquide, l'enchaînement, l'assimilation de voisement et l'argot.

La réduction de la parole désigne l'élimination de lettres et de sons dans le discours. Ce phénomène est très répandu en français parlé : on supprime des syllabes pour parler plus vite, par exemple, *je suis* /ʒə syi/ est souvent prononcé *j'suis* /ʃwi/, et *je ne sais pas* /ʒə nə sɛ pa/ peut être dit *je sais pas* (en omettant la clause ne) et prononcé comme /ʃɛ pa/, /ʃpa/ ou même /ʃpaː/ (en français canadien). Dans ces exemples, on observe également le phénomène d'assimilation de voisement, où le schwa dans je /ʒə/ s'élide et le /ʒ/ suivi de /s/ devient /ʃ/ pour former un son fricatif simple. Selon Adda-Decker et Snoeren (2011), la proportion d'erreurs dues à l'abrègement des mots et des phrases augmente dans le discours informel et conversationnel, principalement avec les marqueurs de discours : *tu sais*, *tu vois* ; les marqueurs de discours

rapporté : *il m'a dit*, *je lui ai dit* ; et les expressions fréquemment utilisées : *c'est pas*, *à cette heure*.

D'autres exemples de réduction de la parole sont la suppression des liquides et l'élision (Tableau 1.2). La suppression des liquides se produit lorsqu'un mot se terminant par un groupe liquide obstruant (une consonne occlusive /b, d, g, k, p, t/ suivie d'une consonne liquide /l, ʁ/) se trouve devant un mot initial consonantique, provoquant l'omission du son liquide, par exemple, *table marron* /tabl maʁɔ̃/ peut être prononcé comme /tab maʁɔ̃/. L'élision est une autre caractéristique du français, où la dernière voyelle de certains mots (p. ex., de, le, ne, que) est omise devant un autre mot commençant par une voyelle.

Les homophones sont plus fréquents en français qu'en anglais (Brousseau et al., 1995), principalement dérivés de formes de mots infléchies vers des variantes de procédure de genre et de nombre (Dufour & Estève, 2008), ce qui représente un grand défi pour la reconnaissance vocale. Comme mentionné par Adda-Decker et Snoeren (2011), il existe des systèmes ASR qui ont rapporté environ 10 % de WER pour le discours éduqué en français et plus de 15 % pour le discours téléphonique occasionnel ; ils ont également souligné qu'environ 30-40 % des erreurs dans les systèmes ASR testés en français avec le discours éduqué proviennent d'homophones ou de quasi-homophones. Le Tableau 1.1 montre quelques exemples d'erreurs courantes causées par des homophones, des élisions, des remplacements de nombre et de genre dans un discours soigneusement préparé.

La liaison apparaît en français parlé et constitue un autre défi majeur pour la reconnaissance vocale. Il s'agit d'un processus phonologique entre certains mots, se terminant par sept consonnes non prononcées (/z, t, n, r, p, v, k/), et le mot suivant commençant par une voyelle, entraînant la prononciation de la consonne muette. Les règles de liaison ne sont pas parfaites, et il est difficile de prédire si une consonne doit être prononcée ou non, car plusieurs facteurs tels que l'orthographe, la syntaxe et la sémantique interagissent. À l'instar de la liaison, il y a aussi l'enchaînement, c'est-à-dire la ré-syllabification de mots à consonne finale devant des

mots à voyelle initiale ; cela s'applique à tous les mots se terminant par une consonne, qu'ils soient ou non le résultat d'une liaison, par exemple :

Tableau 1.1 Homophones français courants confondus par les systèmes ASR.
Adapté de Feng et al. (2021); Ferrand (1999)

Référence	Prononciation	Hypothèse	Prononciation
a	/a/	à	/a/
sait	/sɛ/	c'est	/sɛ/
cesse	/sɛs/	seize	/sɛz/
rentre	/ʁɑ̃tʁə/	rendre	/ʁɑ̃dʁə/
ça avait	/saavɛ/	savaient	/savɛ/
semble que	/sɑ̃bləkə/	somme que	/sɔmkə/
d'avantages	/davɑ̃taʒ/	davantage	/davɑ̃taʒ/
file	/fil/	fil	/fil/
cents	/sɑ̃/	cent	/sɑ̃/
et	/e/	est	/ɛ/
qu'en	/kɑ̃/	quand	/kɑ̃/

Tableau 1.2 Exemples de phénomènes de liaison, d'élision et d'enchaînement

Phrase	Phonèmes	Syllabes	Résyllabification (elision et liaison)
une étoile	/yn etwal /	/y-n e-twal/	/y ne-twal/
un ami	/œ̃ ami/	/œ̃ a-mi/	/œ̃‿na-mi/

La dénasalisation est un autre effet causé par la liaison ; elle se produit avec un nombre limité d'adjectifs se terminant par /ɛ̃/ et /ɔ̃/ qui perdent le son nasal lorsqu'ils forment une liaison avec le mot suivant à initiale vocalique, devenant un son oral. Par exemple, *prochain* /pʁɔʃɛ̃/ sonne /pʁɔʃɛn/ devant une voyelle, comme dans *prochain arrêt* /pʁɔʃɛn aʁɛ/.

Enfin, les différences entre le français utilisé pour entraîner un système ASR (généralement celui de Paris), et la variante française de la personne qui occupe le système posent également une difficulté importante. Le français canadien diffère du français métropolitain de plusieurs façons ; Dumas (2013) en résume quelques-unes :

- Le français canadien a conservé des prononciations qui ont majoritairement disparu en France, comme les deux voyelles longues traditionnelles /ɛː/ et /aː/.
- Affrication des consonnes /t/ et /d/ devant les voyelles /y/ et /i/ devenant /ts/ et /dz/.
- Réduction du pronom il et elle en y et a, respectivement.

1.2.4 Détection et correction des erreurs

Les tâches de détection et de correction des erreurs sont cruciales dans les systèmes ASR. Il existe trois types d'erreurs qui se produisent lors de la génération de transcriptions : les substitutions, les insertions et les suppressions. (Pellegrini & Trancoso, 2010; Zhou, Shi, Feng, & Sears, 2005).

La détection et la correction sont deux processus qui, bien que liés, sont fondamentalement différents. La détection est le précurseur de la correction et cherche à détecter la zone erronée, soit pour exclure les erreurs de la transcription, soit pour trouver une stratégie corrective (Allauzen, 2007).

Errattahi et al. (2018) indiquent dans leur revue qu'il existe deux types de recherche concernant la détection des erreurs dans un système ASR, d'une part, il y a les auteurs qui utilisent uniquement les caractéristiques du décodeur du système comme les scores de confiance et les modèles de langage ; et d'autre part, il y a ceux qui utilisent en plus les caractéristiques de la parole comme les n-grammes, les parties du discours, la syntaxe et la sémantique.

Ringger et Allen (1996) ont indiqué dans leurs recherches que les systèmes ASR étaient de plus en plus encapsulés comme une boîte noire, c'est-à-dire comme des outils qui reçoivent des

entrées et renvoient des résultats bien définis, sans possibilité de modifier ou d'ajuster le fonctionnement interne du système pour améliorer les résultats, ce qui est actuellement vrai. Ceux qui accèdent au système ASR essaient de détecter les erreurs du décodeur pour améliorer le modèle acoustique et augmenter la précision de la reconnaissance vocale. Ceux qui utilisent une boîte noire doivent compter sur des méthodes de détection et de correction postédition (Rudnicky et al., 1994, cité dans Bassil & Semaan, 2012).

La correction des erreurs fait référence à l'ensemble du processus, y compris la détection. Comme l'indiquent Bassil et Semaan (2012), il existe quatre principaux types de techniques de correction d'erreurs :

1) Correction manuelle des erreurs : Un groupe de personnes qualifiées est chargé de réviser les transcriptions et d'effectuer les actions nécessaires pour corriger manuellement les substitutions, insertions et suppressions, parfois par rapport à l'audio original. Il s'agit d'une tâche laborieuse, coûteuse et sujette aux erreurs en vertu de l'œil et de l'oreille humains et des différences de critères entre les transcripteurs.

2) Correction d'erreur fondée sur une hypothèse alternative : L'erreur détectée est remplacée par une séquence de mots alternative, appelée hypothèse, qui est généralement dérivée d'un lexique et est susceptible d'avoir un taux élevé de mots hors vocabulaire. La correction est effectuée pendant le temps de reconnaissance.

3) La correction d'erreurs fondée sur l'apprentissage de formes : Comme son nom l'indique, l'objectif est de trouver des modèles qui sont considérés comme erronés. Le système est entraîné à trouver des erreurs dans les mots d'un domaine spécifique. L'un des inconvénients de cette approche est que des données annotées sont nécessaires pour entraîner le modèle, et comme il s'agit d'un modèle spécifique à un domaine, les erreurs/mots reconnus peuvent être minimes. La correction est effectuée au moment de la reconnaissance.

4) Correction des erreurs après l'édition : Dans cette approche, la sortie du système ASR est utilisée. Une fois que l'audio a été entièrement converti en texte, une couche supplémentaire est ajoutée, indépendamment du système ASR, pour détecter et corriger les erreurs dans la sortie texte finale, sans dépendance du système et sans modifier le modèle acoustique du système ASR.

Conformément à Errattahi et al. (2018), la plupart des recherches se concentrent uniquement sur la détection des erreurs et proposent des corrections possibles à appliquer manuellement ; très peu de recherches abordent le processus de correction. Zhou et al. (2005) précisent que les recherches actuelles sur la détection des erreurs de systèmes ASR se concentrent sur la transcription d'un discours spécifique à un domaine. À notre connaissance, il n'existe pas encore de recherche axée sur la correction des transcriptions automatiques de patients âgés canadiens francophones décrivant des images.

1.3 Tâches de description d'images

Les tâches de description d'images font partie des outils d'évaluation cognitive utilisés pour détecter les signes qui alertent sur la présence de maladies provoquant une démence. Par exemple, dans les TDI, on enregistre un discours semi-spontané dans un contexte restreint (CR) en demandant aux participants de décrire une scène en détail. L'image Cookie Theft est la tâche cognitive descriptive la plus courante (Boschi et al., 2017; Cummings, 2019).

1.3.1 L'image Cookie Theft

L'image *Cookie Theft* fait partie de l'examen diagnostique de l'aphasie de Boston (BDAE), publié pour la première fois en 1972 par Harold Goodglass et Edith Kaplan, puis révisé en 1983, et enfin republié en 2001 (Cummings, 2019).

Généralement, les images utilisées dans les tests cognitifs linguistiques sont en noir et blanc avec des scènes qui peuvent être familières dans le cadre socioculturel du patient. Par exemple, la scène de l'image *Cookie Theft* montre une mère et ses deux enfants dans la cuisine. La mère

semble distraite de son environnement alors que de nombreuses situations l'entourent. Elle fait la vaisselle et l'évier déborde. Deux enfants sont à l'arrière en train de voler des biscuits ; un garçon est sur un banc sur le point de tomber et prend des biscuits d'un jarre. La fille est debout, la main tendue, attendant de recevoir un biscuit.

Les chercheurs et les cliniciens utilisent cette image pour inciter à la production d'une parole connectée chez les patients souffrant d'un grand nombre de maladies neurodégénératives (Gross et al., 2010; Ash and Grossman, 2015; Ash et al., 2011, 2012, 2013; Drummond et al., 2015; Fraser et al., 2015; Tsermentseli et al., 2016; Wilson et al., 2010, cités dans Boschi et al., 2017). La liste des affections comprend le syndrome corticobasal, la variante comportementale de la démence frontotemporale, la variante sémantique de l'aphasie progressive primaire, les patients ayant subi des accidents vasculaires cérébraux de l'hémisphère droit et de l'hémisphère gauche, ainsi que les personnes âgées atteintes de troubles cognitifs légers et de la maladie d'Alzheimer (Agis et al., 2016; Ash et al., 2016; Kavé and Levy, 2003; Mueller et al., 2016, cités dans Cummings, 2019).

La tâche est généralement réalisée entre un enquêteur et le patient. On fournit aux patients une image avec une scène et on leur demande de la décrire. Idéalement, cette tâche nécessite peu d'instructions et une intervention minimale de l'intervieweur. Le rôle de l'enquêteur devient pertinent pour indiquer les zones négligées de l'image et demander au patient d'élaborer davantage la description, si nécessaire. En général, ces conversations sont enregistrées pour être examinées et notées ultérieurement.

Même si les tâches de description d'images sont souvent utilisées pour compléter une évaluation des capacités de parole dans le diagnostic des maladies neurodégénératives, elles ne sont pas considérées comme une mesure clinique valide du langage à cause du manque de données normatives avec des patients sains (Cummings, 2019). Par conséquent, de nombreuses recherches se sont attachées à proposer de nombreuses variables linguistiques pour analyser la production vocale ; cependant, la manière de comparer les données reste ambiguë du fait de la façon dont ces variables sont saisies et mesurées (Boschi et al., 2017).

1.3.2 Contexte restreint

Comme mentionné ci-dessus, la description d'une image suscite un discours semi-spontané, car le vocabulaire utilisé pour une telle description est restreint par les éléments présents dans la scène.

En général, le patient utilisera des phrases simples pour décrire les objets, les personnages et les situations de l'image, et de manière plus ou moins uniforme, ces éléments sont attendus dans toutes les descriptions de la même image pour un patient donné.

Pour l'image *Cookie Theft*, le contexte restreint est donné par des objets tels que : le banc, l'évier, le pot à biscuits, les vêtements des personnages, les meubles de la cuisine et une fenêtre. Les personnages (mère, fils et fille) interagissent avec ces objets pour créer des situations qui peuvent être décrites par des phrases telles que : « le garçon est sur le banc qui est sur le point de tomber », « la fille a la main tendue en attendant un biscuit », « l'évier déborde d'eau », « la mère essuie la vaisselle ».

Bien que le contexte de l'image soit restreint, il est possible que le patient sorte de ce contexte et décrive des situations qui ne figurent pas dans l'image. Dans l'un de nos ensembles de données, un patient commence sa description en disant : « on a fini de souper c'est le temps de la vaisselle, on se dépêche un peu parce qu'on a hâte d'aller écouter la télé ». Inévitablement, cependant, le patient finit par produire des phrases avec le vocabulaire du contexte restreint, par exemple « la femme essuie une assiette » ou « l'eau vient de déborder à ses pieds ».

Par conséquent, le discours obtenu par les patients peut être analysé sur la base des informations attendues du contexte restreint et être noté et comparé en comptant la quantité d'informations correctes qui apparaissent dans la description produite.

CHAPITRE 2

COMPUTER-BASED CORRECTION OF AUTOMATIC TRANSCRIPTIONS OF ELDERLY FRENCH-SPEAKING QUEBECERS

Eric Garcia-Cano[a], Sylvie Ratté[a], Simona Brambati[b]
[a]Department of Software and IT Engineering, École de technologie supérieure, 1100 Notre-Dame St W, Montreal, QC H3C 1K3, Canada
[b]Department of Psychology, Montreal University, Pavillon Marie-Victorin 90, Vincent d'Indy Avenue, Montreal, QC H2V 2S9, Canada

Article soumis pour publication au journal *Computer Speech and Language*. Décembre 2021.

2.1 Introduction and motivation

Dementia is a group of symptoms that severely affects a person's ability to perform daily activities. There are different types of dementia, with Alzheimer's disease (AD) being the most common in older adults, and accounting for 60-70% of all cases (Greenblat, 2021). While language impairment is known to occur early in the disease process (Hernández-Domínguez, García-Cano, Ratté, & Sierra, 2016), minimal attention has been paid to formal language assessment in diagnosing AD (Tóth et al., 2018).

A patient's spontaneous speech and language can be obtained through manual transcriptions and speech recordings. Narrative speech analysis is often done by hand and can be labor-intensive, causing it to become a bottleneck (Cummings, 2019; Fraser, Rudzicz, & Rochon, 2013). The use of automatic speech recognition (ASR) software could result in an analysis that highly efficient and widely available (Fraser, Rudzicz, Graham, et al., 2013). However, the speech of patients with dementia contains a great deal of inconsistencies, which make it less intelligible.

Even with today's technological developments, the performance degradation of ASR systems, particularly when used with older adults, has been noted in multiple research studies (Aman et al., 2012; Le Grand et al., 2012; Tóth et al., 2018; Vipperla et al., 2010; Werner, Huang, & Pitts, 2019; Young & Mihailidis, 2010). These systems also face strong criticisms for the biases present in most of them, even the state-of-the-art ones. ASR systems show deficiencies with speech variations due to factors such as gender, age, speech impairment, race, language, and dialects (Feng et al., 2021).

The above conditions and errors can significantly impact applications such as disease detection, which depend on automatic transcriptions. The present work proposes a model for improving automatic transcriptions within the restricted context of describing the Cookie Theft picture from the Boston Diagnostic Aphasia Examination. In addition, we aim to reduce the time needed to generate datasets from cognitive-descriptive tasks, particularly for the Canadian French language, in order to make the automatic transcription process cheaper and reliable, allowing faster analysis and diagnosis.

2.2 Automatic Speech Recognition Systems

Li et al. (2015) defined Automatic Speech Recognition as the process and technology that convert a spoken speech signal into its corresponding optimal sequence of words and linguistic entities using algorithms implemented on computers.

The performance of an ASR system is typically evaluated by calculating the Word Error Rate (WER), which is the sum of substitutions, insertions and deletions that occur in a recognized word sequence divided by the total number of words originally spoken (Errattahi et al., 2018). According to the documentation of two of the major speech-to-text service providers, Microsoft Azure and Amazon Web Services, a WER of 5% to 10% is considered to be excellent, a WER of 20% is acceptable, and a WER of 30% or more is a signal of poor quality (Farley et al.; Seyfarth & Zhao, 2020).

Spontaneous speech recognition for aging voices is relatively tricky and represents one of the main challenges for this technology (Vipperla et al., 2010). With age, speech tends to become slower, more ambiguous, and less coherent as people structure their sentences and search for their words (Le Grand et al., 2012; Vipperla et al., 2010). Many authors (Fraser, Rudzicz, Graham, et al., 2013; Fraser, Rudzicz, & Rochon, 2013; König et al., 2015; Rudzicz et al., 2015; Shinkawa & Yamada, 2018; Tóth et al., 2018; Vipperla et al., 2010; You et al., 2019; Young & Mihailidis, 2010) have mentioned additional speech disturbances caused by some form of dementia, such as the frequent use of incomplete words, word invention, and repetition of sentences and topics. Multiple studies have reported a WER ranging from 30% to 50% in older and speech-impaired patient automatic transcriptions (Aman et al., 2012; Hakkani-Tür et al., 2010; Lehr et al., 2012; Peintner et al., 2008; Vipperla et al., 2010).

With respect to French, multiple research studies have pointed to the difficulty in transforming it into phonemes and transcribing it with ASR systems (Adda-Decker & Snoeren, 2011; Boyer & Rouas, 2019; Brousseau et al., 1995; Carbajal et al., 2018; Dufour & Estève, 2008). The main challenges are the speech reduction phenomenon and the language's many homophones and silent letters.

Speech reduction refers to the elimination of letters and sounds in speech. e.g., *je suis* /ʒə syi/ (I am) is often pronounced *j'suis*/ʃwi/, and *je ne sais pas* /ʒə nə sɛ pa/ (I don't know) can be shortened to *je sais pas* and pronounced /ʃɛ pa/, /ʃpa/ or even /ʃpaː/ (in Canadian French).

Homophones are more common in French than in English (Brousseau et al., 1995), and are mainly derived from inflected word forms to procedure variants of gender and number, which represents a big challenge for speech recognition. As mentioned by Adda-Decker et Snoeren (2011), there are ASR systems that have reported about 10% WER for educated speech in French and over 15% for casual telephone speech. They also indicate that about 30-40% of the errors in French ASR systems tested with educated speech are derived from homophones or near-homophones.

Since ASR systems are prone to errors, as explained above, error detection and correction tasks are crucial. Detection and correction are two processes which, although related, are fundamentally different. Detection is the precursor to correction and seeks to identify the erroneous area, either to exclude errors from the transcription or to find some corrective strategy (Allauzen, 2007). Error correction for its part refers to the whole process, including detection.

Errattahi et al. (2018) mentioned that most of the research in the field concentrates only on detecting errors and suggests possible corrections to be applied manually, with very few works addressing the correction process. Zhou et al. (2005) specify that current research on ASR error detection focuses on transcribing a domain-specific speech. To the best of our knowledge, there is still no research focusing on correcting automatic transcriptions of Canadian French-speaking older patients describing images.

2.3 Picture description tasks

Picture description tasks are among the cognitive assessment tools used to detect signs that alert to the presence of diseases causing dementia, prompting a semi-spontaneous connected discourse with a simple structure of utterances guided by the restricted context (Boschi et al., 2017; Cummings, 2019).

Figure 2.1 shows the Cookie Theft picture, the most common cognitive-descriptive task. It is part of The Boston Diagnostic Aphasia Examination, first published in 1972 by Harold Goodglass and Edith Kaplan, and then revised in 1983, and finally republished in 2001.

The scene in the picture shows a mother and her two children in the kitchen. The mother seems disconnected from her surroundings, while many dangerous and unrealistic situations are happening. She is washing the dishes without much care, as the sink is overflowing. The two children are in the back stealing cookies; the boy is on a stool, which is about to fall over,

taking cookies from a jar, while the girl is standing with her hand outstretched, waiting to receive a cookie.

The picture description task is usually performed between an interviewer and the patient. The patient is provided an image with the scene and asked to describe it. Generally, these conversations are recorded for later examination and scoring.

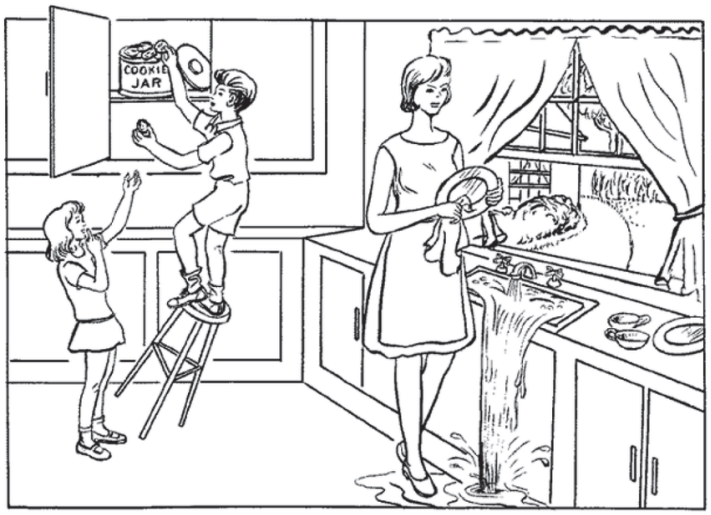

Figure 2.1 The Cookie Theft picture from the
Boston Diagnostic Aphasia Examination

According to Williams et al. (2010, as cited in Cummings, 2019), the descriptions of this image in healthy patients last between 45 seconds and 2 minutes, while in patients with some brain damage, the descriptions range from 13 seconds to 10 minutes.

Since this type of task contains domain-specific vocabulary (restricted context), the resulting discourse has predictable elements, i.e., key phrases or words found in the semantic fields of the image. For example, in the case of the Cookie Theft image, we expect that the discourse will contain subjects such as the mother, the boy and the girl; objects such as the dish, the stool and the sink; actions such as washing the dishes, overflowing or falling; and places such as the

kitchen or the garden. The speech obtained by patients can then be scored and compared by counting the number of correct information elements appearing in the description produced.

2.4 Method

Our approach to correct automatic transcriptions takes advantage of typical patterns of error occurrence in ASR systems. In the first pattern, the vocabulary of a transcript in a restricted context is relatively small. Misrecognized words are thus likely to appear consistently in similar sentences (surrounded by the exact context words). In the second pattern, ASR systems tend to make the same errors with the same words, and in the third, the errors generated by the ASR system are phonetically similar to the words that were originally uttered.

Our correction process consists of four stages. First, there is the construction of a phonemicized correction dictionary (correction model), comprising the most common vocabulary of the task, extracted from correct manual transcriptions. The second stage obtains the automatic transcription of a person's speech recording describing the Cookie Theft picture using a given ASR system. Subsequently, candidate phrases or words to be corrected are detected in the transcriptions. In the final stage, a phonetic comparison is made through a fuzzy search in the dictionary of corrections to apply the relevant corrections.

2.4.1 Datasets

For this work, we analyzed a dataset provided by the *Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal* (CRIUGM), which is composed of 55 Canadian French recordings of 26 healthy young individuals and 29 older patients who perform a description task by indicating everything that happens in the Cookie Theft picture. The goal is to identify whether people can point out and describe the various inconsistencies in the scene; for example, the sink overflowing, the mother washing the dishes but not paying attention, the children trying to steal cookies and the stool falling. All these phrases constitute the task vocabulary of the restricted context since anyone describing this image will use the same vocabulary to do so.

We also have the 55 manual transcriptions of the recordings in the CHAT transcription and coding format, a system created precisely to provide standardized annotations of audio transcriptions of face-to-face conversational interactions. CHAT stands for Codes for the Human Analysis of Transcripts. It is a powerful format with which third-party tools can track several annotation structures, compute automatic indexes, and analyze morphosyntax (MacWhinney, 2014). The duration of the audios in the Lingua corpus ranges from thirty-six seconds to four minutes and forty-five seconds, for a total of one hour and eleven minutes.

Table 2.1 shows a description of each subset.

Table 2.1 Description of the Lingua corpus

Subset	**Min duration**	**Max duration**	**Average duration**	**Total duration**
26 young healthy individuals	00:00:36	00:03:38	00:01:09	00:30:17
29 older patients	00:00:47	00:04:45	00:01:25	00:41:20

One goal was then to extract the specific vocabulary of the task, i.e., as many words and phrases as possible, that may occur during the description of this image. Since we did not have enough information in the Lingua dataset, we decided to augment the data using the manual transcriptions of the Dementia Bank dataset, which contains 551 English transcriptions of the description of the Cookie Theft picture. Table 2.2 shows a description of each subset in this corpus.

The manual transcriptions of the Dementia Bank corpus were automatically translated from English to French with the Google Cloud Translation API; these translated transcriptions were then normalized to have a similar language to the manual transcriptions in French from the Lingua corpus.

Table 2.2 Description of the Dementia Bank Pitt corpus (English)

Subset	Min duration	Max duration	Average duration	Total duration
243 healthy individuals	00:00:17	00:02:48	00:01:02	04:14:02
310 patients with dementia	00:00:23	00:04:28	00:01:16	06:32:56

We then applied one more process, which we call extension, in addition to the data augmentation described above. This process involves taking the most common vocabulary from the restricted context (obtained from the manual transcriptions) and generating a new vocabulary using synonyms. For example, if we have the following 5-grams:

- *le garçon sur le tabouret* (the boy on the stool)
- *le tabouret avec le pot* (the stool with the pot)

and a list of synonyms for the most common words, such as *garçon*, *tabouret* and *pot*.

- *garçon → petit-garçon, frère*
- *tabouret → banc*
- *pot → jarre, boîte*

We can generate ten new 5-grams from the two original ones:

- *le garçon sur le banc*
- *le petit-garçon sur le tabouret*
- *le petit-garçon sur le banc*
- *le frère sur le tabouret*
- *le frère sur le banc*
- *le tabouret avec la jarre*

- *le tabouret avec la boîte*
- *le banc avec la jarre*
- *le banc avec le pot*
- *le banc avec la boîte*

To run our experiments, we divided the Lingua dataset into a development set with 80% randomly chosen manual transcripts (44) and an evaluation set with the remaining 20% (11). From the development set, we generated four variants to build four correction dictionaries:

1) Lingua: Only with the vocabulary extracted from the original 44 manual transcriptions in French.
2) Lingua augmented: The data augmented with the vocabulary from the translated Dementia Bank manual transcriptions.
3) Lingua extended: The vocabulary extracted from the 44 manual transcriptions plus the vocabulary extension with synonyms.
4) Lingua augmented + extended: The vocabulary of the original transcriptions in French and the translated transcriptions from English with the extension of the synonyms.

The processes of transcription normalization and extraction of the most common vocabulary from the restricted context will be discussed in more detail in the following sections.

2.4.2 Correction dictionary

This is the first stage of our method, the pipeline (Figure 2.2) for preparing the correct manual transcriptions to build our automatic transcription correction model. Essentially, we took the 44 development CHAT files and extracted the transcriptions to normalize them.

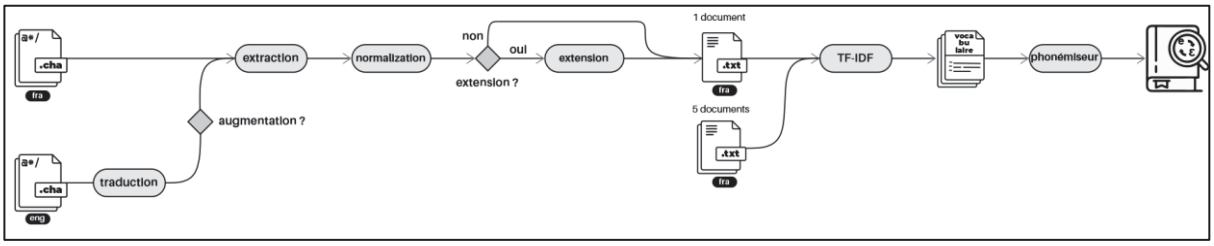

Figure 2.2 Correction dictionary creation pipeline

We also took the manual transcriptions from the Dementia Bank and translated them. After translation, we corrected words and phrases that were taken too literally from English and converted them for French words and phrases that are more commonly used. For example:

- *jeune garçon* (young boy) → *petit garçon* (little boy)
- *jeunes* (young people) → *enfants* (children)
- *sécher la vaisselle* (dry the dishes) → *essuyer la vaisselle* (wipe the dishes)

In the end, we obtained a single document that containing the text from all the manual transcriptions, to which we applied the following normalizations:

- lowercase conversion
- deletion of empty words (mostly interjections)
- deletion of numbers
- deletion of punctuation marks
- deletion of left and right trailing spaces
- deletion of multiple empty spaces

For this work, we did not remove stop words and filled pauses (hum, uh, eh), which are usually the first words to be removed in a text normalization, and there was a good reason for this. One of the goals of this work was to generate better quality automatic transcripts to be employed in machine learning models to facilitate the early diagnosis of dementia or aid in its follow-up.

Certain words are especially important in identifying signs of disease in speech, depending on the type of dementia. For example, patients with semantic dementia produce proportionally fewer nouns and more verbs and pronouns. In contrast, patients with progressive nonfluent aphasia tend to omit functional words, such as determiners and auxiliaries (Fraser, Rudzicz, & Rochon, 2013). In addition, patients with the logopenic dementia variant present an increase of filled pauses and false starts (Boschi et al., 2017). Since these and other morphosyntactic features are relevant in machine learning models, we decided not to eliminate these words to allow the generation of useful transcriptions.

After normalization, we compared our single document (containing all development manual transcriptions) with other documents (5) with information from completely different contexts to highlight and extract the specific vocabulary of the task, using the term frequency-inverse document frequency (TF-IDF) method.

This method, by definition, allows to obtain the importance of a term in a document and adjust it for its importance for all the documents in the corpus. To achieve this, it measures how many times a term appears in a document among the number of times all the terms appear in that document (TF). It then calculates how common a term is in all the analyzed documents (IDF), obtaining an importance score of a term among all the documents in the corpus, eliminating terms that are too common and selecting those that are more descriptive.

After the TF-IDF, we obtain n-grams (1-gram to 5-grams), all of the Cookie Theft description task's most common words and sentences, according to our development manual transcriptions.

Finally, we get the phonemes of each n-gram with the International Phonetic Alphabet (IPA) representation and create a correction dictionary (Table 2.3). These steps are repeated in each of the development datasets defined in section 2.4.1. Table 2.4 lists the number of common vocabulary (n-grams) obtained from each data set.

With the extracted n-grams, i.e., the vocabulary of the task, we will create a dictionary of corrections. The objective of this dictionary is to find, in the automatic transcriptions, possible erroneous sentences which sound the same as the sentences or words we expect to find in this task.

To search for these similar-sounding sentences, it is necessary to obtain the phonetic transcription of the n-grams in order to perform a fuzzy search based on the sound that the words make and not on their spelling. We proceed as such because the ASR system interprets sounds and not words. Thus, we try to find the sounds that ASR has confused and correct them with the proper words.

Table 2.3 Examples of task-specific vocabulary extracted from the correct manual transcriptions

1-grams		**2-grams**		**3-grams**	
Text	**Phonemes**	**Text**	**Phonemes**	**Text**	**Phonemes**
biscuits	biskyi	jarre à	ʒaʁ a	fait la vaisselle	fɛ la vɛsɛl
déborde	debɔʁd	le tabouret	lə tabuʁɛ	la mère est	la mɛʁ ɛ
garçon	gaʁsɔ̃	son évier	sɔ̃n evje	il va tomber	il va tɔ̃be
tabouret	tabuʁɛ	la maman	la mamɑ̃	l'évier déborde	levje debɔʁd

To store the correction dictionary, we use the fuzzy-set data structure created, described, and implemented by Chiacchieri, Axiak, et Altenhoff (2018). It stores all the phonetized n-grams and will be used for phonetic comparison, as explained in section 2.4.4.

Table 2.4 Number of n-grams extracted from the manual development transcriptions to generate the four correction dictionaries

n-gram	Lingua	Lingua extended	Lingua augmented	Lingua augmented + extended
1	455	472	455	472
2	1438	2079	6086	8262
3	2852	3991	7748	11000
4	3794	5124	8774	11993
5	4306	5736	9462	12147
Total	12845	17402	32525	43874

2.4.3 Error detection

Once we have the correction dictionary, we need to proceed with the correction pipeline. In short, we first get the automatic transcription using an ASR system, which usually results in a CSV file.

For this project, we tested the performance of two of the leading commercial ASR systems, Google Cloud Speech-to-Text and Azure Speech-to-Text, in addition to the Applied Research Center and Expertise Centre for Information Technologies (CRIM, its French acronym) system called Vesta.

We performed the automatic transcription of the 55 audios from the Lingua corpus and calculated the average WER. The ASR system with the best performance for our task and our corpus was Azure. We chose the system for our error detection and correction experiments since it has a significant WER, which needs correction. Still, it is not a completely erroneous transcription, like those obtained with Vesta.

Table 2.5 shows the results, and as expected, the performance of these systems was poor, following the standards for an ASR system explained in section 2.2.

The ASR system with the best performance for our task and our corpus was Azure. We chose the system for our error detection and correction experiments since it has a significant WER, which needs correction. Still, it is not a completely erroneous transcription, like those obtained with Vesta.

Table 2.5 Results of ASR systems evaluation

ASR system	Avg. WER
Azure	38.4%
Google Cloud	50.9%
Vesta	93.9%

We applied the same normalizations we use in the manual transcriptions to create the dictionary to the automatic transcriptions, thus obtaining the automatic transcription to be corrected. We transcribed the 55 audios (development and evaluation) and performed the error detection and correction experiments for both sets.

To obtain the candidate words or phrases to be corrected, we again used the TF-IDF method for the same reasons explained in the previous section, but this time, comparing the automatic transcription with the manual transcriptions to highlight the out-of-context vocabulary in the automatic transcription.

A complementary method we used to aid in error detection consisted in constructing a bag of words, with all the words from the correct transcriptions of the development set (44), such that for automatic transcriptions, all n-grams containing at least one word not found in the bag of words were flagged as a possible error.

Thus (Figure 2.3), we would get n-grams that were probably wrong and that had to be checked against the correction dictionary.

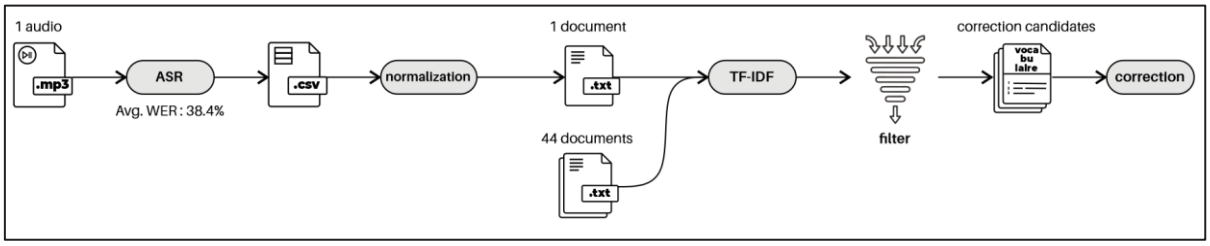

Figure 2.3 Automatic transcriptions error detection pipeline

2.4.4 Phonetic comparison and correction

Once the sentences or words that are probably erroneous are obtained from the automatic transcription, we obtain their phonetic representation. We then proceed to check in the correction dictionary to see if there is a sentence or a word that has a sound that is similar to what we are trying to correct. Section 2.4.2 mentioned that the dictionary is a fuzzy-set data structure that stores phonetic strings and efficiently calculates similarity scores to find approximate sound matching. Thus, the idea is to search for possible matches in the dictionary and sort the matches according to a similarity score.

The fuzzy-set uses a cosine similarity, whose formula is:

$$similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} \tag{2.1}$$

To store the terms of the correction dictionary and subsequently calculate similarities, the fuzzy-set converts each term into a vector using 3-grams and counting how many times each 3-gram appears in the term, as shown in Table 2.6.

The fuzzy-set then calculates the magnitude $\|A\|$ by taking the square root of the sum of squares of the vector, generating a list containing pairs in the format ($\|A\|$, term), for example:

(3.464, ‘ʒaʁ a biskyi’) and (3.873, ‘levje ki debɔʁd’). The vectors generated above are also stored in the data structure, with respect to the position of the term in the list, and therefore, we know that the vector in Table 2.6 corresponds to the first element of the list, which is (3.464, ‘ʒaʁ a biskyi’). These structures are collectively referred to as the reverse index.

Table 2.6 Transformation of the term
jar à biscuits (cookie pot) /ʒaʁ a biskyi/
to vector

Grams	Count	Count^2
-ʒa	1	1
ʒaʁ	1	1
aʁ_	1	1
ʁ_a	1	1
_a_	1	1
a_b	1	1
_bi	1	1
bis	1	1
isk	1	1
sky	1	1
kyi	1	1
yi-	1	1
	Sum:	12
	$\|\|A\|\| = \sqrt{sum} =$	3.464

To find the best correction suggestions for a candidate, we start from the candidate’s phonetic representation (query string) and take its 3-grams to perform a reverse index search in the dictionary. First, a list is created with all the dictionary terms that contain at least one occurrence of the 3-grams of the query string. Second, for each matched term, the Levenshtein similarity between that term and the candidate is calculated. Finally, the matching terms are sorted by similarity, and we keep the one with the best score.

We repeat this process for each candidate to generate a list of candidates and their possible corrections, sorted by similarity scores (Table 2.7). Once this list of corrections has been completed, we automatically apply only the ones whose similarity scores are equal to or greater than 80%, obtaining the final corrected transcription.

Table 2.7 Table of candidates to be corrected with suggested corrections

Candidate	Candidate phonemes	Suggested correction	Suggested correction phonemes	Similarity
la petite fille a tant	la pətit fij a tɑ	la petite fille attend	la pətit fij atɑ	1.0
la lpo de biscuit	la lpo də biskyi	le pot de biscuits	lə po də biskyi	0.937510
tabouret sur le clavier	tabuʁɛ syʁ lə klavje	tabouret sur lequel il	tabuʁɛ syʁ ləkɛl il	0.904761
il déband	il debɑ	il déborde	il debɔʁd	0.888888
pixie une assiette	piksi yn asjɛt	puis une assiette	pyiz yn asjɛt	0.857142

2.5 Results and discussion

As explained in section 2.4.1, we generated four development datasets, with which we built four different correction dictionaries, following the process explained in section 2.4.2.

For the 55 audios of the Lingua corpus, we obtained an automatic transcription using Azure Speech-to-Text. We then performed the error detection and correction process for each automatic transcription, as explained in sections 2.4.3 and 2.4.4, four times.

With the error detection method explained in section 2.4.3, we detected a total of 820 errors in the evaluation sets (656 true errors (true positives) and 164 false errors (false positives)),

obtaining a detection precision of 80%. When running error detection on the 44 transcriptions of the development sets, we detected a total of 4942 errors, of which 4821 were true errors and 121 were false errors, yielding a detection precision of 97.5%.

Table 2.8 shows the results of the automatic correction for the 44 ASR transcriptions of the audios in the development set. As can be seen, the WER improvement is superior to the overall results shown in Table 2.9. This outcome was particularly expected because, even though the output of an ASR system is unpredictable and we do not know in advance the amount and type of errors that may appear, the correction dictionaries were built with sentences found in the transcribed audios.

Table 2.8 Results of the correction process, on the 44 automatic transcriptions of the audios whose manual transcriptions were used to build the correction dictionaries (development set)

Dataset	Avg. Initial WER	Avg. Final WER	Avg. Improvement
Lingua (DEVEL)	39.5%	37.2%	2.3%
Lingua extended (DEVEL)	39.5%	36.6%	2.9%
Lingua augmented (DEVEL)	39.5%	35.1%	4.4%
Lingua augmented + extended (DEVEL)	39.5%	33.7%	5.8%

Table 2.8 also shows that techniques used to artificially increase the dataset play a role, as the average improvement increases as the vocabulary in the correction dictionary increases. Thus, the dataset augmented with translations from the Dementia Bank corpus and extended with synonyms is the best.

Table 2.9 shows the results of the automatic corrections of the 11 ASR transcriptions of the audios in the evaluation set. The results do not compare with those obtained in the previous exercise. However, it must be considered that the dataset is too small – only one hour long – with recordings lasting from thirty-six seconds to four minutes and forty-five seconds (see

section 2.4.1). Nevertheless, consistent with the results obtained previously, we observe that the WER improves as the vocabulary in the correction dictionary increases, with the augmented and extended dictionary performing best.

Table 2.9 Results of the automatic correction process, on the 11 ASR transcriptions of the audios whose manual transcriptions were not used to build the correction dictionary (evaluation set)

Dataset	Avg. Initial WER	Avg. Final WER	Avg. Improvement
Lingua (EVAL)	34.0%	35.0%	-1%
Lingua extended (EVAL)	34.0%	34.8%	-0.8%
Lingua augmented (EVAL)	34.0%	33.3%	-0.3%
Lingua augmented + extended (EVAL)	34.0%	33.7%	0.3%

By analyzing the results of the automatic correction of each transcription in the evaluation set, using each correction dictionary; in our best model (Lingua augmented + extended), the WER improves by 45% of the transcriptions, does not change in 18%, and increases in 36%.

We identified two main reasons explaining the results in our evaluation set. The first was that we sometimes had several correction suggestions with the same similarity score for a correction candidate. The algorithm could not discern the best replacement. For example, for the following correction candidate, we had three options with an 81% phonetic similarity using Levenshtein:

pixie une assiette /piksi yn asjɛt/ (*pixie* a dish) :

→ *essuie une assiette* /esyi yn asjɛt/ (wipes a dish)
→ *cassé une assiette* /kase yn asjɛt/ (broke a dish)
→ *puis une assiette* /pyiz yn asjɛt/ (then a dish)

“pixie” is not a French word, but it was the sound that the ASR recognized. The correct choice should be the last one, but it will not be applied since it appears last in the list. The application of any of the other three options will increase the WER.

The second reason is that it is necessary to have a larger amount of data available to create the correction dictionary and to have a larger number of relevant n-grams to make corrections. For example, the following error *la polaire* /la pɔlɛʁ/ should be corrected to *n’a pas l’air* /na pa lɛʁ/ (he/she doesn’t look). Still, we have no n-gram phonetically similar to *n’a pas l’air* in the correction dictionary because this expression was not uttered in the development transcriptions we use. The best suggestion for this error in our dictionary is *la porte* /la pɔʁt/ (the door), with a similarity of 87%; nevertheless, it is not a good correction.

Knowing the limitations of our dataset, we decided to perform some semi-automatic correction exercises on the ASR transcriptions of the evaluation set. We detect errors automatically and present a user, the context, and the list of candidates to correct along with the suggestions from our best correction dictionary (as described in sections 2.4.3 and 2.4.4), with the user deciding which corrections to apply. As such, we avoid making automatic erroneous corrections, even if the similarity score is high, as in the examples mentioned above. Table 2.10 shows that, with our model and using the semi-automatic correction approach, we managed to improve the WER by up to 6.4% for one of the transcriptions in the evaluation set and obtained an average WER improvement of 1.9%.

Table 2.10 Results of the semi-automatic correction process on the evaluation set transcriptions

Lingua augmented + extended			
Transcription	**WER_i**	**WER_f**	**Improvement**
1	27.7%	26.5%	1.2%
2	32.9%	31.7%	1.2%
3	40.2%	37.8%	2.4%
4	34.1%	31.8%	2.4%
5	58.8%	57.8%	1.1%
6	45.7%	43.4%	2.3%
7	20.6%	18.7%	1.9%
8	37.9%	36.9%	1.0%
9	31.4%	30.6%	0.8%
10	27.7%	21.3%	6.4%
11	16.8%	16.2%	0.6%
Average	**34.0%**	**32.2%**	**1.9%**

2.6 Conclusions

From the results described above, this automatic transcription correction model in a restricted context is promising. Furthermore, the results obtained by correcting the automatic transcriptions (whose output is completely unpredictable) of the audios used to create the correction dictionary indicate that improving the WER of the transcriptions can be done by having good and relevant correction terms.

This is also true for the evaluation dataset if we take the correction suggestions and manually apply relevant ones. The proposed mechanisms for extracting correction terms and identifying erroneous areas are adequate. However, a larger amount of data is needed to improve the WER of automatic transcriptions that have not been seen before.

While it is true that our aim is to have an automatic transcription correction model that will allow us to generate faster and more reliable datasets, it is also true that our model can be used as a tool to detect erroneous areas and suggest corrections to be applied manually by the user.

To our knowledge, this is the first research addressing the correction of automatic transcriptions in a restricted context of Canadian French-speaking older patients. This is therefore our most important contribution, and we hope it will lay the groundwork for future research.

CHAPITRE 3

DISCUSSION

Notre modèle de correction des transcriptions automatiques tente de tirer parti de deux constats sur les erreurs commises par les systèmes ASR. Le premier est que le même système ASR fera presque toujours la même erreur avec le même son ; et le second est que les erreurs du système ASR sont en général phonétiquement similaires aux mots qui ont été initialement prononcés. De même, nous profitons du fait que le discours semi-spontané produit par la description d'une image génère du vocabulaire dans un contexte restreint.

En analysant individuellement les erreurs trouvées par notre modèle et les corrections suggérées et effectuées, nous pouvons comprendre les résultats obtenus. Comme mentionné précédemment, la combinaison de plusieurs facteurs influence les erreurs générées par le système ASR.

Ces erreurs se réduisent pratiquement à une substitution entre homophones. Les homophones en français sont courants et principalement dus au genre et au nombre des noms et à la conjugaison des verbes. Cependant, le phénomène de réduction de la parole, essentiellement causé par les lettres muettes, la liaison, l'élision et le voisement, fait que les sons combinés de deux ou plusieurs mots (réduits) deviennent un homophone ou quasi homophone.

3.1 Homophones causés par le phénomène de réduction de la parole

Certains des cas les plus intéressants de détection et de correction d'erreurs se sont produits au vu des homophones à cause de la réduction de la parole, entre autres :

la *lpo* de biscuits /la lpo də biskyi/	→	*le pot* de biscuits /lə po də biskyi/
en train *de bordel* robinet /ɑ̃ tʁɛ̃ də bɔʁdɛl ʁɔbinɛ/	→	en train *de déborder le* robinet /ɑ̃ tʁɛ̃ də debɔʁde lə ʁɔbinɛ/
que les distraits /kɛl ɛ distʁɛt/	→	qu'elle est distraite /kɛl ɛ distʁɛt/
quatorze heures ajouter /kwatɔʁz œʁz aʒute/	→	*que tu as autre chose à r*ajouter /kə ty az otʁ ʃoz a ʁaʒute/
pixie une assiette /piksi yn asjɛt/	→	*puis‿u*ne assiette /pyiz yn asjɛt/

Les trois premiers résultent de l'élision du e caduc d'un monosyllabe (le, que, de), qui, combiné au son du mot précédent ou suivant, forme un homophone. La quatrième erreur provient d'une combinaison de réduction informelle du discours et la vitesse à laquelle la personne parle, qui a en fait énoncé « qu't'as autre chose à rajouter » /kwtaz otʁ ʃoz a ʁaʒute/. Le dernier cas est dû à l'effet de liaison, mais comporte une composante dialectale, car le mot puis /pyi/ en français canadien se prononce /pi/ qui, lorsqu'il est combiné avec le phonème /z/ et les conditions acoustiques de l'enregistrement, le système ASR génère le mot pixie, qui, soit dit en passant, n'est pas un mot français.

D'autres erreurs sont également dues au dialecte canadien du français, qui, combiné aux conditions acoustiques du locuteur et de l'enregistrement, a produit des phrases qui n'ont pas pu être efficacement corrigées, notamment :

- « à la polaire » /a la pɔlɛʁ/ qui devrait être « elle a pas l'air » /ɛl a pa: lɛʁ/
- « belly fenêtres sont ouvertes » /bɛlɪ fənɛtʁ sɔ̃t uvɛʁt / qui devrait être « puis les fenêtres sont ouverts » /pi le fənɛtʁ sɔ̃t uvɛʁ/
- « maison belge a dit » /mɛzɔ̃ bɛlʒ a di/ qui devrait être « maison bien il y a des » /mɛzɔ̃ bɛn il i a de/

Dans le premier cas, le locuteur a produit la phrase (en français québécois) « elle a pas l'air » que le système ASR, probablement entraîné au dialecte français parisien, détecte comme « à la polaire », principalement en vertu de la différence de prononciation du mot « pas » et a l'omission de la clause « ne ». Dans les deux autres cas, quelque chose de similaire se produit, compte tenu des différences de prononciation des mots « puis » et « bien » /bjɛ̃/ ou /bɛ̃/ ; dans ce dernier avec une possible dé-nasalisation du son /ɛ̃/.

3.2 Homophones causés par le genre, le nombre et la conjugaison

Une grande partie des erreurs détectées par notre modèle provient des homophones en genre et en nombre des noms et de la conjugaison des verbes.

Deux des mots importants du contexte restreint ont des homophones qui ont été correctement détectés et corrigés à plusieurs reprises et de différentes manières, c'est le cas du nom « mère » et du verbe « déborder ». Le nom « mère » /mɛʁ/ est souvent remplacé par le nom « mer » /mɛʁ/ :

la mer semble /la mɛʁ sɑ̃bl/	→	la mère semble /la mɛʁ sɑ̃bl/
la mer est vraiment /la mɛʁ ɛ vʁɛmɑ̃/	→	la mère est vraiment /la mɛʁ ɛ vʁɛmɑ̃/
la mer fait la vaisselle /la mɛʁ fɛ la vɛsɛl/	→	la mère fait la vaisselle /la mɛʁ fɛ la vɛsɛl/

Pour le verbe « déborder », différentes variantes ont été trouvées, toutes bien corrigées :

il débande mais /il debɑ̃d mɛ/	→	il déborde mais /il debɔʁd mɛ/
son évider board pendant /sɔ̃n evide bɔʁd pɑ̃dɑ̃/	→	son évier déborde pendant /sɔ̃n evje debɔʁd pɑ̃dɑ̃/

des bordes de l'évier → débordait de l'évier
/de bɔʁd də levje/ /debɔʁdɛ də levje/

Voici deux autres exemples des verbes qui ont été mal interprétés par des homophones incorrects, mais corrigés :

qui t'a qui tend la main → qui tend qui tend la main
/ki ta ki tɑ̃ la mɛ̃/ /ki tɑ̃ ki tɑ̃ la mɛ̃/

la petite fille a tant → la petite fille attend
/la pətit fij a tɑ̃/ /la pətit fij atɑ̃/

Au sujet des homophones causés par le genre et le nombre des noms, ils constituent la principale source de faux positifs dans la détection des erreurs. Notre méthode de détection repose entièrement sur le vocabulaire contenu dans l'ensemble de développement. Ainsi, une phrase contenant un mot pluriel, singulier, masculin ou féminin qui n'a pas été prononcé dans cet ensemble sera potentiellement signalée comme une erreur. Les mots suivants (Tableau 3.1) sont des exemples d'homophones incorrects générés par l'ASR et signalés comme des erreurs par notre modèle :

Tableau 3.1 Homophones incorrects générés par le système ASR

Mots générés par l'ASR	Mots dans le dictionnaire de correction
tabourets	tabouret
frères	frère
poignets	poignées
nettoyées	nettoyés
cachées	cachés
dégâts	dégât
étagères	étagère
panneaux	panneau

Dans certains cas, le système ASR génère le mauvais mot, par exemple, dans « **tabourets** va se renverser » où, de toute évidence, la forme singulière serait correcte ; cependant, il y a d'autres cas où une phrase correcte est générée, comme « les **panneaux** de l'armoire », donnant lieu à un faux positif, qui ne devrait pas être corrigé, car il est correct.

3.3 Erreurs dues à des caractéristiques discursives

Enfin, il existe un autre groupe d'erreurs très complexes dues aux caractéristiques de la voix et du discours des personnes. Outre les défis imposés par la langue française aux systèmes ASR, l'intensité vocale, la vitesse de la parole et l'articulation jouent un rôle très important dans la reconnaissance vocale. Voici des exemples d'erreurs dans lesquelles tous ces facteurs sont impliqués :

- qu'elle s'entraîne essuyer /kɛl sɑ̃tʁɛn esyije/
- tabouret sur le clavier /tabuʁɛ syʁ lə klavje/
- un tabou gasquet pour /œ̃ tabu gaskɛ puʁ/
- à biscuits re jar /a biskyi ʁə ʒaʁ/
- les armoires à madrid /lez aʁmwaʁz a madʁid/
- e examen /ə ɛgzamɛ̃/

En écoutant l'audio du premier cas, la phrase correcte est « qu'elle est en train d'essuyer » /kɛl ɛt ɑ̃ tʁɛ̃ desyije/, cependant, il est clair que la personne prononce la lettre s dans le mot « est » ce qui produit la substitution que l'ASR a faite.

Dans le deuxième cas, la phrase prononcée était « tabouret sur lequel il » /tabuʁɛ syʁ ləkɛl il/, dans le troisième, c'était « un tabouret qui est pas » /œ̃ tabuʁɛ ki ɛ pa/ et dans le quatrième « à biscuits un jar » /a biskyiz œ̃ ʒaʁ/. Pour les trois, les locuteurs sont des adultes âgés avec des caractéristiques évidentes d'une voix vieillissante et qui, à la place de l'erreur, baissent l'intensité de leur discours.

Au contraire, les deux derniers cas ont été trouvés dans des discours de jeunes locuteurs qui, à l'endroit de l'erreur, font de la réduction excessive de la parole, ce qui génère des erreurs pratiquement non corrigibles. L'avant-dernière phrase devrait être « les armoires à l'intérieur » /lez aʁmwaʁz a lɛ̃teʁjœʁ/ et la dernière phrase, le cas le plus extrême, « euh je veux dire ouais » / œ ʒə vø diʁ uɛ/.

CONCLUSION ET RECOMMANDATIONS

Bien que les systèmes de reconnaissance vocale fassent partie de notre quotidien, ce sont des outils qui commettent encore de nombreuses erreurs, ce qui empêche leur utilisation intensive dans le domaine scientifique. L'objectif principal de cette recherche était de créer un modèle informatique de correction automatique de la transcription pour les francophones effectuant une tâche de description d'image. Ceci afin de réduire le temps nécessaire pour générer des ensembles de données fiables pouvant être utilisés dans différents domaines scientifiques qui nécessitent des transcriptions pour réaliser des expériences d'analyse et de compréhension des données, réduisant ainsi le goulot d'étranglement qui existe encore en vertu de la dépendance aux transcripteurs humains. C'est le cas des études pour le diagnostic précoce de la maladie d'Alzheimer et d'autres types de démence, pour lesquelles notre travail pourrait être d'une grande aide à l'avenir.

Notre modèle se compose de deux pipelines, l'un chargé de générer un dictionnaire de corrections à partir de transcriptions manuelles correctes, auquel une normalisation est appliquée, afin d'extraire le vocabulaire le plus pertinent du contexte. Pour l'élaboration de ce dictionnaire de corrections, nous utilisons les ensembles de données du Lingua et de *Dementia Bank*.

Le deuxième pipeline est celui de la détection et de la correction des erreurs. Avec notre méthode, nous avons obtenu une précision de détection de 80 %. Tel que mentionné par Adda-Decker et Snoeren (2011), la plupart des erreurs que nous avons identifiées sont dues à des homophones, que ce soit du fait du genre et du nombre des noms, des conjugaisons des verbes ou des phénomènes de réduction de la parole. Les tendances qui suivent les systèmes ASR, expliquées au début de la discussion, sont confirmées par le pourcentage d'erreurs correctement signalées.

Pour corriger les erreurs détectées, nous disposons de deux méthodes, l'une automatique et l'autre semi-automatique. La première applique la meilleure suggestion de correction à chaque erreur détectée et la seconde permet à l'utilisateur, compte tenu du contexte, de décider les suggestions à appliquer. Nous avons obtenu les meilleurs résultats avec la deuxième approche, en obtenant une amélioration moyenne du WER de 1,9 % et, dans certains cas, jusqu'à 6,4 %.

Notre mécanisme de correction automatique doit encore être amélioré, car il essaie toujours de corriger toute phrase marquée comme une erreur, puisqu'il ne dispose d'aucun mécanisme pour décider si une suggestion de correction doit être appliquée ou non. Ceci, ainsi que la quantité limitée de données, sont sans doute les raisons de la faible performance de cette méthode.

En guise de réflexion, nous aimerions mentionner que, même si la plus ancienne recherche consultée dans cette étude sur les défis des systèmes ASR pour les voix âgées et francophones date de 10 ans, tous les problèmes mentionnés dans la littérature depuis lors sont toujours valables. C'est pourquoi il est impératif de disposer de modèles de correction et de détection des erreurs plus nombreux et plus performants pour ces systèmes.

4.1 Travaux futurs

Afin d'améliorer les résultats de notre modèle de correction automatique, plusieurs pistes pourraient être pertinentes. En premier lieu, il est nécessaire d'augmenter l’ensemble de données français, afin d'accroître le vocabulaire du dictionnaire de correction.

Par ailleurs, il est conseillé de modifier la méthode d'extension des données, afin non seulement d'utiliser des synonymes, mais également d'augmenter artificiellement les données du dictionnaire de correction en variant les noms en genre et en nombre.

Une autre technique qui pourrait être intéressante de mettre en œuvre, afin d'améliorer le dictionnaire de correction, serait la création de phrases alternatives qui expriment la même action du contexte restreint. C'est-à-dire, qu'en plus de transformer une phrase comme « le

tabouret bascule » en « le banc bascule » ou « les tabourets basculent », avoir aussi des variantes comme « le tabouret qui glisse », « le tabouret qui est en train de chavirer » et « le tabouret est en déséquilibre », qui sont des manières différentes d'exprimer la même idée.

De surcroît, disposer d'un analyseur syntaxique et d'un étiquetage morphosyntaxique de chaque n-gramme aiderait à deux choses : premièrement, pouvoir étendre le vocabulaire des corrections en variant les verbes et pas seulement les noms, et deuxièmement, éliminer les faux positifs dus aux homophones de genre et de nombre de la liste des erreurs détectées, ce qui à son tour, pourrait améliorer les résultats de la méthode de correction automatique.

Enfin, il est encore nécessaire d'améliorer la transformation des mots en phonèmes du français canadien, afin d'améliorer la recherche floue de suggestions de correction en tenant compte des caractéristiques particulières de ce dialecte de la langue française.

BIBLIOGRAPHIE

Adda-Decker, M., & Snoeren, N. D. (2011). Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics, 39*(3), 261-270.

Allauzen, A. (2007). Error detection in confusion network. Dans *Eighth Annual Conference of the International Speech Communication Association*.

Alzheimer's Association. (2021). *2021 Alzheimer's disease facts and figures* (nº 3).

Aman, F., Vacher, M., Rossato, S., & Portet, F. (2012). Contribution à l'étude de la variabilité de la voix des personnes âgées en reconnaissance automatique de la parole [Contribution to the study of elderly people's voice variability in automatic speech recognition]. Dans *JEP-TALN-RECITAL 2012, Atelier ILADI 2012: Interactions Langagières pour personnes Agées Dans les habitats Intelligents* (pp. 49--59). ATALA/AFCP.

Bassil, Y., & Semaan, P. (2012). Asr context-sensitive error correction based on microsoft n-gram dataset. *arXiv preprint arXiv:1203.5262*.

Boschi, V., Catricala, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology, 8*, 269.

Boyer, F., & Rouas, J.-L. (2019). End-to-end speech recognition: A review for the french language. *arXiv preprint arXiv:1910.08502*.

Brousseau, J., Drouin, C., Foster, G. F., Isabelle, P., Kuhn, R., Normandin, Y., & Plamondon, P. (1995). French speech recognition in an automatic dictation system for translators: the transtalk project. Dans *Eurospeech*.

Carbajal, M. J., Bouchon, C., Dupoux, E., & Peperkamp, S. (2018). A toolbox for phonologizing French infant-directed speech corpora.

Chiacchieri, G., Axiak, M., & Altenhoff, A. (2018). fuzzyset. Repéré à https://github.com/axiak/fuzzyset

Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in alzheimer's dementia. *Pragmatics and Society, 10*(2), 153-176.

Dufour, R., & Estève, Y. (2008). Correcting ASR outputs: specific solutions to specific errors in French. Dans *2008 IEEE Spoken Language Technology Workshop* (pp. 213-216). IEEE.

Dumas, D. (2013). La prononciation du français québécois [The pronunciation of Quebec French]. Repéré le 2021/09/24 à https://usito.usherbrooke.ca/articles/th%C3%A9matiques/dumas_1

Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science, 128*, 32-37.

Farley, P., Bullwinkle, M., Lachenmann, A., Orlov, A., Coulter, D., Dempsey, J., & Christiani, T. (2021/02/12). Evaluate and improve Custom Speech accuracy. Repéré le 2021/09/15 à https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-evaluate-data

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Ferrand, L. (1999). 640 homophones et leurs caractéristiques. *L'année psychologique, 99*(4), 687-708.

Fraser, K. C., Rudzicz, F., Graham, N., & Rochon, E. (2013). Automatic speech recognition in the diagnosis of primary progressive aphasia. Dans *Proceedings of the fourth workshop on speech and language processing for assistive technologies* (pp. 47-54).

Fraser, K. C., Rudzicz, F., & Rochon, E. (2013). Using text and acoustic features to diagnose progressive aphasia and its subtypes. Dans *Interspeech* (pp. 2177-2181).

Greenblat, C. (2021). Dementia. Repéré le 17/08/2021 à https://www.who.int/news-room/fact-sheets/detail/dementia

Hakkani-Tür, D., Vergyri, D., & Tur, G. (2010). Speech-based automated cognitive status assessment. Dans *Eleventh Annual Conference of the International Speech Communication Association*.

Hernández-Domínguez, L., García-Cano, E., Ratté, S., & Sierra, G. (2016). Detection of Alzheimer's disease based on automatic analysis of common objects descriptions. Dans *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning* (pp. 10-15).

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., . . . Robert, P. H. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1*(1), 112-124.

Le Grand, J., Aman, F., Vacher, M., Rossato, S., & Portet, F. (2012). Utilisation de la Reconnaissance Automatique de la Parole pour l'aide à l'autonomie des personnes âgées. *Actes de l'Université d'été de la E-santé*, 19-21.

Lehr, M., Prud'hommeaux, E., Shafran, I., & Roark, B. (2012). Fully automated neuropsychological assessment for detecting mild cognitive impairment. Dans *Thirteenth Annual Conference of the International Speech Communication Association*.

Li, J., Deng, L., Haeb-Umbach, R., & Gong, Y. (2015). Robust automatic speech recognition: a bridge to practical applications.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., & Ogar, J. (2008). Learning diagnostic models using speech and language measures. Dans *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4648-4651). IEEE.

Pellegrini, T., & Trancoso, I. (2010). Improving ASR error detection with non-decoder based features. Dans *Eleventh Annual Conference of the International Speech Communication Association*.

Ringger, E. K., & Allen, J. F. (1996). Error correction via a post-processor for continuous speech recognition. Dans *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings* (Vol. 1, pp. 427-430). IEEE.

Rudzicz, F., Wang, R., Begum, M., & Mihailidis, A. (2015). Speech interaction with personal assistive robots supporting aging at home for individuals with Alzheimer's disease. *ACM Transactions on Accessible Computing (TACCESS), 7*(2), 1-22.

Seyfarth, S., & Zhao, P. (2020, 2020/10/05). Evaluating an automatic speech recognition service. Repéré sur AWS Machine Learning Blog à https://aws.amazon.com/es/blogs/machine-learning/evaluating-an-automatic-speech-recognition-service/

Shinkawa, K., & Yamada, Y. (2018). Topic repetition in conversations on different days as a sign of dementia. Dans *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth* (pp. 641-645). IOS Press.

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., . . . Kálmán, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research, 15*(2), 130-138.

Vipperla, R., Renals, S., & Frankel, J. (2010). Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*, 1-10.

Weiner, J., Angrick, M., Umesh, S., & Schultz, T. (2018). Investigating the Effect of Audio Duration on Dementia Detection Using Acoustic Features. Dans *INTERSPEECH* (pp. 2324-2328).

Werner, L., Huang, G., & Pitts, B. J. (2019). Automated speech recognition systems and older adults: a literature review and synthesis. Dans *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, pp. 42-46). SAGE Publications Sage CA: Los Angeles, CA.

You, Y., Ahmed, B., Barr, P., Ballard, K., & Valenzuela, M. (2019). Predicting dementia risk using paralinguistic and memory test features with machine learning models. Dans *2019 IEEE Healthcare Innovations and Point of Care Technologies,(HI-POCT)* (pp. 56-59). IEEE.

Young, V., & Mihailidis, A. (2010). Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology, 22*(2), 99-112.

Zhou, L., Shi, Y., Feng, J., & Sears, A. (2005). Data mining for detecting errors in dictation speech recognition. *IEEE transactions on speech and audio processing, 13*(5), 681-688.