Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Peut-on retrouver votre âge à partir de la transcription de votre parole ?

Vanessa Gaudray Bouju[1] Menel Mahamdi[1] Iris Eshkol-Taravella[1] Angèle Barbedette[2]
(1) MoDyCo (Université Paris Nanterre), adresse, 92000 Nanterre, France
(2) ERTIM (INALCO), 2 Rue de Lille, 75007 Paris, France
v.gaudraybouju@gmail.com, ieshkolt@parisnanterre.fr,
angele.barbedette@gmail.com

RÉSUMÉ

L'identification et la classification des groupes sociaux à partir du langage constitue une préoccupation sociolinguistique majeure. Dans cet article, nous présentons une recherche de classification des locuteurs basée sur leur âge. Pour ce faire, nous exploitons un corpus de données du français oral, où chaque locuteur est associé à des métadonnées, dont son âge au moment de l'enregistrement. Notre objectif est de développer des méthodes d'apprentissage automatique capables de prédire la tranche d'âge d'un locuteur à partir de son discours transcrit de l'oral, allant de l'apprentissage supervisé à l'ingénierie de prompts sur des grands modèles de langage. Cette tâche n'est pas seulement un défi technique, elle soulève également des questions fondamentales sur la nature de la variation linguistique et sur les liens entre le langage et la société. En effet, en identifiant les corrélations entre certains traits linguistiques et l'âge, notre projet contribue à enrichir notre compréhension des mécanismes sous-jacents à la variation du langage et à ses implications dans la construction de l'identité sociale. Son autre apport est de questionner les traits linguistiques classiquement imputés à une tranche d'âge afin de montrer leurs limites.

ABSTRACT

Can age be predicted from transcribed recordings ?

Identifying and classifying social groups from language features has always been topical for sociolinguistical studies. This article aims at presenting a language speaker classification model based on their age. For this purpose, we take advantage of a spoken french dataset. Metadata such as age at time of recording are associated to each speaker. Our goal is to develop machine learning methods able to predict the age of a speaker from a written transcript. Not only is this task a technical challenge, but it is also sheds light on fundamental questions linked to linguistic variation, and links between language and society. As a matter of fact, by identifying correlations between certain linguistic features and age, our work is a contribution to the task of understanding the underlying mechanisms involved in language variation. It also underlines its implications in social identity constitution. This work also questions how some linguistic features are usually associated with a certain age range in order to show the limits of this approach.

MOTS-CLÉS : âge, sociolinguistique, classification, traits linguistiques, LLM.

KEYWORDS: age, sociolinguistic, classification, linguistic features, LLM.

ARTICLE : **Accepté à** TALN.

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 599–613.

1 Introduction et état de l'art

La relation entre le langage et le profil sociologique d'un locuteur est un domaine clé de la sociolinguistique, qui étudie les variations linguistiques en fonction des facteurs sociaux (âge, genre, classe sociale, origine ethnique, niveau d'éducation, etc.). Selon (Bourdieu, 1982) la valeur d'un discours dépend du contexte social et du pouvoir symbolique du locuteur. Certains usages linguistiques sont plus légitimes que d'autres et permettent d'accéder à des positions dominantes. (Chuquet, 1990) indique que l'usage linguistique varie selon le milieu et la situation où se trouve le locuteur : milieu socioculturel, classe d'âge, milieu professionnel, etc. (Gadet, 1996) propose de regrouper la variation linguistique selon l'usage et selon les usagers. Parmi les types de variation proposés figure une variation sociale dite diastratique.

Les sociolinguistes cherchent à repérer et analyser des formes ou phénomènes linguistiques représentatifs des facteurs sociologiques étudiés. Les travaux des années 60-70 ((Labov, 1966) ; (Labov, 1972) ; (Wolfram, 1969) ; (Trudgill & et al., 1974)) s'intéressent aux phénomènes phonologiques et montrent comment les variables phonologiques peuvent varier en fonction du statut social du locuteur. Ainsi, (Labov, 1966) constate que les classes supérieures prononcent le *r* final plus systématiquement que les classes populaires. (Bernstein, 1971) propose un modèle où la langue varie selon le milieu social en distinguant deux codes : (1) le code restreint utilisé dans les classes populaires qui est plus implicite, contextuel et basé sur des références partagées et (2) le code élaboré utilisé dans les classes moyennes et supérieures, un code plus explicite, structuré qui favorise l'expression d'idées abstraites. Le genre est un autre facteur sociologique étudié dans la littérature linguistique. Ainsi, (Coates, 1993) observe que les femmes utilisent plus de stratégies de coopération dans la conversation. (Tannen, 1990) distingue le "rapport talk" (conversation relationnelle des femmes) et le "report talk" (style plus informatif et compétitif des hommes).

Enfin, l'âge en tant que facteur de variation du langage est au centre des travaux de la sociolinguistique contemporaine, étant reconnu comme un facteur majeur de variation linguistique (Wagner, 2012), qu'elle soit due aux spécificités biologiques et physiologiques ou à des tendances générationnelles ((Cheshire, 2008) ;(Wagner, 2012)). De nombreuses études ont démontré que les différentes cohortes d'âge présentent des schémas distincts de participation sociale, de valeurs et de modes de communication. Ainsi, la notion d'âge est traitée du point de vue de l'appartenance à un groupe social dont les membres ont des âges similaires ainsi que des particularités socio-culturelles et socio-économiques communes ((Hockett, 1950) ; (Eckert, 2017a) ; (Sankoff & Blondeau, 2007)). Deux types de langages sont souvent opposés dans la littérature : le langage des jeunes et celui des personnes âgées, qui ont plus tendance à utiliser le langage sous sa forme standard (Gerstenberg, 2015).

Dans le domaine du TAL, le lien entre le langage et le profil sociologique du locuteur se manifeste à travers les travaux sur la prédiction des caractéristiques sociologiques d'un locuteur (âge, genre, classe sociale, origine ethnique, etc.) à partir de ses écrits ou de sa parole. Les réseaux sociaux fournissent des données riches pour étudier ce lien. (Nguyen *et al.*, 2013) analysent par exemple des tweets pour prédire l'âge et l'origine sociale. (Rao *et al.*, 2010) utilisent des modèles probabilistes pour identifier des traits sociologiques à partir de messages sur les réseaux sociaux. (Sap *et al.*, 2014) extraient des caractéristiques linguistiques sur Twitter, Reddit, Facebook pour prédire l'âge, le genre et l'idéologie politique. On peut aussi citer la tâche de détection de l'âge à partir de la communication en ligne : les chats et forums (Tam & Martell, 2009) ou les posts sur divers réseaux sociaux ((Simaki *et al.*,

2016) ; (Pentel, 2015a) ; (Pentel, 2015b) ; (Nguyen *et al.*, 2011) ; (Demmelmaier & Westerberg, 2021) ; (Van de Loo *et al.*, 2016)), afin de détecter les prédateurs dans la lutte pour la protection des mineurs (Van de Loo *et al.*, 2016) ou d'étudier les préférences de certaines communautés (Alroobaea *et al.*, 2020). L'identification de l'âge du locuteur à partir de la parole ((Bonastre *et al.*, 2000) ; (Przybocki & Martin, 1999) ; (Aman, 2014)) ou la classification de locuteurs par l'âge (Naini & Homayounpour, 2006) est essentiellement fondée sur des critères acoustiques et phonétiques.

Les avancées récentes en Intelligence Artificielle et le succès incontournable des LLM posent également la question de leur capacité à identifier l'âge d'un locuteur. Des expériences ont déjà montré que les LLM sont capables d'être adaptés aux tâches de classification (Zhang *et al.*, 2024). Par ailleurs, l'utilisation d'un LLM pour une tâche de détection de l'âge soulève en arrière-plan des questions de nature sociolinguistique, en raison des stéréotypes que ces modèles peuvent reproduire ; les données d'entraînement, souvent issues d'Internet, peuvent contenir des préjugés et des formes de discriminations présentes dans la société ((Ducel *et al.*, 2024) ; (Gallienne & Poibeau, 2023) ; (Liang *et al.*, 2021) ; (Bolukbasi *et al.*, 2016)). Les biais sociaux dans les modèles de langage sont une problématique majeure, car ils peuvent affecter la manière dont les modèles génèrent du texte et reflètent, voire amplifient, les inégalités sociales existantes.

La recherche présentée pose la question du lien entre le profil sociologique du locuteur et son discours en se focalisant sur la question de l'âge. Si ce lien a été démontré à travers les travaux en sociolinguistique, les propriétés proposées ne sont pas nombreuses. Nous cherchons à vérifier cette hypothèse en exploitant les outils du TAL. Si des propriétés linguistiques spécifiques peuvent être associées à des profils sociologiques, il devient alors envisageable de prédire ces derniers de manière automatisée à partir du discours d'un locuteur.

2 Données

2.1 Corpus

Les données utilisées pour ce travail sont composées de transcriptions du français oral et de métadonnées associées comprenant l'âge exact du locuteur au moment de l'enregistrement. Elles proviennent de trois corpus réalisés à de proches intervalles : ESLO2, Enquêtes sociolinguistiques à Orléans ((Baude & Dugua, 2011) ; (Eshkol-Taravella *et al.*, 2011)) récolté à partir de 2008 et composé des entretiens avec des orléanais portant sur leur vie à Orléans ; MPF, Multicultural Paris French (Gadet & Guerin, 2016) comprenant des entretiens récoltés à partir de 2010 avec des jeunes de banlieue parisienne ; enfin, LangAge (Ismail *et al.*, 2022), constitué d'entretiens avec des personnes âgées récoltés en 2005, 2012 et 2015. Les questions posées lors des entretiens de chaque corpus portent sur la vie personnelle des interrogés.

Les tours de parole (tdp) des différents locuteurs ont été extraits grâce à leurs identifiants afin de pouvoir être concaténés lors des expériences. Les transcriptions de MPF et de LangAge ont été alignées selon les conventions d'ESLO, les trois corpus ayant au départ des conventions différentes.

2.2 Classes d'âge

(Cheshire, 2008) et (Wagner, 2012) mettent en lumière deux approches théoriques distinctes pour

examiner la corrélation entre l'âge et l'usage de la langue, distinguant l'usage générationnel de l'usage spécifique à l'âge de l'individu. (Eckert, 2017b) élargit cette réflexion en définissant trois types d'âge - chronologique, biologique et social - soulignant ainsi la complexité de cette variable dans l'étude des pratiques linguistiques, qui doivent être appréhendées en relation avec d'autres variables telles que le genre ou la catégorie socio-professionnelle. Selon (Labov, 1966), les usages individuels se stabilisent au début de l'âge adulte, soit après l'adolescence. Cette période est caractérisée par une adoption plus prononcée de la norme linguistique standard chez les adultes, en contraste avec les variations plus marquées observées chez les adolescents. La notion d'âge semble être liée également à celle de la normativité. L'écart par rapport à la norme est un des critères caractéristiques du parler des jeunes. (Wagner, 2012) affirme que les adultes ont plus tendance à utiliser le langage sous sa forme standard, ce qui est encore plus marqué chez des locuteurs âgés qui adoptent un langage plus conservateur selon (Gerstenberg, 2015) et (Gerstenberg & Voeste, 2015). La recherche sur les adultes plus âgés est relativement limitée, bien que certaines études notent des marqueurs linguistiques liés aux changements cognitifs et physiologiques associés au vieillissement ((Kemper *et al.*, 1992) ; (Stine-Morrow & Payne, 2016)). Ainsi, l'analyse de l'âge en linguistique met en lumière trois phases clés de la vie adulte, chacune susceptible de générer des évolutions linguistiques distinctes : l'adolescence et le début de l'âge adulte, l'âge adulte intermédiaire et l'âge adulte avancé.

Nous considérons dans ce travail que c'est autour de 30 ans que le parler des locuteurs peut être amené à évoluer, à mesure que les responsabilités de la vie d'adulte s'accumulent (Wagner, 2012). L'autre point pivot d'un point de vue sociologique semble se situer autour de 60 ans, au moment du passage de la vie active à la retraite. Une répartition en trois classes (**-30**, **30-60**, **60+**) tenant compte de ces considérations sociolinguistiques a été proposée et testée dans nos expériences.

Le jeu de données est constitué de 90 transcriptions du corpus ESLO, 67 transcriptions du MPF et 56 transcriptions de LangAge (34978 tours de parole) avec 32383 tdp pour la classe -30, 65450 tdp pour les 30-60 et 89887 tdp pour les 60+ (Figure 1 ; voir la Table 1 en Annexe pour la répartition précise).

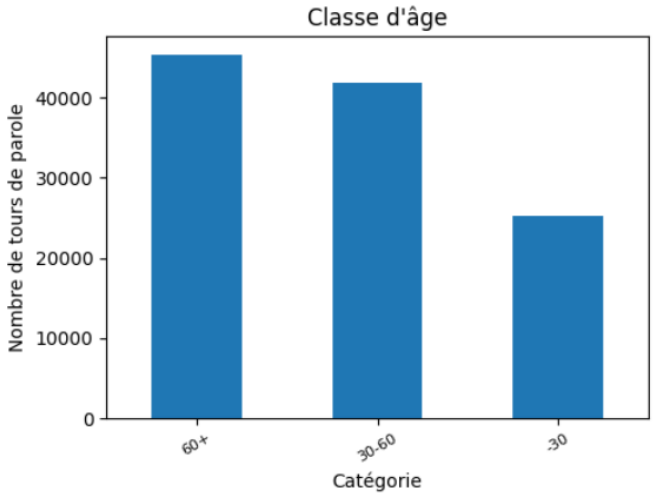

FIGURE 1 – Répartition des données en classes d'âge

3 Prédiction automatique de l'âge

3.1 Méthodologie

Nous cherchons à tester l'hypothèse, fondée sur les travaux en sociolinguistique, d'un lien entre l'âge d'un locuteur et sa parole. Pour cela, nous avons testé divers modèles capables de classer automatiquement les locuteurs dans nos trois classes prédéfinies (-30, 30-60, 60+).

Deux approches d'apprentissage ont été choisies. Tout d'abord, l'apprentissage de surface, en raison de la performance des modèles déjà employés dans la littérature (Tam & Martell, 2009) et parce qu'ils peuvent prendre en entrée des traits linguistiques. Afin de savoir dans quelle mesure ces traits sont utiles, nous comparons les modèles de surface avec des modèles profonds de type transformers. En partant de l'hypothèse que la taille de l'input textuel et la représentation vectorielle des données influenceraient les résultats, nous avons fait varier ces paramètres, en ne gardant que les plus performants. D'abord ont été testées les méthodes de vectorisation classiques (notamment TF-IDF) que nous avons appliquées sur un ensemble de fenêtres de contexte, allant du tour de parole isolé au document entier, augmentant la taille de la fenêtre de 10 en 10. Une fois les fenêtres fixées, nous ajoutons les traits linguistiques aux données vectorisées. Ces différents tests ont conduit aux configurations finales représentées par la Figure **??**. Ces modèles de classification sont également comparés à un modèle de régression, qui constitue une autre approche pour aborder la question de l'âge. Enfin, les LLM ont été utilisés à leur tour, dans l'optique de tester leurs performances et analyser leurs réponses.

3.2 Apprentissage supervisé

3.2.1 Apprentissage de surface

La première approche repose sur l'apprentissage de surface. En nous basant sur l'état de l'art, nous avons testé les classifieurs (SVM, KNN, XGBoost) utilisés dans les travaux précédents ((Pentel, 2015a) ; (Demmelmaier & Westerberg, 2021)). Les SVM se sont révélés les plus performants. Pour la fenêtre du document entier, une méthode de validation croisée à 10 plis a été choisie.

Traits linguistiques Les traits proposés pour renforcer l'apprentissage sont fondés sur des travaux en sociolinguistique et nos propres hypothèses. De manière générale, on considère que le parler des jeunes est caractérisé par un lexique spécifique, des énoncés plus courts et un registre plus familier ; les personnes plus âgées semblent pour leur part utiliser un registre plus formel, des énoncés plus longs et complexes et chercher davantage leurs mots. Les traits utilisés peuvent être regroupés selon les niveaux linguistiques suivants :

1. Longueur
   — longueur moyenne des énoncés et des mots : nous supposons que les locuteurs les plus âgés produisent des phrases et utilisent des mots plus longs.
2. Morphosyntaxe
   — adverbes en *-ment* (Charton & Torres-Moreno, 2011), car ils ont de même une longueur importante et relèvent souvent d'un registre plus formel de la langue, considéré être plus représentatif des locuteur d'un certain âge ;

— pronoms relatifs (Mekki, 2022) et connecteurs logiques (Charton & Torres-Moreno, 2011).

3. Lexique
   — formes lexicales caractéristiques de parler des jeunes : tdp terminant par *quoi*, mots en verlan récupérés à partir d'une annexe dédiée de Wiktionary, néologismes ou mots à la graphie incertaine, identifiés dans les corpus ESLO et MPF par & ;
   — disfluences : nombre d'amorces dans un énoncé (indiquées par un tiret dans les transcriptions) et répétition de mots, considérées plus fréquentes chez les locuteurs âgés.

Au total, ces critères ont été testés à travers 10 traits. Ils ont été convertis en données numériques (longueur pour les traits concernés, fréquence pour les autres), puis mis à l'échelle pour obtenir des valeurs entre 0 et 1 (min-max scaling), fournies en entrée au SVM. La pertinence des critères linguistiques a été vérifiée à l'aide d'un modèle de régression logistique avec lequel des valeurs de significativité p pour chacun des critères évalués et pour chaque classe ont été obtenues. Les critères semblant significatifs (valeur de p inférieure au seuil standard de 0.05) pour la classification ont été sélectionnés (connecteurs logiques, *quoi* en fin d'énoncé et nombre de répétitions par tdp).

Résultats Les différents tests ont montré que les meilleurs résultats sont obtenus avec un SVM à noyau polynomial, une représentation TF-IDF comptant unigrammes et bigrammes, sans stop words, à l'échelle du document entier (contenant tous les tdp d'un locuteur). La Table 1 montre les scores obtenus sans prendre en compte les traits linguistiques (traits-), en prenant en compte l'ensemble des traits (traits+), et en prenant seulement en compte ceux qui se sont révélés plus significatifs (traits++).

	-30			30-60			60+			Total (pondéré)		
	P	R	F	P	R	F	P	R	F	P	R	F
TRAITS-	0.97	0.78	0.86	0.74	0.85	0.79	0.91	0.93	0.92	0.88	0.87	0.87
TRAITS+	0.93	0.68	0.79	0.55	0.51	0.53	0.76	0.89	0.82	0.74	0.73	0.73
TRAITS++	0.85	0.80	0.83	0.67	0.77	0.71	0.91	0.85	0.88	0.82	0.82	0.82

TABLE 1 – Tableau des scores (SVM et tests des traits)

On remarque que l'ajout de traits linguistiques n'améliore pas les résultats mais au contraire les détériore, les meilleurs résultats étant obtenus sans traits. On peut mettre en doute la pertinence des traits sélectionnés, ou bien leur incompatibilité avec les scores de TF-IDF. Par ailleurs, les traits se rattachent plutôt aux classes extrêmes et n'aident donc pas à reconnaître la classe des 30-60, qui voit ses performances plus dégradées par l'ajout de traits que les deux autres classes. Enfin, il faut prendre en compte la taille des données, puisque prendre un document entier revient à réduire le nombre d'échantillons disponibles. Face à ces constats, deux expériences sont menées : (1) recourir à l'apprentissage profond pour utiliser les autres types d'informations contenues dans ces modèles en guise de comparaison (2) poursuivre l'apprentissage de surface avec d'autres types de représentations.

3.2.2 Apprentissage profond

Le modèle pré-entrainé CamemBERT (Martin *et al.*, 2020) a été choisi pour aborder la classification et la régression (partie 3.3) avec de l'apprentissage profond. Une fenêtre de 30 tdp à classer est fournie en entrée afin de respecter la limite de 512 tokens pouvant être traités tout en maximisant les performances. L'ajustement des paramètres du modèle sur la tâche de classification en tranches

d'âge a été fait selon la méthode d'optimisation Adam. Les principaux hyperparamètres utilisés sont le nombre d'epochs (8), la longueur maximum de la phrase (200), la taille du batch (12) et le taux d'apprentissage (2e-5). Une couche de sortie de 3 neurones a enfin été connectée à CamemBERT pour prédire les trois classes -30, 30-60 et 60+. L'entraînement est effectué sur 70 % des données et 30 % servent au test.

Les scores pour chacune des trois classes ne descendent pas en dessous de 0.69 pour le R et la P. La classe intermédiaire reste ici aussi celle qui comporte le plus de mauvaises classifications.

-30			30-60			60+			Total (pondéré)		
P	R	F	P	R	F	P	R	F	P	R	F
0.86	0.93	0.89	0.69	0.83	0.76	0.97	0.70	0.82	0.84	0.82	0.82

TABLE 2 – Tableau des scores (CamemBERT)

3.2.3 Approche hybride à partir d'embeddings de documents

Afin d'optimiser la représentation des documents, les expériences précédentes sont combinées pour obtenir une représentation vectorielle sémantiquement riche mais suffisamment manipulable pour y ajouter nos traits linguistiques, le tout appliqué au modèle le plus performant. Pour ce faire, des embeddings de documents (d'une fenêtre de 30 tdp) sont construits à partir de CamemBERT. L'embedding du premier token de chaque séquence (CLS), qui encapsule la représentation de la séquence entière, est ensuite extrait. Nous testons ensuite 3 configurations : embeddings seuls (CLS), embeddings concaténés à des traits linguistiques pré-selectionnés (CLS_traits++) et embeddings concaténés à tous les traits linguistiques (CLS_traits+). Les données sont divisées similairement aux expériences avec l'apprentissage profond (70 % pour l'entraînement et 30 % pour le test). La pré-sélection de traits est effectuée via le test chi2 grâce auquel 5 traits ont été retenus : la présence d'argot, la longueur moyenne d'un énoncé en mots, la longueur moyenne d'un énoncé en caractères, le nombre moyen connecteurs logiques par énoncé et le nombre moyen de répétitions par énoncé.

Les meilleurs résultats sont obtenus avec un SVM (noyau linéaire) couplé à la configuration CLS_traits++, suivie par CLS_traits+. Les scores sont globalement supérieurs à l'apprentissage profond, et le repérage de la classe 30-60 est comparable aux meilleurs résultats des modèles de surface. De plus, les traits linguistiques se révèlent cette fois-ci décisifs, en particulier pour réduire les écarts de performance entre les classes : on compte 12 points d'écart entre la classe la mieux prédite et la classe la moins bien prédite pour le modèle SVM TF-IDF, contre 10 points pour CLS_traits++.

	-30			30-60			60+			Total (pondéré)		
	P	R	F	P	R	F	P	R	F	P	R	F
CLS	0.89	0.77	0.83	0.74	0.80	0.77	0.87	0.87	0.87	0.83	0.83	0.83
CLS traits++	**0.91**	0.76	**0.83**	0.75	**0.81**	0.78	0.88	0.89	0.88	**0.84**	**0.84**	**0.84**
CLS traits+	0.89	0.77	0.83	0.75	0.80	0.78	0.88	0.89	0.88	0.84	0.84	0.84

TABLE 3 – Tableau des scores (embeddings CLS + SVM + traits)

3.3 Régression

L'âge étant une valeur continue et linéaire, nous avons également souhaité aborder cette question à travers la méthode de la régression. Cette approche entend notamment pallier le problème posé par les locuteurs à la frontière de deux classes : un locuteur de 57 ans a sans doute une façon de parler plus proche de certains individus de la classe des 60+ que de ceux de la classe 30-60. Le modèle de classification utilisant CamemBERT a été adapté afin de correspondre à la tâche de régression. La fonction d'activation de la dernière couche du modèle a été supprimée pour que les sorties correspondent directement aux âges prédits. Les autres paramètres sont restés identiques. Les données ont été réparties avec 70 % dédiés à l'entraînement, 15 % à la validation et 15 % à l'évaluation.

66 % des échantillons ont été bien classés, en considérant une marge d'erreur de 5 ans. La MAE (erreur absolue moyenne) est de 7.46 ans et la médiane des erreurs est de 6 ans. La RMSE (racine de l'erreur quadratique moyenne), qui est pour sa part de 9.71 ans, montre néanmoins que certaines prédictions ont des écarts importants. Enfin, le R^2 (coefficient de détermination) est de 0.79, ce qui signifie que le modèle explique 79 % de la variance de l'âge. Toutefois, si nous transformons les prédictions pour obtenir à nouveau les trois classes d'âge (cf Table 6), les résultats ne parviennent pas à surpasser ceux des modèles précédemment testés.

3.4 Discussion

L'ensemble des expériences réalisées jusqu'alors démontre que la tâche de prédiction de l'âge à partir de transcriptions de l'oral est possible. Les modèles parviennent à saisir certains critères discriminants, qui ne nous sont néanmoins pas toujours accessibles. Il demeure que la tâche présente une certaine complexité et les traits proposés dans la littérature se révèlent suffisants pour rendre compte des variations à l'œuvre. Même si les meilleurs résultats sont obtenus avec un SVM sans traits linguistiques (F1 de 0.87 à l'échelle du document entier), ces derniers se sont avérés utiles lorsqu'associés à un autre type de représentation (embeddings de documents). Afin de compléter notre analyse, nous nous tournons vers les LLM, qui pourraient encapsuler des informations plus subtiles sur les spécificités linguistiques des classes étudiées.

4 Détection de l'âge à l'aide de LLM

Les LLM sont devenus incontournables dans le domaine du TAL et présentent des résultats prometteurs dans les tâches d'analyse linguistique ((di Buono *et al.*, 2024) ; (Ettinger *et al.*, 2023) ; (Gan *et al.*, 2024)). L'objectif est donc double : mesurer leur performance sur la tâche de détection de l'âge (classification) et accéder aux critères sur lesquels ils se basent pour effectuer cette tâche (analyse linguistique). Les prompts nécessitant des indications complexes, nous choisissons le modèle Mistral 7B Instruct, fine-tuné pour les instructions. En nous inspirant des méthodes employées dans la littérature, nous partons d'un prompt initial (Zero-shot Prompt), observons les réponses obtenues et affinons le prompt en ajustant les paramètres jusqu'à obtenir un résultat complet et satisfaisant du point de vue du format de la réponse. Les paramètres gardés pour le dataset final sont une température et un top p à 0.7, et la longueur maximale de la réponse limitée à 1024 tokens.

Le prompt final (cf Figure 1 en Annexe) est composé :

— d'une introduction,

— d'une instruction pour la classification,
— d'une instruction pour l'analyse linguistique,
— de l'énumération des différentes catégories d'âge,
— d'une transcription (30 tours de parole),
— des instructions complémentaires sur le format attendu de la réponse.

Le modèle est testé sur la totalité du dataset et évalué avec les mesures traditionnelles (Table 4). Les performances sont nettement inférieures aux modèles précédents. Cependant, contrairement à ces derniers, la classe des 30-60 a un rappel largement supérieur aux deux autres classes, ce qui semble indiquer qu'elle est souvent prédite "par défaut".

-30			30-60			60+			Total (pondéré)		
P	R	F	P	R	F	P	R	F	P	R	F
0.54	0.09	0.16	0.43	0.79	0.55	0.76	0.57	0.65	0.61	0.55	0.52

TABLE 4 – Tableau des scores (Mistral 7B Instruct v3)

Pour fouiller les analyses linguistiques du modèle, l'outil d'extraction de mots clés (!YAKE) est utilisé. Pour chaque catégorie de traits linguistiques et pour chaque classe prédite par le modèle, les n-grammes les plus représentatifs sont extraits (Table 5). On constate que le modèle associe systématiquement le langage des plus jeunes avec l'emploi d'un registre "informel" et des constructions "simples" et "courtes". À l'inverse, les âges adulte et avancé sont associés à des structures syntaxiques complexes, aux archaïsmes et aux registres de langue formels. Les critères sémantiques évoqués sont également distincts : les jeunes parleraient de leur quotidien (parents, études) tandis que les adultes évoqueraient leur vie professionnelle, et les plus âgés leurs expériences passées. Cela expliquerait les mauvaises performances du modèle pour détecter la classe des -30, ceux-ci pouvant aussi avoir une vie professionnelle. Par ailleurs, certains traits relevés sont attribués à toutes les tranches d'âge (e.g. le registre courant, la répétition de "oui"), ce qui met en doute leur pertinence. Les analyses du modèle sont donc équivoques : si certains critères rejoignent nos observations ainsi que ce qui est reconnu dans la littérature (longueur, registre...), ils sont parfois poussés à l'extrême, en associant par exemple les contractions et abréviations avec les classes plus jeunes, et les temps du passé aux plus âgés.

Classe prédite	**Lexique**	**Morphologie**	**Syntaxe**	**Sémantique**	**Pragmatique**
-30	courant, informel, familier, Wesh, argot, simple	passé et présent simple, formes courtes et abrégées, article défini	constructions syntaxiques simples, courtes ou incomplète, langage courant et informel, conjonctions simples, contraction	courant, vie quotidienne, parents, intérêt, études, expérience professionnelle, éducation, école, activité courante	"oui oui oui", courant, familier, enthousiaste et passionné, conversations informelles, directes et courtes
30-60	décontracté, formel ou informel, varié, riche, courant, complexe, techniques, éducation	formes verbales passées, courtes, composées et complexes, régulières et courantes, contractions	structures complexes et subordonnées, mature, registre formel	expérience professionnelle, expériences passées, maturité, courant, travail	"oui oui oui, oui oui", courant, expérience professionnelle ou de vie, compréhension, maturité
60+	courant, formel, riche et varié, ancien et archaïque, expérience, époque	formes verbales passées voire archaïques, irrégulières	structures syntaxiques complexes, subordonnées complexes, temps passés, formes courantes, simples, courtes, formelles	Seconde Guerre mondiale et événements historiques, expériences et vie passées, expérience professionnelle souvent longue	expériences de vie et professionnelles passées, courant, "oui oui oui", avancé, événements historiques

TABLE 5 – Caractéristiques linguistiques associées aux classes d'âge prédites par Mistral

On souligne également un ensemble de limitations techniques : une qualité des analyses très variable,

des réponses ne respectant pas toujours les instructions, ainsi que la présence d'hallucinations (le modèle a parfois donné des critères phonétiques et évoquait l'accent des locuteurs). Cela peut être dû aux exigences du prompt concernant le format des réponses, problème évoqué par (Gan *et al.*, 2024), et qui pose un vrai défi pour l'utilisation des LLM sur des grands volumes de données. Des améliorations seraient donc possibles avec un LLM fine-tuné pour notre tâche.

5 Discussion

Les différentes expériences (Table 6) mettent en évidence la difficulté à détecter l'âge d'un locuteur à partir de transcriptions orales. De manière générale, la classe médiane des 30-60 ans reste la plus difficile à détecter. Certains modèles parviennent à de relativement bons résultats (jusqu'à 0.87 de F1-score avec un SVM), et ce sans traits linguistiques. La contribution de ces derniers dans la performance des modèles semble dépendre du choix initial de la représentation des données, car des résultats satisfaisants ont été obtenus avec des approches hybrides. Enfin, le LLM testé présente des scores plus faibles que les autres modèles - les traits linguistiques proposés rejoignent ceux de la littérature mais sont parfois exacerbés, ce qui peut indiquer des biais quant à la perception de l'âge.

	-30			30-60			60+			Total (pondéré)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baseline (classe maj)	0	0	0	0.40	1	0.57	0	0	0	0.16	0.40	0.23
SVM+TFIDF (sans traits)	**0.97**	0.78	**0.86**	0.74	**0.85**	**0.79**	0.91	**0.93**	**0.92**	**0.88**	**0.87**	**0.87**
SVM+TFIDF (traits sélec)	0.85	0.80	0.83	0.67	0.77	0.71	0.91	0.85	0.88	0.82	0.82	0.82
CamemBERT (mono-tâche)	0.86	**0.93**	**0.86**	0.69	0.83	0.76	**0.97**	0.70	0.82	0.84	0.82	0.82
CLS traits++ (traits sélec)	0.91	0.76	0.83	**0.75**	0.81	0.78	0.88	0.89	0.88	0.84	0.84	0.84
CamemBERT (régression)	0.87	0.70	0.78	0.72	0.84	0.78	0.85	0.76	0.80	0.79	0.78	0.78
Mistral 7B Instruct	0.54	0.09	0.16	0.43	0.79	0.55	0.76	0.57	0.65	0.61	0.55	0.52

TABLE 6 – Performances des modèles de classification et de régression pour la prédiction de l'âge

La question des biais propres aux corpus d'entraînement doit être relevée : la répartition des classes d'âges se recoupe en partie avec celle des corpus. Il n'est donc pas exclu que les modèles aient saisi au cours de l'apprentissage des informations propres aux corpus plus qu'à l'âge en lui-même, ce qui expliquerait notamment les performances des TF-IDF sans traits. Cependant, la majorité des données proviennent d'ESLO2, représenté dans toutes les classes d'âge et à proportion comparable avec les deux autres corpus (voir Table 1 en Annexe). De même, puisqu'il est très probable que les modèles se soient basé en majorité sur le lexique, il faut rappeler que ce dernier varie en fonction des thématiques abordées par les locuteurs. Les entretiens étant semi-dirigés, il est possible que certains sujets soient plus récurrents que d'autres (par exemple, l'opinion du locuteur sur la ville d'Orléans). Toutefois, la majorité des questions étant ouvertes et portant sur le vécu des locuteurs, on considère que les sujets évoqués font partie des critères pertinents pour déterminer leur âge. Il serait peut-être donc judicieux de filtrer les données en fonction des thématiques dans de futures expériences.

6 Conclusion

La difficulté des modèles à déterminer l'âge d'un locuteur à partir de la transcription de sa parole soulève plusieurs interrogations qui dessinent différentes pistes pour poursuivre cette recherche.

Les traits ayant aidé à la classification sont ceux liés à la longueur et au lexique (argot et connecteurs logiques), ce qui rejoint les résultats des travaux évoqués en 3. Nous souhaiterions réfléchir à des traits plus généralisables qui soient moins liés au lexique, afin de réduire les biais de corpus discutés précédemment. Les biais liés aux corpus pourraient quant à eux être interrogés à travers des expériences de clustering au niveau des documents.

Par ailleurs, le plafond de verre atteint par les modèles et la difficulté à trouver des traits vraiment discriminants pose la question de la tâche en elle-même. Le profil d'un individu est loin d'être entièrement déterminé par son âge et mériterait d'être croisé avec d'autres données, comme la catégorie socio-professionnelle, qui permettrait une analyse plus complète et nuancée.

Les expériences menées avec les LLM gagneraient à être poursuivies. L'approche suivie ici, sans apport informationnel, a montré les limites de ces modèles, dont les justifications s'appuient en partie sur des biais ou des stéréotypes. Il serait intéressant de spécialiser le modèle par finetuning, ou bien en affinant les prompts, ou encore en lui fournissant des exemples (*few-shot learning*) ainsi que plus de connaissances en sociolinguistique.

Enfin, une expérience avec des annotateurs humains a été débutée afin de comparer leur performance avec celle des modèles et d'étudier les critères qu'ils emploient pour déterminer l'âge d'un locuteur.

Références

ALROOBAEA R., ALMULIHI A. H., ALHARITHI F. S., MECHTI S., KRICHEN M. & BELGUITH L. H. (2020). A deep learning model to predict gender, age and occupation of the celebrities based on tweets followers. In *CLEF (Working Notes)*.

AMAN F. (2014). *Reconnaissance automatique de la parole de personnes âgées pour les services d'assistance à domicile*. Thèse de doctorat, Université de Grenoble.

BAUDE O. & DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus*, **10**, 99–118. HAL : hal-01162479.

BERNSTEIN B. (1971). *Class, Codes and Control : Volume 1 - Theoretical Studies Towards a Sociology of Language*. London : Routledge & Kegan Paul.

BOLUKBASI T., CHANG K. W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, p. 4349–4357.

BONASTRE J.-F., DELACOURT P., FREDOUILLE C., MEIGNIER S., MERLIN T. & WELLEKENS C. (2000). Différentes stratégies pour le suivi du locuteur, reconnaissances des formes et intelligence artificielle. In *RFIA 2000*.

BOURDIEU P. (1982). *Ce que parler veut dire : L'économie des échanges linguistiques*. Paris : Fayard.

CHARTON E. & TORRES-MORENO J.-M. (2011). Modelisation automatique de connecteurs logiques par analyse statistique du contexte. *Canadian Journal of Information & Library Sciences*, **35**(3).

CHESHIRE J. (2008). *Age- and Generation-Specific Use of Language*, In *Volume 2 : An International Handbook of the Science of Language and Society*, p. 1552–1563. De Gruyter Mouton.

CHUQUET H. (1990). *Pratique de la traduction : anglais-français*. Editions OPHRYS.

COATES J. (1993). *Women, Men and Language : A Sociolinguistic Account of Gender Differences in Language*. London : Longman.

DEMMELMAIER G. & WESTERBERG C. (2021). *Data Segmentation Using NLP : Gender and Age*. Thèse de doctorat, Uppsala Universitet.

DI BUONO M. P., SPAHIU B. & BARBU MITITELU V. (2024). Evaluating large language models for linguistic linked data generation. In G. SÉRASSET, H. G. OLIVEIRA & G. V. OLESKEVICIENE, Éds., *Proceedings of the Workshop on Deep Learning and Linked Data (DLnLD) @ LREC-COLING 2024*, p. 66–75, Torino, Italia : ELRA and ICCL.

DUCEL F., NÉVÉOL A. & FORT K. (2024). La recherche sur les biais dans les modèles de langue est biaisée : état de l'art en abyme. *Revue TAL : traitement automatique des langues*, **64**(3), 119–143.

ECKERT P. (2017a). *Age as a sociolinguistic variable*. Wiley.

ECKERT P. (2017b). Age as a sociolinguistic variable. In *The handbook of sociolinguistics*, p. 151–167. F. Coulmas.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2011). Un grand corpus oral " disponible " : le corpus d'Orléans 1 1968-2012. *Revue TAL : traitement automatique des langues*, **53**(2), 17–46. HAL : halshs-01163053.

ETTINGER A., HWANG J., PYATKIN V., BHAGAVATULA C. & CHOI Y. (2023). "you are an expert linguistic annotator" : Limits of LLMs as analyzers of abstract meaning representation. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 8250–8263 : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.553.

GADET F. (1996). Niveaux de langue et variation intrinsèque. *Palimpsestes. Revue de traduction*, **10**, 17–40.

GADET F. & GUERIN E. (2016). " Construire un corpus pour des façons de parler non standard : 'Multicultural Paris French' ". *Corpus*, (15), p. 285–307. HAL : halshs-01658279.

GALLIENNE R. & POIBEAU T. (2023). Quelques observations sur la notion de biais dans les modèles de langue. *18e Conférence en Recherche d'Information et Applications*.

GAN Y., POESIO M. & YU J. (2024). Assessing the capabilities of large language models in coreference : An evaluation. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 1645–1665, Torino, Italia : ELRA and ICCL.

GERSTENBERG A. (2015). 14 langues et générations : enjeux linguistiques du vieillissement. In *Manuel de linguistique française*, p. 314–333. De Gruyter.

GERSTENBERG A. & VOESTE A. (2015). *Language development : The lifespan perspective*, volume 37. John Benjamins Publishing Company.

HOCKETT C. F. (1950). Age-grading and linguistic continuity. *Language*, **26**, 449.

ISMAIL E. E. S., GERSTENBERG A., SPAGNOLO M. L., SCHULZ F. & VANDENBROUCKE A. (2022). L'âge avancé en perspective longitudinale et ses outils : Langage, un corpus au pluriel. *SHS Web of Conferences*, **138**, 10003.

KEMPER S., KYNETTE D. & NORMAN S. (1992). Age differences in spoken language. *Everyday memory and aging : Current research and methodology*, p. 138–152.

LABOV W. (1966). *The Social Stratification of English in New York City*. Washington, D.C. : Center for Applied Linguistics.

LABOV W. (1972). *Language in the Inner City : Studies in the Black English Vernacular*, volume 3. University of Pennsylvania Press.

LIANG P. P., WU C., MORENCY L.-P. & SALAKHUTDINOV R. (2021). Towards understanding and mitigating social biases in language models. In *ICML 2021*. arXiv preprint.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MEKKI J. (2022). Caractérisation de registres de langue par extraction de motifs séquentiels émergents. *Thèse de doctorat, Université de Rennes*.

NAINI A. S. & HOMAYOUNPOUR M. (2006). Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods. In *2006 8th International Conference on Signal Processing*, volume 1 : IEEE.

NGUYEN D., SMITH N. A. & ROSE C. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop*, p. 115–123.

NGUYEN D., SMITH N. A. & ROSÉ C. P. (2013). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, p. 115–123.

PENTEL A. (2015a). Automatic age detection using text readability features. In *EDM Workshops*.

PENTEL A. (2015b). Effect of different feature types on age based classification of short texts. In *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, p. 1–7 : IEEE.

PRZYBOCKI M. A. & MARTIN A. F. (1999). The 1999 nist speaker recognition evaluation, using summed two-channel telephone data for speaker detection and speaker tracking. In *Sixth European Conference on Speech Communication and Technology*.

RAO D., YAROWSKY D., SHREEVATS A. & GUPTA M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, p. 37–44.

SANKOFF G. & BLONDEAU H. (2007). Language change across the lifespan : /r/ in montreal french. *Language*, **83**(3), 560–588.

SAP M., PARK G., EICHSTAEDT J., KERN M., STILLWELL D., KOSINSKI M., UNGAR L. & SCHWARTZ H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1146–1151.

SIMAKI V., MPORAS I. & MEGALOOIKONOMOU V. (2016). Evaluation and sociolinguistic analysis of text features for gender and age identification. *American Journal of Engineering and Applied Sciences*, **9**(4), 868–876.

STINE-MORROW E. A. & PAYNE B. R. (2016). Age differences in language segmentation. *Experimental Aging Research*, **42**(1), 83–96.

TAM J. & MARTELL C. H. (2009). Age detection in chat. In *2009 IEEE International Conference on Semantic Computing*, p. 33–39 : IEEE.

TANNEN D. (1990). *You Just Don't Understand : Women and Men in Conversation*. New York : William Morrow.

TRUDGILL P. & ET AL. (1974). *The Social Differentiation of English in Norwich*, volume 13. CUP Archive.

VAN DE LOO J., DE PAUW G. & DAELEMANS W. (2016). Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, **5**(1), 46–60.

WAGNER S. E. (2012). Age grading in sociolinguistic theory. *Language and Linguistics Compass*, **6**(6), 371–382.

WOLFRAM W. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Center for Applied Linguistics. Google-Books-ID : kudZAAAAMAAJ.

ZHANG Y., WANG M., REN C., LI Q., TIWARI P., WANG B. & QIN J. (2024). Pushing the limit of llm capacity for text classification.

Annexe

	document			30 tdp		
	-30	30-60	60+	-30	30-60	60+
ESLO	11	46	27	374	1416	996
MPF	29	0	0	486	0	0
LANGAGE	0	3	42	0	49	1127
Total	40	49	69	860	1465	2123

TABLE 1 – Répartition des échantillons par fenêtre

INSTRUCTIONS

Tu reçois la transcription des paroles d'une personne inconnue qui a répondu à des questions pendant un entretien.
En te basant sur le contenu de la transcription, détermine la tranche d'âge à laquelle appartient la personne et explique ton choix.
Tu as 3 options:
- moins de 30 ans
- entre 30 et 60 ans
- 60 ans et plus
L'explication doit mentionner quelles propriétés linguistiques t'ont aidé dans ta décision. Tu dois prendre en compte la morphologie, le lexique, la syntaxe, la sémantique et la pragmatique.
Pour chaque propriété linguistique, donne des citations exactes du texte. Tu peux aussi donner d'autres critères linguistiques.
Tu n'es pas obligé de donner une réponse pour toutes les propriétés linguistiques : si tu n'en trouves pas, laisse le champ de réponse vide.
Ecris la réponse en français et au format JSON. Exemple :

{ "option" :"30-60" , "explication_syntaxe": {"explications": " ", "citations": [] }, "explication_sémantique" : {"explications": " ", "citations": [] }}.

DEBUT TRANSCRIPTION
[TRANSCRIPTION]
FIN TRANSCRIPTION

FIGURE 1 – Schéma de prompt