Évaluation des systèmes de reconnaissance automatique de la parole pour les personnes atteintes d'aphasie : un cadre d'analyse de performance pertinent au milieu clinique

par

Julien DUPUIS DESROCHES

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

MONTRÉAL, LE 9 SEPTEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

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

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

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

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

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

LE 26 AOÛT 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à exprimer ma sincère reconnaissance à mes directeurs de recherche, Mme Sylvie Ratté et M. Pierre André Ménard, pour leur encadrement, leurs conseils et leur soutien tout au long de ce projet.

Je souhaite également remercier les membres du jury, M. Luc Duong et M. Tony Wong, pour le temps qu'ils ont consacré à l'évaluation de ce mémoire ainsi que pour leurs commentaires constructifs.

Mes remerciements vont aussi au ministère de l'Économie, de l'Innovation et de l'Énergie (Programme de soutien aux organismes de recherche et d'innovation, FR 56260) ainsi qu'à l'École de technologie supérieure pour leur soutien financier.

Enfin, et surtout, un immense merci à mes parents, Michelle et Jacques, à Andréa, ainsi qu'à mes amis, pour leur soutien constant tout au long de mon parcours académique.

:)

Évaluation des systèmes de reconnaissance automatique de la parole pour les personnes atteintes d'aphasie : un cadre d'analyse de performance pertinent au milieu clinique

Julien DUPUIS DESROCHES

RÉSUMÉ

Ce mémoire présente un nouveau cadre d'évaluation conçu pour mesurer la performance des systèmes de reconnaissance automatique de la parole (RAP) lorsqu'ils sont appliqués à la parole aphasique. L'étude débute par une revue exhaustive de la littérature existante, identifiant les principaux défis liés à l'utilisation des technologies de RAP pour la parole désordonnée et mettant en lumière les lacunes des méthodes d'évaluation actuelles. En s'appuyant sur les mesures d'erreur traditionnelles telles que le taux d'erreurs de mots (WER) et le taux d'erreurs de caractères (CER), le cadre proposé introduit une analyse plus fine adaptée aux phénomènes spécifiques de la parole aphasique, y compris les paraphasies, les distorsions sémantiques et les disfluences. À l'aide de grands ensembles de données annotées et de modèles RAP de pointe, ce cadre a été appliqué à des tests empiriques afin de révéler les faiblesses propres à chaque modèle. Les résultats montrent que, bien que les systèmes RAP obtiennent généralement un WER acceptable sur la parole non aphasique, leur performance se détériore considérablement sur les échantillons de parole aphasique, en particulier pour la transcription des disfluences telles que les pauses remplies. Dans l'ensemble, cette recherche fournit des outils précieux pour les chercheurs et les cliniciens cherchant à améliorer les applications de la RAP dans le traitement et l'évaluation de l'aphasie.

Mots-clés: Aphasie, Reconnaissance automatique de la parole, Cadre d'évaluation

Evaluating Automatic Speech Recognition Systems for Aphasic Speech : A Framework for Clinically-Relevant Performance Analysis

Julien DUPUIS DESROCHES

ABSTRACT

This thesis presents a novel evaluation framework designed to assess the performance of automatic speech recognition (ASR) systems when applied to speech affected by aphasia. The study begins with a comprehensive review of existing literature, identifying key challenges in applying ASR technologies to disordered speech and highlighting gaps in current evaluation methods. Building on traditional error rate metrics such as Word Error Rate (WER) and Character Error Rate (CER), the proposed framework introduces a more granular analysis tailored to the specific phenomena of aphasic speech, including paraphasias, semantic distortions, and disfluencies. Using large annotated datasets and state-of-the-art ASR models, the framework was applied in empirical testing to reveal model-specific weaknesses. Results show that while general-purpose ASR systems achieve acceptable WER on non-aphasic speech, their performance degrades significantly on aphasic speech samples, particularly in transcribing disfluencies such as filled pauses. Overall, this research offers valuable tools for both researchers and clinicians aiming to improve ASR applications in aphasia treatment and assessment.

Keywords: Aphasia, Automatic Speech Recognition, Evaluation Framework

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

APROCSA	Auditory-Perceptual Rating of Connected Speech in Aphasia
ASR	Automatic Speech Recognition
BERT	Représentations encodées bidirectionnelles à partir de transformeurs (Bidirectional Encoder Representations from Transformers)
CER	Taux d'erreurs sur les caractères (Character Error Rate)
CNN	Réseau neuronal convolutif (Convolutional Neural Network)
CTC	Classification temporelle connexionniste (Connectionist Temporal Classification)
CW	Modèle CrisperWhisper de Nyra Health
E2E	Bout en bout (End-to-End)
eGeMAPS	Ensemble de paramètres acoustiques minimalistes de Genève étendu (extended Geneva Minimalistic Acoustic Parameter Set)
L2	Modèle Whisper Large-V2
L3	Modèle Whisper Large-V3
LLM	Grand modèle de langage (Large Language Model)
MFCC	Coefficients cepstraux en fréquences de Mel (Mel Frequency Cepstral Coefficient)
MoE	Mélange d'experts (Mixture of Experts)
NLP	Natural Language Processing
PAA	Personne atteinte d'aphasie / Personnes atteintes d'aphasie
PC	Participant Contrôle / Participants Contrôles
PER	Taux d'erreurs sur les phonèmes (Phoneme Error Rate)
PwA	Person with Aphasia / People with Aphasia
RAP	Reconnaissance automatique de la parole
QA	Quotient d'aphasie
RN	Réseau neuronal
RNP	Réseau neuronal profond (Deep Neural Network)
S2T	Parole en texte (Speech-to-Text)

SVM	Machine à vecteurs de support (Support Vector Machine)
TALN	Traitement automatique du langage naturel
VAD	Détection de l'activité vocale (Voice Activity Detection)
WAB	Western Aphasia Battery
WAB-R	Western Aphasia Battery – Revised
WER	Taux d'erreurs sur les mots (Word Error Rate)

INTRODUCTION

Les systèmes de reconnaissance automatique de la parole (RAP) ont connu des avancées significatives au cours des dernières années, ouvrant de nouvelles possibilités d'application dans divers domaines, notamment dans le secteur de la santé. L'aphasie, un trouble sévère du langage causé par des dommages cérébraux, continue d'avoir un impact profond sur la qualité de vie des personnes atteintes, en entravant leur capacité à communiquer efficacement. Lam & Wodchis (2010) a démontré que l'impact sur la qualité de vie est plus négatif que celui du cancer ou de la perte de membres. Avec le vieillissement de la population mondiale, l'incidence de l'aphasie devrait augmenter (Hachioui *et al.*, 2017), soulignant la nécessité d'outils plus efficaces pour le diagnostic et le traitement. Grâce à leurs capacités croissantes, les systèmes de RAP offrent une opportunité prometteuse d'améliorer les flux de travail cliniques, en particulier pour aider les orthophonistes à diagnostiquer l'aphasie et à assurer le suivi personnalisé des traitements et de la réhabilitation.

0.1 Énoncé du problème

Malgré les avancées des technologies de reconnaissance automatique de la parole (RAP), leur adoption en orthophonie reste limitée. En pratique clinique, les orthophonistes doivent souvent s'appuyer sur des annotations manuelles ou des observations en temps réel pour analyser la parole de leurs patients, ce qui peut être fastidieux et sujet à interprétation.

De plus, bien que les systèmes de RAP aient des applications en milieu clinique, leur intégration est limitée en raison de leur instabilité et de leur manque de précision.

Par ailleurs, les outils d'évaluation de la performance de ces systèmes se concentrent majoritairement sur des mesures globales comme le taux d'erreurs sur les mots (WER), sans refléter adéquatement les aspects linguistiques pertinents pour l'analyse des troubles du langage. Il manque donc des

métriques adaptées à ces contextes spécifiques, qui permettraient une évaluation plus fine et utile en orthophonie.

0.2 Objectifs

L'objectif principal de ce mémoire est de développer un outil et des mesures d'évaluation avancés pour aider les cliniciens et les développeurs à évaluer les systèmes de RAP en fonction de leurs besoins spécifiques en pratique clinique. Cela inclut la mise en place d'un cadre permettant de mesurer la performance des systèmes de RAP avec un plus grand détail, au-delà des analyses traditionnelles. Pour illustrer le potentiel de ce cadre, nous démontrerons également son application à des systèmes de RAP de pointe couramment utilisés, ainsi que l'impact de la rédactique (*prompting*) sur ces modèles.

0.3 Contributions

Cette recherche présente un cadre à source libre pour évaluer les systèmes de RAP sur des caractéristiques de la parole spécifiques pertinentes pour la pratique clinique. Ce cadre permet d'évaluer la performance des modèles non seulement en termes globaux, mais aussi en fonction de caractéristiques linguistiques précises, tels que les répétitions, les pauses, ou les erreurs phonétiques, qui sont particulièrement pertinentes en orthophonie.

Par ailleurs, cette recherche fournit une évaluation comparative de quelques modèles de RAP couramment utilisés. Ces modèles sont testés sur un corpus composé de conversations avec des personnes atteintes d'aphasie, ce qui permet de mieux cerner leur robustesse et leurs limites face à la parole altérée.

0.4 Organisation du document

Ce mémoire s'organise en trois grandes parties. La première section présente un aperçu des thèmes clés liés à l'aphasie et à la RAP. Elle commence par une discussion sur l'aphasie, incluant ses différents types, symptômes et impacts sur la communication. Elle se poursuit par une introduction à la technologie RAP, en exposant ses concepts fondamentaux et ses avancées récentes. La revue explore aussi le rôle de l'intelligence artificielle dans la détection et l'évaluation de l'aphasie, en mettant en évidence les méthodes existantes et leur efficacité. Enfin, cette section examine l'application de la RAP à la parole aphasique, en abordant les défis et les adaptations possibles pour améliorer la précision de la reconnaissance.

La section suivante présente l'article central du mémoire, qui présente une étude de recherche originale qui évalue la performance des systèmes de RAP sur la parole aphasique selon des caractéristiques linguistiques pertinentes en milieu clinique. L'étude introduit un cadre d'évaluation qui va au-delà des métriques traditionnelles telles que le WER, en incorporant des mesures plus détaillées basées sur des phénomènes liés à la parole aphasique. Les résultats sont analysés en relation avec les métriques RAP standards, offrant des pistes pour l'adaptation ou l'amélioration des systèmes actuels en vue d'une utilisation plus efficace dans les soins liés à l'aphasie.

Enfin, la dernière section du mémoire résume les principales conclusions du travail et discute des perspectives futures. Elle met en évidence le potentiel de la RAP et de l'IA dans l'évaluation de l'aphasie tout en soulignant les limites actuelles. La conclusion propose aussi des pistes pour des recherches futures et des améliorations possibles des outils d'évaluation.

CHAPITRE 1

REVUE DE LITTÉRATURE

Ce chapitre offre une compréhension fondamentale de l'aphasie, des méthodes d'évaluation traditionnelles, ainsi que du rôle émergent des outils automatisés dans ce domaine. Il présente également un aperçu des avancées récentes dans la détection et la classification de l'aphasie, ainsi qu'une étude approfondie des recherches en RAP appliquées à la parole de personnes atteintes d'aphasie (PAA).

1.1 Aperçu de l'aphasie

L'aphasie est un trouble neurologique qui nuit à la capacité de communiquer, affectant à la fois la production et la compréhension du langage. Elle survient généralement à la suite de dommages aux centres du langage dans le cerveau, souvent causés par un accident vasculaire cérébral ou un traumatisme crânien. La figure 1.1 illustre les zones du cerveau impliquées dans la compréhension et la production du langage. Les aires de Wernicke et de Broca sont associées avec la compréhension du langage et la production de parole respectivement. La gravité et les caractéristiques spécifiques de l'aphasie peuvent varier considérablement (NIDCD, 2017). L'aphasie affecte uniquement le traitement du langage ; Privitera, Ng, Kong & Weekes (2024) précise que, seule, « l'aphasie n'affecte ni l'intelligence, ni l'empathie, ni la conscience culturelle, ni l'introspection, ni l'apprentissage, ni les processus cognitifs utilisés pour la communication non verbale » [traduction libre]. L'aphasie a un impact négatif considérable et sur la qualité de vie des personnes qui en sont atteintes (Lam & Wodchis, 2010).

De manière générale, l'aphasie peut être classée en plusieurs types (National Aphasia Association, s.d.) :

- Aphasie de Broca : Caractérisée par une parole non fluide et laborieuse, avec une compréhension relativement préservée. Les personnes atteintes de ce type d'aphasie produisent souvent des phrases courtes et grammaticalement incomplètes.

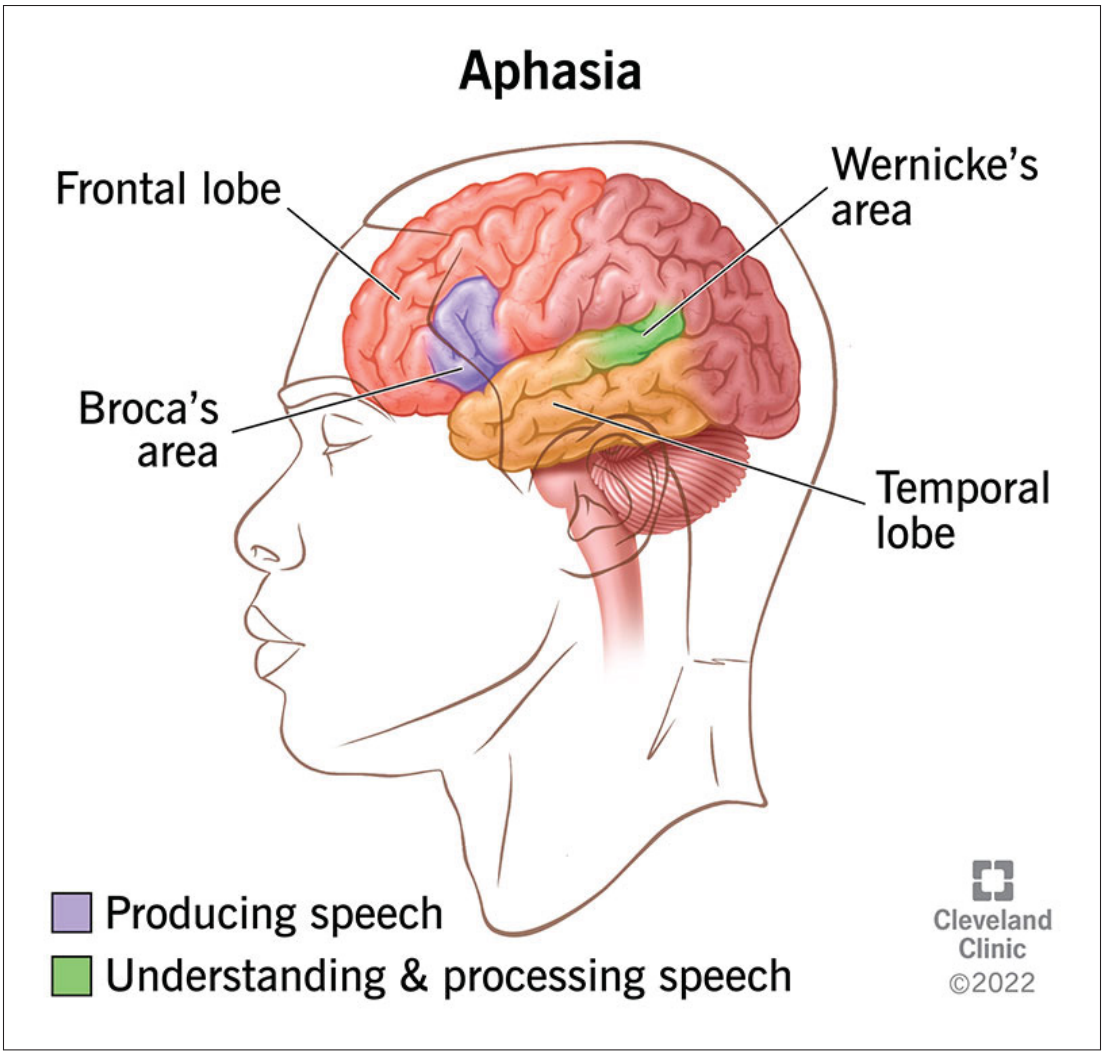

Figure 1.1 Zones du cerveau associées à la compréhension et à la production du langage
Tirée de Cleveland Clinic (2022)

- Aphasie de Wernicke : Marquée par une parole fluide, mais souvent dénuée de sens, avec des troubles importants de la compréhension. La parole peut inclure des mots inventés ou incohérents.
- Aphasie globale : Forme sévère d'aphasie qui affecte les capacités expressives et réceptives du langage, entraînant de grandes difficultés de communication.

Parmi les autres formes d'aphasie, on retrouve l'aphasie de conduction, où la parole reste fluide, mais la capacité de répétition est altérée, et l'aphasie anomique, caractérisée par des difficultés à retrouver les mots, ce qui entraîne des pauses fréquentes et des circonlocutions. Le diagramme de la figure 1.2 illustre les différents diagnostics d'aphasie selon les symptômes observés.

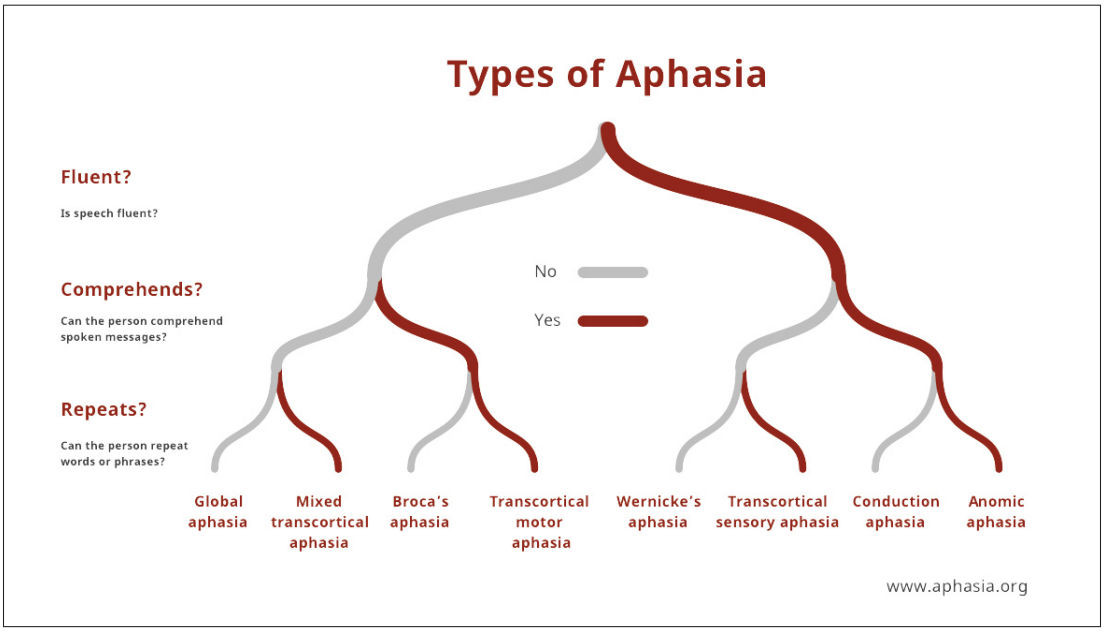

Figure 1.2 Représentation des différents types d'aphasie selon les symptômes
Tirée de National Aphasia Association (s.d.)

L'aphasie de Broca est le type d'aphasie non fluente le plus courant (NIDCD, 2017). Les personnes atteintes de cette forme présentent une parole fragmentée ou désorganisée avec un nombre supérieur à la moyenne de dysfluences, c'est-à-dire des interruptions dans le flux régulier de la parole, telles que des hésitations ou des répétitions.

1.2 Traitement et évaluation clinique de l'aphasie

Le traitement le plus courant de l'aphasie est la rééducation par orthophonie (Fridriksson & Hillis, 2021). Les séances de rééducation sont généralement effectuées individuellement avec un orthophoniste, en présentiel ou à distance. La thérapie à distance constitue une solution potentielle pour réduire les coûts et améliorer l'accessibilité aux orthophonistes. Les autres méthodes de

traitements incluent des interventions pharmaceutiques et des méthodes de stimulation cérébrale non invasive. Ces méthodes sont pratiquement toujours faites en ajout à la rééducation par orthophonie (Fridriksson & Hillis, 2021).

Bien que complètes, ces méthodes de traitement traditionnelles exigent beaucoup de ressources, tant en temps qu'en expertise, pour être administrées et interprétées avec précision. Malgré leur efficacité, elles présentent plusieurs défis (Mahmoud, Kumar, Li, Tang & Fang, 2021) :

- **Longues à administrer :** La nature détaillée de ces évaluations les rend longues à réaliser, nécessitant souvent plusieurs séances avec le patient.
- **Exigeantes en ressources :** Le besoin de cliniciens spécialisés et de matériel particulier limite l'accessibilité de ces tests, surtout dans les régions éloignées ou les milieux défavorisés.
- **Subjectives :** Bien que standardisée, l'interprétation des résultats peut varier d'un clinicien à l'autre, entraînant d'éventuelles incohérences dans le diagnostic et la planification des traitements. La fidélité interjuges est un indicateur important à considérer dans l'évaluation de ces outils.
- **Fréquence d'évaluation limitée :** En raison des ressources nécessaires, il est souvent difficile d'effectuer un suivi fréquent des progrès du patient, ce qui peut retarder l'ajustement des plans de traitement.

La qualité des évaluations est donc essentielle pour garantir l'efficacité du traitement. Des évaluations imprécises ou incomplètes peuvent conduire à des diagnostics erronés ou à des plans de traitement inadéquats. Une évaluation précise et accessible est ainsi importante pour le soutien des soins adaptés et personnalisés.

Historiquement, l'évaluation de l'aphasie repose principalement sur des tests standardisés administrés par des cliniciens formés. Ces évaluations comprennent souvent des tâches telles que la dénomination d'images, la répétition de phrases, la compréhension de lecture et l'analyse du discours conversationnel (Hachioui *et al.*, 2017; Wilson, Eriksson, Schneck & Lucanie, 2018). Le « Western Aphasia Battery » (WAB) et sa version révisée, le « Western Aphasia Battery-Revised » (WAB-R), sont parmi les cadres d'évaluation de l'aphasie les plus utilisés. Ils

permettent de classer les types d'aphasie en fonction d'une combinaison de critères tels que la fluidité de la parole, la compréhension, la répétition et la capacité de dénomination (Kertesz, 2012, 2022). Ces tests produisent un score appelé « quotient d'aphasie » (QA), qui mesure les capacités langagières d'une personne. D'autres tests existent pour évaluer l'aphasie dans différentes langues, comme le test allemand « Aachen Aphasia Test » (AAT) et le test chinois « Chinese Rehabilitation Research Center Aphasia Examination » (CRRCAE) (Mahmoud *et al.*, 2021).

Une méthode plus récente, appelée « Auditory-Perceptual Rating of Connected Speech in Aphasia » (APROCSA), propose un cadre détaillé d'évaluation du langage à travers un ensemble de 27 caractéristiques linguistiques et prosodiques (Casilio, Rising, Beeson, Bunton & Wilson, 2019). Chaque caractéristique est évaluée sur une échelle de 0 (non-présent) à 4 (presque toujours présent) ; les évaluations sont effectuées sur des extraits de discours d'une durée d'environ 5 minutes. Tandis que le WAB permet un diagnostic quantitatif à travers plusieurs modalités, l'APROCSA permet aux cliniciens d'évaluer l'utilisation plus naturelle du langage et d'identifier les ruptures conversationnelles, telles que l'anomie (difficulté à retrouver les mots) et les néologismes (mots inventés), offrant ainsi une perspective complémentaire.

Ces défis soulignent le besoin de méthodes d'évaluation plus efficaces et accessibles, que les outils automatisés de traitement de la parole pourraient permettre. Pour surmonter les limites des évaluations traditionnelles, des approches computationnelles utilisant le traitement automatique de la parole et l'apprentissage automatique ont attiré l'attention ces dernières années.

1.3 Détection et classification de l'aphasie

Une grande partie des recherches récentes en apprentissage machine appliquées à l'aphasie se concentre sur trois tâches principales (Azevedo *et al.*, 2024) :

1. Détecter la présence de l'aphasie,
2. Classifier la sévérité de l'atteinte,
3. Classifier le type d'aphasie.

Cette section présente certains travaux récents dans ce domaine, illustrant la diversité des méthodologies employées et la dépendance fréquente aux systèmes de RAP. Un ensemble de données couramment utilisé pour la recherche en aphasie est AphasiaBank (MacWhinney, Fromm, Forbes & Holland, 2011). Cette base de données regroupe plusieurs séances de conversations enregistrées dans divers contextes, impliquant, en anglais, un total de 476 PAA et 352 participants contrôles (PC), c'est-à-dire, sans diagnostic d'aphasie. AphasiaBank contient aussi des ensembles de données pour autres langues qui ont cependant beaucoup moins d'ampleur ; par exemple, le corpus français contient que 12 PAA. Les études suivantes reposent en majorités sur des données provenant de ces ensembles.

En utilisant une combinaison de caractéristiques acoustiques et d'un système de RAP, Barberis, De Clercq, Tamm, Van hamme & Vandermosten (2024) ont mis en œuvre une détection binaire de l'aphasie dans la langue néerlandaise à l'aide d'un classificateur à vecteurs de support (SVM). Avec un ensemble de données incluant 62 PAA et 57 PC, ils ont atteint une précision de 86,6%. Nivedha, Chandrasekar & Jothi (2023) ont utilisé un réseau neuronal convolutionnel (CNN) pour détecter la sévérité de l'aphasie, réalisant une classification binaire entre des QA faibles et élevés à l'aide du réseau ResNetV2. Sur un ensemble de données comprenant 91 PAA, les résultats montrent une précision de 98,1% et un F1 de 97,4%.

Des études plus récentes ont utilisé des approches basées sur les transformeurs. Qin, Lee, Kong & Lin (2022) ont affiné un modèle de langage préentraîné pour la détection de l'aphasie. Les meilleurs résultats ont été obtenus avec des modèles ajustés utilisant des représentations encodées bidirectionnelles (BERT), atteignant des précisions et des F1 de 98% pour les locuteurs cantonais et 94% pour les locuteurs mandarinophones. L'ensemble de données utilisé comprenait 92 PAA et 92 PC pour le cantonais et 9 PAA et 9 PC pour le mandarin. Tang, Chen, Chang, Watanabe & MacWhinney (2023) ont mis en œuvre une architecture de classification temporelle connexionniste (CTC) pour effectuer simultanément la détection de l'aphasie et la reconnaissance de la parole. Leur étude affiche une précision de 97,3%. Cong, LaCroix & Lee (2024) ont démontré l'utilisation de grands modèles de langue (LLM) pour extraire des caractéristiques pour la détection de l'aphasie. Avec les caractéristiques ressorties et l'utilisation de SVM pour la

classification binaire pour la détection d'aphasie, une précision et un F1 de 92,0% à été obtenue sur un ensemble de données comprenant 441 PAA et 341 PC. Qin, Wu, Lee & Kong (2020b) ont utilisé une méthode basée sur les CNN pour atteindre une précision de 82,4% et un F1 de 82,2%. Leur ensemble de données était composé de 91 PAA.

Des caractéristiques acoustiques ont également été utilisées pour détecter l'aphasie. Qin, Lee & Kong (2020a) ont utilisé des caractéristiques suprasegmentales, notamment la durée des pauses, et ont constaté une forte corrélation avec le QA. Cette même étude intègre aussi les caractéristiques eGeMAPS (Geneva Minimalistic Acoustic Parameter Set), soit un ensemble de paramètres qui décrivent des aspects auditifs de la voix, tels que la hauteur et le timbre. L'utilisation de ces paramètres révèle des résultats variés, l'intensité sonore étant la caractéristique ayant la plus forte corrélation avec le QA. L'ensemble de données utilisé comprenait 92 PAA et 118 PC.

D'autres études ont porté sur la classification de la sévérité et du type d'aphasie. Day *et al.* (2021) ont identifié des caractéristiques issues de transcriptions avec l'aide de techniques de traitement automatique des langues naturelles (TALN). Ils ont exploré l'utilisation de méthodes de regroupement telles que K-Means et les réseaux neuronaux (RN) pour la classification. Les résultats permettent de distinguer différents niveaux de sévérité avec une précision de 73% sur un ensemble de 238 PAA. En utilisant des coefficients cepstraux en fréquences de Mel (MFCC) avec un réseau neuronal profond (RNP), Herath *et al.* (2022) ont obtenu une précision et un F1 allant jusqu'à 99% pour la classification de la sévérité de l'aphasie en anglais sur un ensemble de 54 PAA. Enfin, Wagner, Zusag & Bloder (2023) ont tiré parti des progrès en transcription et en alignement temporel pour extraire des caractéristiques permettant de classer les différents types d'aphasie. En utilisant un classificateur SVM classique, la classification atteint une précision et un F1 de 90,6% sur un ensemble de 352 PAA et 272 PC.

Les tableaux 1.1 et 1.2 résument respectivement les résultats ci-dessus concernant la détection et la classification de l'aphasie. Le tableau 1.3 résume la répartition des données pour chaque étude.

Tableau 1.1 Résumé des résultats pour la détection de l'aphasie

Auteur	Année	Langue	Architecture	Précision	F1
Barberis et al.	2024	Néerlandais	SVM	86,6%	—
Cong et al.	2024	Anglais	LLM, SVM	92%	92%
Nivedha et al.	2023	Anglais	CNN, ResNet V2	98,1%	97,4%
Qin, Lee et al.	2020	Cantonais	CNN	82,4%	82,2%
Qin et al.	2022	Cantonais	LLM, BERT	98%	98%
		Mandarin		94%	94%
Tang et al.	2023	Anglais	CTC, Attention	97,3%	—

Tableau 1.2 Résumé des résultats pour la classification de l'aphasie

Auteur	Année	Langue	Architecture	Classification	Précision	F1
Day et al.	2021	Anglais	TALN, K-Means, RN	3 classes (Léger, Modéré, Sévère)	73,0%	—
Herath et al.	2022	Anglais	MFCC, RNP	4 classes (Globale, Broca, Wernicke, Anomie)	99%	99%
Wagner et al.	2023	Anglais	CTC, Encodeur-Décodeur, SVM	4 classes (Contrôle, Broca, Wernicke, Anomie)	90,6%	90,6%

Tableau 1.3 Résumé des données pour la détection et la classification de l'aphasie

Auteur	Année	Langue	Données (nb. de personnes)	
			PAA	PC
Barberis et al.	2024	Néerlandais	62	57
Cong et al.	2024	Anglais	441	341
Nivedha et al.	2023	Anglais	91	—
Qin, Lee et al.	2020	Cantonais	92	118
Qin et al.	2022	Cantonais	92	92
		Mandarin	9	9
Tang et al.	2023	Anglais	20,1 (heures)	10,2 (heures)
Day et al.	2021	Anglais	238	—
Herath et al.	2022	Anglais	54	—
Wagner et al.	2023	Anglais	352	272

Dans l'ensemble, presque toutes ces études reposaient sur une transcription manuelle ou sur un composant de RAP pour obtenir ces résultats. Plusieurs études soulignent la dépendance à l'exactitude des transcriptions et la nécessité d'améliorer les performances des systèmes RAP lorsqu'ils sont appliqués à la parole de PAA.

1.4 Reconnaissance automatique de la parole

Les premiers systèmes RAP reposaient souvent sur des modèles statistiques tels que les modèles de Markov cachés combinés à des modèles de mélanges gaussiens (Levinson, 1986). Toutefois, le domaine a connu une transformation majeure avec l'avènement de l'apprentissage profond, menant à des améliorations substantielles en matière de précision. Les systèmes de RAP modernes utilisent fréquemment des RNP, incluant les CNN, les réseaux neuronaux récurrents (RNN) et les réseaux de type transformeur (Barberis *et al.*, 2024; Torre, Romero & Álvarez, 2021). Les modèles RAP récents, tels que Whisper (Radford *et al.*, 2022) et wav2vec 2.0 (Baevski, Zhou, Mohamed & Auli, 2020), se montrent robustes lorsqu'ils sont appliqués à la parole d'individus sans troubles du langage. La Figure 1.3 illustre de façon détaillée l'évolution des approches développées au cours des trois dernières décennies pour améliorer les systèmes de RAP. Les premières approches pour les années jusqu'à 2011 se reposent surtout sur des

modèles acoustiques et des modèles de base tels que les modèles de Markov cachés. Autour de l'an 2012, on marque l'arrivée des méthodes qui se reposent sur l'apprentissage profond et on remarque plusieurs gains de performance importants dans très peu de temps. Dans les années les plus récentes, on remarque de plus en plus des approches faisant l'utilisation de mécanismes d'attention, tels que les réseaux LSTM (Long Short Term Memory) et les RNN.

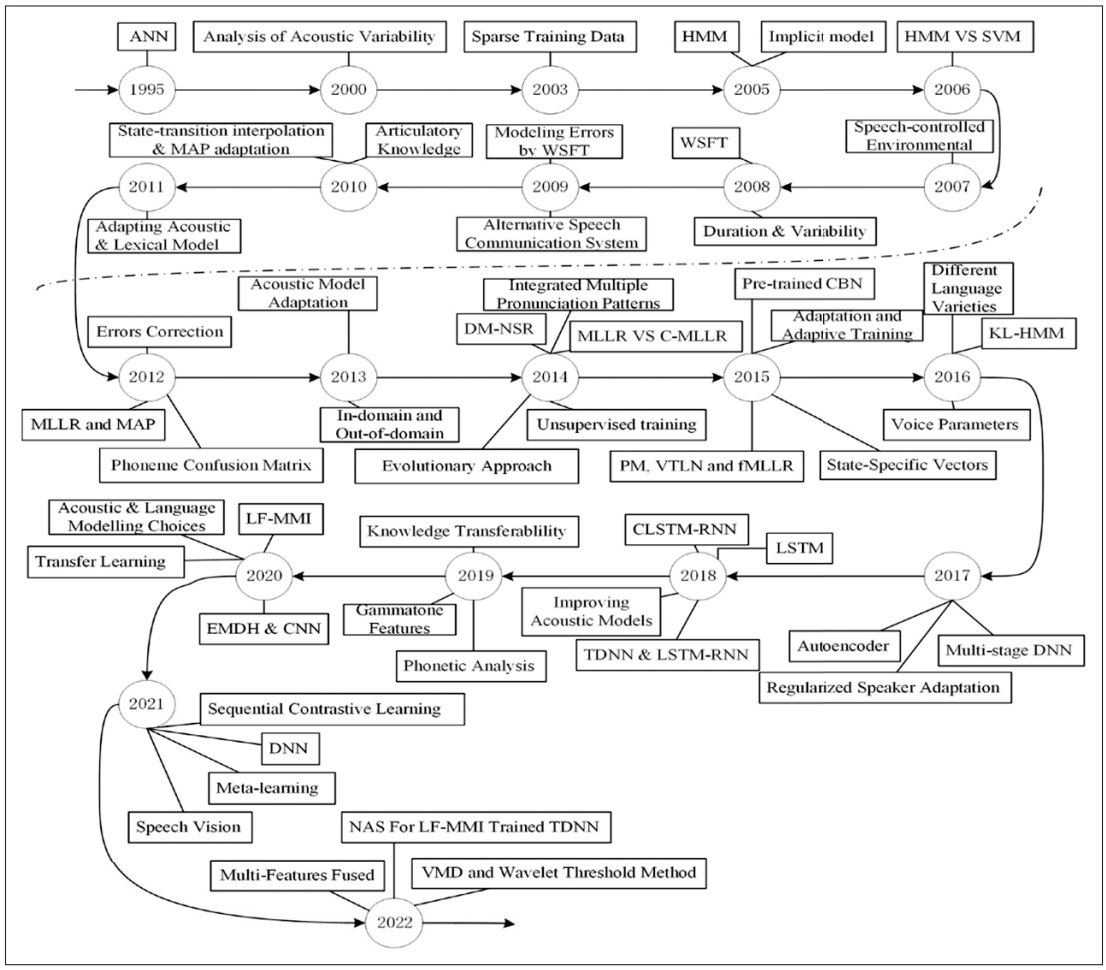

Figure 1.3 Évolution détaillée des axes de recherche pour la RAP
Tirée de Qian & Xiao (2023)
(La ligne pointillée indique la transition avant et après l'introduction de l'apprentissage profond)

Malgré ces avancées, la technologie de RAP rencontre encore d'importants défis, en particulier lorsqu'elle est confrontée à des schémas de parole atypiques, comme ceux observés chez les

personnes ayant des troubles du langage. La parole affectée par l'aphasie présente des obstacles uniques liés à des prononciations altérées, des hésitations, des erreurs sémantiques et des paraphasies (Perez, Aldeneh & Provost, 2020). Les systèmes de RAP standards, principalement entraînés sur la parole de personnes non atteintes d'aphasie, voient souvent leurs performances diminuer lorsqu'ils traitent la parole de PAA. Cette difficulté est amplifiée par le manque de grands ensembles de données annotées de parole de PAA, pourtant essentiels pour entraîner des modèles d'apprentissage profond robustes (Le, Licata & Mower Provost, 2018; R & A, 2024).

Pour surmonter ces obstacles, la recherche actuelle se concentre sur le développement de modèles RAP spécialisés pour la parole désordonnée. Une approche populaire consiste à ajuster des modèles RAP préentraînés à l'aide de petites quantités de données de parole de PAA. Ces modèles récents sont particulièrement robustes pour la parole typique grâce à leur entraînement sur des corpus volumineux et diversifiés, souvent multilingues. Perez *et al.* (2023) ont évalué plusieurs modèles préentraînés afin de générer des représentations vocales pour la partie encodeur de leur modèle de bout en bout (E2E). Les modèles Wav2Vec2, HuBERT et WavLM ont été testés, avec les meilleurs résultats obtenus avec WavLM. Torre et al. (2021) ont basé leur mise en œuvre sur l'architecture Wav2Vec, en ajustant le modèle XLSR-53 avec les données d'AphasiaBank en anglais et en espagnol. Les mesures de taux d'erreurs sur les caractères (Character Error Rate ou CER), de taux d'erreurs sur les mots (Word Error Rate ou WER) et de taux d'erreurs sur les phonèmes (Phoneme Error Rate ou PER) représentent respectivement la proportion de caractères, de mots et de phonèmes mal transcrits comparativement à la transcription de référence. En anglais, les résultats du CER varient de 13,4% pour la parole avec des symptômes d'aphasie légers à 46,0% pour la parole avec des symptômes d'aphasie sévères, et les résultats du WER vont de 22,3% à 55,5%, respectivement. Les auteurs de ces études soulignent que le manque de données annotées constitue l'un des principaux obstacles à des transcriptions automatiques précises.

Perez *et al.* (2020) ont proposé une approche fondée sur un mélange d'experts (MoE), utilisant une combinaison de réseaux de reconnaissance vocale spécialisés selon le degré de sévérité

de l'aphasie. Tous les réseaux partagent les mêmes couches initiales pour l'extraction des caractéristiques acoustiques.

Le Tableau 1.4 compare les performances des modèles de RAP récents appliqués sur la parole de PAA.

Tableau 1.4 Résumé des résultats pour la reconnaissance automatique de la parole de PAA en anglais

Auteur	Année	Méthodologie	WER	CER	PER
Perez et al.	2020	E2E, WavLM Encodeur	14,1 (léger) - 45,2 (très sévère)	—	—
Perez et al.	2023	MoE, RN	—	—	37,03
Torre et al.	2021	Wav2Vec, XLSR-53	22,3 (léger) - 55,5 (très sévère)	13,3 (léger) - 46,0 (très sévère)	—

Les cadres technologiques de RAP combinent généralement plusieurs techniques de traitement, comme la détection d'activité vocale (VAD) pour le découpage, des modèles de transcription de parole, ainsi que des réseaux neuronaux pour l'alignement temporel. Ces pipelines produisent des transcriptions segmentées et alignées dans le temps. Les systèmes RAP modernes reposent sur des réseaux neuronaux profonds, en particulier les architectures de type transformeur, qui ont permis une amélioration notable de la précision par rapport aux méthodes antérieures.

L'un des modèles de RAP les plus utilisés aujourd'hui est Whisper, développé par OpenAI (Radford *et al.*, 2022). Lancée en 2022, cette famille de modèles se décline en plusieurs tailles et prend en charge la transcription multilingue ainsi que des tâches de traduction. Plusieurs cadres logiciels offrent désormais des interfaces et des fonctionnalités supplémentaires pour l'utilisation de ces modèles. WhisperX, introduit par Bain, Huh, Han & Zisserman (2023), est une extension qui s'appuie sur Whisper et propose des horodatages au niveau du mot ainsi que diarization, c'est-à-dire l'attribution d'identités aux locuteurs pour chaque segment de transcription. Ces fonctions sont particulièrement utiles pour l'analyse de la parole aphasique, où les informations temporelles et les disfluences contiennent des indices cliniques importants.

Des recherches supplémentaires sont nécessaires concernant la performance des systèmes RAP face aux variations dialectales. Comme le note Wassink, Gansen & Bartholomew (2022), ces systèmes peuvent afficher des performances inégales selon les dialectes associés à certaines communautés ethnolinguistiques, ce qui pourrait entraîner des biais dans les contextes cliniques. Cela met en évidence les limites des systèmes de RAP lorsqu'ils sont confrontés à des formes de parole différentes de celles sur lesquelles ils ont été entraînés, comme c'est souvent le cas avec les la parole de PAA.

1.5 Hallucinations

Un aspect important dans l'étude des systèmes RAP est la production d'erreurs de transcription. Dans les systèmes plus anciens, ces erreurs résultaient souvent d'un vocabulaire ou d'un lexique limité, menant à des transcriptions inexactes ou incomplètes. Par exemple, un nom propre comme «Pelletier» pouvait être transcrit incorrectement comme «pelle tiers».

Un aspect émergent de l'étude de modèles basés sur les transformeurs concerne, par contre, l'apparition d'hallucinations. Ceux-ci sont des segments de texte inventés qui modifient le sens ou ne correspondent tout simplement pas à l'enregistrement d'origine. Les modèles RAP peuvent générer du contenu halluciné, surtout lorsque l'entrée acoustique est dégradée ou contient de longues pauses. Contrairement aux simples erreurs lexicales, ces hallucinations surviennent souvent parce que les modèles s'appuient excessivement sur les patrons statistiques du langage lorsque le signal d'entrée manque de clarté ou de fluidité, comme c'est souvent le cas dans la parole de PAA. Par exemple, un modèle pourrait insérer une phrase syntaxiquement plausible, mais absente, comme «je suis allé au parc avec mon chien hier.», alors que le locuteur n'a prononcé que «je . . . parc . . . chien».

Koenecke, Choi, Mei, Schellmann & Sloane (2024) ont constaté qu'avec le modèle Whisper de base, des hallucinations apparaissent dans environ 1% des transcriptions. Les auteurs ont jugé qu'environ 38% des hallucinations avaient un impact négatif significatif. Ces hallucinations comprennent notamment des incitations à la violence, des faussetés et des références à des

sources inexistantes. Ils ont également observé que les locuteurs atteints d'aphasie étaient plus susceptibles de provoquer des hallucinations, en raison de la fréquence accrue des pauses dans leur discours, comme l'illustrent les distributions dans la Figure 1.4. La figure illustre bien le plus grand nombre de pauses présents dans le discours de PAA comparé à celui de PC et que la proportion de pauses est plus élevées pour les discours qui contient des hallucinations que pour ceux qui n'en contiens pas.

Dans une étude plus large sur l'utilisation des LLMs dans le domaine médical, 84,7% des médecins ont rapporté que les modèles généraient des hallucinations pouvant affecter la santé des patients (Kim *et al.*, 2025). La même étude a toutefois révélé que 90% des médecins considèrent ces outils comme utiles, et que plus de 50% d'entre eux les utilisent quotidiennement. Cela montre que malgré la présence d'hallucinations, ces outils représentent un ajout pertinent aux flux de travail cliniques et connaissent une adoption rapide.

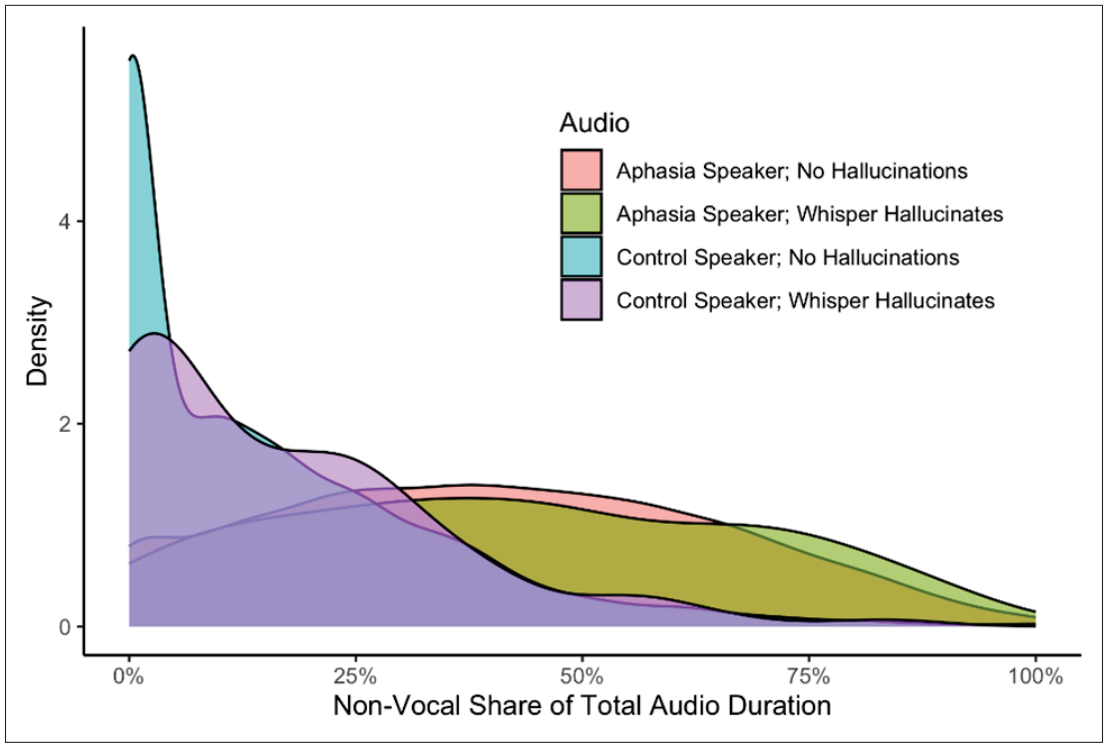

Figure 1.4 Proportion relative de locuteurs avec et sans hallucinations selon la durée des pauses non remplies
Tirée de Koenecke *et al.* (2024)

1.6 La reconnaissance de la parole pour le traitement de l'aphasie

Les modèles d'IA actuels sont généralement capables de détecter l'aphasie et de différencier les types avec un bon niveau de précision (Azevedo *et al.*, 2024). L'attention se tourne maintenant vers les aspects de traitement et de réadaptation, en misant sur des technologies qui soutiennent le travail des orthophonistes, notamment des systèmes de RAP. L'un des principaux avantages de la RAP est la possibilité de soutenir le travail des orthophonistes en fournissant une évaluation et une rétroaction instantanées sur la parole. Cela ouvre aussi la voie à des options de rééducation à domicile, asynchrone ou autogérée.

Les recherches récentes s'orientent vers l'utilisation de systèmes RAP pour produire des transcriptions utiles en contexte clinique. Ces transcriptions permettent l'extraction automatique de mesures quantitatives telles que la longueur moyenne des énoncés, la fréquence et la durée des pauses, ou encore la proportion d'erreurs grammaticales (Le *et al.*, 2018). Des travaux comme celui de Le & Provost (2016) soulignent aussi l'importance de personnaliser les modèles acoustiques et linguistiques selon les types d'énoncés ou les niveaux de sévérité, en tenant compte de la grande variabilité des profils de parole aphasique.

Les caractéristiques extraites des transcriptions de systèmes RAP peuvent être utilisées pour prédire des mesures cliniques telles que le QA du Western Aphasia Battery - Revised (WAB-R), ou pour automatiser certaines sous-épreuves des batteries d'évaluation de l'aphasie Mahmoud *et al.* (2023). Malgré tout, une revue menée par Azevedo *et al.* (2024) n'a trouvé aucun article ayant intégré l'IA, y compris la RAP, dans les dispositifs de communication améliorée et alternative pour l'aphasie. La revue définit ces dispositifs comme des stratégies et systèmes conçus pour compenser les troubles du langage et les difficultés de communication d'une personne. Ces dispositifs peuvent être simples ou d'une technologie avancée.

Même si les récents progrès en RAP appliquée à l'aphasie sont encourageants, plusieurs défis restent. La parole aphasique pose des défis particuliers aux systèmes RAP, notamment en raison de structures grammaticales atypiques, d'énoncés fragmentés et de l'utilisation de néologismes ou de mots tronqués. Les mesures d'évaluation les plus utilisées, comme le WER ou le CER,

donnent une interprétation assez limitée des erreurs de transcription quand il s'agit de parole de PAA. Ces métriques, utiles dans un contexte général, ne prennent pas en compte les particularités linguistiques propres aux troubles du langage. Ce mémoire cherche à combler cette lacune en proposant un cadre d'évaluation plus précis, pensé pour répondre aux besoins du milieu clinique. Le chapitre suivant présente une étude originale qui explore cette question en évaluant la performance de systèmes de RAP à partir de phénomènes de la parole observés chez les PAA.

CHAPITRE 2

EVALUATING ASR FOR APHASIA : A FRAMEWORK FOR CLINICALLY RELEVANT TRANSCRIPTION PERFORMANCE

Julien Dupuis Desroches , Pierre André Ménard , Sylvie Ratté

Département de génie logiciel et TI, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article soumis à la revue « Aphasiology », juillet 2025

2.1 Abstract

2.1.1 Background

The clinical adoption of automated speech recognition (ASR) systems in speech pathology for people with aphasia (PwA) and similar conditions is primarily limited by performance and stability issues. The possibility for detailed evaluation of such systems is critical, as clinical assessment of patients and their following treatments are highly dependent on their accuracy and precision.

2.1.2 Aims

This study aims to address the limitations of conventional ASR evaluation metrics, such as Word Error Rate (WER), by introducing a more granular evaluation approach. The goal is to develop and apply a detailed framework for analyzing ASR system performance, specifically on speech phenomena relevant to clinical aphasia assessment, including disfluencies, grammatical errors, and filler words.

2.1.3 Methods & Procedures

We present a novel evaluation framework and accompanying software tool to assess ASR models on aphasic speech. Using annotated transcripts from the AphasiaBank database, the

framework measures performance across multiple linguistic phenomena by mapping ASR-generated transcriptions against reference CHAT-encoded transcripts. Key metrics such as Character Error Rate (CER) are computed per-phenomenon using Levenshtein distance, enabling a fine-grained analysis of transcription accuracy for features critical to clinical diagnosis. To demonstrate the variability in performance, we applied the framework to three different ASR models, highlighting fluctuations across speech phenomena.

2.1.4 Outcomes & Results

Evaluation of the Whisper Large-V3 model on the English AphasiaBank dataset revealed significant variability in transcription accuracy across speech phenomena. CER ranged from 25.28% on unannotated words to 87.82% on filled pauses, a delta of over 62 percentage points. Prompting reduced the CER for filled pauses from 87.82% to 44.04%, demonstrating that task-specific tuning can yield substantial gains.

Compared to control participants, PwA transcripts obtained CERs that were up to 20 percentage points higher, depending on the speech phenomenon. The largest gaps appeared in phonological and grammatical errors.

In a multi-language evaluation, transcription performance on French aphasic speech was notably worse, with CERs averaging 32% higher than their English counterparts. Morphological errors in French reached up to 43.33%, compared to 29.60% in English, and semantic errors showed deltas exceeding 20 percentage points.

2.1.5 Conclusions

These indicators can help better understand the underlying behaviours of ASR models, thus offering better insights into their reliability. With this framework, ASR systems can be evaluated to detect any weaknesses and highlight areas for improvement, ensuring their suitability for clinical use. By offering a granular analysis of these factors, the tool empowers clinicians and

researchers to make informed decisions about integrating ASR systems into speech pathology workflows.

2.2 Introduction

A third of people who have experienced a stroke are affected by aphasia (Frew, 2017; Frederick, Jacobs, Adams-Mitchell & Ellis, 2022). Today, over 100,000 individuals in Canada and over 1,000,000 individuals in the United States are living with the disorder, a prevalence of approximately three people per thousand. (Aphasia Institute, s.d.; Dickey *et al.*, 2010; NIDCD, 2017). A study comparing 60 different diseases and 15 conditions, such as cancer, Alzheimer's disease, depression, and the loss of limbs, found that aphasia has the most significant negative impact on quality of life (Lam & Wodchis, 2010). Stroke occurrence rates are currently rising ; in Canada, strokes are occurring at a rate of once every five minutes (Heart and Stroke Foundation, 2022).

As the number of people living with aphasia continues to grow, the demand for timely and accessible speech-language assessment and therapy services will increase accordingly. However, clinical resources are limited, and traditional assessments require trained professionals to manually transcribe and analyze speech samples. Automating parts of this workflow using speech technologies such as ASR helps address this gap by enabling more scalable, cost-effective, and frequent assessments. To be clinically meaningful, however, ASR systems must be evaluated for their ability to handle the atypical and diverse speech profiles of people with aphasia (PwA).

Clinical interventions for PwA have traditionally been carried out without the aid of ASR or detailed transcriptions. However, the move toward automating aspects of clinical evaluation and treatment require highly detailed transcriptions that "appropriately account for various error types and linguistic characteristics specific to aphasic speech" (Wang, Cheng, Sufi, Fang & Mahmoud, 2024). Recent advances in the performance of automatic speech recognition (ASR) models have highlighted their focus on capturing intended speech versus verbatim speech. This leads to a tendency to automatically correct speakers' utterances, thereby removing semantic, syntactic, or

terminological errors that are necessary to assess a patient's state (Romana, Koishida & Provost, 2023). Most commercial ASR systems are optimized for fluent, normative speech and evaluated using metrics like Word Error Rate (WER), which fail to capture the clinical impact of errors in PwA speech.

Therefore, it is important to properly evaluate and compare ASR models on individual speech phenomena when applied to aphasic speech. This study provides a comparative evaluation of recent ASR models on a large dataset of aphasic speech. An open-source Python software that implements all the evaluation methods and metrics used in this article is publicly available.

2.2.1 Objectives

To build our proposed framework for evaluating ASR systems on aphasic speech, we set the following objectives :

- Introduce a novel framework for evaluating the performance of ASR models on aphasic speech.
- Using the proposed framework, identify improvements that can offer increased performance of ASR for different speech phenomena in the context of aphasia.

2.2.2 Automatic Speech Recognition Applied to Aphasia

Many efforts have already been made to apply ASR systems to aphasic speech. Researchers created an ASR model fine-tuned on aphasic speech for the Dutch language, finding that the most relevant features to distinguish PwA from the control group are speech rate, long pauses, grammatical complexity, repetitions within words, and semantic errors (Barberis *et al.*, 2024).

Some systems focus on capturing intended speech and, therefore, handle disfluencies by eliminating filler words and recurring utterances, prioritizing clarity but omitting clinically relevant aspects of speech. In contrast, verbatim speech transcriptions capture all articulated utterances and are more useful in the case of clinical assessment (Romana *et al.*, 2023). The choice between these approaches depends on whether clinical analysis or general comprehensibility

is prioritized. Novel approaches to disfluency handling have also emerged. CrisperWhisper, introduced by Zusag, Wagner & Thallinger (2024), transcribes speech verbatim with accurate word-level timestamps and improves timestamp accuracy around disfluencies and pauses by improving the tokenizer.

The out-of-domain adaptation approach trains models initially on more abundant non-aphasic speech data and then adapts them using smaller amounts of aphasic data, addressing the persistent challenge of data scarcity in this field (Le & Provost, 2016; Le, Licata, Persad & Provost, 2016). This approach leverages transfer learning techniques, where a model trained on general speech data is fine-tuned with aphasic speech data. Deep Neural Networks (DNNs) and multi-task learning have been particularly effective in developing ASR systems for aphasic speech using both in-domain and out-of-domain speech data (Cong *et al.*, 2024; Ilias & Askounis, 2022).

2.2.3 Need for Detailed Evaluation Based on Specific Phenomena

Speech disfluencies are interruptions or irregularities in the flow of speech, such as repetitions, fillers, and pauses, that deviate from typical fluency patterns. Impairments to working memory and increased cognitive load tend to lead to higher rates of disfluencies (Yuan, Cai, Bian, Ye & Church, 2021). Studies have shown that disfluencies negatively impact the performance of speech recognition systems (Lea *et al.*, 2023). Concurrently, disfluencies represent important phenomena that can be used in the diagnosis and evaluation of aphasia and other impairments (McGregor & Hadden, 2020; Romana, Niu, Perez, Roberts & Provost, 2022; Yuan *et al.*, 2020). Of course, many diagnostic tests for aphasia and other speech impairments, such as word list repetition, reading aloud, or comprehension tasks, rely on capturing accurate speech output, including errors and disfluencies (Cong *et al.*, 2024). As current ASR models often tend to omit these irregularities, this can lead to misleading or incomplete transcriptions and therefore impact any clinical applications. This highlights why it is essential to measure ASR system performance specifically in relation to such phenomena.

The evaluation of ASR models, particularly in the context of aphasic speech, remains primarily focused on WER. This metric calculates the rate of words incorrectly transcribed in relation to the ground truth. The Open Automatic Speech Recognition Leaderboard (Srivastav *et al.*, 2023), a popular online dashboard that compares the performance of many state-of-the-art transcription models, provides only WER metrics across various datasets. When Szymański *et al.* (2020) analyzed claims of low WER achievements for some ASR models, it was found that real-life WERs were much higher than what had been claimed. They suspected that an "optimistic bias toward ASR accuracy" contributed to published results being unfaithful to actual performance. Reasons for this bias include the overuse of unrealistically clean or simple datasets, and in-domain training as well as the use of methods like oracle segmentation where segments of speech are cleanly separated before being processed. One of the final recommendations stipulated that researchers should focus on “constructing new ASR quality measures, based on more richly annotated data, to better evaluate various aspects of transcription quality.” We therefore recognize that a more detailed analysis may be more resistant to biased outcomes than a singular WER figure.

While WER provides a general measure of transcription accuracy, it fails to capture the nuanced performance of ASR systems on specific speech phenomena relevant to aphasia assessment. This is because WER treats all word-level errors equally without accounting for the linguistic or clinical significance of those errors. For example, omitting a filled pause or disfluency may affect the WER score minimally but can obscure critical diagnostic indicators. For example, in a sentence where filled pauses and false starts account for 20% of words, a transcription omitting all these occurrences, but that is otherwise accurate, would produce a WER of 0.2. While such a result could be considered acceptable depending on the scenario, the transcription itself would give the impression of entirely fluent speech and would provide no pertinent clinical indicators. An accurate transcription of disfluencies, pauses, fillers, and other speech markers is essential in clinical applications, as these features inform diagnostics and treatment planning. Clinical metrics and protocols such as the Assessment Protocol for Communication Samples in Aphasia (APROCSA) rely on linguistic markers, including lexical diversity, syntactic complexity, and

disfluency patterns, to characterize aphasia severity and type (Casilio *et al.*, 2019; Ezzes *et al.*, 2022). However, these metrics are only as reliable as the transcripts they are based on. If ASR output omits or distorts clinically relevant features, the derived metrics may be misleading, undermining the validity of subsequent assessments or treatment decisions.

Open-source tools are particularly well suited to the study and treatment of aphasia due to their adaptability by the research community. Researchers and clinicians can use these tools to better capture the idiosyncrasies of aphasic speech, such as disfluencies, atypical pause patterns, and altered prosody (Wang *et al.*, 2024). Also, the collaborative nature of open-source projects often leads to the rapid improvement and expansion of functionalities, keeping pace with the research community's evolving needs.

Building on existing research, we developed an open-source evaluation framework specifically designed for the performance assessment of ASR systems when applied to aphasia. This framework implements a novel evaluation method that assesses performance on specific linguistic phenomena relevant to aphasia evaluation.

2.3 Materials and Methods

This section outlines the study's methodology, detailing the data sources, processing steps, evaluation framework, and experimental setup.

This research focuses on evaluating the performance of ASR systems when applied to aphasic speech. The methodology introduces a comprehensive pipeline encompassing the parsing of conversational speech data, transcription using ASR models, error analysis, and performance benchmarking. A dedicated evaluation framework is developed and empirically validated on a large-scale aphasic speech dataset using a selection of ASR systems.

2.3.1 Data

The study utilizes the AphasiaBank (MacWhinney *et al.*, 2011) database, accessed in September 2024. AphasiaBank offers a diverse range of speech samples recorded in various clinical settings. The study incorporates all available English and French protocol corpora, encompassing PwA and control groups. This ensures a diverse dataset regarding recording environments, equipment, speaker demographics, aphasia types, and severity levels. The PwA subset includes transcriptions from participants with a wide range of aphasia subtypes, including Broca's, Wernicke's, conduction, and anomic aphasia, thereby increasing the likelihood of capturing a broad spectrum of speech phenomena, many of which vary in frequency depending on the subtype.

2.3.1.1 CHAT Format

Transcriptions from the AphasiaBank database employ the CHAT format (MacWhinney, 2000), a structured text file standard designed for detailed linguistic and conversational analysis. CHAT provides extensive coding specifications for representing different features and types of speech, making it well suited to the goals of our evaluation. Therefore, it was selected as the underlying format for this project. To facilitate the use of this data in computational workflows, we developed a custom conversion library that transforms between CHAT files contents and an object-oriented data structure. This intermediate representation enables efficient querying and manipulation for a range of evaluation tasks. To streamline parsing, the PyLangAcq library (Lee, Burkholder, Flinn & Coppess, 2016) was integrated into our pipeline. An example of a CHAT coded sentence is "<I want> [/] I want &-uh cereal &+m (.) please." ; here we can observe a scoped repetition ("<...>") for the words "I want," a filled pause ("uh"), a phonetic fragment ("m"), and a pause ("(.)"). Figure 2.1 is a visual representation of the type of output the custom library provides.

In the framework's implementation, tokens are the objects that make up utterances ; these can include words, punctuation, and word fragments, as well as special tokens to indicate pauses, specific sentence terminations, and types of errors. Each token can have associated metadata to

denote speech phenomena. Annotation data can also be attributed to entire utterances for some kinds of tokens.

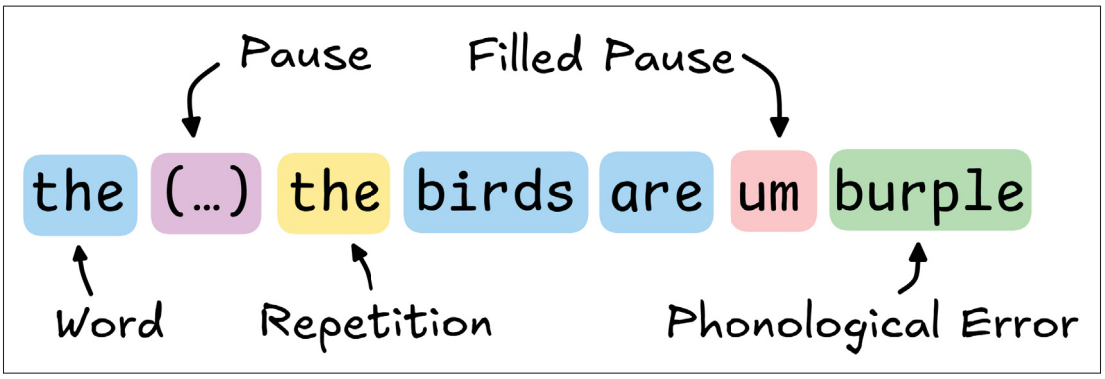

Figure 2.1 Visual representation of CHAT-coded transcription as tokens

2.3.1.2 Datasets

To compare the performance of ASR systems across different groups of speakers, we draw on available English language data from the AphasiaBank database. The dataset is divided into two primary groups :

- PwA subset : This subset comprises the 32 corpora included in the protocol English PWA database from AphasiaBank and includes 996 recorded media files, totalling approximately 307 hours of audio. The dataset represents 476 individuals. The average session duration is 18 minutes and 20 seconds. Individuals in this subset represent a wide range of aphasia types and severities.
- Control subset : This subset comprises the 10 corpora included in the protocol English control database from AphasiaBank, totalling 386 media files, equivalent to approximately 128 hours of recorded audio. The dataset represents 352 individuals.

The final dataset, which includes both the media file and its corresponding CHAT transcript, consists of 996 files for the PwA set and 386 for the control set. For all evaluations in this study, only participant utterances were considered. Interviewer utterances were ignored to measure ASR system performance on disfluent speech more accurately.

A secondary comparative evaluation is performed using the available French language data from the AphasiaBank database. This dataset is comparatively limited and contains a total of approximately 3.5 hours of audio from 12 PwA without control subset.

2.3.1.3 Excluded Data

Several English language media files were excluded due to errors encountered during their parsing. Eleven media files were excluded from the PwA subset because they lacked an associated CHAT file. Also, 20 files in the PwA subset and 4 files in the control subset were excluded from evaluation due to errors in the reference CHAT file or non-conformance with the CHAT format specifications, such as undefined word error codes.

2.3.2 Evaluation and Metrics

A dedicated library was developed to evaluate ASR performance, allowing fine-grained error analysis across different speech phenomena. The evaluation involves parsing both the reference and ASR-generated CHAT files. Transcripts are structured into utterances, which are further segmented into tokens. Utterances are usually formed of a main clause along with any dependent clauses, approximating the form of a traditional sentence, although they may often be incomplete as they represent natural speech. Each token is classified based on speech phenomena such as repetitions, grammatical errors, and jargon.

2.3.2.1 Word Error Rate and Character Error Rate

WER and CER are computed over all tokens produced for each participant, both overall and for each speech phenomenon present. Reference transcripts are evaluated against ASR transcriptions using Levenshtein distance. Levenshtein distance is the lowest number of operations required to transform one text string into another. The available operations are deletions, substitutions, and insertions. For example, to go from "bag" to "bat", the "g" needs to be substituted with a "t", resulting in a Levenshtein distance of 1. In the case of CER, each character is checked to see if

an operation is to be applied ; in the case that the truth and prediction character is not the same, this means an operation needs to be applied and the distance is increased by one. In the case of WER, this process is done on a per word basis. Levenshtein distance calculations are done using the implementation provided in the RapidFuzz library (Bachmann *et al.*, 2025).

WER and CER are then calculated using the operations extracted from the Levenshtein distance, as shown in equations 2.1 and 2.2. For WER, the total number of words is equal to the number of words in the reference transcript. For CER, the total number of characters is equal to the number of characters in the reference transcript. These calculations output a positive ratio between 0 and infinity. An error rate of 0 represents the comparison of two identical text strings. An error rate larger than 1 signifies that there are more errors than there are words or characters in the reference text. For example, a word level Levenshtein distance of 10 for a sentence of 5 words would produce a WER of 2.

$$WER = \frac{Deletions + Substitutions + Insertions}{TotalNumberOfWords} \tag{2.1}$$

$$CER = \frac{Deletions + Substitutions + Insertions}{TotalNumberOfCharacters} \tag{2.2}$$

The Levenshtein distance can be calculated in either direction (Example : text A vs. text B, or text B vs. text A) with the difference in output being the inversion of the number of insertion and deletion operations. As an inversion in these numbers, however, has an effect on error rate calculations, typically the reference text is compared to the predicted text in the manner that a missing character or word in the prediction is counted as a deletion and an added character or word in the prediction is counted as an insertion. For example, turning "banana" into "ba" provides a CER of 0.67 while turning "ba" into "banana" produces a CER of 2 ; both cases have a Levenshtein distance of 4. This is the method employed in the current framework. When calculating the Levenshtein distance for characters, all types of characters are included, including spaces ; however when calculating the Levenshtein distance for words, spaces are used as the

separators between words and are therefore not directly tallied in the operations as they are for CER.

Figure 2.2 provides an example of a WER calculation, and Figure 2.3 provides an example of a CER calculation.

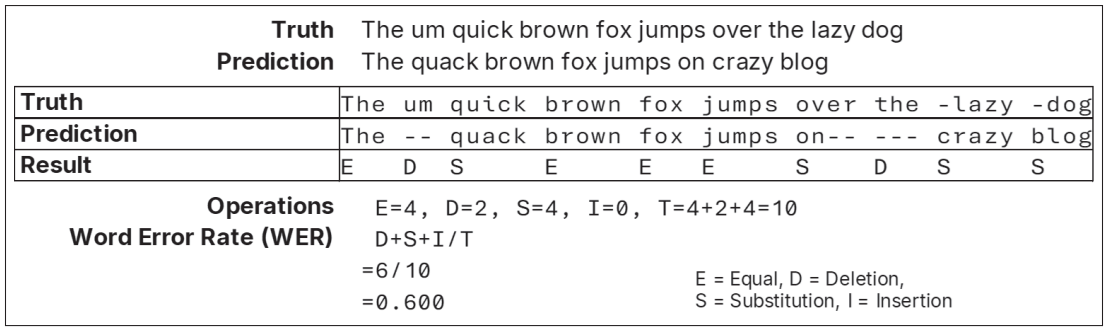

Figure 2.2 Example of WER calculation

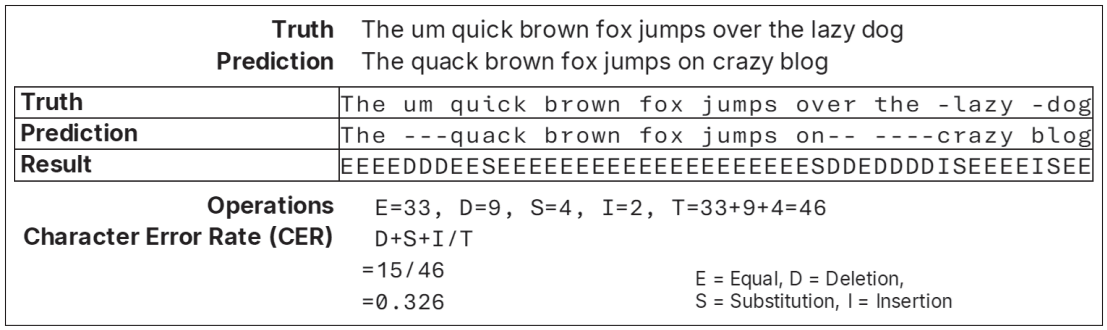

Figure 2.3 Example of CER calculation

Referencing the original coded transcripts, each edit operation is attributed to the appropriate speech phenomena based on the associated token. Error rates are then calculated on a per-phenomenon basis. Figure 2.4 shows an example of CER calculation applied to filled pauses. In this example, we can observe that only characters that are marked as being part of filled pauses such as "um" or "er" are being included in the CER calculation. The filled pause annotation information is provided by the reference CHAT file where filled pauses are coded with an ampersand-hyphen (&-) prefix.

When aggregating metrics across multiple sessions, the tool returns two outputs :

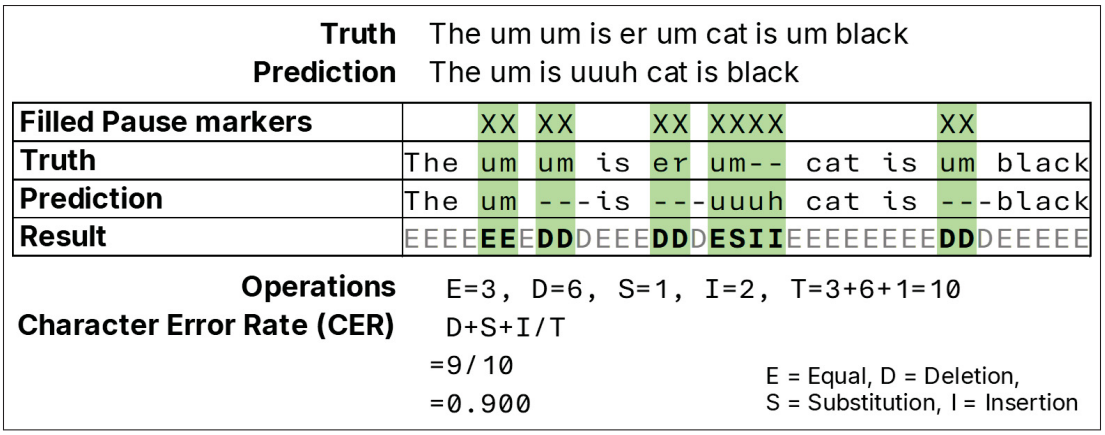

Figure 2.4 Example of CER calculation applied to specific phenomena

- Averaged results, where each session is equally weighted. WER and CER are both summed and divided by the number of sessions regardless of their respective length. Equation 2.3 shows the equation for averaging CER results. For example, if two sessions are averaged, with the first having a CER of 0.3 and the second a CER of 0.5, the final averaged result would be a CER of 0.4.

$$AveragedResults = \frac{\sum_{i=1}^{n} CER_i}{n} \quad , \; n = number\ of\ sessions \tag{2.3}$$

- Weighted results, where weighing is done by the number of words or characters; therefore, longer sessions have more impact and vice versa. The underlying primitive values of the WER and CER calculations, meaning the Levenshtein edit operations, are summed, and the metrics are then calculated based on the entirety of the data. This is effectively the global WER and CER over all transcriptions. Equation 2.3 shows the equation for weighted aggregation of CER results. Using the same example of two sessions with CERs of 0.3 and 0.5, if the first session has 1 deletion, 2 substitutions and 0 insertions over 10 characters total, while the second session has 10 deletions, 30 substitutions and 10 insertions over 100 characters total, the final weighted result for CER would be 0.48. The second session, which contains significantly more content than the first, has a greater impact on the final result.

$$WeightedResults = \frac{\sum_{i=1}^{n}(Deletions_i + Substitutions_i + Insertions_i)}{\sum_{i=1}^{n}(TotalNumberOfCharacters_i)} \quad (2.4)$$
$$, \; n = number \; of \; sessions$$

This study prioritizes weighted results for ASR benchmarking, as they better reflect system performance across varied datasets. However, averaged results may be more relevant for highlighting possible variations between sessions as well as analyzing specific cohorts or session-based trends.

2.3.2.2 Choice of Character Error Rate over Word Error Rate

WER is the more commonly used metric in ASR-related research. This is mainly related to the goal of achieving perfect, error-free performance. CER, however, provides more information as it is less prone to phonetic errors. For example, the words "rapture" and "capture" sound similar apart from their beginnings and would produce a CER of ~ 0.14, while they would produce a WER of 1.00. WER provides a binary judgment and is therefore useful to know if a word is right or wrong, but CER provides information on how close those words might be. The sensitivity of CER makes it more responsive to speech disruptions, such as repetitions ("ba-ba-banana") or phonetic errors ("panana"). Therefore, to better capture information related to disfluencies, CER seems to be a better indicator than WER. While the evaluation tool supports WER and CER metrics, the discussed results will focus mainly on CER.

One caveat is that word spelling might not always directly correspond to its pronunciation. For example, depending on the speaker, "pupil" and "people" sound similar but give a CER of 0.8. This means that in some word instances, CER may provide results that are not reflective of the actual phonetic similarities of spoken words. The metric is therefore more suitable when applied to larger amounts of aggregate data.

2.3.2.3 Effect on Clinical Indicators

The proposed metrics focus on specific speech phenomena. In a clinical context, these performance indicators are essential as the transcription accuracy of these phenomena can be directly related to clinical indicators used in the context of aphasia. For example, quantitative linguistic measures (utterance length and complexity, phonological and semantic errors, etc.) are often used for aphasia detection and assessment tasks (Casilio *et al.*, 2019). Many of these metrics rely wholly on the accurate transcription of aphasic speech, including the disfluencies known to present challenges for ASR systems. The proposed metrics therefore serve both to properly evaluate ASR systems intended for clinical indicator extraction, and to identify which of these indicators can be reliably derived from a given ASR model.

2.3.3 Experimental Setup

This study serves as a proof-of-concept for the evaluation tool, benchmarking ASR performance on aphasic speech. To show the applicability of the evaluation framework, we applied it to the predictions of 3 ASR models to demonstrate the variations in transcription quality for each annotated phenomenon in the dataset.

Figure 2.5 visualizes the experimental setup architecture, with the two main steps being the processing of multimedia files into CHAT-formatted transcript predictions (Audio to CHAT pipeline) and, finally, the evaluation of the newly created transcripts against the reference transcripts (Evaluation Framework).

2.3.3.1 Audio to CHAT Pipeline

A processing pipeline was designed to convert raw audio recordings into structured CHAT transcripts. The pipeline consists of several stages :

- Preprocessing : All audio files are converted to 16 kHz WAV format to ensure standardization of input. Audio files are trimmed in duration to the timestamps available in the associated true

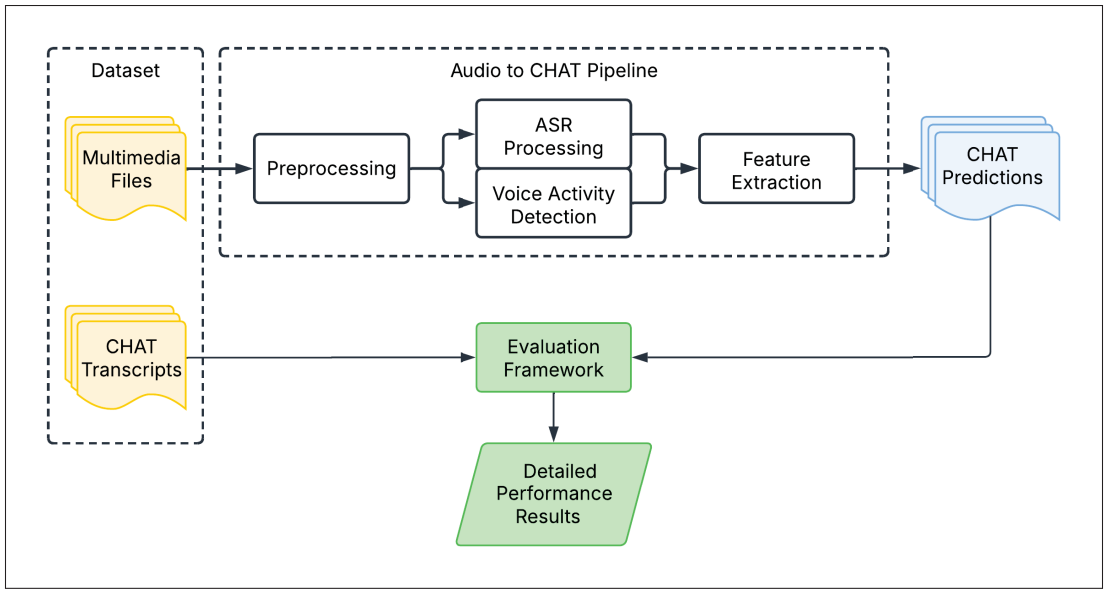

Figure 2.5 Simplified Experimental Architecture Diagram

transcript CHAT file. This is done as some media files are not transcribed in their entirety, therefore only the transcribed portions are kept. Using the reference transcript, the audio is segmented into each utterance and then processed sequentially to bypass speaker diarization.

- ASR Processing : The system detects the language (if unspecified), performs speech segmentation, transcription, and time alignment. This step is performed with the WhisperX library, which is compatible with many ASR models. The segmentation model, used to portion audio inputs into 30 seconds or less chunks, is the Pyannote segmentation-3.0 model (Bredin, 2023; Plaquet & Bredin, 2023). Forced time alignment is performed using Wav2Vec2 (Baevski *et al.*, 2020) with a Connectionist Temporal Classification (CTC) algorithm ; this step adds word level timestamps to the transcription. To avoid the varying performance impact of the diarization model on the evaluation, speaker identities were applied to automatic transcription segments by referring to the original transcriptions.
- Annotation Encoding : Additional processing extracts repetitions and disfluencies, such as filled pauses, and part-of-speech tags and applies the proper CHAT based annotations to each token.

The final output is a CHAT transcript with coded speech features. While the pipeline annotates basic linguistic elements such as filled pauses like "um" and "uh", more complex semantic-level annotations such as semantic errors, like using the word "mother" instead of "father", were outside the scope of this study and remain a challenge for future works.

2.3.3.2 Prompting

Models were evaluated both with and without prompting to observe the effect of prompting on ASR performance. Prompting, in this context, refers to the use of a short textual prefix provided to the model prior to transcription. This technique is made possible through the architecture of large language-based ASR models such as Whisper, which allows for text conditioning at inference time (Radford *et al.*, 2022; OpenAI, n.d.).

The prompt used ("um, hm, okay, so, mm, and, uh, mhm") was chosen to specifically target filled pauses. These words were selected based on their frequency in the aphasic speech dataset and their relevance to disfluency analysis. The list was built manually by inspecting common fillers in the data. Other prompts were briefly tested during preliminary experimentation ; the current version was selected for its minimal length and stronger performance. The goal was to evaluate whether explicitly biasing the model toward recognizing filled pauses would improve ASR performance related to the phenomenon.

2.3.3.3 Model Selection

For the purposes of setting a general baseline for ASR performance, a publicly available local version of OpenAI's Whisper Large-v2 (L2) and Large-v3 (L3) models are used (Radford *et al.*, 2022). No fine-tuning or additional training is done, and the model is used as-is. To evaluate the effect of prompting on model performance on specific types of speech, models are evaluated both with and without specific prompting.

The analysis also includes Nyra Health's CrisperWhisper (CW) model, a model designed for verbatim transcription of speech containing disfluencies (Zusag *et al.*, 2024). This model is

a fine-tuned version of Whisper Large-v2, trained on a selection of verbatim speech datasets. Importantly, AphasiaBank data has not been used for this model's training. Therefore, there is no risk of data leakage or bias, ensuring an accurate evaluation of the model's capabilities.

For reproducibility, in the appendix, Table II-2 presents the specific model versions used in the study while Table II-1 presents the ASR model configuration applied.

2.4 Results

This section presents the results of our evaluation of ASR models in transcribing disfluent speech, focusing on different speech phenomena and the impact of prompting. We evaluate the model's performance on aphasic speech by analyzing WER and CER.

Below is a list of the selected phenomena categories and an example for each :

- Empty Speech : "we got little things over here", meaningless speech
- Filled Pauses : "um, uh"
- Grammatical Errors : "whatever I'm think up"
- Jargon : unintelligible speech, invented words
- Phonological Fragments : "w", sound not connected to any word
- Repetitions : "I wanted I wanted to"
- Retracing : "I wanted I thought I wanted to"
- Phonological Errors : "panana" instead of "banana"
- IPA Transcriptions : "feɪʒə" (aphasia), words which include phonetic characters
- Semantic Errors : "Mother" instead of "Father"
- Morphological Errors : "child" instead of "children"
- Unannotated words : words that are not covered by an annotation of other phenomena

Table 2.1 summarizes WER results for selected speech phenomena when evaluating ASR models and prompt use across the entire corpus available on AphasiaBank. Using the same parameters, Table 2.2 summarizes CER results.

Tableau 2.1 Word Error Rate (%) of speech phenomena
for different models and prompt combinations
Best results in bold
NP = Without Prompting, P = With Prompting

Phenomenon	**L2**		**L3**		**CW**	
	NP	P	NP	P	NP	P
Empty Speech	44.94	42.32	44.63	41.34	**38.08**	42.97
Filled Pauses	57.49	44.38	59.07	45.18	**42.64**	44.02
Grammatical Errors	47.00	42.79	46.97	43.80	**41.01**	43.17
Jargon	56.03	53.95	55.55	53.60	**47.79**	52.14
Phonological Fragments	65.43	58.06	64.22	58.33	**51.44**	54.64
Repetitions	50.74	45.87	47.78	**43.67**	44.23	46.38
Retracing	50.45	43.72	49.38	43.76	**42.27**	45.74
Phonological Errors	62.03	57.42	60.91	57.66	**53.86**	57.61
Unannotated words	38.31	36.33	37.57	36.69	**35.78**	41.46
Overall	40.40	38.04	39.76	38.91	**37.34**	42.11

Table 2.1 shows that the CW model without prompting achieves the lowest WER across most speech phenomena, including Empty Speech (38.08%), and Phonological Errors (53.86%). Prompting improves performance for L2 and L3 in several cases, such as for Filled Pauses, where L2 drops from 57.49% to 44.38% and L3 from 59.07% to 45.18%. Phonological Fragments also see improvements with prompting, decreasing from 65.43% to 58.06% for L2 and from 64.22% to 58.33% for L3. In contrast, prompting tends to degrade performance for CW, for instance increasing overall WER from 37.34% to 42.11%. The highest WERs are found in Phonological Fragments (65.43%) and Phonological Errors (62.03%), while Unannotated Words show the lowest error rate overall.

Table 2.2 shows that the CW model without prompting achieves the lowest CER for most phenomena, including Empty Speech (23.21%), Grammatical Errors (26.72%), and Jargon (37.48%). Prompting improves CER for L2 and L3 in several categories ; for example, Filled Pauses improve from 86.38% to 41.85% for L2, and from 87.82% to 44.04% for L3. However, prompting increases CER for CW in most cases, such as for Empty Speech (from 23.21% to 30.97%) and also overall (from 27.80% to 34.15%). The highest CERs overall are observed

Tableau 2.2 Character Error Rate (%) of speech phenomena for different models and prompt combinations
Best results in bold
NP = Without Prompting, P = With Prompting

Phenomenon	L2		L3		CW	
	NP	P	NP	P	NP	P
Empty Speech	33.43	29.48	33.60	28.38	**23.21**	30.97
Filled Pauses	86.38	41.85	87.82	44.04	36.97	**35.91**
Grammatical Errors	33.91	29.48	33.85	28.91	**26.72**	31.23
Jargon	45.50	43.03	44.94	41.62	**37.48**	42.41
Phonological Fragments	72.52	70.47	72.75	69.16	65.72	**63.98**
Repetitions	44.19	33.90	42.17	**31.35**	31.73	35.19
Retracing	50.18	39.70	50.28	37.81	**37.47**	42.45
Phonological Errors	66.38	65.02	65.57	**63.83**	64.98	67.09
Unannotated words	25.99	23.96	25.28	22.64	**22.61**	31.49
Overall	30.82	29.38	30.45	29.21	**27.80**	34.15

in Phonological Fragments (72.75%) and Filled Pauses (87.82%), while Unannotated Words consistently show lower error rates.

A comparison of L2 and L3 across Tables 2.1 and 2.2 shows that L3 provides a slight edge over L2 on some speech phenomena, particularly when prompting is applied. However, the differences in both WER and CER between the two model versions remain small overall and are unlikely to be significant. Performance across most categories is comparable, suggesting that both model versions behave similarly in terms of transcription accuracy for aphasic speech.

Transcriptions from PwA and control subsets were compared using the Large-V3 model without prompting to analyze the different outcomes between the two groups. These results are presented in Table 2.3 Across all phenomena, error rates are higher for the PwA group. The largest differences in CER are observed for Grammatical Errors (19.73%) and Phonological Errors (20.39%). The smallest differences appear for Filled Pauses (0.48%) and Phonological Fragments (1.47%).

Tableau 2.3 Word and Character Error Rates (%) on Selected Speech Phenomena Comparing Control and Aphasia Groups on Large-V3 Model (L3) Without Prompt Best CER results in bold

Phenomenon	**PwA**		**Control**		**PwA - Control (Δ)**	
	CER	WER	CER	WER	CER	WER
Filled Pauses	87.82	59.07	**87.34**	46.15	0.48	12.92
Grammatical Errors	33.85	46.97	**14.12**	20.78	19.73	26.19
Phonological Fragments	72.75	64.22	**71.28**	49.4	1.47	14.82
Repetitions	42.17	47.78	**30.64**	28.78	11.53	19.00
Retracing	50.28	49.38	**44.87**	36.69	5.41	12.69
Phonological Errors	65.57	60.91	**45.18**	50.00	20.39	10.91
Unannotated words	25.28	37.57	**15.44**	23.56	9.84	14.01
Overall	30.45	39.76	**18.42**	24.54	12.03	15.22

As shown in Table 2.4, the evaluation of ASR performance on French aphasic speech revealed consistently higher transcription error rates compared to English aphasic speech across all measured linguistic phenomena. CER was used as the primary metric. Overall, with the L3 model, French aphasic transcriptions exhibited a 32% higher CER than their English equivalents.

This discrepancy was evident across a wide range of clinically relevant speech phenomena. For instance, semantic errors in French reached CERs of 62.50% (L3) and 58.64% (CW), compared to markedly lower rates in English. Morphological errors showed even greater disparity, with CW yielding a CER of 64.86% in French, versus 26.13% in English. Similar patterns were observed for phonological errors, retractions, and repetitions.

2.5 Discussion

First, section 2.5.1 compares ASR models and the effect of prompting. Then, section 2.5.2 examines performance differences between individuals with PwA and a control group. Finally, section 2.5.3 presents an evaluation of the ASR models when applied to French.

Tableau 2.4 Character Error Rates (%) on Selected Speech Phenomena for French and English Aphasia Groups
Best CER results in bold

Phenomena	French		English		French - English (Δ)	
	L3	CW	L3	CW	L3	CW
Repetitions	47.64	51.56	42.17	**31.73**	5.47	19.83
Retracing	57.76	48.78	50.28	**37.47**	7.48	11.31
Phonological Errors	75.51	72.32	65.57	**64.98**	9.94	7.34
Non-words	87.17	77.01	77.08	**73.42**	10.09	3.59
IPA Transcriptions	78.70	**70.39**	79.24	77.97	-0.54	-7.58
Semantic Errors	62.50	58.64	38.19	**35.66**	24.31	22.98
Morphological Errors	43.33	64.86	29.60	**26.13**	13.73	38.73
Unannotated words	37.97	45.59	25.28	**22.61**	12.69	22.98
Overall	40.33	43.86	30.45	**27.80**	9.88	16.06

2.5.1 Model and Prompting

Observing CER metrics at a per-phenomenon level provides a more nuanced view of model performance. As expected, without any prompting, CW outperforms L2 and L3 on speech features that benefit from verbatim transcription, such as filled pauses and phonological fragments. The impact of prompting is particularly evident in the performance shifts of L2 and L3, as detailed in Table 2.2. The given prompt was chosen to determine if performance could be improved and measured in terms of filled pauses. Results show a significant improvement for this feature with a 49.85% reduction of CER between L3 without prompting (87.82%) and L3 with prompting (44.04%).

Unexpectedly, this prompt also had the effect of improving CER for many other features being measured, resulting in better overall performance. For L3, prompting decreased CER for repetitions from 42.17% to 31.35%, and overall CER decreased from 30.45% to 29.21%. This improvement may be attributed to the prompt guiding the ASR models to focus more on detecting hesitations and disfluencies, leading to better alignment with the conversational patterns of aphasic speech. Additionally, explicit prompts may compensate for this bias since

filled pauses often appear in natural speech but are underrepresented in training data (Wang *et al.*, 2024).

A more detailed comparison is shown in Figure 2.6, which presents the distribution of CER across models. In the left-hand plot, which shows results without prompting, there is a clear performance gap between the baseline models (L2 and L3) and the verbatim-optimized model (CW). However, the right-hand plot, illustrating results with prompting, shows a marked convergence in CER distributions. This larger overlap indicates that, for specific speech phenomena, prompting enables baseline models to approach the performance of a model explicitly fine-tuned for verbatim transcription (CW). Notably, this is achieved without the need for additional training on domain-specific aphasic data.

These findings suggest that carefully designed prompts can serve as a lightweight strategy to enhance ASR performance on disfluent or clinical speech. Future work should investigate whether similar improvements can be observed when prompting is applied to models already fine-tuned on aphasic speech.

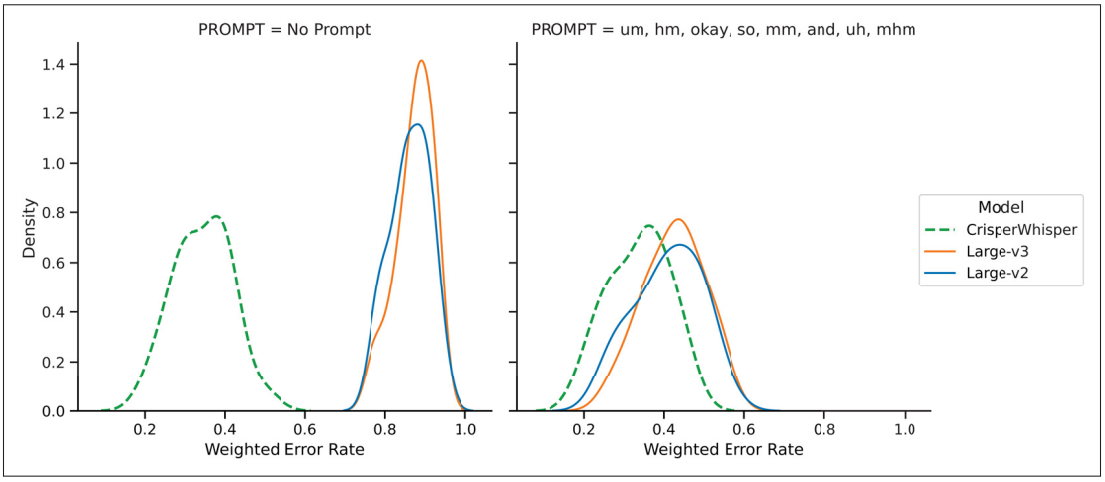

Figure 2.6 Filled Pauses CER Distribution for ASR Models with and without Prompting on English PWA subset

In contrast to the gains observed in L2 and L3, the application of prompting to the CW model led to notable declines in performance, as shown in Table 2.2 Surprisingly, when looking at

overall performance with prompting, the L3 model achieves comparable performance to the CW model without prompting. This could possibly be explained by the fact that the CW model was trained only on samples of healthy speech, and the authors only made use of the AphasiaBank corpus for evaluating hallucinations (Zusag *et al.*, 2024). This result suggests that prompting can serve as a lightweight alternative to fine-tuning for specific transcription tasks. Conversely, CW's decline in performance when prompted indicates that the model is heavily fine-tuned for strict verbatim transcription, leaving little room for adaptation when external linguistic cues are introduced.

Further inspection of Table 2.2 reveals that the poorest performance across all models was observed in two categories : phonological errors and phonological fragments. The reason for this seems to be twofold. Firstly, these phenomena mainly consist of transcribed sounds. Therefore, they appear in transcripts as either incomplete words or deliberate misspellings. For example, for the word "banana", it could appear as an error like "babana" or a fragment like "bana". Most transcription models, on the other hand, are trained to provide intended speech, coherent text with real words and without any faults Romana *et al.* (2023). Therefore, when transcribing, models will commonly skip transcribing these phenomena entirely or output a near-sounding real word. Secondly, in some corpora, these phenomena sometimes include IPA characters to denote specific sounds. This has a somewhat negative impact on performance, as these will automatically produce errors when they occur, since typical ASR models will never include any phonetic characters.

Figure 2.7 visualizes the data from Table 2.2 using a radar chart. L2 was omitted to improve the readability of the figure, given its strong similarity to L3. The CER scale is inverted ; therefore, better-performing models extend further to the edges of the chart. The larger the covered area, the better the overall performance.

This visualization is helpful for quickly comparing different models and parameters. It is evident that CW, without prompting, outperforms the other model combinations. The most evident

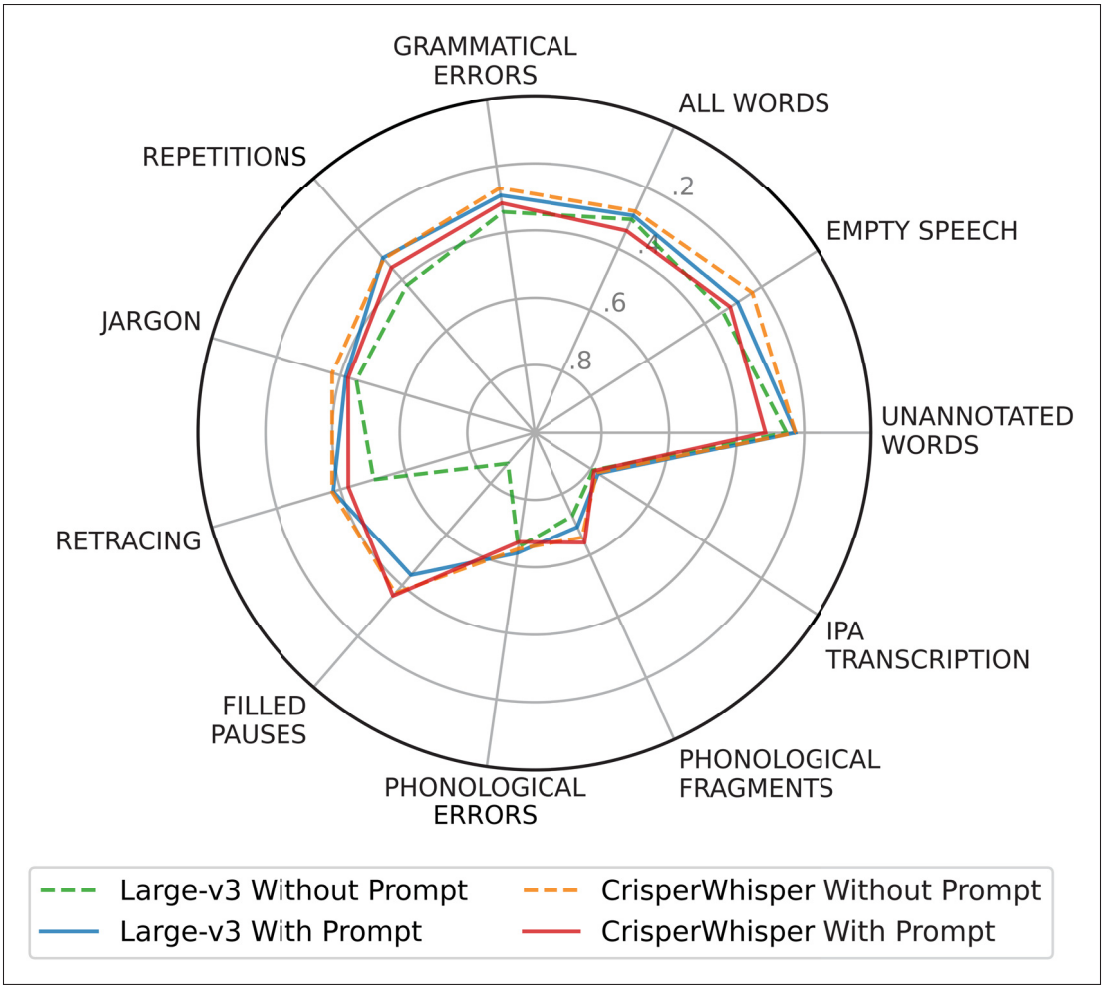

Figure 2.7 Radar chart of selected phenomena CER performance for different models and prompts on English PWA subset

difference is the effect of prompting on the L3 model, with a significant improvement in the filled pauses category as well as moderate improvements across all other phenomena.

2.5.2 PwA versus Control

Comparing results for the PwA dataset against the control dataset, we observe a significant decline in transcription performance across all phenomena when processing speech from individuals with aphasia, as shown in Table 2.3. The largest degradation was found in grammatical errors,

where CER increased by 140% in the PwA group compared to the control group. This decline aligns with expectations, given the additional complexity and variability present in disordered speech. Among all categories, filled pauses showed the smallest change in CER between the two groups. This can be attributed to the fact that, without prompting, the model frequently failed to transcribe filled pauses in both PwA and control datasets, suggesting that such disfluencies are often omitted by default, regardless of speaker condition.

Again, using radar charts, in Figure 2.8 and Figure 2.9, we can more easily observe the expected drop in performance when comparing PwA conversational data versus control data. The graphs show a consistent drop in performance across all phenomena.

Overall, the evaluation framework provides users with a structured and interpretable way to assess ASR performance on aphasic and conversational speech. By organizing CER and WER metrics across a range of clinically and linguistically relevant speech phenomena, the framework provides a detailed error profile for ASR models. This structured output enables direct comparisons across models, model conditions and parameters (e.g., prompted vs. unprompted), and speaker groups (e.g., PwA vs. control). It allows users to pinpoint specific weaknesses in model performance, such as difficulties with filled pauses or phonological fragments, and to track how these vary with different prompting strategies or model configurations. In doing so, the framework supports more informed ASR model selection, fine-tuning decisions, and future adaptations for clinical or accessibility-focused applications.

2.5.3 French Language Evaluation

While L3 slightly outperformed CW in French overall, both models struggled with the language, showing substantial degradation in performance compared to their English results. Notably, while the CW model showed improved results in English, it actually underperformed in French, with an 8.70% relative decline compared to L3. This suggests that newer model refinements are not necessarily optimized for multilingual robustness.

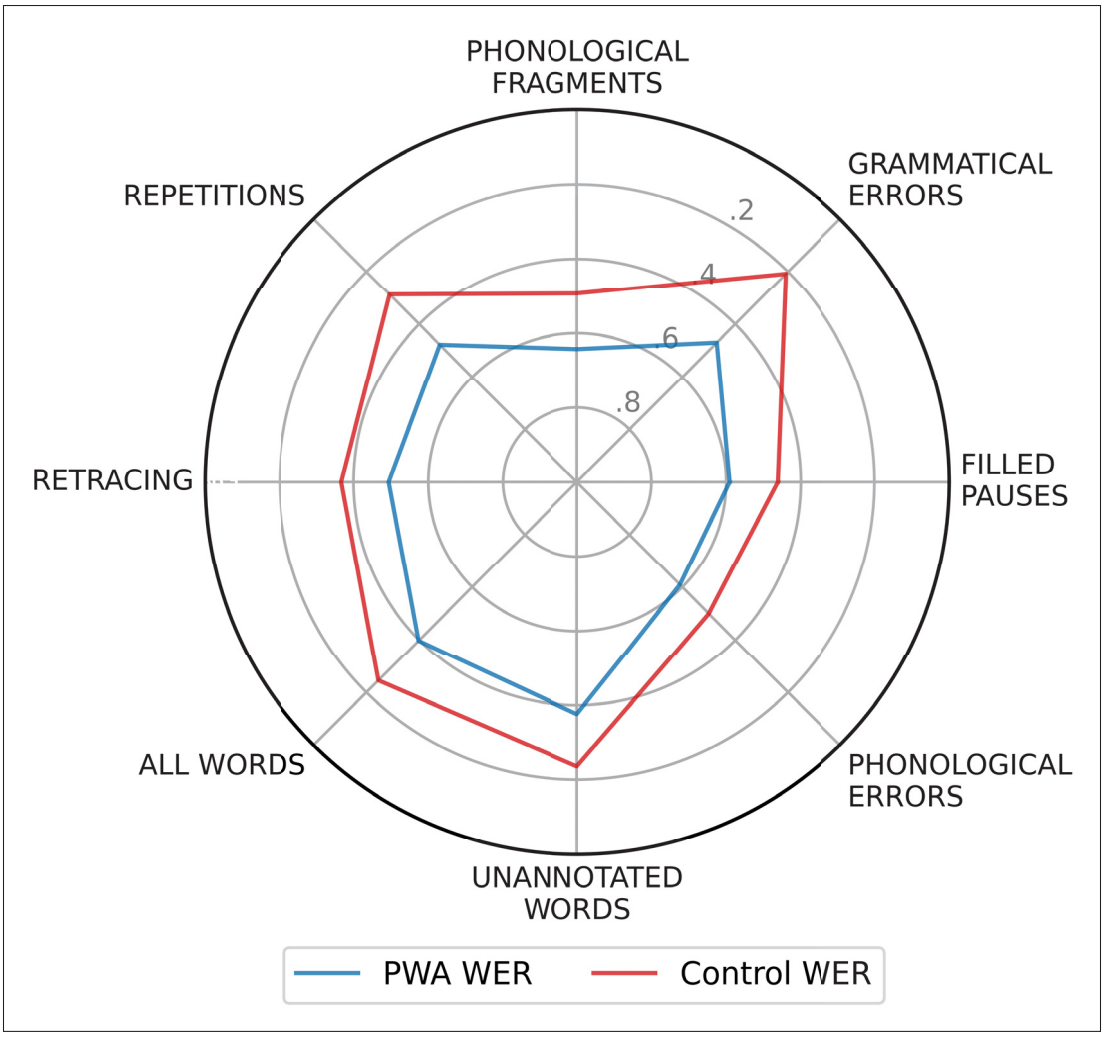

Figure 2.8 Radar charts of selected phenomena WER performance for English control vs. PwA groups

These results demonstrate clear limitations in the capacity of current ASR systems to accurately transcribe aphasic speech in French. The increased error rates, particularly in linguistically and diagnostically relevant categories such as semantic and morphological errors, point to a lack of generalizability across languages. While aphasic speech is inherently challenging to transcribe due to its disfluencies and atypical linguistic patterns, the significantly higher CER in French relative to English suggests that model performance is also affected by factors such as data scarcity.

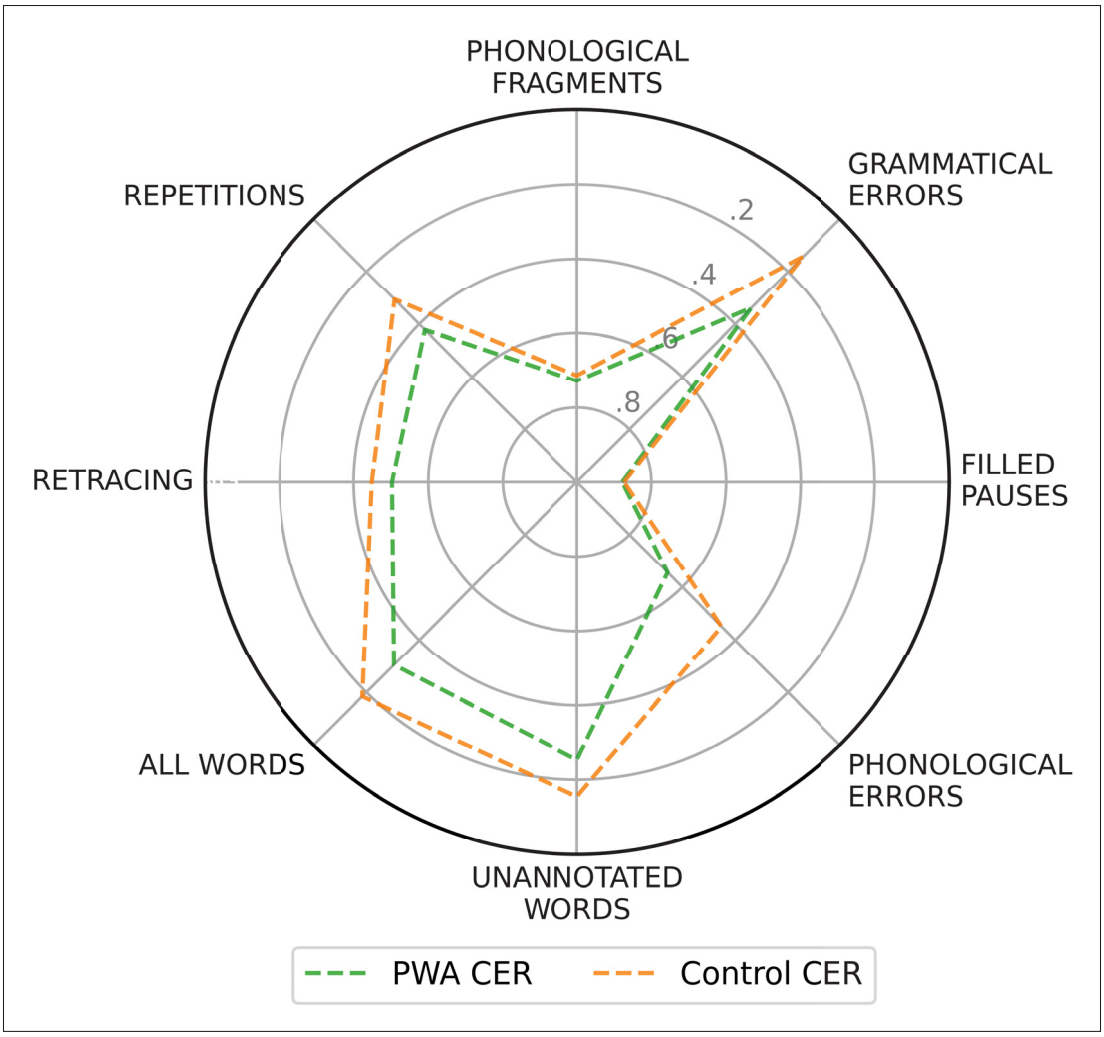

Figure 2.9 Radar charts of selected phenomena CER performance for English control vs. PwA groups

The French subset of the AphasiaBank corpus used in this study was substantially smaller and less richly annotated than the English corpus. Models trained or even simply fine-tuned predominantly on English data may learn patterns that are insufficient or even misleading when applied to French, especially when faced with aphasic speech.

The difference between French and English highlights the bias of existing models : models trained on large multilingual corpora perform unevenly across different languages. The weakness

of models on French-language data is precisely the type of conclusion that can be drawn using the evaluation framework.

These results highlight the importance of developing more inclusive multilingual ASR tools for clinical use. The evaluation framework introduced in the present study contributes directly to this goal by enabling a more detailed, phenomenon-specific analysis of ASR performance on aphasic speech. Clinicians and researchers working in francophone contexts should interpret ASR outputs with caution and consider the limitations identified here when integrating these tools into clinical workflows.

2.5.4 Limitations and Further Work

It should be noted that this work carries certain methodological and scope-level limitations.

Firstly, the proposed evaluation framework relies primarily on WER and CER metrics, which have their own limitations. In particular, WER may not be applicable to languages without explicit word boundaries in their writing systems, such as Japanese or Chinese, where text is not naturally separated by spaces.

Additionally, WER and CER metrics do not always accurately reflect the similarity or distance between words and characters. These metrics treat all errors as equal, without accounting for how closely an incorrect output resembles the target in form or sound. As previously mentioned, CER can offer more detailed insights than WER, especially by capturing partial errors within a single word. However, both metrics fall short when applied to conversational data, where pronunciation and orthography often diverge. For instance, the English word colonel is pronounced as "/kûrˈnəl/" (like *kernel*), bearing little resemblance to its written form. In such cases, even a phonetically accurate transcription may be penalized heavily by CER due to orthographic mismatch. This highlights a fundamental limitation; standard error metrics may not fully capture transcription quality in cases where phonetic accuracy matters more than spelling fidelity, especially in conversational or disordered speech contexts.

Regarding data availability, the number of phenomena measured by the given framework depends on the CHAT-coded transcripts. Even across hundreds of hours of conversational data, some speech phenomena have relatively low representation. This is because many datasets choose to encode only a limited number of features, and some features are generally uncommon or uncommonly used. Given that the results for these features may therefore not be representative of the actual performance of ASR systems, the low number of examples must be considered.

Another important limitation has to do with how datasets like AphasiaBank are used. Since it's one of the largest and most widely available resources for speech from PwA, many models are both trained and tested on it. If the separation between training and evaluation data isn't carefully maintained, there's a real risk of bias. Some models might seem to perform well simply because they've seen similar data before, not because they're truly robust. Over time, as the field increasingly focuses on improving performance on aphasia-specific benchmarks, there is a risk of models becoming overfit to the characteristics of these datasets. While the proposed evaluation framework does not contribute to this directly, the inclusion of more granular metrics may unintentionally increase the incentive to optimize for specific annotated phenomena rather than broader generalizability.

Finally, because the evaluation methodology relies on comparing a coded reference transcript to an non-coded ASR output, the framework has limited ability to detect insertions introduced by the ASR system. Although manual verification revealed minimal evidence of such errors, models prone to hallucinations or insertion mistakes could still affect the reliability of the results.

Further research can address some of the current limitations and enhance the framework's utility. Firstly, ASR transcript parsing can be improved to allow for automatic encoding of specific CHAT codes. The Batchalign2 (Liu, MacWhinney, Fromm & Lanzi, 2023) project has already made some efforts to automate a number of CHAT annotations, but there is still a need for a more comprehensive solution. The framework could then leverage the improved CHAT annotations to provide more accurate feature measurements.

Furthermore, integrating speech-to-phoneme models into the pipeline may provide an improved method for measuring certain phenomena. However, the applicability of such tools is currently limited by performance, especially when applied to a broad range of accents and aphasic speech.

2.6 Conclusion

This project provides a comprehensive evaluation framework capable of revealing detailed performance indicators for ASR systems, as demonstrated by the results presented in the previous sections. The proposed framework is well suited to guide improvements for ASR systems in the context of disfluent speech, particularly for people with aphasia or other cognitive impairments. Resulting metrics from our comparison showed areas, such as filled pauses, where models drop in performance and how we can measure specific improvements when targeting these phenomena.

This advances conventional ASR evaluation methods and reduces the unique challenges posed by speech with high variability and disfluencies. This framework enables a deeper understanding of current model limitations by providing a fine-grained analysis of ASR errors across different speech phenomena. It serves as a tool for guiding future development efforts. Importantly, the ability to systematically evaluate ASR performance on phenomena specific to aphasic speech has important implications for clinical contexts. It can help ensure that models used in diagnosis or therapy provide reliable, representative outputs. This increases the trustworthiness of ASR-based tools used by clinicians, researchers, and PwA. Beyond improving accessibility, such tools could support more accurate assessments and tailored interventions.

Future research can build upon this framework by exploring additional error metrics that consider phonetic similarity and semantic proximity, as well as specific improvements such as enhancing ASR transcript parsing to support automatic encoding of CHAT codes.

2.7 Data availability statement

All speech data used in this study is sourced from AphasiaBank (MacWhinney *et al.*, 2011), specifically the English and French protocol corpora.

2.8 Disclosure Statement

The authors report there are no competing interests to declare.

2.9 Funding

This work was supported by Gouvernement du Québec (Ministère de l'Économie, de l'Innovation et de l'Énergie, Programme de soutien aux organismes de recherche et d'innovation) under Grant FR 56260.

CHAPITRE 3

ARCHITECTURE LOGICIELLE ET IMPLÉMENTATION

3.1 Introduction

Ce chapitre décrit l'implémentation technique du cadre d'évaluation développé pour évaluer la performance des systèmes de RAP sur la parole de PAA. Écrit en Python, le projet intègre les systèmes de RAP, la prise en charge des formats de transcription clinique et l'analyse des erreurs dans une architecture modulaire et extensible. Il a été conçu pour permettre des comparaisons robustes entre les transcriptions générées par les modèles RAP et les données annotées par les experts. L'architecture respecte la structure et les conventions du format CHAT.

3.2 Technologies utilisées

Le cadre a été développé en Python en utilisant une combinaison de librairies à sources ouvertes. La transcription est gérée par WhisperX ; il s'agit d'une enveloppe construite autour de Whisper d'OpenAI (et d'autres modèles compatibles) qui ajoute un alignement précis au niveau des mots et une diarization optionnelle des locuteurs.

Les autres librairies utilisées sont les suivantes :

- **pandas** : pour la manipulation des données (Pandas, 2020),
- **re** : pour la normalisation basée sur les expressions régulières et la recherche de motifs,
- **pytest** : pour les tests unitaires (Krekel *et al.*, 2004),
- **pyLangAcq** : pour ouvrir et parcourir les fichiers CHAT (Lee *et al.*, 2016),
- **python-iso639** : pour gérer les codes de langue conformément à la spécification CHAT (Lee, 2024),
- **ffmpeg** et **python-magic** : pour le prétraitement et la segmentation des médias avant la transcription (Tomar, 2006; Hupp, 2022).

L'architecture logicielle n'a pas été décrite comme un exercice d'application de patrons classiques, car elle ne s'appuyait pas sur des patrons formalisés comme Strategy ou Factory. Elle repose plutôt sur une structuration orientée objet avec une hiérarchie claire.

Le développement a suivi une approche modulaire, où chaque composant du système (lecture des fichiers CHAT, modélisation des données, transcription automatique, et évaluation) a été conçu comme un module indépendant avec des responsabilités bien définies. Cette séparation permet de modifier, réutiliser ou remplacer certaines parties du pipeline (par exemple, le système de RAP ou la stratégie d'évaluation) sans affecter l'ensemble de l'architecture. Une attention particulière a été portée à la maintenabilité du code ; l'approche est soutenue par une suite de tests unitaires et une interface en ligne de commande pour l'exécution du pipeline.

3.3 Architecture

Le cadre se compose de trois éléments principaux : la modélisation des annotations de type CHAT, le pipeline de transcription de RAP et le cadre d'évaluation. Chaque module fonctionne de manière indépendante et communique à travers des objets Python structurés, ce qui facilite le dépannage, l'extension et les tests. La figure 3.1 illustre l'architecture de haut niveau. Les composantes `Modélisation` et `Évaluation` sont décrits plus en détail à la figure 3.2.

3.3.1 Spécification CHAT et modélisation des données orientée objet

Le cadre logiciel repose sur une couche d'analyse personnalisée intégrant la bibliothèque PyLangACQ afin de traiter les transcriptions en format CHAT, qui constitue le format standard utilisé, entre autres, dans le corpus AphasiaBank. PyLangACQ est utilisé principalement pour ouvrir les fichiers CHAT et les parcourir, en extrayant les *utterances* (énoncés) sous forme d'objets contenant des attributs de base tels que l'identifiant du locuteur. La bibliothèque prend en charge les opérations de lecture, de segmentation des *utterances*, et une validation syntaxique minimale du format CHAT. Toutefois, la bibliothèque n'offre que des fonctionnalités de base, insuffisantes pour répondre aux besoins analytiques propres à ce projet.

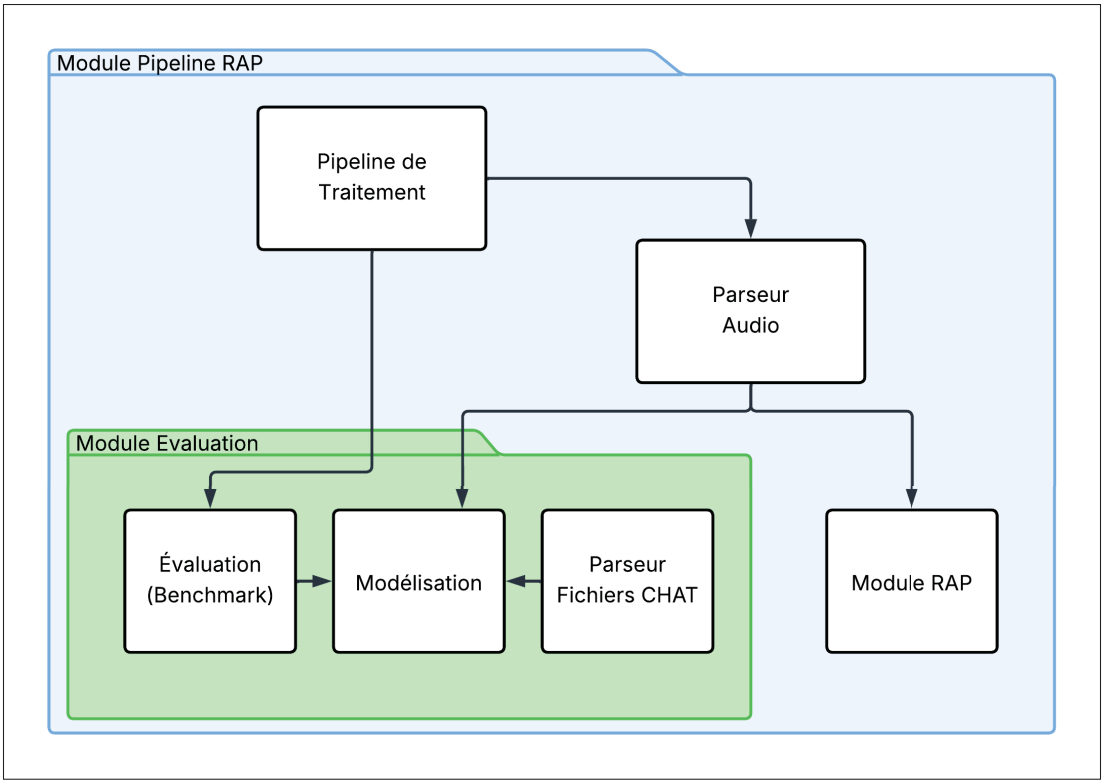

Figure 3.1 Schéma d'architecture

Afin d'assurer une structuration adaptée à nos objectifs d'évaluation, nous avons développé une hiérarchie orientée objet spécialisée. Chaque *utterance* est convertie en une structure d'objet qui encapsule à la fois les tokens individuels et leurs annotations. Ce processus repose sur un parcours systématique des objets retournés par PyLangACQ, qui servent essentiellement de point d'entrée vers les données brutes.

Nous avons appliqué des vérifications supplémentaires lors du chargement. Notamment, tout fichier contenant des codes non reconnus par la spécification CHAT est automatiquement ignoré par le parseur. Ces règles permettent d'exclure automatiquement les fichiers invalides, qui sont alors ignorés lors des étapes d'analyse.

L'architecture orientée objet est nécessaire pour lier chaque token aux phénomènes spécifiques (erreur sémantique, pause, reformulation, etc.), pour suivre ces tokens à travers le processus d'évaluation, et pour permettre des opérations complexes telles que l'alignement entre *tokens* de

référence et transcrits. Le modèle devait refléter fidèlement la structure hiérarchique du format CHAT (sessions > *utterances* > *tokens*), tout en étant suffisamment flexible pour supporter des extensions, les filtres contextuels et le suivi par locuteur. L'objectif était donc de fournir un outil capable de tirer avantage des données richement annotées, facilitant à la fois le calcul des métriques d'erreur et l'intégration dans un pipeline de transcription automatisée.

Le modèle structuré permet un alignement direct et une comparaison fine entre les *tokens* annotés au format CHAT et les sorties générées par un système de RAP, tout en préservant les métadonnées cliniques et la structure linguistique. La Figure 3.2 présente un diagramme de classes simplifié illustrant la hiérarchie des objets, y compris l'héritage entre *tokens* ainsi que l'agrégation des *tokens* dans les *utterances* et des *utterances* dans les sessions. L'organisation reflète la structure imposée par la spécification CHAT. Chaque niveau (Ex. : Session, Utterance, Token, Pause) est représenté par une classe dédiée pour respecter la granularité des transcriptions et permettre une navigation et une extension faciles. Pour des raisons de lisibilité, tous les attributs et les méthodes ne sont pas affichés, tels que les méthodes d'analyse de la classe session qui ne sont pas utilisés dans ce projet. Le modèle de données comprend principalement les classes `Token`, `Marker`, `Utterance`, `Session`, `Participant`, et `Benchmark` ; ils sont présentés dans la section suivante.

3.3.1.1 Token

Les objets de type `Token` représentent une unité linguistique minimale, avec un support pour des annotations de plusieurs types. Plusieurs classes héritent de la classe `Token` pour modéliser des types de jetons plus spécifiques. Ces classes incluent :

- `Word` : mots normaux ou spécifiques tels que les épellations, par exemple, «`apple@k`» (a-p-p-l-e),
- `PostCode` : codes ajoutés à la fin d'un énoncé pour en qualifier le contenu ou la qualité, par exemple `[+ gram]` pour «*Grammatical Error*»,
- `SatelliteMarker` : marqueurs indiquant qu'un énoncé constitue un préfixe ou un suffixe ; «‡» pour un préfixe ou «„» pour un suffixe,

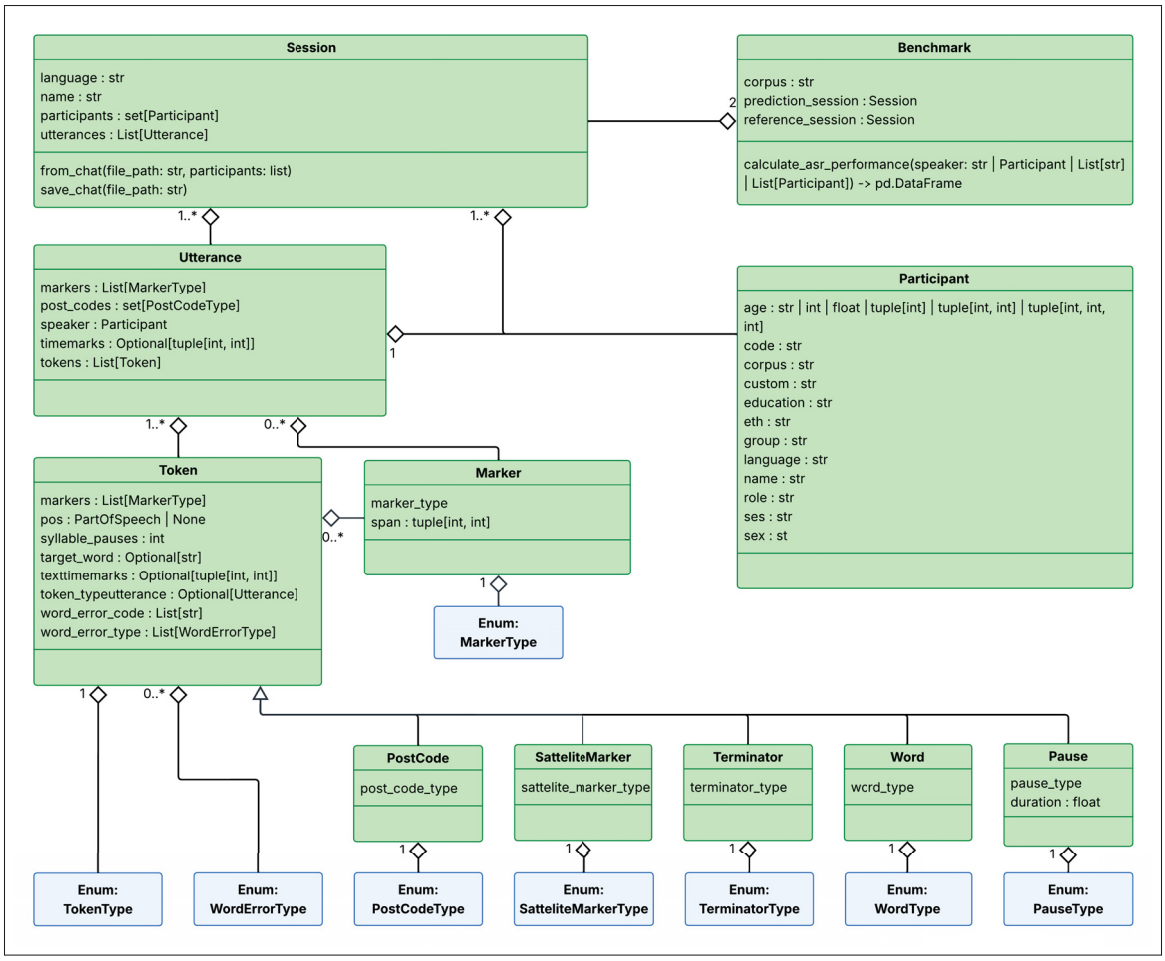

Figure 3.2 Diagramme UML de classes illustrant la modélisation des transcriptions CHAT

- `Terminator` : ponctuations terminales indiquant, par exemple, une interruption : «`+/.`»,
- `Pause` : pauses notées dans la transcription, par exemple «`(.)`» pour une pause courte.

Chaque instance de la classe `Token` possède un champ indiquant son type général, et chaque classe dérivée possède également un champ définissant son type spécifique (par exemple : le type de `PostCode`). Tous les types sont définis à l'aide de valeurs Enum, qui servent de base au système d'annotation des *tokens*. Pour référence, l'ensemble des valeurs Enum est présenté à l'annexe I.

3.3.1.2 Marker

Les objets de type `Marker` décrivent toutes les annotations qui s'appliquent à des groupes de *tokens*, tels que les répétitions ou les reformulations. Ces annotations couvrent donc plusieurs *tokens* à la fois. Comme pour `Token`, chaque instance de la classe `Marker` possède un champ `marker_type` défini par une valeur de l'énumération `MarkerType`. L'ensemble des valeurs de `MarkerType` est présenté à l'annexe dans le tableau I-4

3.3.1.3 Utterance

La classe `Utterance` est un conteneur qui regroupe une séquence d'objets `Token`, avec attribution du locuteur et des métadonnées temporelles. Comme dans les fichiers de type CHAT, les *utterances* sont généralement constitués d'une proposition principale et de propositions dépendantes, se rapprochant ainsi de la forme d'une phrase traditionnelle, bien qu'ils puissent souvent être incomplets, car ils représentent la parole naturelle.

3.3.1.4 Session

La classe `Session` regroupe l'ensemble des objets `Utterance` d'une session de conversation. Cet objet offre des méthodes pratiques pour charger ou sauvegarder les transcriptions au format CHAT. Il fournit aussi diverses fonctions d'analyse, par exemple la durée de la session, le nombre de mots par minute, ou encore le comptage des pauses remplies.

3.3.1.5 Participant

L'objet `Participant` représente un locuteur dans la conversation. Chaque participant peut avoir plusieurs attributs, tels que l'âge, le sexe, ou le niveau de scolarité. Dans les ensembles de données AphasiaBank, le champ custom est fréquemment utilisé pour indiquer le QA du participant.

3.3.1.6 Benchmark

L'objet `Benchmark` contient la logique d'évaluation. Il peut être utilisé avec deux objets `Session`, soit une référence et une prédiction. L'appel à un objet `Benchmark` retourne les résultats détaillés de l'évaluation de RAP tel que présenté dans la section 2.4.

3.3.2 Pipeline de transcription

Le processus de transcription est principalement géré à l'aide de la bibliothèque WhisperX. Les fichiers audio sont transcrits à l'aide de modèles tels que Whisper Large-v3 et CrisperWhisper. WhisperX enrichit la sortie en y ajoutant des horodatages au niveau des *tokens*. Une étape de normalisation et d'analyse est ensuite appliquée à la transcription, incluant notamment la conversion des nombres en leur forme écrite (par exemple, « 200 » devient « deux cents »), la transcription de certains symboles (par exemple, « & » devient « et »), ainsi que la suppression de symboles spéciaux (tels que les traits d'union, les deux points, etc.). Enfin, les transcriptions sont converties en un format CHAT orienté objet, tel que décrit dans la section précédente.

Dans l'implémentation actuelle, le module de RAP WhisperX est défini comme une classe qui hérite d'une classe abstraite de base, laquelle spécifie l'interface requise. Cette conception permet d'intégrer facilement d'autres systèmes RAP au pipeline existant.

Bien que WhisperX offre la possibilité de faire de la segmentation et la diarization, cette fonctionnalité n'a volontairement pas été utilisée dans l'étude présentée au chapitre précédent. Des expériences préliminaires ont montré que la diarization provoquait une variabilité importante dans les sorties de RAP, ce qui compliquait l'évaluation des performances. Afin d'isoler les sources d'erreurs et de centrer l'évaluation uniquement sur la qualité de la transcription, tous les résultats présentés dans ce mémoire ont été générés à partir de fichiers audio déjà segmentés par locuteur.

Le pipeline de transcription prend en charge le traitement par lots de sessions de conversation et enregistre toutes les transcriptions au format CHAT dans une structure de répertoires organisée,

assurant ainsi la traçabilité entre les fichiers audio, les prédictions RAP et les annotations de référence.

3.4 Évaluation

L'évaluation constitue une composante essentielle du cadre développé pour l'analyse des performances des systèmes de RAP appliqués à la parole aphasique. Elle repose sur la classe `Benchmark`, qui permet de comparer de manière structurée les transcriptions générées automatiquement avec les transcriptions de référence annotées selon le format CHAT.

3.4.1 Alignement des sessions

Le processus commence par le chargement et l'alignement de deux objets `Session` : une session de référence produite manuellement et une session prédite par un modèle de ASR. Les énoncés produits par le locuteur ciblé sont concaténés, ce qui permet d'effectuer une évaluation globale sur la session tout en conservant l'information structurée des *tokens* d'origine.

3.4.2 Métriques d'erreur

Les taux d'erreur sont calculés à l'aide de la distance de Levenshtein, avec la bibliothèque `rapidfuzz`. Deux métriques principales sont considérées : le CER, basé sur les caractères, et le WER, basé sur les mots. Pour chaque session, le système effectue une évaluation non seulement globale, mais aussi catégorielle : les erreurs sont attribuées à chaque type de token tel qu'annoté dans la transcription de référence. Cela inclut les types de *tokens* définis par `TokenType`, tels que les erreurs sémantiques, les paraphasies phonologiques, les pauses remplies, et les néologismes.

Les *tokens* de type `Word` peuvent également être analysés selon leur `WordType` (ex. : mot partiel, pause, néologisme) et leur `WordErrorType` (ex. : erreur morphologique ou phonologique). Ces catégories sont cruciales pour l'évaluation de la parole aphasique, car elles correspondent à des phénomènes linguistiques souvent ciblés dans le diagnostic et la planification des interventions.

Le cadre prend aussi en compte d'autres annotations présentes dans les transcriptions CHAT, notamment les marqueurs de discours représentés par `MarkerType`. Ceux-ci incluent notamment les répétitions, les reformulations.

3.4.3 Comparaison multisessions

Le cadre permet l'évaluation sur plusieurs sessions, en combinant les résultats issus de plusieurs objets `Benchmark`. Chaque session est analysée individuellement, puis les résultats sont regroupés selon la session, le corpus, ou le type de *token*. Les métriques peuvent être moyennées ou pondérées selon le nombre total d'unités (caractères ou mots), et des mesures de variance sont calculées afin d'estimer la stabilité des performances entre les sessions.

Cette agrégation repose sur une utilisation extensive de la bibliothèque `pandas`, qui permet de structurer les résultats d'évaluation sous forme de tableaux (DataFrames) facilement filtrables, groupables et exportables. Chaque ligne représente un type de *token* ou une session, et chaque colonne encode les opérations d'édition de la distance de Levenshtein (*equal, insertion, deletion, substitution*), le total d'unités évaluées, et le taux d'erreur résultant. Les fonctions de groupement (`groupby`), de pivotement (`pivot`), et de fusion de tables facilitent l'analyse comparative.

Le système prend également en compte l'information démographique et clinique (ex. : âge, sexe, quotient d'aphasie), lorsqu'elle est disponible dans les métadonnées du corpus. Cette intégration permet une analyse exploratoire des corrélations possibles entre les erreurs de transcription et les caractéristiques des locuteurs. Dans les cas où l'information n'est pas disponible dans les métadonnées, l'attribut tient simplement une valeur nulle (*None*).

3.4.4 Filtrage par locuteur et phénomène linguistique

L'évaluation peut être restreinte à un locuteur donné, ce qui est essentiel dans les corpus conversationnels avec plusieurs locuteurs. Des règles de correspondance sont intégrées pour gérer les variations dans les codes de locuteur (ex. : `PAR`, `INV`) entre les fichiers de référence et de prédiction. Une classe utilitaire, `TokenFilter`, permet de filtrer les énoncés à inclure dans

l'évaluation selon la présence de certains types de *tokens*. Cela permet d'analyser, par exemple, uniquement les énoncés contenant des erreurs sémantiques, des reformulations ou des pauses longues, afin de cibler les phénomènes linguistiques les plus problématiques pour un modèle donné.

3.5 Validation

Une suite de tests unitaires a été développée à l'aide du cadre Pytest afin de garantir la fiabilité des composantes centrales du système, en particulier le modèle de données conçu pour représenter et manipuler les transcriptions cliniques au format CHAT.

Le processus de validation s'est concentré sur la vérification du traitement et de la structuration correcte des données CHAT. Les tests unitaires ont confirmé que les transcriptions étaient interprétées correctement, et que tous les éléments pertinents, tels que les tours de parole, et les annotations, étaient reconstruits fidèlement, conformément aux spécifications du format CHAT.

Ces tests ont également permis d'évaluer la cohérence interne de l'architecture orientée objet. Une attention particulière a été accordée à l'intégrité des relations entre objets, en s'assurant par exemple, à ce que les objets `Tokens` soient correctement associés aux objets `Utterances`, et que ces derniers soient bien reliés aux objets `Participant` et `Session` correspondantes.

De plus une suite de tests unitaires est effectuée pour l'objet `Benchmark`. Ces tests assurent la validité des mesures d'évaluations en sortie du cadre d'évaluation en vérifiant l'exactitude des calculs de taux d'erreur.

3.6 Conclusion

La conception modulaire du cadre le rend adaptable aux besoins futurs. Parmi les améliorations potentielles, on peut envisager l'ajout d'autres systèmes de transcription RAP, tels que Google Speech-to-Text ou Microsoft Azure AI Speech, l'intégration de formats de données cliniques supplémentaires comme ELAN ou les TextGrids de Praat.

Ce chapitre a présenté la conception et l'implémentation du cadre d'évaluation, développé en Python, pour l'analyse de la performance de la reconnaissance vocale automatique sur la parole aphasique. Le système intègre un pipeline de transcription basée sur WhisperX, une modélisation orientée objet du format CHAT, ainsi qu'un outil d'évaluation. En prenant en compte des phénomènes cliniquement pertinents, ce cadre constitue une base solide pour la recherche et le développement futurs en RAP appliquée à l'aphasie.

CONCLUSION ET RECOMMANDATIONS

Ce mémoire avait pour objectif de concevoir, implémenter et tester un cadre d'évaluation adapté à l'analyse des performances de systèmes de RAP lorsqu'ils sont appliqués à la parole aphasique. La reconnaissance de la parole atypique liée à des troubles du langage, comme l'aphasie, présente des défis majeurs, notamment en raison de la présence fréquente de disfluences, de fragments phonologiques, d'erreurs grammaticales et d'hésitations. Les modèles de RAP actuels, souvent optimisés sur des données de locuteurs non aphasiques, ont des difficultés à s'adapter à ces formes atypiques de parole. C'est dans ce contexte que s'inscrit notre contribution.

Une revue critique de la littérature a permis d'identifier les approches existantes, souvent centrées sur des métriques globales comme le WER, insuffisantes pour décrire précisément les erreurs commises par les modèles dans des contextes pathologiques. En réponse à cette lacune, nous avons proposé un cadre d'évaluation plus granulaire, basé sur l'annotation détaillée des transcriptions de référence fournies en format CHAT, et une segmentation des erreurs par phénomène linguistiques. Le cadre repose sur une architecture combinant l'extraction des métriques, l'alignement des transcriptions, et le regroupement des erreurs par types de phénomènes linguistiques.

Ce cadre a été appliqué à une série d'expérimentations avec des modèles RAP récents. Nous avons comparé leurs performances selon différents paramètres : présence ou absence de rédactique, groupes de locuteurs, et type de phénomène observé. Les résultats montrent que certains phénomènes, tels que les pauses remplies ou fragments phonologiques, posent des difficultés constantes aux modèles de base, tandis que la rédactique permet dans certains cas de réduire significativement les erreurs, en particulier sans recourir à un adaptation coûteux.

Au-delà de l'analyse comparative, le cadre fournit un outil concret et exploitable : une table de données structurée contenant les résultats CER/WER par phénomène, par modèle et par condition expérimentale. Ce format permet des analyses croisées, un repérage rapide des

faiblesses des modèles, et peut être intégré dans des pipelines d'évaluation plus larges. Ainsi, ce cadre peut servir à la fois à des chercheurs en TALN souhaitant adapter leurs modèles à des cas pathologiques, et à des cliniciens ou orthophonistes cherchant à mieux comprendre comment un modèle transcrit, ou omet, certains phénomènes linguistiques pertinents dans leurs évaluations.

4.1 Perspectives de recherche et recommandations

Plusieurs évolutions de ce travail peuvent être envisagées afin d'approfondir l'évaluation et l'adaptation des systèmes RAP à la parole aphasique. Une première piste consiste à intégrer des modèles préalablement adaptés sur des données aphasiques. Dans ce mémoire, l'approche adoptée reposait sur la rédactique comme moyen léger d'amélioration sans recourir à un entraînement supervisé. Il serait toutefois pertinent de comparer cette approche à des modèles explicitement entraînés sur des corpus de parole aphasique, afin d'évaluer les bénéfices d'un apprentissage ciblé sur des profils de parole atypique.

Par ailleurs, les limites des métriques classiques comme le CER et le WER soulignent la nécessité d'explorer des métriques complémentaires. Celles-ci pourraient inclure des mesures phonémiques, telles que le taux d'erreur sur les phonèmes, ou des métriques perceptuelles ou acoustiques qui tiennent compte de la similarité sonore entre mots, même en présence de divergences orthographiques importantes. Une telle approche permettrait d'évaluer plus justement la qualité des transcriptions dans des cas où la phonologie est plus pertinente que l'orthographe, notamment dans la transcription de fragments ou d'erreurs phonologiques. Cela impliquerait toutefois des données de PAA annotées avec des informations phonétiques, ce qui serait relativement difficile à obtenir.

Enfin, afin de faciliter l'adoption de ce cadre par un public non technique, une interface graphique ou une API pourrait être développée. Celle-ci permettrait aux orthophonistes, chercheurs en sciences du langage ou professionnels de la santé de charger leurs propres transcriptions, de

lancer automatiquement l'évaluation, et d'obtenir un rapport structuré sur les performances du modèle selon les phénomènes ciblés. Une telle interface favoriserait la diffusion du cadre au-delà du milieu de la recherche computationnelle, et renforcerait son utilité dans des contextes cliniques concrets.

ANNEXE I

DÉTAILS DE L'IMPLÉMENTATION LOGICIEL

Tableau-A I-1 Description du Enum TokenType

Name	**Value**
PUNCTUATION	0
PAUSE	1
WORD	2
WORD_ERROR	3
POST_CODE	4
TERMINATOR	5
UNINTELLIGIBLE	6
UNTRANSCRIBABLE	7
UNIDENTIFIABLE	8
SAT_MARKER	9
PHONOLOGICAL_FRAGMENT	&+
FILLER	&-
NON_WORD	&~
GESTURE	&=
LAZY_OVERLAP	+<

Tableau-A I-2 Description du Enum WordType

Name	Value
NORMAL	1
ADDITION	@a
BABBLING	@b
CHILD_INVENTED	@c
DIALECT	@d
ECHOLALIA	@e
FAMILY_SPECIFIC	@f
FILLED_PAUSE	@fp
GENERAL_SPECIAL	@g
INTERJECTION	@i
SPELLING	@k
LETTER	@l
LETTER_PLURAL	@lp
NEOLOGISM	@n
ONOMATOPOEIA	@o
PHONOLOGICAL	@p
META_LINGUISTIC	@q
SECOND_LANGUAGE	@s
SINGING	@si
SIGNED_LANGUAGE	@sl
SIGN_AND_SPEECH	@sas
TEST_WORD	@t
IPA_TRANSCRIPTION	@u
WORD_PLAY	@wp
EXCLUDED_WORDS	@x
USER_DEFINED	@z

Tableau-A I-3 Description du Enum PauseType

Name	Value
SHORT	(.)
MEDIUM	(..)
LONG	(...)
DURATION	(:)

Tableau-A I-4 Description du Enum MarkerType

Name	Value
REPETITION	/
RETRACING	//
REFORMULATION	///
OVERLAP_FOLLOWING	>
OVERLAP_PRECEDING	<
FALSE_START_WITHOUT_RETRACING	/-
UNCLEAR_RETRACING_TYPE	/?
STRESSING	!
CONTRASTIVE_STRESSING	!!
BEST_GUESS	?
EXCLUDED	e

Tableau-A I-5 Description du Enum TerminatorType

Name	Value
TRAILING_OFF	+...
TRAIL_OFF_TO_A_QUESTION	+..?
QUESTION_WITH_EXCLAMATION	+!?
INTERRUPTED	+/.
INTERRUPTED_QUESTION	+/?
SELF_INTERRUPTION	+//.
SELF_INTERRUPTION_WITH_QUESTION	+//?
TRANSCRIPTION_BREAK	+.
QUOTATION_FOLLOWS	+/."
QUOTATION_PRECEDES	+."
QUOTED_UTTERANCE	+"
QUICK_UPTAKE	+^
SELF_COMPLETION	+,
OTHER_COMPLETION	++

Tableau-A I-6 Description du Enum PostCodeType

Name	Value
GRAMMATICAL	gram
EXCLUDED	exc
EMPTY_SPEECH	es
JARGON	jar
CIRCUMLOCUTION	cir
PERSEVERATION	per

Tableau-A I-7 Description du Enum WordErrorType

Name	Value
GENERIC	(empty)
SEMANTIC	s
PHONOLOGICAL	p
NEOLOGISM	n
MORPHOLOGICAL	m
DYSFLUENCY	d

Tableau-A I-8 Description du Enum SatelliteMarkerType

Name	Value
PREFIX	‡
SUFFIX	„

ANNEXE II

DÉTAILS DES MODELS RAP - ANNEXE DE L'ARTICLE

Tableau-A II-1 ASR Parameters Configuration

Parameter	**Value**
beam_size	5
best_of	5
patience	1
length_penalty	1
repetition_penalty	1
no_repeat_ngram_size	0
temperatures	[0.0, 0.2, 0.4, 0.6, 0.8, 1.0]
compression_ratio_threshold	2.4
log_prob_threshold	-1.0
no_speech_threshold	0.6
condition_on_previous_text	False
prompt_reset_on_temperature	0.5
initial_prompt	"um, hm, okay, so, mm, and, uh, mhm" or None
hotwords	None
prefix	None
suppress_blank	True
suppress_tokens	[-1]
without_timestamps	True
max_initial_timestamp	0.0
word_timestamps	True
prepend_punctuations	"'“¿([{-
append_punctuations	"',!?:”)]\}
suppress_numerals	True
max_new_tokens	None
clip_timestamps	None
hallucination_silence_threshold	None

Tableau-A II-2 Model Snapshots Used in the Study

Model	**Snapshot Date**	**Commit Hash**
openai/whisper-large-v2	Feb 29, 2024	ae4642769ce2ad8fc292556ccea8e901f1530655
openai/whisper-large-v3	Jun 10, 2024	3d0618527a343f8ad58c34d26542213f0444e901
nyrahealth/faster_CrisperWhisper	Sep 2, 2024	6a042898d09e64133d76ef09b8bb25440debaa17

BIBLIOGRAPHIE

Aphasia Institute. (s.d.). What is Aphasia ? – Aphasia Institute. Repéré à https://www.aphasia.ca/family-and-friends-of-people-with-aphasia/what-is-aphasia-2/.

Azevedo, N., Kehayia, E., Jarema, G., Dorze, G. L., Beaujard, C. & Yvon, M. (2024). How artificial intelligence (AI) is used in aphasia rehabilitation : A scoping review. *Aphasiology*, 38, 305-336. doi : 10.1080/02687038.2023.2189513.

Bachmann, M., layday, thomasryde, Kokkinou, G., Schreiner, H., Fihl-Pearson, J., dheeraj, Vioshim, pekkarr, Górny, M., Sherman, M., odidev, glynn, H, T., tr halfspace, Schütz, R., Renkamp, N., Tang, K., Moine, H. L., Rosin, G., Hess, D., Clauss, C., V., B. & Delfini. (2025). rapidfuzz/RapidFuzz : Release 3.12.0. Repéré à https://zenodo.org/records/14673428.

Baevski, A., Zhou, H., Mohamed, A. & Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. Repéré à https://arxiv.org/abs/2006.11477.

Bain, M., Huh, J., Han, T. & Zisserman, A. (2023). WhisperX : Time-Accurate Speech Transcription of Long-Form Audio. *Interspeech 2023*, pp. 4489–4493. doi : 10.21437/Interspeech.2023-78.

Barberis, M., De Clercq, P., Tamm, B., Van hamme, H. & Vandermosten, M. (2024). Automatic recognition and detection of aphasic natural speech. *Interspeech 2024*, pp. 1990–1994. doi : 10.21437/Interspeech.2024-517.

Bredin, H. (2023, 8). pyannote.audio 2.1 speaker diarization pipeline : principle, benchmark, and recipe. *INTERSPEECH 2023*, pp. 1983-1987. doi : 10.21437/Interspeech.2023-105.

Casilio, M., Rising, K., Beeson, P. M., Bunton, K. & Wilson, S. M. (2019). Auditory-Perceptual Rating of Connected Speech in Aphasia. *American Journal of Speech-Language Pathology*, 28, 550. doi : 10.1044/2018_AJSLP-18-0192.

Cleveland Clinic. (2022, 10). Aphasia : Causes, Symptoms & Treatment. Repéré à https://my.clevelandclinic.org/health/diseases/5502-aphasia.

Cong, Y., LaCroix, A. N. & Lee, J. (2024). Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14, 15573. doi : 10.1038/S41598-024-66576-Y.

Day, M., Dey, R. K., Baucum, M., Paek, E. J., Park, H. & Khojandi, A. (2021). Predicting Severity in People with Aphasia : A Natural Language Processing and Machine Learning Approach. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2299-2302. doi : 10.1109/EMBC46164.2021.9630694.

Dickey, L., Kagan, A., Lindsay, M. P., Fang, J., Rowland, A. & Black, S. (2010). Incidence and profile of inpatient stroke-induced aphasia in Ontario, Canada. *Archives of physical medicine and rehabilitation*, 91, 196-202. doi : 10.1016/J.APMR.2009.09.020.

Ezzes, Z., Schneck, S. M., Casilio, M., Fromm, D., Mefferd, A. S., de Riesthal, M. & Wilson, S. M. (2022). An Open Dataset of Connected Speech in Aphasia with Consensus Ratings of Auditory-Perceptual Features. *Data*, 7(11). doi : 10.3390/data7110148.

Frederick, A., Jacobs, M., Adams-Mitchell, C. J. & Ellis, C. (2022). The Global Rate of Post-Stroke Aphasia. *Perspectives of the ASHA Special Interest Groups*, 7, 1567-1572. doi : 10.1044/2022_PERSP-22-00111.

Frew, A. (2017). Different Strokes Recovery triumphs and challenges at any age. 2017 Stroke Report.

Fridriksson, J. & Hillis, A. E. (2021). Current Approaches to the Treatment of Post-Stroke Aphasia. *Journal of stroke*, 23, 183-201. doi : 10.5853/jos.2020.05015.

Hachioui, H. E., Visch-Brink, E. G., de Lau, L. M., van de Sandt-Koenderman, M. W., Nouwens, F., Koudstaal, P. J. & Dippel, D. W. (2017). Screening tests for aphasia in patients with stroke : a systematic review. *Journal of Neurology*, 264, 211-220. doi : 10.1007/S00415-016-8170-8.

Heart and Stroke Foundation. (2022). Stroke in Canada is on the rise | Heart and Stroke Foundation. Repéré à https://www.heartandstroke.ca/what-we-do/media-centre/news-releases/stroke-in-canada-is-on-the-rise.

Herath, H. M. D. P. M., Weraniyagoda, W. A. S. A., Rajapaksha, R. T. M., Wijesekara, P. A. D. S. N., Sudheera, K. L. K. & Chong, P. H. J. (2022). Automatic Assessment of Aphasic Speech Sensed by Audio Sensors for Classification into Aphasia Severity Levels to Recommend Speech Therapies. *Sensors*, 22(18). doi : 10.3390/s22186966.

Hupp, A. (2022, 6). python-magic (Version 0.4.27). Repéré à https://pypi.org/project/python-magic/0.4.27/.

Ilias, L. & Askounis, D. (2022). Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Frontiers in Aging Neuroscience*, Volume 14 - 2022. doi : 10.3389/fnagi.2022.830943.

Kertesz, A. (2012). Western Aphasia Battery–Revised. *PsycTESTS Dataset*. doi : 10.1037/T15168-000.

Kertesz, A. (2022). The Western Aphasia Battery : a systematic review of research and clinical applications. *Aphasiology*, 36, 21-50. doi : 10.1080/02687038.2020.1852002.

Kim, Y., Jeong, H., Chen, S., Li, S. S., Lu, M., Alhamoud, K., Mun, J., Grau, C., Jung, M., Gameiro, R., Fan, L., Park, E., Lin, T., Yoon, J., Yoon, W., Sap, M., Tsvetkov, Y., Liang, P., Xu, X., Liu, X., McDuff, D., Lee, H., Park, H. W., Tulebaev, S. & Breazeal, C. (2025). Medical Hallucinations in Foundation Models and Their Impact on Healthcare. Repéré à https://arxiv.org/abs/2503.05777.

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H. & Sloane, M. (2024, 6). Careless Whisper : Speech-to-Text Hallucination Harms. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, (FAccT '24), 1672–1681. doi : 10.1145/3630106.3658996.

Krekel, H., Oliveira, B., Pfannschmidt, R., Bruynooghe, F., Laugher, B. & Bruhin, F. [Version 8.3. Contributors include Holger Krekel, Bruno Oliveira, Ronny Pfannschmidt, Floris Bruynooghe, Brianna Laugher, Florian Bruhin, and others.]. (2004). pytest 8.3. Repéré à https://github.com/pytest-dev/pytest.

Lam, J. M. & Wodchis, W. P. (2010). The relationship of 60 disease diagnoses and 15 conditions to preference-based health-related quality of life in Ontario hospital-based long-term care residents. *Medical care*, 48, 380-387. doi : 10.1097/MLR.0B013E3181CA2647.

Le, D. & Provost, E. M. (2016). Improving Automatic Recognition of Aphasic Speech with AphasiaBank. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 08-12-September-2016, 2681-2685. doi : 10.21437/INTERSPEECH.2016-213.

Le, D., Licata, K., Persad, C. & Provost, E. M. (2016). Automatic Assessment of Speech Intelligibility for Individuals with Aphasia. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 24, 2187-2199. doi : 10.1109/TASLP.2016.2598428.

Le, D., Licata, K. & Mower Provost, E. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100, 1-12. doi : https://doi.org/10.1016/j.specom.2018.04.001.

Lea, C., Huang, Z., Narain, J., Tooley, L., Yee, D., Tran, D. T., Georgiou, P., Bigham, J. P. & Findlater, L. (2023). From User Perceptions to Technical Improvement : Enabling People Who Stutter to Better Use Speech Recognition. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, (CHI '23). doi : 10.1145/3544548.3581224.

Lee, J. L. (2024, 4). python-iso639 2024.4. Repéré à https://pypi.org/project/python-iso639/2024.4.27/.

Lee, J. L., Burkholder, R., Flinn, G. B. & Coppess, E. R. (2016). *Working with CHAT transcripts in Python* (Rapport n°TR-2016-02).

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech & Language*, 1, 29-45. doi : 10.1016/S0885-2308(86)80009-2.

Liu, H., MacWhinney, B., Fromm, D. & Lanzi, A. (2023). Automation of Language Sample Analysis. *Journal of Speech, Language, and Hearing Research*, 66(7), 2421–2433. doi : 10.1044/2023_JSLHR-22-00642.

MacWhinney, B. (2000). *The CHILDES project : Tools for analyzing talk : Transcription format and programs, Vol. 1, 3rd ed.* Lawrence Erlbaum Associates Publishers.

MacWhinney, B., Fromm, D., Forbes, M. & Holland, A. (2011). Aphasiabank : Methods for studying discourse. *Aphasiology*, 25, 1286-1307. doi : 10.1080/02687038.2011.589893.

Mahmoud, S. S., Kumar, A., Li, Y., Tang, Y. & Fang, Q. (2021). Performance Evaluation of Machine Learning Frameworks for Aphasia Assessment. *Sensors*, 21(8). doi : 10.3390/s21082582.

Mahmoud, S. S., Pallaud, R. F., Kumar, A., Faisal, S., Wang, Y. & Fang, Q. (2023). A Comparative Investigation of Automatic Speech Recognition Platforms for Aphasia Assessment Batteries. *Sensors*, 23(2). doi : 10.3390/s23020857.

McGregor, K. K. & Hadden, R. R. (2020). Brief Report : "Um" Fillers Distinguish Children With and Without ASD. *Journal of Autism and Developmental Disorders*, 50, 1816-1821. doi : 10.1007/S10803-018-3736-1.

National Aphasia Association. (s.d.). What is Aphasia ? Repéré à https://aphasia.org/what-is-aphasia/.

NIDCD. (2017, 3). What Is Aphasia ? - National Institute on Deafness and Other Communication Disorders. Repéré à https://www.nidcd.nih.gov/health/aphasia.

Nivedha, E., Chandrasekar, A. & Jothi, S. (2023). An optimal hybrid AI-ResNet for accurate severity detection and classification of patients with aphasia disorder. *Signal, Image and Video Processing*, 17(8), 3913-3922. doi : 10.1007/s11760-023-02620-0.

OpenAI. (n.d.). Docs - Speech to text, Prompting. Repéré à https://platform.openai.com/docs/guides/speech-to-text/prompting#prompting.

Pandas. (2020, 2). pandas-dev/pandas : Pandas. Zenodo. Repéré à https://doi.org/10.5281/zenodo.3509134.

Perez, M., Aldeneh, Z. & Provost, E. M. (2020). Aphasic Speech Recognition Using a Mixture of Speech Intelligibility Experts. *Interspeech 2020*, pp. 4986–4990. doi : 10.21437/Interspeech.2020-2049.

Perez, M., Le, D., Romana, A., Jones, E., Licata, K. & Provost, E. M. (2023). Seq2seq for Automatic Paraphasia Detection in Aphasic Speech. Repéré à https://arxiv.org/abs/2312.10518.

Plaquet, A. & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *Interspeech 2023*, pp. 3222–3226. doi : 10.21437/Interspeech.2023-205.

Privitera, A. J., Ng, S. H. S., Kong, A. P. H. & Weekes, B. S. (2024). AI and Aphasia in the Digital Age : A Critical Review. *Brain Sciences*, 14, 383. doi : 10.3390/BRAINSCI14040383.

Qian, Z. & Xiao, K. (2023). A Survey of Automatic Speech Recognition for Dysarthric Speech. *Electronics*, 12, 4278. doi : 10.3390/electronics12204278.

Qin, Y., Lee, T. & Kong, A. P. H. (2020a). Automatic Assessment of Speech Impairment in Cantonese-Speaking People with Aphasia. *IEEE Journal on Selected Topics in Signal Processing*, 14, 331-345. doi : 10.1109/JSTSP.2019.2956371.

Qin, Y., Wu, Y., Lee, T. & Kong, A. P. H. (2020b). An End-to-End Approach to Automatic Speech Assessment for Cantonese-speaking People with Aphasia. *Journal of Signal Processing Systems*, 92, 819-830. doi : 10.1007/S11265-019-01511-3.

Qin, Y., Lee, T., Kong, A. P. H. & Lin, F. (2022). Aphasia Detection for Cantonese-Speaking and Mandarin-Speaking Patients Using Pre-Trained Language Models. *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 359-363. doi : 10.1109/ISCSLP57327.2022.10037929.

R, R. & A, C. (2024). GTSO : Gradient tangent search optimization enabled voice transformer with speech intelligibility for aphasia. *Computer Speech & Language*, 84, 101568. doi : https://doi.org/10.1016/j.csl.2023.101568.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. Repéré à https://arxiv.org/abs/2212.04356.

Romana, A., Niu, M., Perez, M., Roberts, A. & Provost, E. M. (2022). Enabling Off-the-Shelf Disfluency Detection and Categorization for Pathological Speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2022-September, 1916-1920. doi : 10.21437/INTERSPEECH.2022-10971.

Romana, A., Koishida, K. & Provost, E. M. (2023). Automatic Disfluency Detection from Untranscribed Speech. Repéré à https://arxiv.org/abs/2311.00867.

Srivastav, V., Majumdar, S., Koluguri, N., Moumen, A., Gandhi, S. et al. (2023). Open Automatic Speech Recognition Leaderboard. Hugging Face. Repéré à https://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J. & Carmiel, Y. (2020). WER we are and WER we think we are. *Findings of the Association for Computational Linguistics : EMNLP 2020*, pp. 3290-3295. doi : 10.18653/v1/2020.findings-emnlp.295.

Tang, J., Chen, W., Chang, X., Watanabe, S. & MacWhinney, B. (2023). A New Benchmark of Aphasia Speech Recognition and Detection Based on E-Branchformer and Multi-task Learning. Repéré à https://arxiv.org/abs/2305.13331.

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, 2006(146), 10.

Torre, I. G., Romero, M. & Álvarez, A. (2021). Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Applied Sciences*, 11(19). doi : 10.3390/app11198872.

Wagner, L., Zusag, M. & Bloder, T. (2023). Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2023-August, 3013-3017. doi : 10.21437/INTERSPEECH.2023-1653.

Wang, Y., Cheng, W., Sufi, F., Fang, Q. & Mahmoud, S. S. (2024). A Systematic Review of Using Deep Learning in Aphasia : Challenges and Future Directions. *Computers*, 13(5). doi : 10.3390/computers13050117.

Wassink, A. B., Gansen, C. & Bartholomew, I. (2022). Uneven success : automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140, 50-70. doi : 10.1016/J.SPECOM.2022.03.009.

Wilson, S. M., Eriksson, D. K., Schneck, S. M. & Lucanie, J. M. (2018). A quick aphasia battery for efficient, reliable, and multidimensional assessment of language function. *PLoS ONE*, 13. doi : 10.1371/JOURNAL.PONE.0192773.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z. & Church, K. (2020). Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. *Interspeech 2020*, pp. 2162–2166. doi : 10.21437/Interspeech.2020-2516.

Yuan, J., Cai, X., Bian, Y., Ye, Z. & Church, K. (2021). Pauses for Detection of Alzheimer's Disease. *Frontiers in Computer Science*, 2. doi : 10.3389/FCOMP.2020.624488/FULL.

Zusag, M., Wagner, L. & Thallinger, B. (2024). CrisperWhisper : Accurate Timestamps on Verbatim Speech Transcriptions. *Interspeech 2024*, pp. 1265–1269. doi : 10.21437/Interspeech.2024-731.