Des données aux systèmes : étude des liens entre données d'apprentissage et biais de performance genrés dans les systèmes de reconnaissance automatique de la parole

Mahault Garnerin

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEURE DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Sciences du Langage & Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Mahault GARNERIN

Thèse dirigée par **Claudine MOÏSE**
co-dirigée par **Laurent BESACIER**
et co-encadrée par **Solange ROSSATO**

préparée au sein des **Laboratoires LIDILEM & LIG**
et de l'**ED LLSH**

DES DONNÉES AUX SYSTÈMES

Étude des liens entre données d'apprentissage et biais de performance genrés dans les systèmes de reconnaissance automatique de la parole

Thèse soutenue publiquement le **16 mars 2022**,
devant le jury composé de :

François PORTET
Professeur des Universités, Université Grenoble Alpes, LIG, Président du jury
Jean-François BONASTRE
Professeur des Universités, Université d'Avignon, LIA, Rapporteur
Frédérique SEGOND
Directrice de recherche, INRIA, Rapporteure
Maria CANDEA
Professeure des Universités, Université Sorbonne Nouvelle, CLESTHIA, Examinatrice
Claudine MOÏSE
Professeure des Universités, Université Grenoble Alpes, LIDILEM, Directrice de thèse
Laurent BESACIER
Principal Scientist, NAVER Labs Europe & Professeur, Université Grenoble Alpes, Co-Directeur de thèse
Solange ROSSATO
Maîtresse de Conférences, Université Grenoble Alpes, LIG, Co-Encadrante de thèse

Résumé

Certains systèmes issus de l'apprentissage machine, de par leurs données et les impensés qu'ils encapsulent, contribuent à reproduire des inégalités sociales, alimentant un discours sur les "biais de l'intelligence artificielle". Ce travail de thèse se propose de contribuer à la réflexion collective sur les biais des systèmes automatiques en questionnant l'existence de biais de genre dans les systèmes de reconnaissance automatique de la parole ou ASR (pour *Automatic Speech Recognition*).

Penser l'impact des systèmes nécessite une articulation entre les notions de biais (ayant trait à la constitution du système et de ses données) et de discrimination, définie au niveau de la législation de chaque pays. On considère un système comme discriminatoire lorsqu'il effectue une différence de traitement sur la base de critères considérés comme brisant le contrat social. En France, le sexe et l'identité de genre font partie des 23 critères protégés par la législation.

Après une réflexion théorique autour des notions de biais, et notamment sur le biais de prédictif (ou biais de performance) et le biais de sélection, nous proposons un ensemble d'expériences pour tenter de comprendre les liens entre biais de sélection dans les données d'apprentissage et biais prédictif du système. Nous nous basons sur l'étude d'un système HMM-DNN appris sur des corpus médiatiques francophones, et d'un système end-to-end appris sur des livres audio en anglais. Nous observons ainsi qu'un biais de sélection du genre important dans les données d'apprentissage contribue de façon assez partielle au biais prédictif du système d'ASR, mais que ce dernier émerge néanmoins lorsque les données de parole regroupent des situations d'énonciation et des rôles de locuteurs et locutrices différents. Ce travail nous a également conduite à questionner la représentation des femmes dans les données, et plus généralement à repenser les liens entre conception théorique du genre et systèmes d'ASR.

Abstract

Machine learning systems contribute to the reproduction of social inequalities, because of the data they use and for lack of critical approches, thus feeding a discourse on the "biases of artificial intelligence". This thesis aims at contributing to collective thinking on the biases of automatic systems by investigating the existence of gender biases in automatic speech recognition (ASR) systems.

Critically thinking about the impact of systems requires taking into account both the notion of bias (linked with the architecture, or the system and its data) and that of discrimination, defined at the level of each country's legislation. A system is considered discriminatory when it makes a difference in treatment on the basis of criteria defined as breaking the social contract. In France, sex and gender identity are among the 23 criteria protected by law.

Based on theoretical considerations on the notions of bias, and in particular on the predictive (or performance) bias and the selection bias, we propose a set of experiments to try to understand the links between selection bias in training data and predictive bias of the system. We base our work on the study of an HMM-DNN system trained on French media corpus, and an end-to-end system trained on audio books in English. We observe that a significant gender selection bias in the training data contributes only partially to the predictive bias of the ASR system, but that the latter emerges nevertheless when the speech data contain different utterance situations and speaker roles. This work has also led us to question the representation of women in speech data, and more generally to rethink the links between theoretical conceptions of gender and ASR systems.

Remerciements

Ce manuscrit, s'il se présente comme le résultat de l'expérience individuelle qu'est le doctorat est en fait une production collective, et n'existerait pas sans toutes les personnes qui m'ont aidée à nourrir cette réflexion, mais aussi soutenue durant ces dernières années.

Je tiens donc tout d'abord à remercier Solange Rossato, Claudine Moïse et Laurent Besacier de m'avoir accompagnée, depuis le mémoire de master. Merci d'avoir accepté de me suivre sur ce sujet, et merci pour vos précieux retours et conseils et qui m'ont rendue, et me rendent encore aujourd'hui, infiniment reconnaissante de vous avoir eu·es dans ma direction de thèse.

Merci également à Jean-François Bonastre, Maria Candea, François Portet et Frédérique Segond d'avoir accepté de faire partie du jury de cette thèse.

Je remercie ensuite l'ensemble de l'équipe GETALP du LIG, et tout particulièrement le bureau 325, petit havre de douceur et de rires dans l'aventure de la thèse. Merci également aux membres du LIDILEM, notamment celles et ceux de l'Axe Genre et à toute l'équipe des doctorant·es de Claudine. Je garde le souvenir de beaux moments de recherche et de vie à vos côtés.

Enfin, je tiens également à remercier mes proches, famille et ami·es, qui ont été un soutien sans faille. Merci pour vos mots et votre présence, qui m'ont été plus que précieuses.

À mon père,

The presumption that what is male is universal is a direct consequence of the gender data gap. Whiteness and maleness can only go without saying because most other identities never get said at all. But male universality is also a cause of the gender data gap : because women aren't seen and aren't remembered, because male data makes up the majority of what we know, what is male comes to be seen as universal. It leads to the positioning of women, half the global population, as a minority. With a niche identity and subjective point of view. In such a framing, women are set up to be forgettable. Ignorable. Dispensable - from culture, from history, and from data. And so, women become invisible.

Caroline Criado-Perez, *Invisible Women*

Sommaire

Note sur l'écriture inclusive

Lors de l'écriture de cette thèse nous avons souhaité utiliser l'écriture inclusive. La majorité du temps, la double flexion est utilisée (*les locuteurs et les locutrices*), ainsi que l'accord de proximité. Nous avons choisi d'utiliser le point médian de manière très ponctuelle, lorsqu'il permettait un allégement conséquent pour la lecture.

Pour autant, au fur et à mesure de l'avancée de la rédaction du manuscrit, et notamment celle du chapitre 9, nous nous sommes rendue compte de la tension existant entre notre réflexion sur la possibilité de sortir d'un système binaire du genre et la manière dont notre écriture réaffirmait cette binarité. Ce document témoigne donc lui aussi, par son écriture, de la difficulté que l'on peut rencontrer à remettre en question les systèmes catégoriels qui organisent nos sociétés.

Introduction

Avec l'émergence du *Big Data*, les données sont devenues le nouvel axiome de la production scientifique, notamment en informatique et en intelligence artificielle (IA). Pour résoudre un problème, il suffirait de réunir suffisamment de données et de trouver l'algorithme adéquat pour traiter ces informations, les données étant considérées comme objectives et infalsifiables. Le recours à des systèmes automatiques a d'ailleurs d'abord été basé sur le mythe selon lequel de tels outils nous permettraient d'atteindre l'objectivité absolue, venant ainsi compenser la subjectivité humaine. Cependant, les données ne sont pas indépendantes des contextes dans lesquels elles ont été produites, ni de la manière dont elles ont été collectées, et par voie de conséquence, ne sont pas non plus exemptes des inégalités qui traversent nos sociétés. Certains systèmes, de par leurs données et les impensés sociaux qu'ils encapsulent, contribuent ainsi à reproduire ces inégalités et échouent à atteindre l'objectivité attendue. C'est le cas par exemple du travail de Joy Buolamwini et Timnit Gebru, dans leur étude *GenderShades*, qui montraient comment les modules de reconnaissance automatique de genre pour les systèmes de reconnaissance visuelle obtenaient des taux d'erreurs bien plus importants pour les femmes et les personnes noires, le taux d'erreur maximal étant atteint pour les femmes noires. Étudier et corriger les biais des systèmes est ainsi devenu un enjeu de la recherche en IA, mais penser l'impact de ces technologies doit se faire de manière située, dans un cadre de référence, qu'il soit légal ou social. En effet, l'utilisation de ces systèmes implique une articulation entre les notions de biais (ayant trait à la constitution du système et de ses données) et de discrimination, définie au niveau de la législation de chaque pays. Le caractère discriminatoire d'un système surgissant lorsque la différence de traitement est effectuée sur la base de critères considérés comme brisant le contrat social. En France, le sexe et l'identité de genre font partie des 23 critères protégés par la législation.

Cette thèse se propose de contribuer à la réflexion collective sur les biais des systèmes automatiques en questionnant l'existence de biais de genre dans les systèmes de reconnaissance automatique de la parole ou ASR (pour *Automatic Speech Recognition*). Le genre est une notion complexe et les différents développements qu'il a connus ont contribué à en faire un terme flou et polysémique, mais il permet de rendre compte à la fois du système normatif du genre construit autour des catégories d'"homme" et de "femme" dans lequel s'inscrivent les personnes et qui influence implicitement nos technologies, mais également de l'identité de genre de ces personnes, à savoir le discours qu'elles portent sur elles-mêmes ainsi que leurs pratiques et caractéristiques vocales. De fait, le genre s'incarne dans la voix, dans un processus complexe et interactionnel, qui fait de la production d'une voix genrée

une pratique sociale et culturelle qui ne peut être réduite uniquement au produit d'un donné biologique (un tractus vocal féminin ou masculin). Notre intérêt pour la question de l'existence de disparités de performances des systèmes de reconnaissance automatique de la parole en fonction du genre vient également de l'existence historique d'une remise en question de la voix et par extension de la parole des femmes. Le développement des technologies de la parole s'est fait dans un cadre largement masculin, où la question de la variation de genre ne se posait pas. Les systèmes de reconnaissance de la parole se sont d'ailleurs attachés à caractériser les sons de parole indépendamment des caractéristiques vocales relevant de l'identité du locuteur ou de la locutrice, se focalisant sur le contenu linguistique. Ces dernières années, les performances des systèmes ont fait un bond en avant considérable avec les nouveaux paradigmes à base de réseaux de neurones. Ces réseaux artificiels sont une modélisation inspirée du traitement de l'information par le cerveau humain. Un réseau de neurones se pense donc comme un réseau de noeuds de calculs reliés entre eux, dont les paramètres sont appris à partir de quantités importantes de données, données qui peuvent encapsuler des phénomènes sociaux, et donc susceptibles d'intégrer des inégalités de traitement en fonction du genre.

Cette thèse s'est donc attachée à explorer l'existence de biais genrés dans les systèmes d'ASR, ce qui nécessite de circonscrire la notion de biais, d'en questionner les impacts et donc de revenir à une conception socio-technique des systèmes (Chapitre 1). Questionner les biais genrés amène également à définir le genre en tant que concept et comment celui-ci s'articule avec les systèmes technologiques (Chapitre 2). Nous nous intéresserons également à définir l'ASR comme objet technique et à le replacer dans le contexte historique de son développement (Chapitre 3). L'ensemble de ces réflexions nous permet d'élaborer notre problématique de recherche (Chapitre 4) et nous présentons ensuite notre cadre méthodologique à travers nos données, nos outils ainsi que nos hypothèses de travail et nos méthodes d'analyse (Chapitre 5). Les résultats sont présentés dans les Chapitres 6 à 8 et nous proposons un travail de mise en perspective dans notre Chapitre 9 avant de conclure.

Première partie

Approche théorique

Chapitre 1

Éthique et IA

En 2017, le Premier Ministre Édouard Philippe confiait à Cédric Villani et une équipe de différents responsables dans les secteurs du numérique et de la recherche, une mission parlementaire pour définir une stratégie nationale et européenne suite au développement de l'intelligence artificielle (IA). Cette mission a donné lieu à un rapport intitulé *Donner un sens à l'intelligence artificielle* (Villani *et al.*, 2018). Ce choix de titre est le reflet de l'émergence de la question de l'explicabilité des systèmes intelligents qui investissent de plus en plus les différentes activités décisionnelles, rythmant aussi bien nos vies administratives que quotidiennes. Ces nouveaux systèmes automatiques, issus pour leur majorité d'un apprentissage supervisé, sont de plus en plus utilisés pour toutes sortes de tâches : du GPS à la traduction automatique, du tri de CV à l'attribution de prêts.

Face à une utilisation croissante de systèmes automatiques, les recherches concernant l'éthique et l'explicabilité de ces systèmes se sont démultipliées, notamment du fait de différences de traitement en fonction du genre ou de la race. En 2018, l'ancienne étudiante du MIT Joy Buolamwini montrait dans une étude intitulée *Gendershades* que les systèmes de reconnaissance faciale ne fonctionnaient pas sur elle, car elle avait la peau noire (Buolamwini et Gebru, 2018). Dès lors, de nombreux travaux se sont intéressés à l'évaluation de ces systèmes et à la détection de leurs éventuels biais et limites, mais aussi à comprendre leur fonctionnement pour prévenir ces derniers. Poser la question de l'explicabilité des systèmes implique de poser la question de leur conception et de leur développement ainsi que celle de la posture épistémologique qu'ils représentent. Il s'agira donc dans ce chapitre de discuter dans un premier temps des différentes épistémologies de la science qui ont conduit aux débats actuels sur l'explicabilité des systèmes, pour ensuite nous intéresser à l'impact de ces systèmes à travers les définitions multiples du concept de "biais". Enfin nous recenserons les différentes recommandations ayant émergé de ces débats, dans l'optique de créer des systèmes plus expliquables et équitables, ce qui nous amènera dans un dernier temps à réfléchir aux dynamiques de pouvoir derrière les développements de ces technologies.

1.1 La question épistémologique en IA

Lorsqu'on parle d'IA et de systèmes automatiques, beaucoup critiquent l'aspect "boîte noire" des systèmes d'IA, à l'instar de l'algorithme d'Admission Post Bac. En 2017, il a

donné lieu à une mise en demeure du ministère de l'enseignement supérieur par la CNIL du fait de son opacité [1] et a finalement été remplacé par Parcoursup. Les polémiques autour de l'utilisation de ces systèmes sont dues à l'impact grandissant qu'ils ont sur nos vies et nos trajectoires, aussi bien professionnelles que personnelles. Cependant, cette question de l'explicabilité en soulève une autre, à savoir : qu'est-ce qui est modélisé dans un système et comment un processus automatique, et donc logique, censé être défait de toute subjectivité humaine, en arrive à produire voire reproduire des discriminations ? Poser la question de l'impact social des systèmes nécessite de prendre en compte leur place dans les contextes socio-historiques de leur développement et de sortir d'une vision positiviste et déterministe des données à leur origine. C'est donc une question aussi bien technique qu'épistémologique que de savoir ce qu'apprend, modélise et produit un système d'IA.

1.1.1 Universalisme et données infalsifiables

Le recours à des systèmes automatiques a d'abord été basé sur le mythe selon lequel de tels outils, en explicitant l'ensemble des règles de décisions sur lesquels ils reposent, nous permettraient d'atteindre l'objectivité absolue, venant ainsi compenser la subjectivité humaine. Ce discours a émergé tout au long des développements technologiques et on le retrouvait déjà sur les avantages de l'introduction de la machinerie dans le domaine industriel : la machine permettait de produire plus, des pièces standardisées, avec moins de défauts. De simple outil, la machine est devenue une entité à laquelle l'être humain se compare, ou qui le complète : la machine est précise, puissante là où l'humain est intelligent et adaptatif. L'ensemble de ces développements a amené à l'idée commune que l'être humain serait faillible, car doué d'une sensibilité, là où la machine, basée sur des lois physiques et mathématiques, et dans le cadre des systèmes d'IA, sur des données, serait libre de toute interprétation subjective et de toute influence sociale. Cette conception de l'objectivité technologique va souvent de pair avec une appréhension positiviste des sciences dites "dures" ou "exactes" au nom d'une rigueur et d'une méthodologie qui seraient absentes des sciences dites "molles" ou, pour poursuivre notre analogie, "humaines". Si nous ne nous attarderons pas ici sur les débats philosophiques et épistémologiques sur ce qu'est une science, il est intéressant de noter que cette pensée a amené à une conception des données comme infalsifiables, contrairement aux hypothèses et aux théories dont la possibilité de falsification était un présupposé à la validité de la méthode scientifique (Popper, 1973). Les données seraient donc une observation objective du monde, porteuses d'une réalité absolue et universelle et donc dénuées de toute valeur sociale.

Cette position a ensuite été renforcée par l'émergence du *big data*. L'accès facilité à des grandes collections de données a permis de nombreuses découvertes auparavant non

1. `https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000035647959/`

observables, notamment du fait de « the capacity to collect and analyze data with an unprecedented breadth and depth and scale » (Lazer *et al.*, 2009, p.722). Ces progrès ont fait naître une nouvelle posture épistémologie selon laquelle les données en elles-mêmes seraient suffisantes à la découverte scientifique. On en retrouve l'exemple dans les propos provocateurs tenus par le directeur en chef du magazine Wired, Chris Anderson, dans un article intitulé *The End of Theory*. À propos de l'émergence du *Big Data*, qu'il nomme le *Petabyte Age*, il écrit :

> « This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do ? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. » [2]

Si l'abandon complet de la théorie n'a pas reçu l'adhésion de la communauté de recherche, il n'en reste pas moins que l'idée selon laquelle les données seraient suffisantes à produire du savoir, si rassemblées en quantité suffisante, est restée dans l'imaginaire collectif (Frické, 2015). On a vu apparaître de nombreuses méthodes dites *data-driven*, notamment dans le cadre de l'entreprise où l'on cherche à faire émerger une "culture de la donnée" car celle-ci serait "source de valeur" (Pingflow, 2019). Les décisions ne sont alors plus basées sur l'expérience et l'intuition des managers, mais sur une vaste collection de données pensées comme représentant une intelligence collective. En 2007, l'informaticien Jim Gray, ayant reçu le prix Turing en 1998, annonçait ainsi l'entrée de la science dans un nouveau paradigme, basé sur les données (Hey *et al.*, 2009). On peut en effet concevoir l'histoire des sciences comme découpée en paradigmes, à l'instar de la pensée de Thomas Kuhn ou d'Edgar Morin. Selon Thomas Kuhn (1962), chaque nouveau paradigme est porté par une rupture technologique ou théorique nécessitant de repenser la pratique de la science. Le premier paradigme scientifique était celui de la science empirique, qui décrivait les phénomènes naturels par l'observation. Puis le paradigme théorique est apparu, et avec lui la création de modèles et de généralisations, comme les lois de Kepler en astronomie ou de Newton en physique. Avec la complexification des modèles, nous avons basculé dans le troisième paradigme, celui de la computation, ouvrant la possibilité de simuler des systèmes complexes. Enfin, le quatrième paradigme, celui dans lequel nous serions aujourd'hui, unifie théorie, expérience et simulation grâce aux données et à leur exploration. Cette posture est adoptée par une partie de la communauté scientifique, comme en témoigne le commentaire de Hans Rosling, médecin et statisticien à l'origine de la Fondation Gapminder [3], dans une vidéo de 2010 :

2. `https://www.wired.com/2008/06/pb-theory/`
3. `https://www.gapminder.org` Dernière consultation le 22/06/21.

> « The data deluge [...] is leading us to an ever greater understanding of life on Earth and the Universe beyond. [...] it may fundamentally transform the process of scientific discovery. The more data there is the more discoveries can be made. [4] »

Avec l'émergence du Big Data, les données sont donc devenues le nouvel axiome de la production scientifique. Pour résoudre un problème, il suffirait de réunir suffisamment de données et de trouver l'algorithme pour traiter cette information. Cette conception des données comme infalsifiables est cependant questionnée par des chercheuses et chercheurs, comme David Bollier qui explique que :

> « As a large mass of raw information, Big Data is not self-explanatory. And yet the specific methodologies for interpreting the data are open to all sorts of philosophical debate. Can the data represent an 'objective truth' or is any interpretation necessarily biased by some subjective filter or the way that data is 'cleaned ?'. » (Bollier *et al.*, 2010, p. 13)

L'argument de David Bollier est de remettre la création et la collecte des données ainsi que leurs utilisations dans le contexte socio-culturel à leur origine. Ainsi les données ne contiendraient pas une réalité universelle, mais pourraient être la base d'un ensemble varié d'hypothèses scientifiques, selon les biais des chercheurs et chercheuses ou la manière dont la tâche a été envisagée.

1.1.2 Données situées et objectivité forte

Cette remise en cause de l'objectivité scientifique absolue, basée sur une interprétation unique des données a été l'objet de nombreux débats dans la philosophie des sciences et la sociologie de la connaissance.

Ludwig Fleck démontrait déjà en 1935, dans son ouvrage intitulé *Genèse et développement d'un fait scientifique*, que la production scientifique n'est pas indépendante du contexte historique et socio-culturel dans lequel elle s'inscrit (Fleck, 1934). Ouvrant ainsi le champ de la sociologie de la connaissance (Scheler, 1993), il a théorisé les liens existants entre cette production scientifique et les normes du "collectif de pensée" à son origine. Cette conception épistémologique a été reprise par Thomas Kuhn, qui démontre comment la notion de paradigme scientifique s'est peu à peu imposée dans les sciences dites "dures", présupposant qu'il existait un cadre théorique unique, jusqu'à ce que ce dernier soit remis en question puis remplacé avec la découverte d'un nouveau paradigme (Kuhn, 1963). Mais les paradigmes ne pouvant pas co-exister, il avance l'idée selon laquelle le choix du paradigme retenu ne peut être seulement fondé sur les données scientifiques et résulte donc d'influences externes (historiques, sociales, politiques) (Gingras, 2020, p. 87).

4. `http://www.gapminder.org/videos/the-joy-of-stats/`

Les épistémologies féministes reprennent la thèse selon laquelle le contexte de production du savoir influe sur la connaissance, thèse affinée notamment par les travaux de Sandra Harding puis de Donna Haraway. Dans sa théorie du point de vue ou *standpoint*, Sandra Harding défend la notion d'une "objectivité forte" qui ne serait non pas absolue et universelle, mais le résultat construit par la somme des travaux conduits par des subjectivités. Pour atteindre l'objectivité, il est donc nécessaire de rendre compte des déterminismes socio-historiques et de la perspective du chercheur ou de la chercheuse (Harding, 1992). Donna Haraway, quant à elle, propose la notion de "savoir situé", et défend la nécessité de reconnaitre les enjeux de pouvoir derrière le savoir et la production de la science. Elle rappelle que le discours scientifique est produit par "quelqu'un.e qui parle, de quelque part" et qu'il est important de rendre compte de cette perspective partielle. Elle s'oppose aux discours universalisants, produits d'un point de vue revendiqué comme universel mais qui correspond souvent à celui d'un homme, blanc, cisgenre et hétérosexuel (Haraway, 1988).

1.1.3 Des exemples en IA

Mais qu'en est-il alors de la question épistémologique en IA ? S'il ne sera pas question ici de retranscrire l'ensemble des débats et positions des chercheuses et chercheurs du domaine, force est de constater que de nombreux systèmes, entrainés sur de vastes quantités de données ont rendu explicite l'existence de ces postures situées. C'est d'ailleurs ce qu'écrit Cathy O'Neil dans son livre *Weapons of Math Destruction* où elle définit un algorithme comme « une opinion intégrée au programme », cette opinion étant une *doxa*, au sens platonicien, c'est-à-dire un système de représentation hautement social et situé, qui influe notre manière de penser et de percevoir le monde (O'Neil, 2016). Les données ne sont pas indépendantes des contextes dans lesquels elles ont été produites et de la manière dont elles ont été collectées. C'est d'ailleurs le propos de danah Boyd et Kate Crawford lorsqu'elles écrivent :

> « Furthermore, researchers must be able to account for the biases in their interpretation of the data. To do so requires recognizing that one's identity and perspective informs one's analysis (Behar & Gordon 1996). » (Boyd et Crawford, 2012, p. 668)

Cette représentation partielle et située des données et l'impact qu'elle peut avoir sur les systèmes se retrouve dans plusieurs exemples ces dernières années. L'enquête réalisée par l'ONG ProPublica en 2016, a observé que le système COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) développé par la société Equivant (anciennement Northpoint), et utilisé dans les cours de justice nord-américaines présentait des résultats discriminatoires envers les personnes noires. En effet, l'outil utilisé pour prédire le taux de récidive des personnes inculpées, sur-évaluait le risque de récidive des

personnes noires et sous-évaluait celui des personnes blanches, impactant directement leurs conditions de remise en liberté (Angwin *et al.*, 2016). Cette distribution de performances différente entre personnes blanches et noires était due à la sur-représentation des personnes noires dans les établissements pénitenciers nord-américains. Mais l'histoire de la criminalisation des personnes racisées, les disparités de patrouillage à travers les quartiers ou le traitement judiciaire qui est fait selon la couleur de peau du ou de la prévenue s'inscrivent dans une longue histoire de racisme complexe et que les données seules ne sont pas capables de mettre en lumière (Barlow, 2005). Le cas de l'étude COMPAS est cependant complexe, une contre-étude ayant d'ailleurs été publiée suite à l'article de ProPublica (Flores *et al.*, 2016). Pour autant, il a permis de mettre en avant le fait que certains systèmes, de par leurs données et les impensés sociaux qu'ils contiennent, contribuent à reproduire des subjectivités et n'atteignent pas l'objectivité impartiale et absolue attendue.

Le travail de Joy Buolamwini et Timnit Gebru dans leur étude *GenderShades* (2018) a lui aussi montré que les personnes noires et les femmes obtenaient des taux d'erreur bien plus importants que les personnes blanches et les hommes dans les modules de reconnaissance automatique du genre basée sur les visages. Le taux d'erreur était maximal pour les femmes noires, démontrant ainsi le fonctionnement intersectionnel de la discrimination. [5] N'ayant pas eu accès aux données d'apprentissage des différents systèmes sous licence commerciale, les chercheuses supposent que ces écarts de taux d'erreur s'expliquent par une représentation inégale de ces catégories. En adoptant un point de vue universaliste, dans lequel un visage est un visage, peu importe s'il appartient à une femme, un homme, une personne de couleur ou non, on passe à côté de tout un ensemble de paramètres pouvant influer le fonctionnement du système et reproduire, à l'insu de ses concepteurs et conceptrices, des discriminations déjà existantes. Il est important de souligner que malgré ces écarts de taux d'erreur, les performances globales du système restaient très bonnes. Ces problématiques se retrouvent dans les systèmes des grands groupes de la *data*, avec notamment la génération automatique de légendes d'images de Google [6] qui proposait de légender un selfie de personnes noires comme étant des "gorilles".

Le cas de l'algorithme de tri de CV d'Amazon nous offre un autre exemple : l'algorithme ne sélectionnait pas les CV contenant le mot "women" ou des mentions d'éta-

5. La notion d'*intersectionnalité* est empruntée à Kimberlé W. Crenshaw, qui l'a proposé pour démontrer l'incapacité de la théorie de l'identité à rendre compte de l'expérience des femmes de couleur, car l'expérience de celles-ci n'était pas prise en compte par les luttes féministes et antiracistes : « [...] je montre que les expériences des femmes de couleur sont souvent le produit des croisements du racisme et du sexisme, et qu'en règle générale elles ne sont pas plus prises en compte par le discours féministe que par le discours antiraciste. Du fait de leur identité intersectionnelle en tant que femmes et personnes de couleur, ces dernières ne peuvent généralement que constater la marginalisation de leurs intérêts et de leurs expériences dans les discours forgés pour répondre à l'une ou l'autre de ces dimensions (celle du genre et celle de la race) » (Crenshaw, 2005, p.54)

6. `https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app`. Dernière consultation le 02/06/2021.

blissements scolaires pour filles, car celles-ci n'étaient pas représentées dans les données d'apprentissage, l'informatique ayant été un domaine principalement masculin sur cette période. [7]

Le problème n'est donc pas le résultat d'un manque de compétences ou de moyens, puisque ces phénomènes se retrouvent même dans les grandes entreprises du domaine, mais bien la conséquence d'un manque dans les formalisations et la manière d'envisager les tâches modélisées. La pensée universalisante et la non-prise en compte des différences entre populations a conduit ces systèmes à ne pas bien fonctionner sur des catégories sociales, de race ou de genre, déjà minorisées dans la société. Il est important de rappeler que ces performances ne peuvent se limiter à l'expression de la subjectivité de ses concepteurs, mais sont le reflet des structures sociales qui traversent les données. Ainsi le cas de COMPAS s'explique par la non-prise en compte des liens entre pauvreté, criminalité et racisme. Le cas d'Amazon est quant à lui le reflet d'une division genrée du travail et notamment de la place de femmes dans l'informatique sur la période de recueil des données d'apprentissage. Ainsi la question qui se pose n'est pas tant d'ordre moral, à savoir si un système est "bon" ou "mauvais", que social à savoir à qui appartient la vérité et l'expérience qui est implémentée dans et par ces systèmes.

C'est d'ailleurs l'objet du travail de Solon Barocas et Andrew D. Selbst, qui mettent en relation les conséquences que peuvent avoir les impensés de l'utilisation du *big data* dans le développement des systèmes et le cadre législatif nord-américain, notamment la loi concernant l'interdiction de la discrimination à l'emploi :

> « Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makes, or simply reflect the widespread biases that persist in society. » (Barocas et Selbst, 2016)

Penser l'impact des technologies de l'intelligence artificielle doit donc se faire de manière située, dans un cadre de référence, qu'il soit légal ou social pour que les questions et les réponses qui y seront apportées soient elles-mêmes inscrites dans une perspective socio-historique pertinente.

1.2 Biais et discrimination

Ce questionnement sur le caractère socio-technique des systèmes et leurs impacts a donné lieu à tout un ensemble de discours sur les "biais de l'intelligence artificielle". Pour autant, le terme de biais recouvre des réalités parfois peu homogènes, ce qu'a souligné Kate Crawford dans son intervention à NeurIPS 2017. Le problème posé par ce flou terminologique réside en ce qu'il empêche de savoir précisément où se situe le problème et

7. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G` Dernière consultation le 02/06/2021.

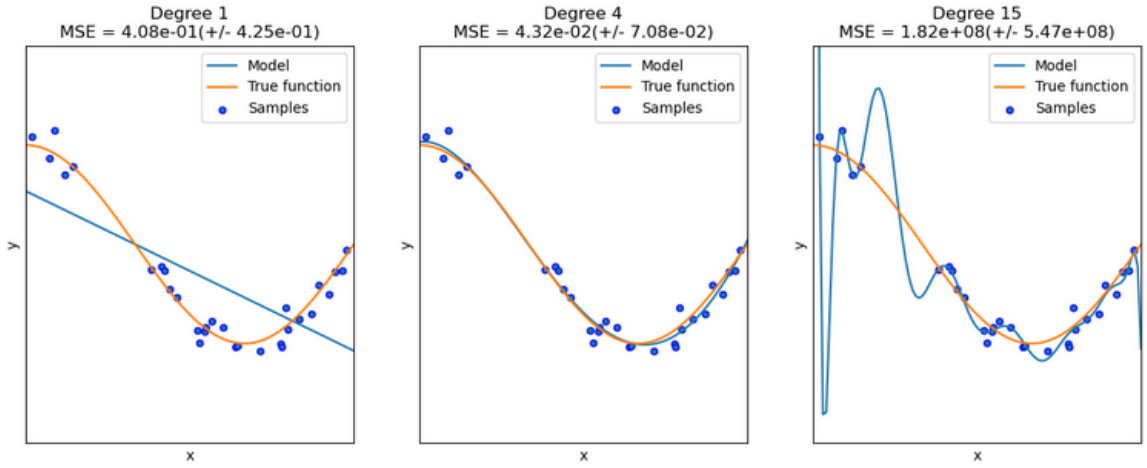

FIGURE 1 – Exemple de sous et de sur-apprentrissage. Tiré de (Pedregosa *et al.*, 2011)

donc, d'y apporter des réponses. Cette polysémie complique aussi les échanges entre communautés de recherche notamment entre celles de l'apprentissage machine et des autres domaines comme le droit ou la linguistique. Il s'agit donc ici de recenser les différentes acceptions du terme de biais, pour positionner notre questionnement et circonscrire les problèmes auxquels s'attaque ce travail de thèse.

1.2.1 Les notions de biais

1.2.1.1 Le biais statistique

Historiquement, la notion de biais a un sens technique en statistiques, où elle décrit les différences systématiques entre un échantillon et une population. Cette définition statistique, qu'on appelle "biais de sélection" ou "biais d'échantillonnage" se retrouve d'ailleurs à la base des notions de sur- et de sous-apprentissage des systèmes issus de l'apprentissage supervisé (cf. Figure 1) : dans le cas du sous-apprentissage, le modèle ne réussit pas à capturer la structure des données et aura un biais fort et une variance faible, là où dans le cas du sur-apprentissage, le système aura trop "collé" aux données et généralisera mal ; on aura donc un biais faible mais une variance forte (Crawford, 2017).

1.2.1.2 Les biais cognitifs

Outre les acceptions de biais dans son sens statistique originel, les biais au sens scientifique peuvent aussi désigner les biais cognitifs, étudiés principalement en psychologie et en sciences cognitives. Un biais est défini dans le Grand Larousse de la Psychologie comme « une distorsion (déviation systématique par rapport à une norme) que subit une information en entrant dans le système cognitif ou en sortant. Dans le premier cas, le sujet opère une sélection des informations, dans le second, il réalise une sélection des réponses. »

Les biais cognitifs sont nombreux et peuvent être des biais sensori-moteurs (ou illusions perceptives), mnésiques, de jugement, ou de raisonnement. Les biais sensori-moteurs sont largement connus, notamment les illusions d'optique comme l'illusion d'Ebbinghaus (cf. Figure 2) ou encore les illusions acoustiques comme la gamme de Shepard qui donne l'impression d'une gamme qui monte ou descend à l'infini.

Les biais mnésiques ont trait au fonctionnement de notre mémoire, avec par exemple l'effet de récence qui implique de mieux se rappeler des évènements ou informations récentes, ou de l'effet de primauté qui amène à mémoriser plus facilement les premiers éléments d'une liste.

Les biais de raisonnement vont altérer nos prises de décision. On pensera par exemple au biais de confirmation, qui implique de favoriser les informations corroborant notre hypothèse plutôt que celles qui la réfutent. Les stéréotypes, ou biais d'association, sont un des exemples les plus courants des biais de jugement et sont mesurés notamment par le test d'association implicite Greenwald *et al.* (1998). Le test se compose de plusieurs tâches de classification d'un ensemble d'items (liste de mots) entre deux concepts (les sciences et les lettres par exemple) et deux attributs (masculin et féminin). Dans notre exemple, la liste d'items pourrait être : mathématiques, géologie, etc. pour les sciences ; littérature, histoire, etc. pour les lettres ; homme, père, etc. pour le masculin ; femme, mère, etc. pour le féminin. Dans le cas où nous associerions inconsciemment un concept et un attribut, nous sommes plus rapides à résoudre la tâche de classification quand le concept et l'attribut requièrent la même réponse que quand ils nécessitent une réponse différente : par exemple, dans le cas des biais d'association masculin/science et féminin/lettres, on classera plus vite les domaines scientifiques tels que les mathématiques ou la physique, quand ceux-ci sont associés à la touche de réponse qui est également celle utilisée pour le masculin. On sera à l'inverse moins rapide si la touche de réponse est également celle utilisée pour le féminin. Ces biais d'association sont le reflet que nos capacités décisionnelles sont basées sur notre connaissance du monde et un ensemble d'attentes que nous avons construites au cours de notre vie. En linguistique, sur des tâches de transcriptions, il a été remarqué que selon les informations que l'on avait d'un locuteur ou d'une locutrice, des mêmes enregistrements ne seront pas transcrits de la même manière pour rester en adéquation avec les représentations des transcripteurs et transcriptrices. C'est ce que montrent Claire Blanche-Benveniste et Colette Jeanjean avec l'exemple de la transcription par Pierre Guiraud du /ske/ comme une réalisation de "ce que", construisant ainsi une particularité du français populaire (Guiraud, 1966), là où il pouvait aussi être envisageable que cette réalisation soit une prononciation rapide de "est-ce que" dans laquelle la voyelle du début serait tombée (Blanche-Benveniste et Jeanjean, 1986, p.144).

Les biais cognitifs se situent en dehors de notre niveau de conscience et à l'instar des illusions perceptives, comme les illusions d'optique ou encore l'effet McGurk (McGurk et MacDonald, 1976), ils sont donc indépendants de notre volonté et de nos capacités cog-

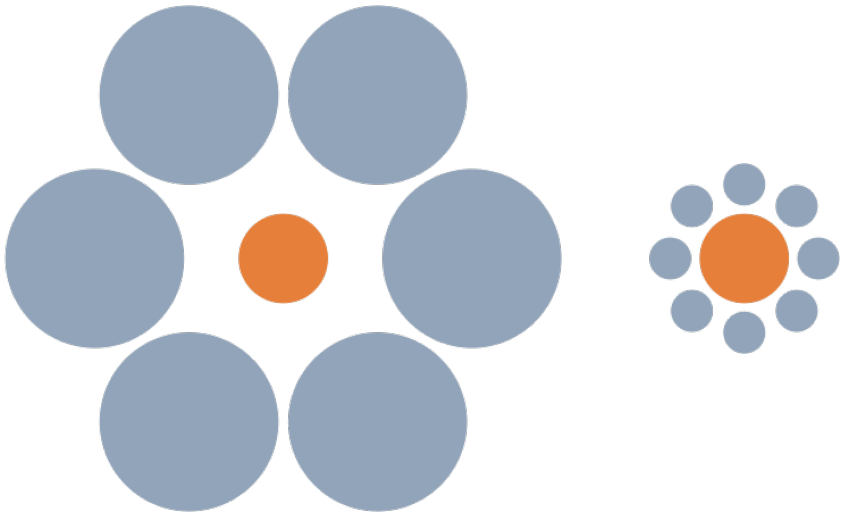

FIGURE 2 – Illusion d'Ebbinghaus. Les deux cercles orange sont de la même taille, mais la variation de taille entre les cercles gris qui les entoure nous fait percevoir le cercle orange de droite comme plus gros. Tiré de *The Illusions Index*.[8]

nitives. Même en étant conscients et conscientes des biais qui nous traversent, il nous est impossible d'agir directement dessus. Cependant à la différence des illusions perceptives, qui, elles, sont acceptées et reconnues comme des limites physiques, les stéréotypes ont un impact social fort, et on cherchera donc à en limiter les conséquences dans la prise de décision et/ou à changer les représentations pour s'adapter à un modèle de société. Si certaines représentations sont difficilement modifiables pour tout un chacun, elles ne sont pas pour autant acceptables comme telles dans la société.

1.2.1.3 Les biais en IA : définir une terminologie commune

Lorsqu'il s'agit des systèmes d'IA, les biais dont il est question sont à la fois statistiques et cognitifs, rendant parfois obscurs les débats sur le sujet quand les terminologies ne sont pas partagées. Dans le rapport réalisé en février 2019 par Telecom ParisTech et la fondation Abeona, les auteurs questionnent l'existence de biais cognitifs et statistiques au sein des systèmes (Bertail *et al.*, 2019). Les biais cognitifs jouent lors des étapes de la conception des outils, de la perception de l'objectif et du choix des variables et sont le produit de l'interaction entre les créateurs et créatrices d'outils avec la société. Les biais statistiques, quant à eux, ont trait aux données, notamment dans leur processus de sélection : dans le cas de l'étude *GenderShades* par exemple, le jeu de données utilisé avait été considéré comme représentatif alors qu'il contenait une majorité d'hommes et de personnes à la peau blanche. La présence de biais cognitifs ou statistiques est donc, dans la majorité des cas, le résultat d'impensés et de questions non-posées.

8. `https://www.illusionsindex.org/ir/ebbinghaus-illusion`

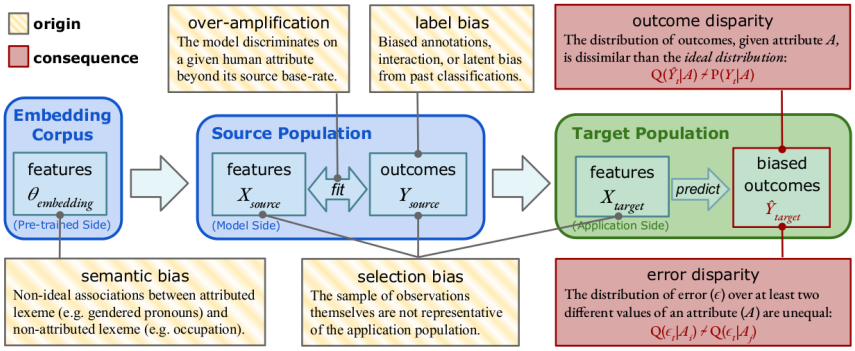

FIGURE 3 – *The Predictive Bias Framework for NLP* : Représentation de l'origine d'un biais dans un pipeline NLP standard supervisé par disparité. L'existence du biais est observable en ŷ, avec des mesures de disparité des résultats (*outcome disparity*) et/ou de disparité de l'erreur (*error disparity*). Tiré de Shah *et al.* (2020).

Concernant plus précisément les systèmes de TAL, Deven Shah, Andrew Schwartz et Dirk Hovy ont récemment proposé un cadre terminologique pour unifier les discussions autour des biais dans les systèmes (Shah *et al.*, 2020). Dans leur typologie, ils identifient les différents types de biais pouvant être introduits et les étapes de développement du système où ils sont susceptibles de survenir. Ils recensent 4 types de biais à savoir les biais d'annotation (*label bias*), les biais de sélection (*selection bias*), les biais de modèle (*over-amplification bias*) et les biais sémantiques (*semantic bias*).

— Le biais d'annotation survient lorsque l'annotateur ou l'annotatrice projette sa vision du monde sur les données. C'est le cas par exemple dans le travail de Marteen Sap *et al.* (2019), sur la détection automatique des discours de haine. Les auteurs ont observé comment l'ignorance des annotatrices et annotateurs concernant les variations dialectales de l'anglais afro-américain les a conduits à considérer comme relevant du discours de haine, certaines constructions typiques du dialecte. Ainsi une mauvaise annotation des données pour une certaine catégorie de population amène le système final à avoir un fonctionnement que l'on pourrait qualifier de "raciste". Les biais d'annotation peuvent être repérés dans les phases de vérification des données, mais la majorité du temps ils ne sont pas détectés avant l'observation de comportements inattendus lors du processus d'évaluation.

— Les biais de sélection quant à eux, correspondent dans ce cas à l'adéquation entre l'échantillon d'apprentissage et la population-test. Si l'écart est trop grand, alors la généralisation du système sera mauvaise et les performances en seront impactées.

— Le biais de modèle consiste en l'exagération de phénomènes fréquents. Le système va donc avoir tendance à amplifier des associations par exemple.

— Enfin le biais sémantique apparaît lorsque les données reflètent des biais d'association comme défini plus haut, à savoir des associations préférentielles qui sont basées

sur des stéréotypes et continuent à les entretenir. C'est notamment le cas soulevé dans les travaux de Tolga Bolukbasi *et al.* (2016) et de Nikhil Garg *et al.* (2018) qui observent comment les *word-embeddings* ou plongements de mots, largement utilisés dans les systèmes de TAL encapsulent tout un ensemble de stéréotypes genrés et ethniques.

Ces différents types de biais, pouvant survenir à différentes étapes du développement, conduisent à avoir un système prédictif biaisé. On parlera alors de biais prédictif qui se mesure via une disparité de traitement ou une disparité d'erreur (*outcome disparity* ou *error disparity*). Ces disparités sont donc des mesures d'écarts à une norme ou distribution attendue.

Il est intéressant de souligner cependant qu'un biais prédictif se compare toujours à une distribution idéale. Mais la définition de cet idéal attendu soulève encore d'autres questions. Est-ce une distribution reflétant la réalité ? Si tel est le cas, cette distribution ne sera donc pas forcément équitable entre les différentes catégories et potentiellement non satisfaisante d'un point de vue social (inégalités de genre, racisme, etc.)... Partant de là, une distribution idéale peut aussi être imaginée comme une distribution équitable sur un ensemble de critères considérés comme protégés, mais il devient alors nécessaire de formuler explicitement cette équité : sur quels critères est-elle mesurée ? et comment ? Aujourd'hui quand on parle de biais, on parle souvent de ces écarts face à la distribution idéale attendue, mais cet attendu reste vague, voire complètement impensé. Prenons l'exemple d'un système d'un assistant vocal comme SIRI, initialement développé par la DARPA pour alléger la charge cognitive des généraux de l'armée des États-Unis (Santolaria, 2016, p. 31). Sachant que les femmes représentent à peine 19% des officiers de l'armée nord-américaine [9], est-ce que l'on considèrera un système d'ASR plus performant sur les voix d'hommes comme adapté à la tâche ou comme "biaisé" ? La distribution idéale des performances doit-elle favoriser le plus grand nombre, à savoir ici les utilisateurs, ou doit-elle être équitablement distribuée entre hommes et femmes, quitte à diminuer les performances sur les hommes ?

D'une manière générale, il est nécessaire de comprendre que l'on ne peut pas parler de biais tout court. La notion de biais recouvre des phénomènes multiples et parfois très différents : les êtres humains ont un ensemble de biais cognitifs (de mémoire, d'associations, etc.), les données peuvent être biaisées du fait d'inégalités sociales résultant d'une construction historique (sexisme, racisme, classisme, homophobie, islamophobie, validisme, âgisme, etc.), les modèles peuvent être biaisés, du fait de leurs données ou de la modélisation statistique de biais cognitifs... Dans leur article s'intéressant aux biais dans les systèmes de TAL, Su Lin Blodgett *et al.* (2020) ont mis en avant ce flou autour de la notion de biais dans les travaux en TAL. Bien qu'ayant permis l'émergence d'un question-

9. 19,14% d'après le rapport de juillet 2021, disponible ici : `https://dwp.dmdc.osd.mil/dwp/app/dod-data-reports/workforce-reports`

nement vif autour de ces enjeux, ces travaux ne définissent que rarement le type de biais qu'ils adressent et l'impact de ces biais sur les utilisateurs et utilisatrices de ces systèmes. Savoir si des biais sont présents dans un système et de quel type de biais il s'agit constitue la première étape d'une réflexion nécessaire sur l'éthique de ces systèmes. La suppression des biais ou l'adaptation des systèmes à leur existence doit se faire au regard de l'usage du système. Le problème central aujourd'hui soulevé par l'existence de biais dans les systèmes d'IA est que ceux-ci peuvent conduire à des performances non-souhaitables voire discriminatoires. Or il semble exister une frontière invisible qui placerait d'un côté les questions techniques, la découverte et la remédiation des biais d'un côté, et les impacts et les questions sociales de l'autre. Si les travaux scientifiques en IA n'ont pas la réponse à ces questions sociales, ils doivent contribuer en revanche à fournir un maximum d'informations aidant à préciser les impacts des différentes prises de décision. Il est donc nécessaire de replacer les systèmes dans leurs usages et de questionner les conséquences et les impacts de l'existence de biais, notamment via une compréhension extensive de la notion de discrimination. Comme l'écrivent Bettina Berendt et Sören Preibusch :

> « [...] discrimination is not the existence of some statistical imbalance (e.g., more men than women have jobs in higher management). It is a property of a decision that may lead to such an imbalance in the population or disadvantage a specific individual (such as a women not gettint a job juste because of her gender. » (Berendt et Preibusch, 2012, p. 344)

1.2.2 Des biais aux discriminations

Dès lors que l'on s'intéresse à l'impact des systèmes et non plus seulement à leur développement, un basculement s'opère entre les notions de biais (ayant trait à la constitution du système et de ses données) et celle de discrimination, décrivant la prise de ces systèmes sur le monde. Bettina Berendt et Sören Preibusch proposent de distinguer les notions de différenciation et de discrimination. La différenciation est entendue comme l'adaptation d'une réponse ou d'un comportement à un individu en fonction d'un ensemble de ses caractéristiques propres (Berendt et Preibusch, 2014). On ne parlera pas forcément de la même manière à un enfant qu'à un adulte, à une personne maitrisant peu une langue qu'à une native. Ces processus cognitifs différenciés sont le gage de notre intelligence, dans le sens où nos actions sont fonction de notre environnement. Mais ce processus d'adaptation est aussi parfois lui-même nommé "processus de discrimination" lorsqu'on s'intéresse à l'acquisition de cette capacité de discernement :

> « Cette phase consiste à apprendre à distinguer les choses les unes des autres et donc, nécessairement, à les classifier, les catégoriser. Cela conduit à des opérations d'inclusion/exclusion de groupe, sur la base de certains critères qui distinguent le même du différent (les fruits des légumes, et au sein des

> fruits, les agrumes, etc.) –ce que Jean Piaget a été l'un des premiers à mettre en évidence dans ses travaux sur l'acquisition de la logique chez l'enfant. » (Fracchiolla et Sini, 2020, p.)

La définition du "processus de discrimination" par Béatrice Fracchiolla et Lorella Sini, décrit ici l'opération de catégorisation qui mène à la différenciation. Le terme discrimination a, dans leur conception, un sens premier, celui de cette opération de catégorisation, et un sens second où, par extension, cette catégorisation trouve une valeur sociale dans les discours qui l'entourent et qui s'avèrent, dans la majorité des cas, péjoratifs envers une catégorie donnée (minoritaire et/ou minorisée).

Aux yeux de la loi, ce caractère discriminatoire surgit lorsque la différence de traitement est effectuée sur la base de critères considérés comme brisant le contrat social. En France, la discrimination constitue un délit inscrit dans le Code Pénal (articles 225-1 à 225-4), mais également dans le Code du travail (art. L 1132-1). Les critères protégés sont au nombre de 23 dans la dernière version du texte, datant de 2016 ; parmi eux, on retrouve le sexe, l'identité de genre, l'âge, l'appartenance religieuse, ethnique ou raciale (vraie ou supposée), la situation économique, la situation de handicap, etc. Les notions de discrimination et de biais restent fortement culturelles et l'ancrage légal souligne qu'un biais ou une discrimination ne peut être perçue ainsi que dans un cadre socio-culturel donné. Un article de Sánchez-Monedero *et al.* (2020) s'intéressant aux systèmes automatiques pour l'embauche soulignait d'ailleurs que les travaux actuels sur les biais de ces systèmes sont majoritairement faits en considérant le cadre socio-légal des États-Unis.

Une manière de s'extraire du flou terminologique est de regarder les incidences concrètes de ces systèmes. Cette proposition faisait l'objet de la deuxième partie de la communication de Kate Crawford (2017), dans laquelle elle proposait de distinguer, non plus des biais ou des discriminations, mais des types de préjudices (*harms*). Sa typologie comporte deux types de préjudices à savoir, les *allocative harms*, qu'on pourrait traduire comme des préjudices d'attribution ou d'allocation, et les *representational harms*, ou préjudices de représentation ou d'image. Le préjudice d'attribution recouvre l'idée que les individus d'un groupe (de genre, de race, de classe, etc.) n'auraient pas accès aux mêmes ressources et/ou traitement que les autres. Le cas du tri de CV par exemple s'inscrit dans un préjudice d'attribution. Le préjudice de représentation, quant à lui, est plus difficilement quantifiable dans le sens où il recouvre tous les fonctionnements et résultats du système qui viendraient impacter négativement la manière dont est perçu un groupe, comme dans le cas de la légende automatique de Google. En ce sens, il est discriminatoire dans le cadre où il vient réactiver tout un ensemble de stéréotypes souvent issus de structures de domination et d'oppression présentes et/ou passées.

1.2.3 Les discriminations en TAL

Dans leur article de 2016, Dirk Hovy et Shannon Spruit ont mis en évidence l'articulation complexe entre données, systèmes de TAL et leurs différentes implications sociales. Le premier point des auteurs est que la majorité des travaux portant sur les impacts des systèmes de TAL se sont d'abord intéressés à des questions éthiques en lien avec le respect de la vie privée et la gestion des données. Ces considérations ayant émergé avec la généralisation du *big data* ont d'ailleurs donné lieu en 2018 au Réglement Général sur la Protection des Données (RGPD) ; appliqué au niveau européen. Mais outre ces considérations légales, Dirk Hovy et Shannon Spruit questionnent les impacts directs des systèmes sur les utilisateurs et utilisatrices notamment à travers les notions d'*exclusion*, de *surgénéralisation* et d'*exposition*. En écho avec le concept de savoir situé de Donna Haraway, ils rappellent que l'objet du TAL, le langage est également situé, dans un espace géographique (Bamman *et al.*, 2014a), mais également un contexte social. La non-prise en compte de cet aspect situé du langage peut conduire à un biais démographique, excluant certaines populations de ces technologies :

> « For instance, standard language technology may be easier to use for white males from California (as these are taken into account while developing it) rather than women or citizens of Latino or Arabic descent. This will reinforce already existing demographic differences, and makes technology less user friendly for such groups, cf. authors like Bourdieu and Passeron (1990) have shown how restricted language, like class specific language or scientific jargon, can hinder the expression of outsiders' voices from certain practices. » (Hovy et Spruit, 2016, p. 593)

Si le biais démographique peut-être atténué de manière algorithmique et relève plutôt du préjudice d'attribution, le phénomène de surexposition (*overexposure*) est le résultat du design de la recherche et des décisions prises par les chercheurs et chercheuses. En effet, le focus fait sur certains groupes de la population peut conduire à des préjudices représentationnels, renforçant des stéréotypes ou contribuant à stigmatiser des populations :

> « If research repeatedly found that the language of a certain demographic group was harder to process, it could create a situation where this group was perceived to be difficult, or abnormal, especially in the presence of existing biases. » (Hovy et Spruit, 2016, p. 594)

Ce travail s'inscrit donc dans un ensemble de problèmes soulevés par l'utilisation grandissante des systèmes de TAL qui appelle la communauté de recherche à réfléchir sur sa méthodologie pour en réduire les impacts négatifs.

1.3 Élaboration de recommandations

1.3.1 Une prise de conscience récente

Historiquement, les travaux concernant l'éthique des systèmes se sont d'abord focalisés sur les aspects relatifs au caractère privé des grandes masses de données. Les principes FAIR Data, proposés par Wilkinson *et al.* (2016) sont des principes pour la gestion des données scientifiques, basés sur quatre caractéristiques fondamentales des données que sont la repérabilité (*foundability*), l'accessibilité (*accessibility*), l'interopérabilité (*interoperability*) et la réutilisabilité (*reusability*). Ces discussions et recommandations veillaient principalement à poser un cadre méthodologique fort pour chercher à atteindre une certaine transparence dans les résultats. Une autre discussion sur la description des données au sein de la communauté du TAL a été initiée par Couillault *et al.* (2014), qui ont proposé une Charte sur l'éthique et les *big data* (*Ethics and Big Data Charter*). Au fur et à mesure que des biais de performance étaient constatés dans les systèmes, une des solutions poursuivies a été de proposer une solution technique à un problème initialement considéré comme lui-même technique. Plusieurs techniques de "débiaisement" (*debiasing*) ont été proposées : c'est le cas notamment du travail de Tolga Bolukbasi *et al.* (2016), de Jieyu Zhao *et al.* (2018) et d'Amanda Bower *et al.* (2018).

1.3.2 Les techniques de débiaisements

L'étude de Tolga Bolukbasi *et al.* (2016) a été largement médiatisée à sa publication. En effet, les auteurs ont travaillé sur les plongements de mots ou *word-embeddings* qui sont une modélisation vectorielle du vocabulaire. Ils sont utilisés comme une forme de dictionnaire et des mots proches sémantiquement auront tendance à avoir des vecteurs proches également. L'intérêt des *word-embeddings* réside également dans le fait qu'ils permettent de modéliser certaines relations entre les mots, notamment via des opérations arithmétiques traduisant des analogies. Par exemple si l'on considère l'équivalence suivante :

$$\overrightarrow{Tokyo} - \overrightarrow{Japon} \approx \overrightarrow{Paris} - \overrightarrow{France}$$

On peut retrouver x = France, en faisant une analogie du type "Tokyo est au Japon ce que Paris est à x". Or ils se sont aperçus que la modélisation des données textuelles à travers les plongements de mots (en anglais dans leur travail) avaient tendance à reproduire des associations sexistes et racistes, comme montré également par Nikhil Garg *et al.* (2018). Le titre de leur article "Man is to Computer Programmer as Woman is to Homemaker ? Debiaising Word Embedding" présentait un résultat d'analogie particulièrement frappant de modélisations du sexisme dans leurs données. Ces représentations étant utilisées dans la résolution de nombreuses tâches de TAL, les auteurs proposent

deux méthodes de débiaisement : ils identifient d'abord le sous-espace vectoriel capturant le biais (celui du genre dans le cas de leur exemple) puis ils proposent soient de neutraliser la valeur du vecteur sur cette direction (i.e. leur assigner une valeur de 0 sur cet axe) ou de les "égaliser", à savoir de rendre équidistants : après une telle opération, le mot "babysitting" devrait être équidistant de "grand-mère" et "grand-père" ainsi que de "fille" et "garçon".

Jieyu Zhao *et al.* (2018) travaillent quant à eux sur de la reconnaissance visuelle. Or ils observent que dans leur corpus d'apprentissage, les données ayant trait à la cuisine ont 33% de chance de plus de faire apparaître des femmes que des hommes, et que cette disparité monte à 68% dans les résultats obtenus sur les données de test. La solution qu'ils proposent est une technique de débiaisement appelée RBA (*Reducing Bias Amplification*), consistant à ajouter des contraintes au modèle prédictif pour que ses sorties reflètent la distribution désirée. Dans les deux tâches adressées par leur travail (*Visual Semantic Role Labeling et Multilabel Classification*), leur méthode permet de réduire les biais statistiques observés.

1.3.3 La question de l'équité

Mais évaluer et réduire les biais d'un système nécessite de pouvoir quantifier une mesure d'équité. La notion de *fairness* a été largement utilisée par la communauté, mais comme écrit par Alexandra Chouldechova : « It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. » (2017, p. 154). La notion d'équité appelle à l'éthique et à un aspect moral, qui peut grandement varier selon les positions éthiques adoptées, qu'elles relèvent d'une posture déontologique ou conséquentialiste, par exemple.

Dans leur article de 2019, Deirde Mulligan *et al.*, montraient comment le flou autour du terme *fairness* avait contribué à ralentir les débats, notamment selon les acceptions qui en sont faites dans chaque discipline (Mulligan *et al.*, 2019).

Dans le cadre de ses travaux sur la justice sociale, François Dubet (2014) distinguait l'égalité des places de l'égalité des chances. L'égalité des places vise à réduire les différences entre des individus appartenant à différents groupes définis par notre variable protégée (le genre, l'âge, la race sociale, etc.), là où l'égalité des chances vise à donner à des individus somme toutes égaux par ailleurs les mêmes chances d'accéder à un service, un système, etc.

Ramenés à un niveau algorithmique, ces questionnements impliquent que les définitions mathématiques de l'équité seront dépendantes de la compréhension qui est faite de la notion d'équité et de l'objectif visé. De nombreuses définitions ont alors été proposées notamment par Pedreshi *et al.* (2008); Dwork *et al.* (2012); Barocas et Selbst (2016); Hardt *et al.* (2016); Kilbertus *et al.* (2017). Dans une étude de 2020 qui s'intéressait aux

différentes acceptions de la *fairness* dans l'IA, Eirini Ntoutsi *et al.* (2020) dénombrent plus de 20 définitions mathématiques différentes.

En écho avec les travaux sociologiques de François Dubet (2014), Le Chen *et al.* (2018) distinguent deux types d'équités, à savoir l'équité de groupe (égalité des places) et l'équité individuelle (égalité des chances). L'équité individuelle, posant l'hypothèse qu'à individus équivalents (hormis la valeur de la variable protégée), résultat équivalent, se teste à l'aide de modèles linéaires mixtes. L'équité de groupe, quant à elle, supposant une similarité des distributions dans les différents groupes définis par la variable protégée, se mesure avec des tests statistiques comme le test U de Mann-Whitney. La question est donc de savoir si nous considérons que les données à propos d'un candidat ou d'une candidate sont le reflet pertinent de ses capacités auquel cas à compétences égales (hormis la variable protégée) on s'attend à un traitement égal, ou si ces données sont influencées par des inégalités structurelles dans nos sociétés, auquel cas on cherchera à rendre similaires les distributions au sein de chaque groupe structurel (de genre, de race, etc.)

Dans le rapport d'AI Now (Whittaker *et al.*, 2018), sont ajoutés à ces deux stratégies, nommées ici respectivement *anti-classification strategies* et *classification parity*, le concept de *calibration strategies* qui vise plutôt à rendre le processus décisionnel équitable à l'aide de post-processing et le concept d'*observational fairness strategies* qui regroupe toutes les notions précitées ainsi que l'analyse de données dans une perspective d'audit.

Les travaux et références sur la notion de *fairness* sont donc nombreux et les différentes définitions parfois inconciliables. Suite à l'apparition de ces multiples techniques pour "débiaiser" les systèmes, d'autres études se sont intéressées à démontrer que ces solutions techniques n'étaient souvent pas suffisantes : c'est le cas notamment d'Hila Gonen et Yoav Goldberg (2019), qui ont travaillé sur les méthodes pour supprimer les biais présents dans les *word-embeddings*. Elles observaient que la définition de biais entendue par les auteurs, les avait fait s'orienter vers des méthodes de débiaisement qui ne débarrassaient pas les représentations vectorielles de la majorité de l'information stéréotypique des associations, mais simplement neutralisaient la position des mots sur l'axe du genre. Au final, la géométrie globale ne s'en trouvait que peu modifiée, et les techniques relevaient donc plutôt d'une solution superficielle, ne prenant en compte qu'une définition du biais très spécifique. Flavien Prost *et al.* (2019) ont, eux, montré que les techniques de débiaisement pouvaient au contraire renforcer les performances discriminatoires du système.

La question se pose également de savoir si les attributs protégés doivent être pris en compte dans le processus global : en effet, selon les approches, on souhaite rendre les systèmes aveugles à la variable protégée, mais la connaissance de celle-ci est également nécessaire pour vérifier si un modèle est équitable ou dans certains cas, pour ajuster les performances en fonction des différents groupes. Dans leur rapport sur l'utilisation de systèmes automatiques dans les processus de recrutement, Miranda Bogen et Aaron Rieke (2018) soulignent également le fait que ces prises en compte sont souvent focalisées sur une

variable protégée unique et ne permettent pas de rendre compte de l'intersectionnalité des discriminations rencontrées par les individus. Elles se limitent souvent aux variables effectivement rendues explicites dans les données, principalement le genre et la "race", alors que l'on sait que les discriminations à l'embauche peuvent aussi se baser sur des attributs comme l'orientation sexuelle ou la situation de handicap des candidats et candidates.

Étudier, mettre à jour et corriger les biais des systèmes est ainsi devenu un "hot topic" de la recherche en apprentissage machine. Depuis 2018, ACM organise la conférence FAccT (anciennement ACM FAT*) pour *Fairness, Accountability and Transparency* qui a pour objectif d'envisager les nouveaux systèmes issus de l'apprentissage machine comme profondément socio-techniques. La création d'organismes de recherche comme le AI Now Institute [10] ou l'Algorithmic Justice League [11] et d'ONG comme le Disruption Network Lab [12], pour médiatiser et alimenter le débat sur les interactions entre technologie, société et politique se sont multipliées. Concernant particulièrement le TAL, ACL a créé en 2017 un atelier intitulé *Workshop on Ethics in Natural Language Processing*, réédité en 2018. Dans le contexte français, l'atelier EteRNAL a vu le jour en 2015, puis a connu une seconde édition en 2020. Le discours émergent sur les biais des systèmes est venu s'inscrire dans un champ de travaux plus large sur "l'éthique" des systèmes de TAL et plus largement d'IA, et a donné lieu à tout un ensemble de travaux proposant solutions et recommandations pour réduire les biais des systèmes ou en contrôler l'impact.

1.3.4 Situer les données

Il devient alors évident que ces questions ne sont pas uniquement des problèmes techniques et ne peuvent se satisfaire d'une solution purement mathématique. Une autre approche du problème des biais a été de prendre le parti d'une description extensive des modèles et des données pour en comprendre le fonctionnement et les mécanismes. Emily Bender et Batya Friedman (2018) ont proposé la notion de *data statement* pour garantir la transparence des données, en définissant un ensemble de méta-données à renseigner pour chaque collection de données utilisées dans le développement de systèmes de TAL. En analogie avec le domaine de l'industrie électronique, dans laquelle chaque composant est accompagné d'une fiche technique détaillant ses caractéristiques, les tests effectués ainsi que l'usage défini, Timnit Gebru *et al.* (2018) proposent une méthodologie de standardisation des corpus avec leurs *datasheets for datasets*. En associant de manière systématique, un ensemble de méta-données prédéfinies, l'idée est de permettre une plus grande transparence sur le contenu de ces corpus (ce qu'ils contiennent, ce qu'ils recouvrent, comment ils peuvent être utilisés), mais également ce qu'ils peuvent encapsuler de social (pourquoi ont-ils été créés ? comment ? quelles questions éthiques peuvent-ils soulever ?) ainsi que

10. `https://ainowinstitute.org/`
11. `https://www.ajl.org/`
12. `https://www.disruptionlab.org/about-the-lab`

des informations plus logistiques comme des informations pour la maintenance de ces ensembles de données. En convoquant d'autres exemples industriels comme l'arrivée de l'automobile, les essais cliniques ou encore l'ingénierie électronique, les autrices et auteurs argumentent pour la nécessité d'une standardisation et d'un cadre de référence dans la communauté de l'apprentissage machine. Ces questions sont d'autant plus cruciales dans un mouvement de démocratisation de l'IA, où des recettes-modèles sont rendues disponibles *off-the-shelf* et peuvent donc être entrainées par des personnes qui ne seraient pas expertes du domaine ou du phénomène modélisé. Margaret Mitchell *et al.* (2019) proposent la notion de *model cards*, qui se veut complémentaire à la proposition de Timnit Gebru *et al.*, et qui appelle à rendre compte des performances des modèles sur différents groupes "culturels, démographiques ou phénotypiques".

1.4 Questionner le pouvoir

Replacer les systèmes issus de l'apprentissage automatique dans leur contexte socioculturel, questionner les notions de biais, de discrimination et d'équité, ainsi que les solutions proposées pour y remédier marque la nécessité de la construction d'un discours interdisciplinaire sur ces nouveaux systèmes. Dans leur article de 2020, Su Lin Blodgett *et al.* (2020) montrent les limites des travaux réalisés jusqu'à maintenant sur les biais dans les systèmes de TAL. Recensant 146 articles s'intéressant aux biais, les auteurs et autrices montrent que les travaux réalisés jusqu'à présent, bien qu'ayant contribué à l'émergence des questionnements sur l'impact des systèmes de TAL, la majorité d'entre eux ne propose pas de position critique ou ne formalise pas explicitement les préjudices et risques encourus via l'utilisation de systèmes "biaisés". Dans le cadre du TAL, ils et elles rappellent que langage et hiérarchie sociale s'articulent dans la production des discours et que questionner les outils de traitement automatique de la langue doit être fait conjointement à une critique des hiérarchies qu'ils cristallisent. Ce lien entre technologie, normes et pouvoir est d'ailleurs l'objet de l'ouvrage de Catherine D'Ignazio et Lauren Klein (2020). En soulignant comment les données contribuent à renforcer un système établi, elles mettent en opposition "l'éthique des données" avec la "justice des données". Leur réflexion rejoint l'épistémologie féministe introduite en 1.1 et déplace la source du problème de l'individu ou du système aux différentiels de pouvoir et aux structures hiérarchiques. À travers la notion de "matrice de domination", empruntée à Patricia Hill Collins (2002) qui décrit comment les systèmes de pouvoir fonctionnent et sont vécus, elles proposent un basculement de typologie en ne considérant non plus les biais mais les oppressions, non plus le concept de *fairness* mais celui d'*equity* et de sortir de l'idéal de la transparence pour questionner notre réflexivité. C'est ce que fait Sasha Costanza-Chock (2018) dans son travail en montrant comment les algorithmes de sécurité des aéroports oppriment les personnes trans ne rentrant pas dans les mesures standards des corps considérés dans la matrice

binaire homme/femme. Dès lors, s'intéresser à la voix des hommes et des femmes et les technologies qui s'y rapportent, requiert dans un premier temps, de comprendre comment le concept de genre articule des structures de pouvoir et comment ces structures s'implémentent dans les systèmes de reconnaissance automatique de la parole.

Conclusion

Cette première partie a cherché à dresser un rapide portrait de la question des "biais" en IA. Après nous être intéressée à l'épistémologie positiviste et universaliste qui sous-tend toute une partie de la recherche en sciences dites "dures" ou "exactes", nous avons essayé de montré que d'autres postures épistémologies, et notamment les épistémologies féministes, posent la question de la représentativité des systèmes, dans le sens où les expériences et les réalités qu'ils modélisent sont les expériences et les réalités d'un groupe social déterminé, à savoir celui d'un homme blanc occidental ayant eu accès aux études supérieures. Ces réflexions épistémologiques ont été remises à l'ordre du jour notamment lorsque le public ainsi que des chercheurs et chercheuses du domaine se sont rendues compte que le postulat universaliste n'était pas directement modélisable par la technique, et que les systèmes technologiques faisaient apparaître des profils d'utilisateurs et d'utilisatrices différentes, en fonction de catégories sociales comme le genre ou la race, donnant lieu à la naissance d'un discours sur les "biais des systèmes".

Nous nous sommes ensuite appliquée à définir le terme de "biais" en montrant sa grande polysémie, et en quoi ses différentes acceptions pouvaient parfois empêcher une prise en compte exhaustive des phénomènes de disparité de performances des systèmes d'IA. Pour sortir de l'impasse terminologique, nous nous sommes également intéressée à la notion de discrimination, en ce qu'elle permet de réinscrire les "biais" et notamment le biais prédictif, dans un rapport au monde et à la société, notamment sur la base du cadre légal de la discrimination. Enfin, nous avons également regardé comment, ces discours sur les "biais" et les discriminations ont émergé dans la communauté du TAL ainsi que les réactions qu'ils ont suscitées. Une tentative a été de fournir des solutions techniques aux problèmes causés par la technique, mais celle-ci se retrouve souvent confronté à ses limites. Une autre a été de produire des bonnes pratiques, concernant principalement la description des données et des systèmes, pour réaffirmer une conception socio-technique des systèmes, permettant de questionner plus largement les dynamiques de pouvoir dans lesquelles ces outils s'inscrivent.

Ainsi nous avons précisé la première partie de notre question *Existe-t-il des biais de genre dans les systèmes d'ASR ?* Les prochains chapitres s'intéresseront donc à creuser les deux autres concepts clés de notre question de recherche, à savoir le genre (Chapitre 2) et les systèmes d'ASR (Chapitre 3).

Chapitre 2

Le genre comme thématique de recherche, débat social et facteur de discrimination

Nous avons montré dans le chapitre précédent les liens forts entre les technologies et les points de vue socio-culturels des personnes de la conception de ces systèmes. En nous inscrivant dans un questionnement sur l'articulation de ces liens avec l'impact de ces systèmes, nous nous proposons d'étudier particulièrement l'existence de discriminations de genre dans les performances de ces systèmes, et de voir où se situent les racines des discriminations : à savoir donc, si elles sont le résultat de disparités sociales et de différentiels de pouvoirs, l'expression "d'angles morts" techniques ou une interaction conjointe de processus cognitifs et de biais statistiques. Mais pour pouvoir parler de différenciation ou de discrimination de genre, encore faut-il définir la conception du genre que nous adopterons ici.

Dans ce chapitre, nous procéderons d'abord à un ancrage de la notion de genre dans son contexte historique expliquant ainsi notre utilisation de la terminologie du genre et non de celle du sexe, pour ensuite nous intéresser à son expression dans la voix. Nous mettrons également en lumière la longue histoire d'inadéquation entre les femmes, particulièrement leur voix, et la technologie, pour inscrire les discriminations de genre observées actuellement dans les technologies du traitement automatique des langues dans une perspective historique et structurelle.

2.1 Histoire d'un concept

2.1.1 Construction de l'opposition sexe/genre : du 'donné' au 'faire'

Le concept de *gender* a d'abord été introduit par le psychiatre John Money en 1955. Travaillant sur l'intersexuation et inspiré par l'anthropologue Margaret Mead, il propose le concept de *gender role* pour nommer l'écart existant entre le rôle social et le sexe biologique d'une personne :

> « by the term, gender role, we mean all those things that a person says or does to disclose himself or herself as having the status of boy or man, girl or woman, respectively. » (Money *et al.*, 1955)

.

Robert Stoller, psychiatre et psychanalyste, utilise quant à lui la notion de *gender identity* correspondant au sentiment d'une personne d'appartenir au sexe féminin ou masculin, indépendamment de son sexe biologique (Stoller, 1964). Ces conceptions du genre en opposition au sexe, fond écho à la dualité nature/culture qui a été à la base du développement de l'anthropologie, et que l'on retrouve dans les travaux de Claude Levi-Strauss [1], mais également de théories théologiques et philosophiques opposant le corps à l'esprit. Le sexe relèverait donc de la "nature" et le genre de la "culture", les deux concepts étant reliés par une relation de correspondance stricte. Cette approche behavioriste, loin de questionner les hiérarchies produites par le système de genre comme le feront par la suite les travaux féministes, consistait donc plus à permettre la réassignation aux catégories normatives homme/femme, des personnes s'échappant de la matrice du genre : à savoir les personnes intersexes et trans. En agissant sur la "nature" de l'homme ou de la femme, à savoir son corps, on le réaligne avec son "esprit", son identité d'homme ou de femme.

Construit comme objet social, le genre a été ensuite mobilisé dans les études ethnométhodologiques [2], notamment celle du sociologue Harold Garfinkel, qui s'intéressait, à partir du cas d'Agnès, une jeune femme trans, aux pratiques employées par les personnes pour réaliser leur identité féminine ou masculine (Garfinkel, 1967). S'intéresser aux pratiques conscientes d'Agnès pour articuler un "être femme" et "passer pour" a permis de rendre visibles les pratiques jusqu'ici naturalisées de l'ensemble des hommes et femmes pour se rendre intelligibles comme telles. Cette conception du genre comme une pratique en interaction est également celle de Suzanne Kessler et Wendy McKenna (1978), qui axeront leur travail sur les processus d'attribution du genre. Il n'est alors plus seulement une pratique de l'individu voulant s'inscrire dans une identité féminine ou masculine, mais le résultat d'un processus interactionnel dans lequel les différentes pratiques genrées produisent des indices qui sont ensuite interprétés pour permettre la catégorisation de la personne en tant qu'homme ou femme par ses interlocuteurs et interlocutrices. L'apport des travaux de Su-

1. On peut notamment citer *Les Structures Elementaires de la Parenté* : « Tout ce qui est universel chez l'homme relève de l'ordre de la nature et se caractérise par la spontanéité, tout ce qui est astreint à une norme appartient à la culture et présente les attributs du relatif et du particulier. » (Levi-Strauss, 1967, p. 10). Cette opposition nature/culture sera par la suite dépassée, notamment avec les travaux de Philippe Descola (2005).

2. L'ethnométhologie, est définit par Jérome Mbiatong comme les : « ... méthodes que les membres d'un groupe utilisent pour donner sens et en même temps accomplir leurs actions de tous les jours, communiquer, prendre des décisions, raisonner. L'ethno-méthodologie cherche à comprendre le monde social tel qu'il est perçu par ceux qui y vivent. Par conséquent, son but est de mettre au jour les procédures qui régissent la «construction sociale de la réalité» par les individus. C'est à ces procédures – les processus interprétatifs de la vie ordinaire – apprises dans la vie courante que renvoie le terme «ethnométhodes» » (Mbiatong, 2019, p.219)

zanne McKenna et Wendy Kessler est de rendre visible les mécanismes de catégorisation du genre comme étant le résultat conjoint d'un faire de l'individu (s'habiller, parler, socialiser comme un homme/une femme) mais également d'une perception. Ce qui implique qu'une fois la catégorisation genrée faite, des indices incohérents (comme un homme avec une voix aiguë ou une femme avec une forte pilosité) ne seront plus suffisamment saillants pour remettre en question le genre de la personne, mais seront à l'inverse intégrés comme des "variations", qui pourront néanmoins être sources d'étonnement ou de moquerie.

Que le genre constitue un donné ou une pratique de l'individu, il est entendu dans les approches psychiatriques et ethnométhodologiques comme une propriété de l'individu et n'a pas encore la signification politique qu'il prendra par la suite.

2.1.2 Genre et rapports de pouvoir

Le développement du concept du genre doit également beaucoup au travail des féministes. La sociologue Ann Oakley, qui s'est inspirée des travaux de John Money (Money *et al.*, 1955) et Robert Stoller (1964), pose également cette distinction du sexe comme terme biologique, là où le genre relève du culturel et recouvre donc des caractéristiques sociales (Oakley, 1972). Le genre devient alors un outil critique de déconstruction de la naturalisation du sexe, à l'instar de la célèbre formule de Simone de Beauvoir : « on ne naît pas femme, on le devient ». Comme indiqué en prologue du volume historiographique des féminismes français de Bibia Pavard, Florence Rochefort et Michelle Zancarini-Fournel « Faire bouger les lignes des rôles attribués à chaque sexe revient à bouleverser l'ordre social et à interroger les inégalités de classe et de race. » (Pavard *et al.*, 2020, p.5). L'idée ici est de remettre en question les attitudes, les valeurs et les comportements associées à chaque sexe et instituées en normes de la féminité et de la masculinité. Dès lors, les femmes ne seraient plus "naturellement" disposées à rester à la maison du fait de caractéristiques biologiques et physiologiques, mais bien du fait d'un choix d'organisation sexuée de la société : le patriarcat. [3]

C'est ce que l'historienne Joan Scott mettra en avant dans un article fondateur, dans lequel elle définit le genre comme catégorie analytique nécessaire aux recherches en histoire :

> « My definition of gender has two parts and several subsets. They are interrelated but must be analytically distinct. The core of the definition rests on an integral connection between two propositions : gender is a constitutive element of social relationships based on perceived differences between the sexes, and

3. Redéfini par Christine Delphy, le terme de patriarcat désigne une structure sociale hiérarchique et inégalitaire, basée sur des différences essentialisées entre hommes et femmes, mais également sur les idéologies et les discours. Le terme trouve sa source dans le féminisme matérialiste et cherche à rendre compte des pratiques sociales matérielles de la domination patriarcale sur les femmes tout en insistant sur le fait qu'elles ne sont pas simplement une conséquence du capitalisme (voir l'entrée "Patriarcat" du Dictionnaire critique du féminisme (Hirata *et al.*, 2000, pp. 141-146)

> gender is a primary way of signifying relationships of power. Changes in the organization of social relationships always correspond to changes in representations of power, but the direction of change is not necessarily one way. As a constitutive element of social relationships based on perceived differences between the sexes, gender involves four elements : first, culturally available symbols that evoke multiple (and often contradictory) representations – Eve and Mary as symbols of woman, for example, in the Western Christian tradition – but also, myths of light and dark, purification and pollution, innocence and corruption. [...] Second, normative concepts that set forth interpretations of the meanings of the symbols, that attempt to limit and contain their metaphoric possibilities. These concepts are expressed in religious, educational, scientific, legal, and political doctrines and typically take the form of a fixed binary opposition, categorically and unequivocally asserting the meaning of male and female, masculine and feminine. [...] The point of new historical investigation is to disrupt the notion of fixity, to discover the nature of the debate or repression that leads to the appearance of timeless permanence in binary gender representation. This kind of analysis must include a notion of politics as well as reference to social institutions and organisation – the third aspect of gender relationships. [...] The fourth aspect of gender is subjective identity. [...] Historians need instead to examine the ways in which gendered identities are substantively constructed and relate their findings to a range of activities, social organizations, and historically specific cultural representations. » (Scott, 1986, pp. 1067-1068)

L'apport de Joann Scott dans ce travail définitoire est de mettre en avant que le genre est non seulement l'articulation de différents éléments, à savoir un marquage de la différence des sexes à travers la construction d'un système symbolique binaire, la présence de normes portant sur l'interprétation de cette symbolique, l'application de ces normes à travers les institutions et la construction des identités genrées, mais surtout un moyen de construire et de définir le pouvoir. En ce sens, le genre, à l'instar de la race et de la classe sociale façonne une réalité sociale construite autour de rapports de domination entre chaque catégorie binaire : homme/femme, blanc/noir, propriétaires/prolétaires donnant l'illusion d'une classification manichéenne simple, bien loin de la complexité des situations réelles.

> « [...] "man" and "woman" are at once empty and overflowing categories. Empty because they have no ultimate, transcendent meaning. Overflowing because even when they appear to be fixed, they still contain within them alternative, denied, or suppressed definitions. » (Scott, 1986, p. 1074)

Postuler la production d'identités sexuées et la naturalisation d'un "être homme" et d'un "être femme" comme bases de rapports de pouvoir constituera un argument clé des discours féministes s'appuyant largement sur des ouvrages de référence de l'anthropologie féministe anglo-saxonne (Rosaldo *et al.*, 1974; Rubin, 1975). L'assignation des femmes à la sphère privée, leur interdiction de travailler et d'avoir accès à l'éducation, de disposer de leur corps, de leur argent sont alors formulées comme des combats politiques à mener. Cette seconde vague de pensée autour du concept de genre, scientifique, mais également militante, ne vient plus simplement mettre en avant la composante comportementale et culturelle de l'identité sexuée, mais également questionner la structure sociale qu'elle façonne et les normes qu'elle impose.

En France, l'utilisation du concept de genre est beaucoup plus tardive. Le terme est connu des chercheuses, l'article de Joan Scott étant traduit en 1988, mais des différences épistémiques et théoriques ralentissent son adoption. Depuis les années 1960, en sociologie, l'épistémologie dominante est l'épistémologie marxiste qui analyse ces rapports de pouvoirs en terme de classes de sexes. La notion de classe de sexe sera théorisée entre autres par Christine Delphy et Danièle Kergoat, qui tout en s'inspirant du marxisme, ont montré la spécificité des rapports hommes/femmes par rapport à la hiérarchie entre bourgeoisie et prolétariat. En anthropologie, Nicole-Claude Mathieu utilisera elle les termes de "sexe social" et "système social des sexes" (Mathieu, 1991). Cependant ces approches ne remettent pas en cause, dans un premier temps, la question de la naturalité des catégories sexuées. La sociologue Colette Guillaumin, en critiquant l'utilisation de ces catégories vues comme anhistoriques, publie *Sexe, Race et pratiques du pouvoir* (1992) qui postule que « dans toute société il y a des faits matériels qui sont la conséquence de rapports sociaux de pouvoir et des faits idéologiques qui sont les formes mentales que prennent ces rapports. » (Parini, 2010). Elle utilisera dans ses écrits le terme de "sexage" pour décrire le processus de socialisation des femmes. La conceptualisation du genre en France doit également beaucoup aux historiennes, qui bien qu'ayant longtemps débattu sur le choix des termes ("Histoire des femmes" ou "Histoire du genre"), ont largement embrassé le concept, et ont été à l'origine d'un projet historique dont « la question majeure devient celle du rapport entre les sexes, compris comme un rapport social qui est à la fois effet et moteur de l'histoire, qui fonctionne à tous les niveaux de réalité et de représentations et dont on peut comprendre les rouages et marquer les spécificités selon les systèmes historiques » (Thébaud, 2005, p. 63), à l'instar des travaux de Michèle Perrot ou de Françoise Thébaud.

2.1.3 Théories queer et postmoderne du genre

Il est difficile de parler de la notion de genre sans évoquer les travaux de Judith Butler. Dans ses ouvrages, la philosophe critique la distinction sexe/genre proposée jusqu'alors :

elle argumente que le corps ne peut pas être interprété et perçu en dehors du contexte social signifiant qui le produit et que donc le sexe est déjà une interprétation faite à l'aune du genre (Butler, 1990). Elle formule ainsi une théorie queer du genre, mobilisant la notion de performativité. En s'inspirant de la théorie des actes de langage d'Austin (1975), Butler suppose que le genre s'apparente aux énoncés performatifs qui réalisent l'action qu'ils dénomment. Le genre est alors perçu non plus comme un seul système de production de différentiations sexuées, mais comme des réalisations individuelles de performance genrée s'inscrivant dans une "citationnalité générale" [4] (Derrida, 1972; Butler, 1997). Il devient ainsi discursif, créé, produit et maintenu par le discours et le langage. Ainsi, ce sont les actions répétées des différentes personnes en interaction qui construisent et rendent intelligible le genre, et ces actions s'inscrivent dans une intertextualité [5] qui contribue à cristalliser des normes. On bascule alors dans une approche constructiviste [6] du genre, plus focalisée sur la réalisation des identités que sur les rapports de pouvoir à l'origine du genre.

Cette théorie queer du genre permet néanmoins, dès lors que le genre et ses catégories ne sont pas fixes et peuvent évoluer, de remettre en cause les catégories binaires du genre, les normes de la masculinité et de la féminité qui y sont associées, ainsi que l'apparente normalité hétérosexuelle. En France, cependant, les théories queer sont arrivées tardi-

4. Contrairement à Austin dont l'approche pragmatique ne concevait le performatif inscrit en langue, la notion de citationnalité générale pose que « chaque mot d'un énoncé performatif est itérable, c'est-à-dire non seulement répétable, mais structurellement accidentel, sans lien indéfectible avec son contexte de production (ou d'énonciation). » (Cotton, 2016, p. 7). Derrida y adjoint alors la conception foucaldienne selon laquelle tout discours est complètement historicisé et construit aussi hors de son contexte énonciatif. Dès lors le genre s'inscrit dans une histoire d'itération de comportements et d'actes de langage appelant et maintenant les représentations des catégories binaires. Judith Butler ira plus loin en mettant également en avant le potentiel subversif créé par cette citationnalité, en jouant justement sur l'écart entre l'attendu (l'intertexte) et la production de l'individu dans une performance de genre.

5. Pour une définition approfondie de la notion de Julie Kristeva, voir l'article de Roland Barthes sur la *Théorie du texte* dans l'Encyclopaedia Universalis : « tout texte est un intertexte ; d'autres textes sont présents en lui à des niveaux variables, sous des formes plus ou moins reconnaissables : les textes de la culture antérieure et ceux de la culture environnante ; tout texte est un tissu nouveau de citations révolues. » (Barthes, 1973).

6. Le constructivisme conçoit la connaissance comme construite par l'interaction entre l'individu et le monde extérieur, ce qui construit une "réalité". En sociologie, le constructivisme social, introduit par Peter Berger et Thomas Luckmann (1966) et s'inspirant notamment des travaux d'Emile Durkheim, postule que la réalité est à la fois construite par subjectivement par l'individu et par le biais des institutions. En ce qui concerne le genre, l'approche constructiviste souligne donc le poids des institutions et de la société dans la création de normes de genre, par la suite adoptées comme une réalité. On peut faire le parallèle avec la définition de société donnée par Emile Durkheim et reprise dans le *Vocabulaire technique et critique de la philosophie* d'André Lalande : « les sociétés humaines présentent un phénomène nouveau, d'une nature spéciale, qui consiste en ce que certaines manières d'agir sont imposées ou du moins proposées du dehors à l'individu et se surajoutent à sa nature propre ; tel est le caractère des « institutions » (au sens large du mot), que rend possible l'existence du langage, et dont le langage est lui-même un exemple. Les sociétés humaines présentent un phénomène nouveau, qui consiste en ce que certaines manières d'agir sont imposées ou du moins proposées du dehors à l'individu et se surajoutent à sa nature propre, tel est le caractère des institutions. » (Lalande, 1926, p.1002). Le système de genre serait donc en ce sens, une institution.

vement notamment car comme l'écrivent Bibia Pavard, Florence Rochefort et Michelle Zancarini-Fournel :

> « ... la *queer theory* est marquée au départ par un relatif désintérêt de la part des féministes françaises. À cela plusieurs explications : Judith Butler accorde une place importante au corps et à la sexualité, ce qui n'a pas toujours constitué des questions centrales pour les théories féministes françaises. Le recours à la psychanalyse et la *French theory* du *French feminism* heurtent par ailleurs les féministes matérialistes. » (Pavard *et al.*, 2020, p.388)

notamment car :

> « L'accent mis sur les discours et les pouvoirs dans le sillage de Michel Foucault et Jacques Lacan, peut contribuer à s'écarter des réalités concrètes, celles de l'oppression et de la domination qui concernent tout autant la classe, la race, l'âge, etc. et des structures sociales inégalitaires et hiérarchiques. » (Pavard *et al.*, 2020, p.388)

La théorie queer, marquée par les travaux de Judith Butler mais également de Sam Bourcier en France, est à l'origine de la pensée non-binaire du genre amenant de plus en plus de personnes à s'identifier en dehors de ses catégories hégémoniques d'homme et de femme. Comme l'écrit Noémie Marignier :

> « Dans un cadre queer il ne s'agit plus de parler du *genre* mais des *genres* : les identités sont multiples (femme, homme, trans'*, intersexe, agenre, etc.) tout en étant produites et opprimées par le même système de genre. Ce système produit donc des normes identitaires binaires (masculines et féminines), et celles-ci peuvent être confirmées, reproduites mais aussi déjouées et subverties. En ce sens, *genres* peut être utilisé pour désigner les identités, c'est-à-dire les rôles sociaux sexués qui sont performés par les individus. » (Marignier, 2016, p. 38)

2.1.4 Le genre en France

Que sa conception soit ethnométhodologique ou constructiviste, le genre permet de rendre visible tout un ensemble de pratiques sociales jusqu'ici masquées par un discours essentialisant et naturalisant, rendant possible une critique politique des hiérarchies qu'il contribue à produire et reproduire. Mais les différents développements que le genre a connus ont contribué à en faire un terme flou et polysémique. Comme l'explique Eric Fassin (2008), le genre est un terme anglo-saxon, on lui préférait en France celui de "rapports sociaux de sexe" (Kergoat, 2005) ou de "système social des sexes" (Mathieu, 1991).

La réticence à l'adoption de la terminologie du genre vient autant de différents théoriques entre les féministes françaises et anglo-saxonnes [7] que d'une impossibilité de la notion à trouver un ancrage institutionnel. Comme l'écrivent Alexandre Duchêne et Claudine Moïse (2011) :

> « Cet ancrage institutionnel était impensable en raison de l'idéologie portée, dans une vision universaliste, par la nation et par la langue française, langue du citoyen, langue homogène qui se veut unificatrice au-delà des différences, ethniques ou sexuées. » (Duchêne et Moïse, 2011, p. 1)

Il a cependant fini par être adopté par une partie de la communauté scientifique au début des années 2000, bien que son utilisation soit encore aujourd'hui soumise à débat. [8]

En dehors des débats universitaires, la terminologie du genre a cependant fini par s'imposer en France, notamment autour des questions des droits des personnes LGBTI. Ainsi depuis 2016, l'identité de genre a été ajoutée dans les facteurs de discriminations listés par le Code Pénal. Dans les discours publics et médiatiques, l'identité de genre fait donc souvent référence aux personnes trans et moins au système hiérarchisé autour de catégories (homme/femme) auxquelles sont associées un ensemble de valeurs, d'attitudes et de comportements.

Dans ce travail, nous faisons le choix de parler de genre au sens large, recouvrant ainsi, l'identité de genre des personnes, à savoir le discours qu'elles portent sur elles-mêmes mais également leur pratique et leurs caractéristiques vocales, tout en reliant ces performances individuelles au système normatif dans lequel elles s'inscrivent, contribuant ainsi au maintien des catégories du genre. Nous nous intéresserons donc aux catégories existantes lors de nos recueils de données, à savoir principalement les catégories homme/femme, mais n'excluons pas non plus l'émergence de nouvelles catégories hors de cette binarité. L'idée ici est de mettre en lumière l'existence de cette catégorisation et de voir ce qui contribue à la définir et/ou la redéfinir, comment les personnes s'y retrouvent ou cherchent à en sortir, et quel rôle jouent ou peuvent jouer les technologies de la parole dans ce processus. Pour reprendre les mots de Joan Scott :

> « To pursue meaning, we need to deal with the individual subject as well as social organization and to articulate the nature of their interrelationships, for both are crucial to understanding how gender works, how change occurs. Finally, we need to replace the notion that social power is unified, coherent, and centralized with something like Foucault's concept of power as dispersed constellations of unequal relationships, discursively constituted in social "fields of force." Within the processes and structures there is room for a concept of

7. La philosophe Geneviève Fraisse, par exemple, reprocha au genre de réactualiser la dichotomie nature/culture en opposant le genre (social) au sexe (biologique) (Fraisse *et al.*, 2008).

8. Pour une perspective historique et politique sur l'adoption du concept, voir Fassin (2008) et Parini (2010).

> human agency as the attempt (at least partially rational) to construct an identity, a life, a set of relationships, a society with certain limits and with language – conceptual language that at once set boundaries and contains the possibility for negation, resistance, reinterpretation, the play of metaphoric invention and imagination. » (Scott, 1986, pp. 1067)

2.2 Voix et genre

Le genre se pense donc à la fois comme système et comme catégorie identitaire. En ce sens, la voix comme marqueur de l'identité est un lieu où le genre se "fait entendre" mais également où il se questionne. Après avoir présenté comment s'articulent les concepts de voix et d'identité, nous nous intéresserons plus particulièrement à l'expression du genre dans la voix, aux discours biologisants qui ont contribué à essentialiser une différence des sexes mais également aux stratégies mises en place par les locuteurs et locutrices pour faire varier les indices du genre dans leur voix.

2.2.1 La voix comme marqueur d'identité

Comme l'écrivent Carmen Llamas et Dominic Watt dans leur ouvrage intitué *Language and Identities* :

> « The connection between language and identity is a fundamental element of our experience of being human. Language not only reflects who we are but in some sense it *is* who we are, and its use defines us both directly and indirectly. We use language in a direct way to denote and describe who a person is through use of naming, kinship terms, description based on appearance, behaviour, background, and so on, and we use language to assign identities indirectly when we base our judgments of who people are on the way they speak. Language-mediated attribution of identity to individuals is so ingrained in human social affairs that we consider a person lacking a name to also lack an identity. » (Llamas et Watt, 2009, p.1)

La voix étant le résultat d'un processus biomécanique (passage de l'air dans le tractus vocal, mouvement des articulateurs) et chaque individu étant physiquement différent, il a semblé logique de considérer que l'unicité de l'individu impliquait l'unicité de la voix : la voix étant conçue comme le résultat d'une anatomie, des corps différents créent des voix différentes. La notion d'empreinte vocale a été proposée par Charles Grey et George Kopp (1944) du Bell Laboratories, qui voyaient dans les spectrogrammes une possibilité d'identifier des espions ennemis à partir d'extraits de communications radio ou téléphonique. Le terme a été popularisé ensuite par Lawrence Kersta, en 1962, qui affirmait que les différences intra-locuteur étaient négligeables face aux différences entre

individus et qu'il serait donc possible d'identifier une personne à partir d'enregistrements de sa voix (Kersta, 1962). Cependant la voix n'est pas une empreinte au sens strict, elle résulte de mouvements et est en cela plus proche de l'écriture. La non-pertinence de cette métaphore a donc été démontrée quelques années plus tard par Richard Bolt *et al.* (1973), qui ont mis en avant les différents facteurs pouvant contribuer à une més-identification (conditions d'enregistrements réelles et non en laboratoire, bruit de l'environnement, état émotionnel du locuteur ou de la locutrice, reconnaissance en milieu fermé ou non, etc.). Si cette conception de la voix comme identificateur unique de l'individu a été rapidement infirmée par la communauté scientifique, il n'en reste pas moins que cette idée a continué à perdurer dans le grand public, notamment du fait de sa relative simplicité et de son apparente logique. Déconstruire cette conception erronée a été l'objectif de nombreux chercheurs et chercheuses, avec notamment le long travail de Jean-François Bonastre, Louis-Jean Boë et de l'AFCP, qui a largement contribué à questionner l'utilisation de la reconnaissance du locuteur dans le cadre judiciaire, réouvrant un dialogue entre la Justice et la sphère académique (Bonastre *et al.*, 2003; Bonastre, 2020). Cette vision déterministe de la voix comme un produit d'une anatomie est également questionnée à travers d'autres travaux : dans le cas de la parole de jumeaux ou de jumelles où les caractéristiques physiques et physiologiques sont aussi identiques que possible, plusieurs études ont montré que les personnes restaient capables, au-delà du niveau du hasard, de différencier leur voix comme appartenant à des individus différents (Nolan et Oh, 1996; Johnson et Azara, 2000).

Mais si la voix n'est pas un facteur d'identification stricte, elle reste cependant un canal d'information important concernant l'identité d'une personne : on projette facilement le genre d'une personne, parfois son âge, voire sa taille, ainsi que son origine ou sa classe sociale à partir de sa voix (Krauss *et al.*, 2002). Ainsi, dans une étude Thomas Shipp et Harry Hollien (1969), les auteurs observent que les personnes sont capables d'estimer, à 5 ans près, l'âge d'un locuteur ou d'une locutrice à partir d'une phrase. Ellen Ryan et Harry Capadano (1978) ont quant à eux rapporté des corrélations entre l'âge réel et l'âge deviné de 0,93 pour les femmes et de 0,88 pour les hommes, après avoir présenté deux phrases à des auditeurs et auditrices. Mais cette perception n'est pas toujours exacte, Gary Neiman et J.A. Applegate (1990) ont observé un âge deviné entre 14 et 31 ans pour des locuteurs et locutrices âgées de 70 à 75 ans. Concernant le genre, une étude a montré que les enfants sont capables, dès l'âge de 6 mois de distinguer une voix d'homme d'une voix de femme (Miller *et al.*, 1982). Les enfants âgés de 6 à 9 ans présentent le même taux de précision dans la reconnaissance du genre que les adultes, d'après Suzanne Benett et Luisa Montero-Diaz (1982). Les adultes sont capables de discriminer des voix d'hommes de voix de femmes à partir d'une seule fricative avec une précision de 72.5% dans l'étude de Frances Ingemann (1968) et 91.5% dans celle de Martin Schwartz (1968). Norman Lass *et al.* ont montré que la robustesse de l'identification du genre avec différents

stimuli obtenant des taux de reconnaissance toujours supérieurs à 91% (1976; 1979; 1980), et même 99% pour des phrases passées en sens inverse. D'une manière générale, quand l'extrait proposé pour la tâche de reconnaissance est plus long qu'une syllabe, on obtient pratiquement toujours un taux de reconnaissance près de 100% (Kreiman et Sidtis, 2011, p. 131). Il arrive bien sûr que l'on se trompe, mais néanmoins les indices vocaux sont la plupart du temps suffisamment robustes pour que ces associations perdurent dans nos imaginaires. La voix est donc une manière pour les personnes de projeter une identité, leurs « physical, psychological, and social characteristics » selon John Laver (Laver, 1980, p.2) ou leur "auditory face" selon Pascal Belin *et al.* (2004).

Ces capacités à reconnaitre le genre, l'âge ou la race d'une personne à partir de sa voix ont contribué à légitimer une conception de l'identité comme un donné inné ou une appartenance fixe à une catégorie sociale, là où elle résulte en fait de processus interactionnels et évolutifs. Nos capacités de perception évoluent au rythme de la société mais permettent de maintenir un effet de fixité dans un cadre social mouvant. Ainsi les attributs vus comme typiquement féminins ou masculins ont une histoire, mais les catégories en tant que telles semblent anhistoriques. On peut prendre pour exemple les différentes interprétations symboliques de la couleur rose ou de la barbe (et plus largement de la pilosité) dont les significations ont évolué à travers le temps (Auzépy et Cornette, 2017; Pastoureau et Simonnet, 2007).

Au-delà de ce donné biologique, la voix indexe également tout un ensemble de traits sociaux, de caractère, ce qui motive, par exemple, le choix de comédiens et comédiennes pour le doublage au cinéma ou le travail de la voix au théâtre. Le choix des voix de synthèse aujourd'hui utilisées pour les assistants vocaux fait aussi appel à ces associations pour produire une image de marque : on cherchera à rendre Siri sympathique, Cortana serviable, ou encore Alexa drôle.[9] Il n'est pas anodin non plus que la majorité des voix par défaut de ces assistants soient en fait des voix d'assistantes. La capacité de la voix à indexer une personnalité est rendue encore plus visible par ces assistants décorporéisés (ou du moins non-humains) qui acquièrent quand même des caractéristiques et une identité sociale, dont les exemples les plus flagrants sont les personnages de Samantha dans *Her* ou HAL dans *2001 : l'Odyssée de l'Espace*.

La voix a donc d'abord été envisagée comme un donné biologique, déterminée par la forme du tractus vocal des personnes et cette vision positiviste a finalement laissé place à une conception socio-biologique de la voix, dans laquelle les individus composent sur une base biologique et anatomique pour indexer leur identité sociale, à un niveau macro, dans des catégories démographiques larges (race, classe, genre) mais également locales, en se positionnant ponctuellement dans les interactions, (en tant qu'expert, supérieur hiérarchique, amie, etc.). La voix, et plus largement le langage, permettent donc de faire

9. On peut à ce sujet se référer à François Perea (2018).

émerger ces identités multiples, tout en contribuant à redéfinir ces catégories identitaires. Comme l'écrit le linguiste John Joseph :

> « Researchers have been analysing how people's choice of languages, and ways of speaking, do not simply *reflect* who they are, but *make* them who they are – or more precisely, allow them to make themselves. In turn, languages they use are made and re-made in the process. » (Joseph, 2009, p.9)

Cette approche interactionnelle de l'identité est également celle défendue par Mary Bucholtz et Kira Hall :

> « We argue for the analytic value of approaching identity as a relational and socio-cultural phenomenon that emerges and circulates in local discourse contexts of interaction rather than a stable structure located primarily in the individual psyche or fixed social categories » (Bucholtz et Hall, 2009, p. 18)

Selon elles, l'identité dans le langage s'articule autour de 5 principes que sont : l'émergence, la positionnalité, l'indexicalité, la relationnalité et la partialité (respectivement *emergence principle*, *positionnality principle*, *indexicality principle*, *relationnality principle* et *partiality principle*), que l'on peut développer comme suit :

- Le principe d'émergence pose que, au lieu d'être un donné statique ou un mécanisme psychologique d'auto-catégorisation, l'identité émerge à travers l'action et l'interaction sociale, et notamment à travers le langage.
- Le principe de positionnalité renvoie aux différentes granularités d'appartenance qui articule l'identité : à savoir les catégories démographiques telles que le genre et la race, mais également des groupes reflétant des positions culturelles et sociales plus locales ainsi que des rôles et *stances* temporaires, propre à une interaction donnée. Une personne se positionne ainsi sur plusieurs ou la totalité de ces trois niveaux aux tailles et aux temporalités différentes.
- Le principe d'indexicalité est peut-être le plus pertinent lorsque l'on s'intéresse à la voix : comme le pose Elinor Ochs dans ses travaux, l'indexicalité décrit le fait que certaines formes linguistiques renvoient à des représentations sociales à travers la construction de liens sémiotiques (Ochs, 1992). Ces liens sémiotiques ne sont pour autant pas univoques et une même caractéristique acoustique peut, selon le contexte, renvoyer à des significations complètement différentes.
- Le principe de relationnalité constitue le coeur de l'apport théorique des autrices. En effet, elles mettent en avant une conception de l'identité comme un phénomène relationnel s'articulant autour de différents axes : la similarité/différence, adéquation/distinction et l'authenticité/dénaturalisation. Ces relations, qu'elles appellent également *tactiques d'intersubjectivité* appuient l'idée selon laquelle les identités ne sont jamais autonomes mais existent et acquièrent un sens au regard des autres positions identitaires possibles.

— Enfin le principe de partialité renvoie à la conception de savoir situé discuté en section 1.1, et pose donc que « identities are constituted by context and are themselves asserted as partial accounts » (Visweswaran, 1994, p. 41).

Si ces apports théoriques sont difficilement transposables directement aux technologies de la parole, ils ont l'avantage de mettre en avant le caractère évolutif de l'identité comme phénomène complexe et interactionnel, s'inscrivant à la fois dans plusieurs temporalités et s'exprimant à travers tout un ensemble de moyens.

> « The interactional view that we take here as the added benefit of undoing the false dichotomy between structure and agency that has long plagued the social theory [...]. On the one hand, it is only through discursive interaction that large-scale social structures come into being, on the other hand, even the most mundane of everyday conversations are impinged upon by ideological and material constructs that produce relations of power. Thus both structure and agency are intertwined as components of micro as well as macro articulations of identity » (Bucholtz et Hall, 2009, p. 26)

L'identité ne se joue donc pas à un seul niveau analytique, mais se code via de multiples canaux et granularités. Et c'est l'interaction complexe de tous ces processus qui la crée. Dès lors, l'identité n'est pas que le corps, ni uniquement la voix, mais celle-ci contribue à la définir. De même, le genre constitue une catégorie identitaire mais n'échappe pas à cette inter-relationnalité qui en fait une catégorie évolutive. Pour reprendre les mots de Marie-Cécile Bertau :

> « So, "voice" is a vocal-auditory event, and it is a concept belonging to a certain socioculturally constructed way of expression. The uttered voice is absolutely individual , coming from a unique body, but this body is located in specific sociocultural contexts and has a history of action, movements, labels, etc. So, the voice too. As for every human expression, the voice is individual and societal, both aspects being the facets of a wholeness... » (Bertau, 2008, pp.101-102)

2.2.2 De la physiologie à l'acoustique : conception biologique de la différence des sexes

Parler de genre en parole peut s'avérer compliqué dès lors que la majorité de la littérature en phonétique s'articule uniquement autour de la variable du sexe. Le fait que la majorité des études se soit intéressée au sexe et non au genre, reflète, comme le souligne Aron Arnold dans son travail de thèse :

> « comment les identités, le lien entre identité et corps, et le lien entre identités et langage sont conçus en phonétique. On peut généralement trouver dans

> ce domaine une vision positiviste et déterministe des identités : elles sont considérées comme allant de soi, aproblématiques, statiques, biologiquement déterminées (p. ex. le sexe, la race) ou socialement déterminées (p. ex. la classe sociale). Le genre, ou plus souvent le sexe, est pensé comme une variable binaire, qui détermine la forme des voix et les manières de produire de la parole. Les questions d'ordre épistémologique ne sont par ailleurs que rarement abordées en phonétique et la manière dont des idéologies de genre influencent la construction des savoirs n'est donc jamais thématisée. Il s'ensuit un ensemble de problèmes dans les façons d'aborder, représenter et décrire le genre et le corps. » (Arnold, 2015, p. 62)

Par la suite, nous utiliserons donc les termes de sexe et de genre, selon la terminologie utilisée par les auteurs et autrices, bien que nous fassions référence tout au long de notre propos au système de catégorisation initialement basée sur des différences sexuées et résultant en un ensemble de normes et de pratiques stéréotypées de la masculinité et de la féminité.

Un des paramètres acoustiques les plus saillants concernant le genre est la fréquence fondamentale (fréquence de vibration des plis vocaux) qui est le corrélat acoustique de la hauteur de la voix et est attribué à une différence de taille du larynx et de la longueur des cordes vocales qui diffèrent d'environ 60% (Kahane, 1978; Hirano *et al.*, 1983; Titze, 1989). Chez les hommes, la longueur des plis vocaux varie entre 17 et 25 mm et entre 13 et 20 mm pour les femmes (Ormezzano, 2000; Schuster *et al.*, 2005). Toutes ces différences ont une répercussion directe sur la fréquence fondamentale (F0), d'environ 1 octave : on se retrouve donc avec des valeurs moyennes de F0 à 130Hz pour les hommes et 220Hz pour les femmes pour l'anglais (Peterson et Barney, 1952; Hollien et Paul, 1969; Horii, 1975) et de 120Hz pour les hommes et 240Hz pour les femmes pour le français (Vaissière, 2006, p. 51). D'autres auteurs ont rapporté une différence dans les patterns de vibrations des plis vocaux (Titze, 1989). James Hillenbrand & Michael Clark (2009) ont montré que cette différence était suffisante pour que les auditeurs distinguent correctement voix d'hommes et voix de femmes anglophones dans 96% des cas.

Mais cette construction de catégories strictes et excluantes des voix d'hommes et de femmes du fait de morphologies distinctes ne correspond pas à une réalité mesurée. Comme l'écrivait Hermann Künzel :

> « We can at times event get the sex of the speaker wrong, which is less surprising than it sounds given that the ranges of average fundamental frequency (the physical correlate of pitch) for men's and women's voices overlap to a considerable degree. » (Künzel, 1989)

D'autres paramètres ont été présentés comme jouant un rôle dans l'identification du genre des locuteurs et locutrices tels que que l'intonation, les fréquences de résonance

(Arnold, 2012) ou les zones de bruits des fricatives sourdes (Fox et Nissen, 2005). Les fréquences de résonances, corrélat acoustique du timbre des voyelles sont souvent citées comme paramètre discriminant pour la perception du genre dans la voix. Elles dépendent de la taille du tractus vocal mais également de l'utilisation des articulateurs par le locuteur ou la locutrice. Le tractus vocal est en moyenne 15% plus court chez les femmes que chez les hommes (Fitch et Giedd, 1999). Cette différence est principalement due à une différence de taille du pharynx, celui-ci étant plus grand chez les hommes du fait de l'abaissement global du larynx à la puberté (Vorperian et Kent, 2007).

Le collectif Calliope a proposé des valeurs de références pour les différents formants des voyelles du français, avec des valeurs différentes pour les hommes et les femmes (Calliope, 1989, p. 84). L'écart entre les valeurs de références pour les hommes et les femmes ne s'exprime pas par une constante et le type de voyelle semble également jouer, questionnant le rôle de la différence anatomique. Simpson propose une hypothèse alternative liée à la différence de taille de cavité buccale : les bouches des femmes étant en moyenne plus petites, leur langue a une plus courte distance à parcourir que les hommes pour atteindre les mêmes lieux d'articulation (Simpson, 2002). Cette différence de distance induirait une différence de vitesse qui impacterait la précision des gestes articulatoires, mais il ne vérifie pas cette hypothèse dans une expérience ultérieure (Simpson, 2009).

Il existe donc des différences physiologiques et anatomiques sur lesquelles ont été construits des discours visant à préserver l'idée qu'il existe un dimorphisme sexuel exclusif et minimisant la réalité du chevauchement entre les distributions des morphologies féminines et masculines et des fréquences associées. Mais il n'existe pas de consensus et ce qui relève de la pratique articulatoire socialement construite et/ou de propriétés morphologiques n'est pas séparable. Comme souligné par Aron Arnold (2015, p. 65), on peut opposer à ce dimorphisme sexuel la notion de continuum soutenue par les travaux de la biologiste Anne Fausto-Sterling (2000).

2.2.3 Genre et pratique vocale

Les différences physiologiques et anatomiques ne permettent pas d'expliquer l'ensemble des variations observées : il a été remarqué par Rendall *et al.* (2004) que les différences de F0 étaient bien plus importantes que ce que les différences en carrure ne suggéreraient, et les auteurs émettent l'hypothèse que la longueur des cordes vocales chez les hommes a évolué de manière dramatique pour constituer une caractéristique sexuelle secondaire différenciant les hommes adultes des adolescents prépubères et des femmes. De même, si la fréquence fondamentale est en partie définie par la longueur des plis vocaux, l'impact de l'utilisation des différents muscles laryngaux, à la source de l'abduction et de l'adduction des plis vocaux reste largement ignoré. Une étude de Kiyoshi Honda *et al.* (1999) a cependant montré le rôle de ces muscles dans les modulations de la F0. D'un point de

vue perceptif, il est également intéressant de noter que, comme le souligne Aron Arnold (Arnold, 2015, p. 74) la perception de l'oreille humaine étant logarithmique, la mesure de fréquence fondamentale en Hertz plutôt qu'en demi-tons amène à l'observation de différences genrées de modulations de la F0 qui ne se retrouvent pas avec l'utilisation d'une échelle logarithmique (Henton, 1989).

Des différences sont également observables sur des enfants prépubères, à un âge où les conduits vocaux ne présentent encore pas de différences morphologiques entre les sexes. Dans l'étude d'Andrea Meditch, le sexe de l'enfant était reconnu à 79% sur des échantillons de 2 minutes de parole produits par des enfants âgés de 3 à 5 ans, ce qui amène l'autrice à conclure que : « At the age studied, children have learned sex-specific markers sufficiently well to allow their sex to be identified by [the respondents of the perceptual test] solely on the basis of speech » (1975, p. 424). Des résultats similaires ont été obtenus dans une étude plus récente avec des enfants âgés de 4 à 12 ans (Perry *et al.*, 2001). Des taux de reconnaissances à 85% ont été obtenus par Amir *et al.* (2012) pour des locuteurs et locutrices âgés de 8 à 18 ans.

Dès lors, on peut supposer que l'indexation du genre dans la voix a également lieu chez des hommes et des femmes adultes via des pratiques articulatoires. L'étude de Traunmüller et Eriksson s'est intéressée à la comparaison des fréquences fondamentales rapportées dans plusieurs études portant sur des communautés linguistiques différentes (voir Tableau 2.1). Les valeurs rapportées pour l'anglais sont celles de Yukio Takefuta *et al.* (1972), avec un F0 moyen de 127Hz pour les hommes et 186Hz pour les femmes. En français, l'étude de Louis-Jean Boë *et al.* (1975) rapporte une moyenne de 118Hz pour les hommes et 207Hz pour les femmes, tandis que celle de Phil Rose (1991) pour le dialecte wù obtient une F0 moyenne de 170Hz pour les hommes et 187Hz pour les femmes. Ce qui est intéressant dans ce travail et qui avait déjà été souligné par Adrian Simpson (2009), c'est que les variations de F0 moyennes sont importantes à l'intérieur des catégories de genres, mais que les écarts femme-homme varient aussi grandement selon la langue, et que ces écarts ne peuvent pas être uniquement imputés à des différences anatomiques, attestant du caractère socio-culturel de ces différences.

Ce rôle de la pratique articulatoire a également été montré par l'étude de Keith Johnson (2006) qui s'intéressait aux fréquences de résonances pour des locuteurs et locutrices de langues différentes (norvégien, anglais de Californie, d'Australie et de Nouvelle-Zélande, polonais, français, coréen, italien, flamand, hébreu, islandais, suédois, espagnol, néerlandais, allemand, hongrois et danois). Après avoir calculé une différence moyenne entre les F1, F2 et F3 des locuteurs et des locutrices de chacune de ces langues (voir Figure 4), et en s'appuyant sur les données démographiques d'une autre étude (Tolonen *et al.*, 2000), il observe que les différences entre la taille moyenne des populations et leurs fréquences formantiques ne sont pas corrélées. Selon l'auteur, les différences de taille et donc de tractus vocal n'expliquent que 6% à 40 % des différences observées : « In this data, Danish men

Étude	**Langue**		**Nb. loc.**	**Genre**	**Âge**	**F0 moy.**	σ
Rappaport (1958)	allemand		190	H	–	129	2.3
			108	F	–	238	1.9
Chevrie-Muller et Gremy (1967)	français		21	H	20-61	145	2.5
			21	F	19-72	226	2.3
Takefuta *et al.* (1972)	anglais		24	H	–	127	3.8
			24	F	–	186	5.4
Chen (1974)	chinois mandarin		2	H	30-50	108	4.1
			2	F	30-50	184	3.8
Boë *et al.* (1975)	français		30	H	–	118	1.8
			30	F	–	207	3.0
Kitzing (1979)	suédois		51	H	21-70	110	3.0
			141	F	21-70	193	2.7
Johns-Lewis (1986)	anglais	(conv.)	5	H	24-49	101	3.4
			5	F	24-49	182	2.7
		(lecture)	5	H	24-49	128	4.35
			5	F	24-49	213	4.5
		(théâtre)	5	H	24-49	142	4.85
			5	F	24-49	239	5.3
Graddol (1986)	anglais	(lecture A)	12	H	25-40	119	3.6
			15	F	25-40	207	3.05
		(lecture B)	12	H	25-40	131	4.55
			15	F	25-40	219	3.9
Krook (1988)	suédois		198	H	20-79	113	2.65
			467	F	20-89	188	2.55
Rose (1991)	wù		4	H	25-62	170	4.1
			3	F	30-64	187	3.8

TABLE 2.1 – Valeurs moyennes de F0 en Hz et variation moyenne de F0 (sd) en demi-tons selon dix recherches qui rapportent les résultats de locuteurs et locutrices adultes. Tiré de Traunmüller et Eriksson (1994).

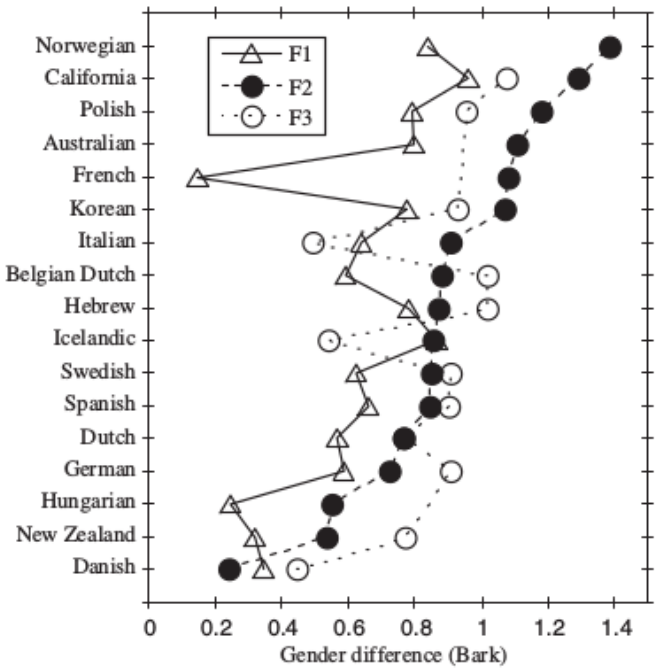

FIGURE 4 – Différence homme/femme pour les F1, F2 et F3 moyens dans 17 langues. Tiré de (Johnson, 2006, p.486)

and women's vowels differ quite a bit less than we would expect given their height difference, while the Australian English men and women differ more than the linear regression line predicts » (Johnson, 2006, p. 487).

Un autre paramètre largement utilisé pour distinguer le sexe dans la voix est celui des fréquences de résonances. Les différences en termes de longueur de tractus vocal et de configuration sont à l'origine de ces résonances qui sont en moyenne 20% plus élevés chez les femmes. Dans son étude de 2012, Aron Arnold postule d'ailleurs que les fréquences de résonnances sont plus importantes que la fréquence fondamentale lorsqu'il s'agit d'indexer le genre.

Les rôles respectifs de la fréquence fondamentale et des fréquences de résonances dans la perception du genre ne sont pas complètement définis. Dans son étude de 2012 sur des locuteurs francophones, Aron Arnold montrait que pour les voix de femmes, seules les fréquences de résonances jouaient un rôle, alors que pour les voix d'hommes, les fréquences de résonances et la fréquence fondamentale jouaient un rôle dans la perception de la masculinité (Arnold, 2012). Dans sa thèse de doctorat portant sur l'identification du genre chez des locuteurs bilingues, Erwan Pépiot a également montré que « les différences acoustiques inter-genres tout comme le processus d'identification du genre par la voix sont fortement dépendants de la langue et donc construits socialement » (Pépiot, 2013, p. 5)

D'une manière générale donc, si des facteurs anatomiques comme la longueur des plis vocaux ou la taille du tractus vocal participent à la production des indices acoustiques

du genre, il semble que les locuteurs et locutrices mobilisent également tout un ensemble de pratiques articulatoires pour se conformer à leurs représentations du genre.

2.3 Voix féminine et technologie : une histoire d'inadaptation

Notre intérêt pour la question de l'existence de disparités de performances des systèmes de reconnaissance automatique de la parole en fonction du genre vient de l'existence historique d'une remise en question de la voix et par extension de la parole des femmes.

Comme expliqué dans le Chapitre 1, toute connaissance est située et le développement des technologies de la parole s'est fait dans un cadre purement masculin, où la question de la variation de genre ne se posait pas. Par la suite, la pensée universaliste majoritaire a laissé impensée l'utilisation de technologies censées convenir à l'ensemble de la population et n'étant adaptée en réalité qu'au groupe socio-démographique à son initiative. On peut d'ailleurs citer l'exemple des films photographiques qui ont été développés pour faire ressortir les visages et les corps blancs, étant de fait inadaptés pour la prise de photos de personnes à la peau noire (Lewis, 2019). Cet impensé se retrouve encore aujourd'hui dans les systèmes de reconnaissance faciale, comme l'ont démontré Joy Buolamwini et Timnit Gebru dans leur étude *Gendershades* (Buolamwini et Gebru, 2018).

La création d'une technologie reste imprégnée de la perception du monde de ses concepteurs et conceptrices et les technologies de la voix n'en sont qu'un exemple supplémentaire. Comme l'écrit Shaye Lynn DiPasquale :

> « The first audio-technology was designed by male inventors to record the male voice. The equipment designers likely did not foresee a time when a female voice would be broadcasted on-air and thus, the equipment was not tested to see if female voices were able to be reproduced clearly. This technical limitation turned into a commonly used excuse for why female voices were unfit for broadcasting. Poor transmission was blamed on the woman and her voice, rather than on the equipment being used. Critics claimed female voices were "unmodulated" and too harsh for radio commentary (Ehrick, 2010, p. 75). Even after the resolution of the reproduction of women's voices by radio equipment in the 1930s, a general dislike for women's voices on-air continued to be blamed on technical deficiencies. » (DiPasquale, 2019, p. 15)

En effet, comme expliqué dans un article du New Yorker, la prolifération des stations radio AM a conduit à la réduction de la bande passante dans les années 1900, pour réduire les interférences de signal. Cette contrainte a conduit les industriels de la radio et du téléphone à limiter leurs appareils pour ne conserver que des signaux compris en 300 et 3400Hz, plage considérée comme minimum pour une bonne retransmission de la

parole. Mais les voix aigues étaient souvent déformées lors de la conversion des ondes sonores en signal (Lacey, 2013). En 1928, John Steinberg de Bell Laboratories est cité par John Rider dans un article expliquant pourquoi les voix de soprano n'étaient pas retransmises correctement à la radio (Rider, 1928). John Steinberg avance que « women are found to talk less distinclty than men », il ajoutera dans un second article intitulé " Understanding women" que : *nature has so designed woman's speech that it is always most effective when it is of soft and well modulated tone.* Cette incapacité des premiers systèmes radiophoniques et téléphoniques à transmettre les voix féminines n'est pas perçu comme une limite des systèmes, mais bien comme une caractéristique féminine :

> « ...the speech characteristics of women, when changed to electrical impulses, do not blend with the electrical characteristics of our present day radio equipment. » (Rider, 1928, p. 334)

En 1933, Harvey Fletcher and Wilden Munson proposent ce que l'on appelle de nos jours les courbes de Fletcher Munson, qui servent aujourd'hui à l'équalisation des sons : ils découvrent que le niveau d'intensité perçu pour un son émis avec la même puissance dépend de sa fréquence. De leur découverte découle l'observation que l'appareil auditif humain est plus sensible aux fréquences contenues en 1000 et 7000Hz (ce qui représente plus ou moins la zone conversationnelle) et qu'ils sont perçus comme ayant une intensité plus forte que des sons ayant la même puissance mais émis à des fréquences en dessous de 1000Hz (voir Figure 5). Cette sensibilité nous aide dans la perception des consonnes, notamment les fricatives, mais aussi les plosives dont le *burst* (perturbation acoustique de courte durée suite au relâchement de l'occlusion) se situe généralement entre 5000 et 7000Hz. Les choix techniques de plafonner les signaux à 3400Hz ont entraîné une perte plus conséquente d'informations pour les voix féminines que masculines (pour lesquelles la majorité de l'information se situe en dessous de 5000Hz). À cela s'ajoute la croyance selon laquelle les femmes parlent plus doucement que les hommes, ce qui incitait les ingénieurs à augmenter le volume lorsqu'une femme prenait place derrière le micro, rendant la voix des femmes stridente ou dure.

Si l'arrivée des stations FM dans les années 1930 permettait une bande passante plus large, celles-ci ne devinrent populaires qu'à partir des années 1970. Les manques des voix féminines se retrouvent encore aujourd'hui dans des discours comme ceux du vice-président des technologies de la parole du groupe ATX, Tom Schalk, qui disait en 2011 à propos des GPS fournis dans les voitures du groupe que : « many issues with women's voices could be fixed if female drivers were willing to sit through lengthy training... Women could be taught to speak louder, and direct their voices towards the microphone » [10] ou encore avec la polémique plus récente autour des propos de Denis Balbir concernant la possibilité qu'une femme commente un match de foot. Le journaliste sportif expliquait que : « Dans

10. `http://techland.time.com/2011/06/01/its-not-you-its-it-voice-recognition-doesnt-recognize-women/`

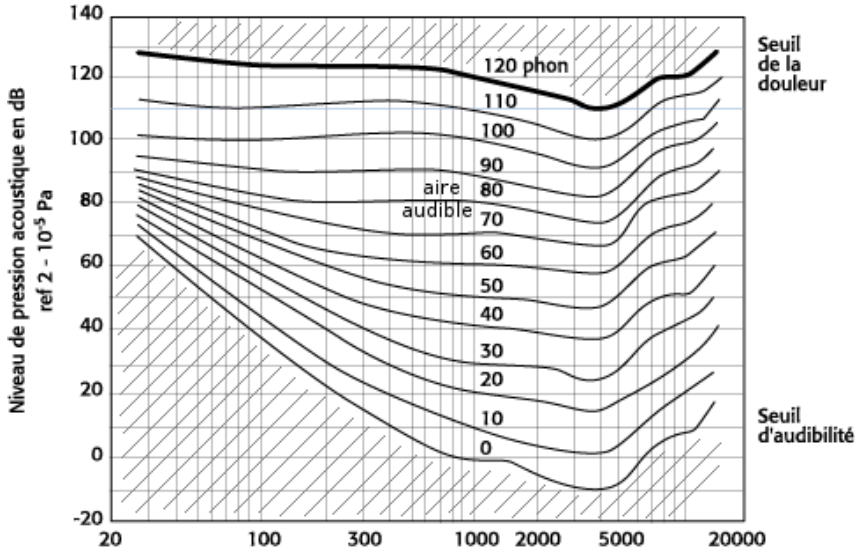

FIGURE 5 – Diagramme de Fletcher Munson. Inspiré de Fletcher et Munson (1933).

une action de folie, elle va monter dans les aigus, ça va être forcément, peut-être, un peu plus délicat à supporter [...] Je sais que l'on va me traiter de misogyne et de sexiste parce que je dis ça, mais ce n'est pas parce que c'est une femme : c'est le timbre de voix qui ne fonctionnerait pas ». [11] [12]

2.4 Le genre en IA

Les systèmes de traitement automatique de la parole et plus généralement les systèmes de TAL au sens large s'inscrivent donc dans un cadre historique de développement qui n'est pas exempt du sexisme de son époque. La question des biais de genre en TAL en est de ce fait d'autant plus pertinente. En effet, les femmes sont assez peu représentées dans ces domaines que ce soit en tant que chercheuses (Wang, 2011) ou dans les données (voir Chapitre 7) et comme l'écrit Smriti Parsheera : « AI artifacts tend to reflect the goals, knowledge and experience of their creators » (Parsheera, 2018, p. 1). On peut donc s'attendre à retrouver dans ces technologies des biais cognitifs et de sélection qui conduiraient à l'existence de biais prédictifs dans les systèmes. Une quantité de travaux se sont donc intéressés à mettre en avant les "biais de genre" présents dans différents systèmes de TAL, nous en présenterons quelques uns dans la section ci-dessous, pour ensuite réfléchir aux implications de l'utilisation de la variable du genre dans le TAL et plus largement, dans l'IA.

11. `https://www.francetvinfo.fr/sports/foot/elles-vont-monter-dans-les-aigus-le-commentateur-de-foot-denis-balbir-se-dit-contre-les-femmes-commentatrices_2988841.html`

12. Il est intéressant néanmoins de souligner l'utilisation d'une prolepse par l'intéressé, soulignant ainsi l'évolution de l'opinion publique et la moindre permissivité quant à l'énonciation de telles "vérités" dans la sphère publique.

2.4.1 Des exemples de discrimination de genre dans le TAL

Jusqu'à présent, la majorité des travaux s'étant intéressés à l'existence de "biais de genre" dans le TAL se sont focalisés sur l'écrit avec notamment les systèmes de traduction automatique et les modélisations intermédiaires telles que les *word embeddings*. Comme nous l'avons déjà vu en section 1.3.2, les données textuelles utilisées pour entraîner les *word embeddings* contiennent tout un ensemble d'associations stéréotypiques qui se retrouvent dans la géométrie de l'espace vectoriel produit. Ainsi Tolga Bolukbasi *et al.* (2016) ont montré que le modèle word2vec entraîné sur le corpus de Google News répondait à l'analogie "man is to computer programmer as woman is to x" avec "x = homemaker". Dans une autre étude, Aylin Caliskan *et al.* (2017) observent les mêmes phénomènes d'association implicite et proposent un test basé sur le IAT pour mesurer le degré de ces associations : le *Word-Embedding Association Test*. Ces associations implicites ont également été mises en avant par le travail de Nikhil Garg *et al.* (2018) qui montre l'évolution des stéréotypes sur une période de 100 ans à partir de plongements de mots. Les *word embeddings* étant une représentation du vocabulaire, ils ne sont pas utilisés de manière isolée mais constituent des ressources pour d'autres tâches du TAL comme la traduction automatique. L'impact de ces disparités de représentations peut donc se retrouver par la suite dans les performances de ces systèmes. Dans les nouveaux systèmes neuronaux, les modèles de langues stochastiques ont été remplacés par des *pretrained language models* (PLMs) ou *contextualized embeddings* qui reprennent des architectures similaires aux *word embeddings* décrits plus haut mais modélisent également le contexte (BERT (Devlin *et al.*, 2019), ELMO (Peters *et al.*, 2018) ou GPT-2 (Radford *et al.*, 2019) en sont les exemples les plus connus). Kurita *et al.* (2019) ont montré via des tests d'implications l'existence de biais d'associations dans ces PLMs.

On peut citer également, dans le domaine de la traduction automatique, le travail de Marcelo Prates *et al.* (2020) qui se sont intéressés au biais prédictif de Google Translate. Les auteurs ont, à partir d'une liste d'emplois de l'*US Bureau of Labor Statistics* (BLS), soumis à Google Translate des phrases types constituées de "he/she is [occupation]" dans 12 langues dont les pronoms sont non-genrés (malais, estonien, finnois, hongrois, arménien, bengali, japonais, turc, yoruba, swahili, basque, chinois), pour les traduire en anglais. Les auteurs ont ensuite observé si le pronom choisi en anglais était féminin, masculin ou neutre (*he, she, it*). La majorité du temps, le choix par défaut était celui du pronom masculin quand bien même cela ne correspondait pas à la distribution réelle de la profession, estimée d'après les statistiques du BLS. Dans une autre étude, Eva Vanmassenhove *et al.* (2018) ont démontré la sur-représentation des énoncés produits par des hommes (et plus précisément des hommes âgés de 50 à 60 ans) dans le corpus Europarl sur lequel se base leur travail. Cette sous-représentation des femmes a un impact sur les traductions proposées par leur système entraîné sur ces données et l'équipe propose donc de four-

nir aux systèmes des annotations de genre. Apporter au système des informations sur le genre augmente significativement ses performances sur plusieurs paires de langues, dont les paires anglais-français, anglais-italien et anglais-danois.

Dans leur étude sur la résolution de co-références, (Rudinger *et al.*, 2018) ont analysé 3 systèmes différents : chaque système présentait une préférence à résoudre les coréférences en faveur d'un pronom masculin plutôt que féminin. Il faut souligner cependant que la majorité du temps, ces associations étaient en adéquation avec les statistiques du BLS.

En ce qui concerne l'ASR, peu de travaux ont exploré l'existence d'une différence de performance genrée au sein des systèmes et aucun consensus n'a été atteint dans la littérature. Martine Adda-Decker et Lori Lamel (2005) ont constaté que les systèmes de reconnaissance vocale étaient plus performants sur les voix féminines dans un corpus d'enregistrements radio et téléphoniques, avec une différence de 2%. Elles ont proposé plusieurs explications à cette observation, comme la présence plus importante de parole masculine non professionnelle dans les données, impliquant un discours moins préparé pour ces locuteurs et donc plus de disfluences (interruption, fillers, etc.) qui peuvent venir compliquer la tâche de transcription. Dans une perspective sociolinguistique variationniste, elles avancent également l'hypothèse d'un langage plus normatif et d'une prononciation plus standardisée pour les femmes, liés à leur rôle traditionnel dans l'acquisition des langues et l'éducation. La même tendance a été observée par Sharon Goldwater *et al.* (2010), qui se sont intéressés à savoir quels mots étaient difficiles à reconnaître pour les systèmes d'ASR. Une différence de performance de 2.8% à 3.1% est rapportée dans leur article. Plus récemment, Rachel Tatman (2017) a découvert un biais prédictif de genre dans le système de sous-titrage automatique de YouTube, avec de meilleurs résultats sur les voix d'hommes, mais ce biais n'a pas été observé dans une deuxième étude évaluant le système Bing Speech et les sous-titres automatiques de YouTube sur un ensemble de données plus important (Tatman et Kasten, 2017). Si la significativité statistique des différences de genre a disparu, des différences de performances liées à la race et au dialecte des locuteurs et locutrices ont été constatées. Cette étude fournit des preuves supplémentaires que : « despite dramatic improvements in the technology, automatic speech recognition systems continue to struggle to maintain high accuracy in the face of well-documented systematic sociolinguistic variation. »(Tatman et Kasten, 2017, p. 937). Le travail de Siyuan Feng *et al.* (2021) s'est intéressé à l'existence de biais de genre (mais également d'âge et d'accent) dans un système d'ASR pour le néerlandais. Leur système obtient de meilleures performances pour les femmes que pour les hommes, en adéquation avec les résultats de Martine Adda-Decker et Lamel (2005) et Sharon Goldwater *et al.* (2010). La parole d'enfant est bien moins reconnue que celle des adolescents et des adultes et les locuteurs et locutrices natives sont bien mieux reconnues que les non-natives, peu importe leur âge. L'accent des Flandres est également mal reconnu par le système.

2.4.2 Perspective critique sur l'utilisation du genre en IA

Le flou autour de la notion de genre, sa démocratisation mais aussi l'hégémonie de l'anglais dans les sciences "dures", ont fait qu'aujourd'hui de nombreux travaux portant sur les systèmes d'intelligence artificielle utilisent les notions de *gender* et *gender bias*. Le genre y est d'abord envisagé comme une propriété de l'individu qui apporte une information sur son comportement et ses caractéristiques (notamment anatomiques et physiologiques, ou acoustiques dans le cadre de la voix) et est alors utilisé de manière quasi synonymique au sexe. Les étiquettes du genre remplacent celles du sexe dans les "sciences dures", sans forcément faire référence aux travaux des sociologues, anthropologues et féministes qui sont à l'origine du concept, comme le dénonçaient déjà les sociologues Judith Stacey et Barrie Thorne, en parlant de cooptation du terme de genre tel qu'utilisé dans le cadre des travaux en sociologie (Stacey et Thorne, 1985).

Si la section précédente présente donc tout un ensemble d'études mettant en avant l'existence de biais prédictifs en fonction du genre dans les systèmes de TAL, plusieurs auteurs et autrices ont questionné l'utilisation même de la catégorisation de genre dans les systèmes. Si celle-ci peut se justifier dans une volonté de contrôle des biais cognitifs et face à une histoire de l'invisibilisation des femmes dans les données (voir Chapitre 7) et de la problématisation des voix féminines (voir Section 2.3), elle pose cependant d'autres questions éthiques : comme souligné par Brian Larson (2017), l'utilisation des catégories de genre en TAL se fait souvent indépendamment de toute pensée critique ou explicitation de la conception théorique du genre adoptée. Or, la technologie, en matérialisant un système de catégorisation, est également le reflet des rapports sociaux de pouvoir et implémente comme expliqué au chapitre précédent, une vision et un point de vue, qui est souvent celui du dominant. Les systèmes d'IA contribuent alors à l'entretien du système de pouvoir décrit par les féministes, et ce dans le sens où il réactualise et légitimise l'existence de ces catégories. En effet, comme l'écrivaient David Bamman *et al.* (2014b) dans leur étude s'intéressant à la variation lexicale en fonction du genre sur Twitter :

> « If we start with the assumption that 'female' and 'male' are the relevant categories, then our analyses are incapable of revealing violations of this assumption... [W]hen we turn to a descriptive account of the interaction between language and gender, this analysis becomes a house of mirrors, which by design can only find evidence to support the underlying assumption of a binary gender opposition. » (Bamman *et al.*, 2014b, p. 148)

Si les catégories utilisées contribuent à pérenniser des systèmes hiérarchiques et des normes de pensée, il est donc nécessaire de réfléchir à l'utilisation de la variable du genre dans les systèmes. C'est également le propos de Catherine D'Ignazio et Lauren Klein,

dans leur chapitre intitulé *What gets counted counts.*[13] En s'appuyant sur les travaux de Bowker et Star sur les systèmes de classifications, les autrices notent que :

> « When a system is in place it becomes naturalized as "the way things are." This means we don't question how our classification systems are constructed, what values or judgments might be encoded into them, or why they were thought up in the first place. » (D'Ignazio et Klein, 2020, p. 104)

Le risque, en utilisant les catégories binaires du genre dans les systèmes de TAL, par exemple dans des systèmes de reconnaissance automatique du genre pour des auteurs et autrices de texte, c'est de cristalliser les stéréotypes autour des représentations de comment les femmes et les hommes écrivent (Koolen et van Cranenburgh, 2017). On retombe dans le paradigme de la différence des premières études sociolinguistiques variationnistes initié par Labov.

En effet, en linguistique dans les années 1970, on commence à questionner la variation de genre pour tenter de définir un parler des femmes en opposition avec celui des hommes. On a cherché à distinguer d'un point de vue lexical et syntaxique les parlers "masculin" et "féminin", émettant l'hypothèse que l'utilisation des variétés standard et de prestige, plus présente chez les femmes, témoignait d'une volonté d'ascension sociale (Trudgill, 1972) ou d'une plus forte insécurité linguistique (Labov, 1966, 1976). Ces études visaient à dresser un parallèle entre une conception hiérarchisée de la société sexuée et les réalisations langagières de ses groupes constitutifs à savoir les hommes d'un côté et les femmes de l'autre. En pré-imposant la grille de lecture de la binarité du genre, les stéréotypes genrés s'en sont trouvés soulignés et renforcés (Aebischer, 1985). Comme l'écrit Louise Mullany :

> « The majority of these studies founds that women were more indirect, co-operative and collaborative than men in their interactional speech styles [...]. Instead of simply dismissing these findings as overgeneralisations, we can reinterpret them as ideological expectations governed by the 'rigid regulatory frame' of the gender differences discourse. Early work thus provides an 'analytic window that constructed and constrained women's linguistic stage, offering a set of features that operated stereotypically and en masse to help index "woman"' (Queen, 2004 :291–2) » (Mullany, 2009, p. 182)

Si l'on considère que l'utilisation de catégories de genre comme catégories d'analyses révèle le cadre normatif de ce qui est attendu comme "féminin" et "masculin" alors l'utilisation de variable de genre dans les technologies ne peut être considérée comme anodine. C'est également le propos de Blodgett *et al.* (2020), qui, au-delà d'une critique portant sur l'absence de définition claire des biais adressés par ces études, soulignent la relative absence de réflexivité des chercheurs et chercheuses sur les structures sociales à l'origine des objets technologiques qu'ils et elles utilisent et évaluent :

13. La formule, empruntée à la géographe féministe, Joni Seager

> « How are social hierarchies, language ideologies, and NLP systems coproduced ? This question mirrors Benjamin's (2020) call to examine how "race and technology are coproduced" —i.e., how racial hierarchies, and the ideologies and discourses that maintain them, create and are re-created by technology. We recommend that researchers and practitioners similarly ask how existing social hierarchies and language ideologies drive the development and deployment of NLP systems, and how these systems therefore reproduce these hierarchies and ideologies. » (Blodgett *et al.*, 2020)

Cristalliser ces distinctions dans des systèmes automatiques qui serviront par la suite à "prédire" vient donc essentialiser les distinctions homme/femme, à un moment où la binarité et le système hiérarchique qu'elle tend à légitimer sont questionnés. À l'inverse, ne plus prendre en considération le genre ne permettrait pas, dans le cas de l'ASR par exemple, de mettre en avant comment les femmes ont été longtemps mises de côté et comment le discours scientifique a contribué à construire le mythe de la complexité du traitement des voix féminines. On se retrouve donc à un point de tension, entre la nécessité de se reposer sur des systèmes de classifications, car ce sont ces catégories sur lesquelles se basent les systèmes et le risque de contribuer à pérenniser des discours essentialisants sur les différences hommes/femmes.

Conclusion

Pour attester de l'existence de biais prédictifs genrés au sein d'un système d'ASR, encore faut-il savoir ce que l'on entend par genre. Dans ce chapitre, nous avons donc retracé l'histoire du concept de genre, montrant son hétérogénéité et ses tensions théoriques, notamment à travers les oppositions entre féminisme marxiste et théorie queer. Quelle que soit la tradition adoptée, le concept de genre permet de mettre en avant la construction des catégories binaires hégémoniques "homme" et "femme" et questionne leur utilisation aussi bien dans la voix que dans les systèmes. Dans la voix, car le discours sur la binarité a contribué à la perception de deux catégories exclusives, là où les réalités acoustiques se recouvrent largement (ce qu'on observe sur les distributions de F0 par exemple). Cette dichotomie voix d'homme/voix de femme a aussi été construite par la technologie, notamment avec l'émergence de la radio, ce qui a permis le développement d'un discours sur la supposée difficulté attachée à la voix des femmes. Mais l'utilisation du genre est aussi questionnée dans les systèmes, car en choisissant certaines catégories d'analyses plutôt que d'autres, on peut à la fois mettre en lumière des inégalités structurelles entre hommes et femmes dans nos sociétés qui se retrouvent dans nos données et dans nos systèmes de TAL, mais on contribue également à la réactualisation d'un système catégoriel binaire aujourd'hui questionné. Ainsi nous sommes amenée à préciser notre question de recherche

qui peut s'en trouver dédoublée : *Existe-t-il des biais prédictifs selon les catégories de genre dans les systèmes d'ASR ? Et quelles sont ces catégories ?*

Chapitre 3

Données et technologies en reconnaissance automatique de la parole

Dans ce chapitre, après avoir présenté la tâche de reconnaissance automatique de la parole, nous présenterons les systèmes actuellement utilisés pour résoudre cette tâche. Dans un second temps, après une brève description de la genèse des technologies de l'ASR, nous regarderons comment, le développement historique des technologies a défini la place accordée à la variation due au locuteur ou à la locutrice dans le développement des systèmes, pour éclairer notre questionnement sur les biais de genre, à la lumière de la conception de l'individu dans les systèmes. Puis nous nous attarderons sur l'impact des données et des campagnes d'évaluation sur l'évolution des technologies de reconnaissance de la parole.

3.1 La reconnaissance automatique de la parole (ASR)

3.1.1 Principe de la reconnaissance automatique de la parole

L'objectif de la recherche en reconnaissance automatique de la parole est le développement d'un système qui transcrit automatiquement la parole naturelle. On peut distinguer trois domaines différents en reconnaissance de la parole, avec des objectifs distincts : la reconnaissance de mots isolés, où les mots sont séparés par des pauses distinctes a été le premier objectif de la recherche et est à la base des systèmes de types serveurs vocaux, utilisés notamment pour les répondeurs téléphoniques. La reconnaissance de la parole continue vise à reconnaître des énoncés produits de façon continue, sans segmentation préalable des mots. Ce type de système peut servir à des tâches de dictée vocale, de sous-titrage, etc. Enfin la compréhension de la parole, où l'objectif n'est pas uniquement la transcription, mais la compréhension du sens, pour des applications comme les assistants virtuels (Siri chez Apple, Alexa chez Amazon ou encore Cortana chez Microsoft et l'assistant Google). Dans ces systèmes la tâche d'ASR est une première brique d'un pipeline de NLU (*Natural Language Undersanding*). L'intérêt porté à la transcription et la compréhension du langage naturel s'inscrit dans une conception plus large de la parole

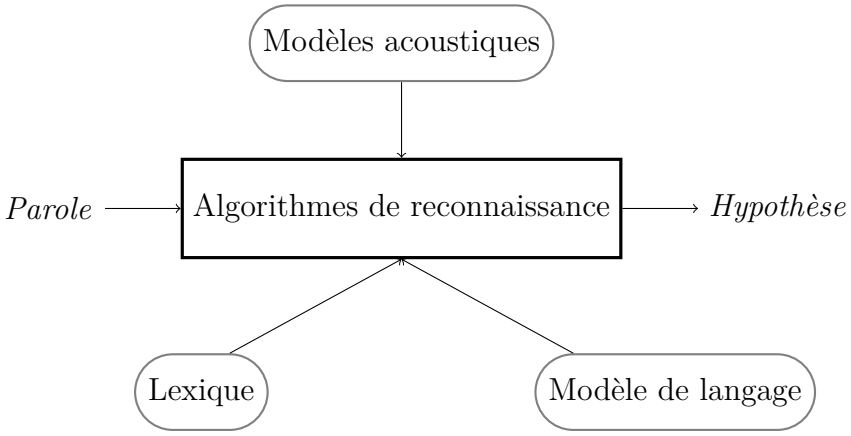

FIGURE 6 – Architecture d'un système de reconnaissance automatique de la parole. Tiré de (Haton *et al.*, 2006, p.4).

comme interface ergonomique pour les interactions homme/machine (car nécessitant un apprentissage moindre, laissant les mains libres, etc.) Dans le cadre de ce travail nous nous intéressons à la reconnaissance automatique de la parole continue.

Trois approches différentes ont amené à la création de modèles d'ASR : initialement, les modèles utilisés étaient principalement des modèles experts, basés sur un ensemble de connaissances phonétiques préalables. Puis, se sont développées les approches stochastiques ou probabilistes, fortement appuyées par des modèles statistiques comme les mixtures de gaussiennes et les chaînes de Markov. Plus récemment, l'approche neuronale a été réinvestie, notamment avec les succès des réseaux de neurones dans les tâches de reconnaissance visuelle. Hormis les premiers systèmes experts, la majorité des modèles d'ASR a été pendant longtemps des modèles probabilistes (Jelinek, 1976). Un tel système peut se représenter comme décrit sur la Figure 6. Le système est découpé en modules : les paramètres acoustiques extraits de la parole sont reconnus grâce au modèle acoustique qui définit l'unité acoustique la plus probablement produite (phonèmes, diphones, syllabes). Cette forme acoustique est mise en lien avec des formes lexicales via le lexique et le modèle de langue. Le lexique recense toutes les prononciations possibles du vocabulaire et le modèle de langue décrit les structures de phrases probables, i.e. les suites de mots possibles à priori. Ainsi, à partir d'un signal de parole, le système de reconnaissance propose une hypothèse de transcription de l'énoncé présenté en entrée.

D'un point de vue mathématique, on décrit le problème de la reconnaissance automatique de la parole comme la transcription d'un signal X en une séquence de mots $\hat{W}$ la plus probable (Bahl *et al.*, 1983, p. 180). Le signal est représenté par une suite d'éléments $X = x_1, x_2, ..., x_y$ et le système de reconnaissance, à l'aide d'un modèle acoustique et d'un modèle de langue cherche à définir la suite de mots $\hat{W}$ la plus probable, en maximisant l'expression suivante :

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X) \tag{3.1}$$

La formule de Bayes permet de transformer l'équation (3.1) en :

$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)} \tag{3.2}$$

où $P(W)$ est la probabilité à priori, donnée par le modèle de langue, d'obtenir la suite de mots W. Le modèle acoustique permet quant à lui d'estimer la probabilité $P(X|W)$, soit la probabilité d'observer ces paramètres acoustiques sachant la suite de mots W. Comme $P(X)$ ne dépend pas de W, maximiser l'équation (3.2) est équivalent à :

$$\hat{W} \approx \underset{W}{\operatorname{argmax}} P(X|W)P(W) \tag{3.3}$$

La quantité $P(X|W)$ représente donc le modèle acoustique et la quantité $P(W)$ le modèle de langue, qui est indépendant du signal. Dans le cas des systèmes avec un grand vocabulaire le modèle acoustique est une combinaison d'un modèle acoustique de phonèmes et d'un dictionnaire de prononciation ou lexique. $P(X|W)$ peut donc se décomposer pour obtenir l'équation suivante :

$$\hat{W} \approx \underset{W}{\operatorname{argmax}} P(X|U)P(U|W)P(W) \tag{3.4}$$

où $P(U|W)$ représente le modèle de prononciation ou lexique et $P(X|U)$ le modèle acoustique phonémique.

Un modèle stochastique peut donc être illustré par la Figure 7. La suite de la section s'intéressera aux différents composants de ce schéma à savoir : les paramètres acoustiques, le modèle acoustique, le lexique et le modèle de langue.

3.1.2 Les systèmes stochastiques

3.1.2.1 Paramètres acoustiques

La première étape consiste donc à extraire les paramètres acoustiques du signal de parole. Pour pouvoir considérer le signal comme stationnaire, on extrait des trames de 20 à 40 ms (usuellement une fenêtre glissante, de type fenêtre de Hamming de 25ms, avec un recouvrement de 10ms). Le signal est donc transformé en une suite de trames et il s'agit ensuite d'extraire les paramètres pour chaque trame. Différentes techniques ont été proposées dans la littérature : *Mel-Frequency Cepstral Coefficients* (MFCC), *Linear Predictive Coding* (LPC), *Linear Prediction Cepstral Coefficients* (LPCC), *Perceptual Linear Prediction* (PLP) ou encore avec les nouveaux systèmes neuronaux les sorties de bancs de filtres (*filterbanks*). Chaque type de paramètres se base sur différentes propriétés du signal de parole : par exemple les coefficients LPC (Makhoul, 1975), analogues à l'analyse de

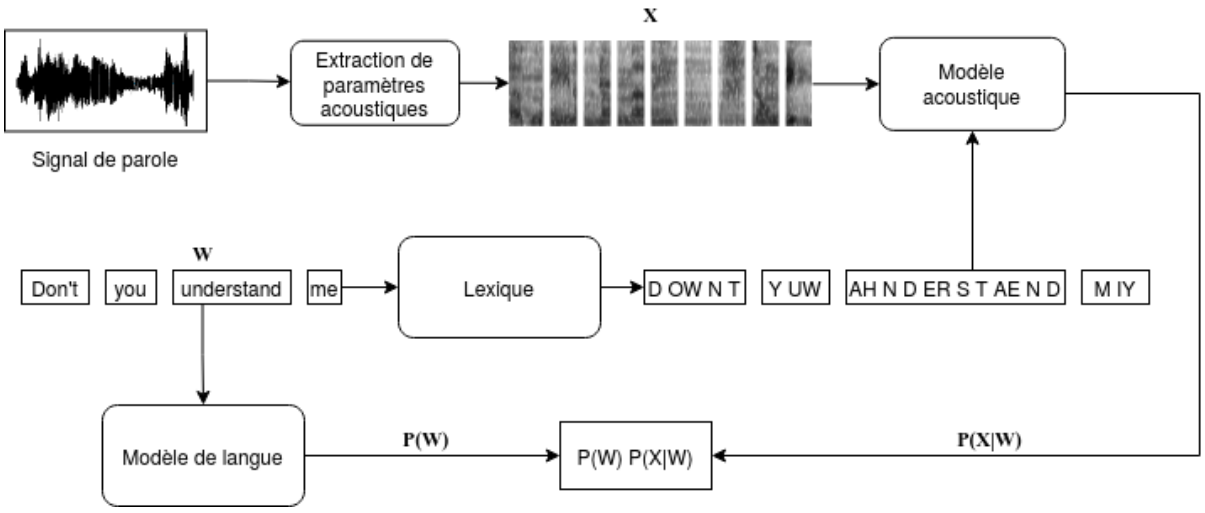

FIGURE 7 – Modèle de reconnaissance automatique de la parole sctochastique. Tiré de (Kamath *et al.*, 2019, p. 380)

Prony, supposent une conception de la production de la parole comme celle proposée par Fant dans sa théorie source/filtre (Fant, 1960). La méthode de LPC modélise l'enveloppe spectrale dans son ensemble.

Les MFCCs eux se basent sur une approche perceptive s'inspirant de la perception non-linéaire de l'oreille (Davis *et al.*, 1952). En effet, ils utilisent l'échelle de perception Mel qui s'appuie les travaux de Stevens (Stevens et Volkmann, 1940). Les MFCCs présentent également l'avantage d'être des coefficients décorrélés. Ils font partie de l'ensemble des méthodes qui utilisent des coefficients cepstraux, obtenus après transposition d'un sonagramme dans le domaine fréquentiel puis dans le domaine des quéfrences (inverse des fréquences), du fait de l'application successive de transformées de Fourier. Les LPCC sont également la transposition dans le domaine cepstral des coefficients obtenus par LPC. La méthode de PLP (Hermansky, 1990) a eu pour but de prendre en compte la perception logarithmique de l'oreille humaine dans l'utilisation de prédiction linéaire en se basant cette fois non pas sur l'échelle des Mel mais sur celle des Bark, plus utilisée en psycho-acoustique. Les coefficients PLP et les MFCCs sont les méthodes les plus largement utilisées dans les systèmes probabilistes.

Il est également d'usage de calculer les dérivées première et seconde des différents paramètres pour représenter la variation ainsi que l'accélération de chacun des paramètres.

3.1.2.2 Modèle acoustique

Les différentes modélisations acoustiques sont aussi dépendantes des théories qui les sous-tendent. Les modélisations basées sur les modèles de Markov cachés [1] (en anglais *Hidden Markov Model* - HMM) ont été et sont encore aujourd'hui largement utilisées. Ce

1. Pour une description approfondie des HMM voir Jelinek (1976) et Rabiner et Juang (1986).

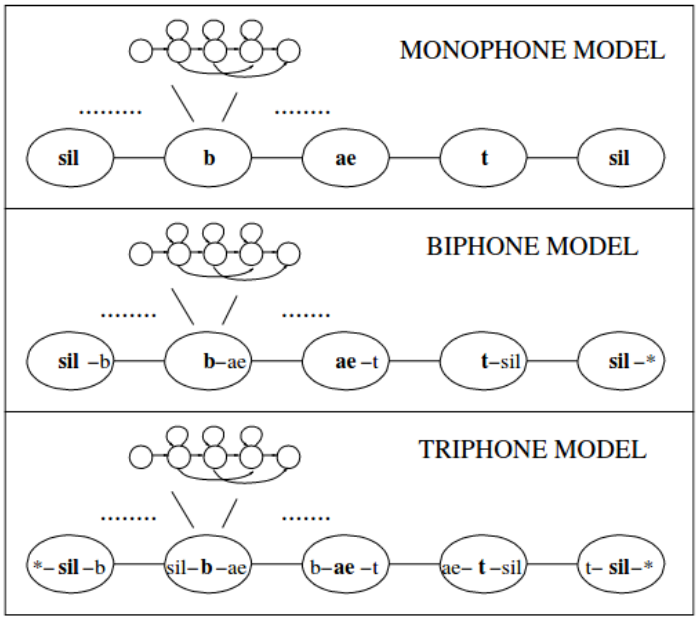

FIGURE 8 – Modèle HMM monophone, biphone et triphone pour le mot “bat” en anglais. sil représente un silence et marque le début et la fin de l'énoncé, il est modélisé comme un phone à part entière. Extrait de (Resch, 2003, p.4)

sont des automates stochastiques qui permettent de déterminer la probabilité d'une suite d'observations. Comme expliqué par Haton et al. :

> « Un HMM est caractérisé par un double processus stochastique : un processus interne, non observable $X(t)$ et un processus externe observable $Y(t)$. Ces deux chaînes se combinent pour former le processus stochastique. Dans le cas de la parole, la chaîne interne $X(t)$ est une chaîne de Markov qui est supposée être à chaque instant dans un état où la fonction aléatoire correspondante engendre un segment élémentaire (de l'ordre de 10ms ou plus) de l'onde acoustique observée représenté par un vecteur de paramètres. Un observateur extérieur ne peut voir que les sorties de ces fonctions aléatoires, sans avoir accès aux états de la chaîne sous-jacente, d'où le nom de modèle caché. » (Haton *et al.*, 2006, p. 85)

En ASR, le signal est l'observation et le phonème (ou partie de phonème) est l'état caché. En adoptant l'hypothèse de Markov d'ordre 1, qui suppose que les états futurs ne dépendent que de l'état présent, on peut alors représenter chaque phonème par un HMM. Il s'agira ensuite de savoir quel HMM a la plus grande probabilité d'avoir émis le signal en entrée. Pour modéliser les phénomènes de coarticulation, des modèles triphones ont été proposés (voir Figure 8) ou un phonème donné est modélisé par 3 états, prenant en compte son contexte. Les fonctions aléatoires utilisées pour estimer les densités de probabilités d'émission des états des HMM sont soit des mixtures de gaussiennes (GMM), soit des réseaux de neurones profonds (DNN).

3.1.2.3 Modèle de langue

Les modèles de langue ont été introduits dans le but d'apporter des informations syntaxiques et sémantiques dans la tâche de reconnaissance automatique de la parole. Les modèles de langues permettent d'obtenir la probabilité d'une suite de mots dans un langue donnée, quantité représentée par $P(W)$, à partir d'un corpus d'apprentissage. En utilisant la règle de probabilité en chaîne, on peut approximer $P(W)$ composée de N éléments par :

$$P(W) \approx \prod_{n=1}^{N} P(w_n|w_0, ..., w_{n-1}) \tag{3.5}$$

On simplifiera l'équation obtenue grâce à l'hypothèse de Markov, qui nous permet d'estimer la probabilité d'occurrence d'un mot non plus en fonction de la totalité de la séquence précédente, mais uniquement de la probabilité d'une sous-séquence de n mots (définie par l'ordre de l'hypothèse de Markov choisi, la plupart du temps 2), d'où leur nom de modèles n-grammes (Jelinek et Mercer, 1980). Ainsi pour un modèle tri-gramme (d'ordre 2), on obtient la formulation suivante :

$$P(w_n|w_0, ..., w_{n-1}) \approx P(w_n|w_{n-2}w_{n-1}) \tag{3.6}$$

Cependant le corpus d'apprentissage ne contient pas forcément l'ensemble de mots et de n-grammes associés, que le système peut être amené à transcrire. Ces évènements inconnus sont pris en compte par des techniques de lissage (Witten et Bell, 1991; Kneser et Ney, 1995). Les modèles de langues sont évalués à l'aide de deux mesures, la perplexité et l'entropie. L'entropie représente le degré d'aléatoire présent dans le modèle, pour avoir un bon pouvoir descriptif des constructions syntaxiques et sémantiques possibles dans une langue. Elle doit être la plus faible possible. La perplexité, quant à elle, estime la qualité de prédiction du modèle sur un corpus de test. Plus la perplexité est faible, meilleur est le modèle de langue. Pour une description de ces mesures se référer à Jelinek *et al.* (1977).

3.1.2.4 Lexique et modèle de prononciation

Le lexique ou dictionnaire phonétisé contient l'ensemble des transcriptions phonétiques possibles pour un mot, prenant ainsi en compte les variations de prononciations et les phénomènes de liaisons. Il contient l'ensemble du vocabulaire d'un système, soit l'ensemble des mots que le système est capable de reconnaitre. Dans le cas des systèmes à grand vocabulaire, soit la majorité des systèmes de reconnaissance de la parole continue, les entrées du lexique ne sont pas directement des mots mais des sous-mots. L'utilisation de lexique aide à l'apprentissage du modèle de prononciation, sous-composant de notre modèle acoustique (voir Équation 3.4) qui permet l'interface entre une séquence de phones et un mot.

3.1.3 L'approche neuronale

Les réseaux de neurones ont été introduits dans les années 1980-1990, mais leurs résultats étaient bien en dessous des modèles HMM-GMM standards. Mais avec l'augmentation des capacités de calcul (notamment grâce aux GPU [2]) et la disponibilité de grands corpus de données comme Librispeech (Panayotov *et al.*, 2015), ils ont été de plus en plus utilisés jusqu'à constituer la majorité des systèmes états de l'art. Cette section s'intéressera à une description basique des principes qui sous-tendent ces réseaux de neurones. [3]

3.1.3.1 Principes des réseaux de neurones artificiels

Les réseaux de neurones artificiels sont une modélisation inspirée du traitement de l'information par le cerveau humain. Un réseau de neurones se pense donc comme un réseau de noeuds de calculs reliés entre eux. Chaque noeud de calcul représente un "neurone" et les liens qui les relient représentent les connexions synaptiques. Un neurone reçoit des informations, via les synapses qui mènent à lui, et ces informations sont additionnées à l'aide d'une somme pondérée, qui sera ensuite transmise à la fonction d'activation du neurone. L'information transite ainsi à travers le réseau jusqu'à fournir une "sortie". Cette proposition du neurone formel date de 1957 avec le perceptron de Frank Rosenblatt (Rosenblatt, 1957). Schématiquement, un neurone peut se représenter comme dans la Figure 9. Les fonctions d'activation utilisées sont des fonctions non-linéaires : la plupart du temps on utilisera les fonctions sigmoïde, ReLU ou tangente hyperbolique. Si le graphe de connexion contient au moins un cycle, on parlera de réseau récurrent (*Recurrent Neural Network* - RNN) et dans le cas contraire de réseau acyclique (*Feed-Forward Neural Network* - FFNN).

Pour apprendre, le réseau va, de manière itérative, modifier les poids de calcul de la somme pondérée de chaque neurone, en fonction de leur contribution à l'erreur (à savoir l'écart entre la sortie du réseau et la sortie attendue). La mesure de l'erreur est définie par une fonction de coût qui peut varier selon les implémentations. La correction des poids se fait ensuite via ce qu'on appelle la rétropropagation du gradient. En réajustant les poids du modèle à chaque itération, le réseau va converger pour atteindre le minimum de la fonction de coût (mais il arrive parfois que celui-ci se bloque sur des minima locaux).

Le modèle initial de Frank Rosenblatt n'avait qu'une seule couche de neurones, par la suite Paul Werbos proposa en 1974 l'utilisation de réseaux multi-couches (ou MLP pour *Multi-Layer Perceptron*), dont la première implémentation sera réalisée en 1986 par David Rumelhart *et al.* (1986). On appelle "réseau de neurones profond" (*Deep Neural Network*

2. Un GPU pour *Graphics Processing Unit*, ou processeur graphique, est une unité de calcul très puissante, du fait d'un structure hautement parallèle, ce qui la rend particulièrement utile pour les fonctions de calcul d'images, le traitement du signal vidéo, etc. Les GPU sont aujourd'hui largement utilisés pour l'entrainement des réseaux de neurones.

3. Pour une revue exhaustive de l'approche neuronale en TAL et des descriptions mathématiques des modèles voir Uday Kamath *et al.* (2019).

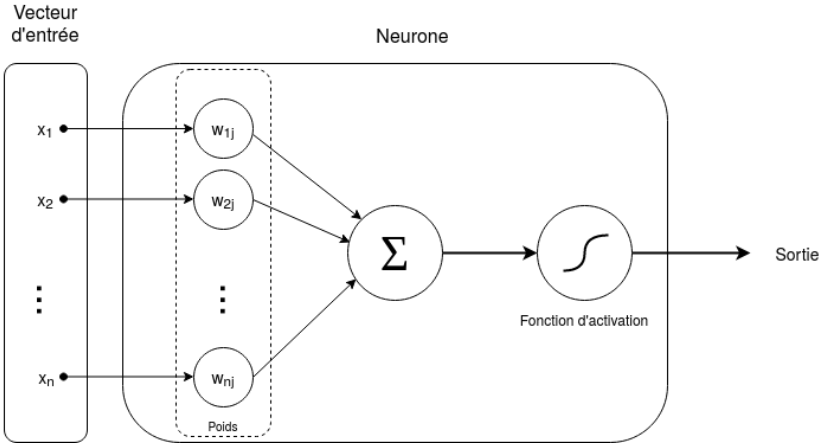

FIGURE 9 – Représentation d'un neurone artificiel. La somme pondérée est calculée après multiplication des entrées avec la matrice de poids. La sortie de cette somme pondérée est alors passée à une fonction d'activation non-linéaire qui produira une sortie. Celle-ci sera ensuite utilisée comme entrée par les neurones suivants si le neurone ne fait pas partie de la couche de sortie.

- DNN), un MLP contenant au moins 3 couches cachées (voir Figure 10). Ce type de réseau de neurones est utilisé dans les approches hybrides. Pour l'entraînement du modèle acoustique, on utilise un DNN acyclique pour estimer les probabilités d'un état HMM et ce sont ces probabilités qui seront utilisées comme entrées du modèle HMM-GMM.

Cependant avec l'arrivée de réseaux de neurones à plusieurs couches, le problème de l'explosion ou de la disparition du gradient apparaît (*exploding or vanishing gradient*). Cela vient du fait que dans l'étape de rétropropagation, le gradient est multiplié par la sortie de chaque couche successive, dans l'optique de rétropropager l'erreur et d'adapter les poids. Mais ces multiplications successives peuvent conduire à un gradient de plus en plus grand (si multiplié par une valeur absolue supérieure à 1) ce qui conduira à un réajustement des poids inadéquats et empêchera l'apprentissage. En revanche, si le gradient est multiplié par un nombre compris entre 0 et 1, il diminuera au fur et à mesure des couches et l'erreur ne sera propagée qu'en très faible partie vers les premières couches du réseau, ce qui rendra l'apprentissage très long voire inexistant. Ces problématiques sont traitées par exemple par l'utilisation de mécanismes d'attention dans le cas de la disparition du gradient ou de "clipping" dans le cas de son explosion.

3.1.3.2 Les réseaux récurrents

Les réseaux acycliques présentent cependant un ensemble de limites, notamment dues à leur incapacité à gérer des entrées de taille variable, ce qui est souvent le cas avec la parole. Pour gérer ces spécificités du signal de parole, on peut soit, comme dans le cas des modèles acoustiques, coupler le réseau avec un HMM, soit utiliser des RNN qui sont plus adaptés à la gestion des séries temporelles. En effet, la majorité des tâches de TAL manipulant des séquences (de paramètres acoustiques, de phones, de caractères, de

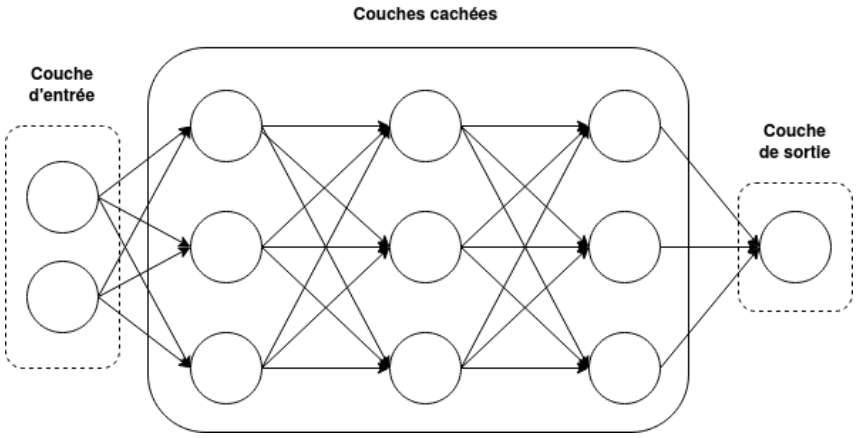

FIGURE 10 – Réseau de neurones profonds avec 3 couches cachées.

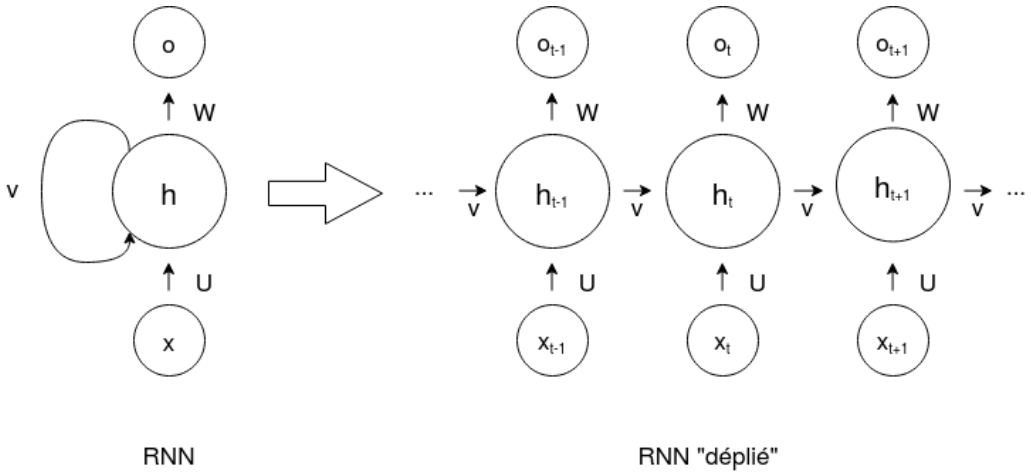

FIGURE 11 – Réseau de neurones récurrent déplié.

mots, etc.), l'approche adoptée est une approche séquence à séquence (souvent abrégée *seq2seq*). L'idée est de venir modéliser la séquence d'entrée via un encodeur sous forme d'états cachés, puis de produire une séquence de sortie à partir de cette modélisation via un décodeur. La présence de cycles dans les RNN permet d'introduire un mécanisme de mémoire des entrées précédentes, mais ils sont plus difficiles à entraîner que les réseaux acycliques (voir Figure 11).

Plusieurs variations des RNN existent pour conserver plus ou moins d'informations en "mémoire", parmi elles les cellules LSTMs (pour *Long Short Term Memory*) qui conservent une partie de la mémoire dans la cellule elle-même et non pas uniquement via la sortie de l'itération précédente et les cellules GRU (pour *Gate Recurrent Unit*) qui sont une autre variation des LSTM. On utilise également des RNN bidirectionnels, qui permettent de parcourir la séquence d'entrée dans les deux sens, prenant ainsi en compte le contexte antérieur mais également ultérieur, ce qui est particulièrement utile dans des tâches notamment de traduction.

3.1.3.3 Attention

En considérant que la mémoire est contenue dans un seul vecteur, les RNN perdent une partie de l'information, notamment du fait de l'aspect séquentiel de la parole ou du texte dans lesquelles certaines informations se retrouvent souvent à des endroits particuliers de la séquence (début ou fin de phrases, par exemple). Ainsi le mécanisme d'attention permet de donner plus de poids à certaines parties de la séquence pour ainsi garder des informations plus "pertinentes" mais aussi réduire le coût computationnel de la mémoire à l'aide d'un vecteur de contexte. Les modèles RNN avec attention ont d'abord été appliqués à la traduction automatique (Bahdanau *et al.*, 2014) avant d'être utilisés en ASR, avec notamment le travail de Jan Chorowski *et al.* (2015) et de Dzmitry Bahdanau *et al.* (2016). Il existe différents types d'attention, selon que le vecteur de contexte est calculé comme la somme pondérée des états cachés de l'encodeur (*soft attention*) ou si un seul état caché est sélectionné via les scores d'attention (*hard attention*). L'attention peut également être calculée à partir de l'ensemble des états cachés ou seulement ceux se trouvant dans une fenêtre plus petite, on parlera alors de *local vs global attention*.

3.1.3.4 La classification temporelle connexionniste (CTC)

Une autre manière de gérer l'aspect séquentiel des données est d'utiliser la classification temporelle connexionniste (CTC). Pour entraîner un système, un modèle acoustique HMM-DNN par exemple, il est nécessaire d'avoir un alignement préalable des unités linguistiques (phones, triphones, etc.) avec le signal. Mais ces alignements peuvent être coûteux à obtenir, particulièrement dans le cas de grands corpus. La classification temporelle connexionniste permet de se passer de cet alignement préalable (Graves *et al.*, 2006). Si nous ne rentrerons pas dans les détails de la formulation mathématique de la CTC, cette approche permet d'obtenir les probabilités d'une séquence de sortie sachant la séquence d'entrées en sommant sur l'ensemble des alignements possibles. [4] Initialement utilisée pour la reconnaissance de phonèmes, la CTC est aujourd'hui également utilisée pour les modèles de langues. Avec l'attention, elles constituent les deux mécanismes qui ont permis le passage de modèles hybrides reposant encore sur des HMM à des systèmes complètement neuronaux.

3.1.3.5 CNN

Les réseaux de neurones convolutifs (*Convolutionnal Neural Network* - CNN) sont un autre type de réseaux de neurones initialement utilisés pour traiter des données sous forme de matrices et sont donc particulièrement utilisés en traitement de l'image (LeCun *et al.*, 1995; Lecun *et al.*, 1998). Les CNN reposent sur l'utilisation de convolutions, une opération mathématique qui peut s'apparenter à un filtrage, chaque filtre permettant de reconnaitre

4. Pour une description détaillée de la modélisation des séquences avec la CTC, voir Hannun (2017)

une caractéristique particulière des données d'entrées, et l'ensemble des filtres constituant le noyau (*kernel*). En plus des couches de convolution, un CNN contient des couches de *pooling* qui permettent de remplacer une valeur de la matrice par le résultat d'une opération sur les valeurs voisines (on utilise souvent le max pooling (Zhou et Chellappa, 1988). Le pooling permet d'éviter les risques de sur-apprentissage et garantit également une bonne robustesse au bruit. Plusieurs systèmes d'ASR à base de CNN ont donc été proposés (Sainath *et al.*, 2013b,a; Qian *et al.*, 2016).

3.1.3.6 Transformer

Si nous ne rentrerons pas de le détail de son architecture, il nous faut également citer le modèle *Transformer* proposé par Ashish Vaswani *et al.* (2017). Cette architecture a la particularité de n'utiliser aucun réseau récurrent ou convolutionnel, réduisant ainsi le temps de calcul, notamment grâce à un mécanisme d'attention appelé "multi-tête". Si son utilisation est majoritaire en traduction automatique, il a donné lieu à de multiples travaux en ASR, comme ceux de Linhao Dong *et al.* (2018), Sheng Li *et al.* (2019) et Hang Le *et al.* (2020) mais est également à l'origine des modèles de langues pré-entrainés BERT (Devlin *et al.*, 2019) et GPT-2 (Radford *et al.*, 2019) très largement utilisés aujourd'hui en TAL.

3.1.4 Systèmes end-to-end

Comme nous l'avons vu précédemment, les approches stochastiques décomposent la tâche de reconnaissance automatique de la parole en modèle acoustique, modèle de langue et lexique. Ce découpage présente l'inconvénient de devoir entraîner chaque modèle séparément. Des approches bout-en-bout ou *end-to-end* (E2E), qui avaient déjà fait leurs preuves sur des tâches de traduction automatique, ont alors été proposées pour estimer directement $P(W|X)$ dans l'équation (3.1), à l'aide de réseaux de neurones.

Dans ces nouvelles architectures *end-to-end*, le signal de parole est d'abord traité par un encodeur qui remplace la paramétrisation du signal. Puis le décodeur permet d'obtenir une transcription de la séquence audio fournie en entrée. Encodeurs et décodeurs sont des réseaux neuronaux composés de différentes couches combinants LSTM, GRU, CNN, etc. en fonction des architectures.

L'utilisation de l'attention ou de la CTC ayant chacune des inconvénients ont amené à l'utilisation d'apprentissage multi-tâches utilisant les deux techniques. L'attention avait déjà démontré son efficacité dans des approches end-to-end sur de la traduction automatique mais ces modèles avaient parfois du mal à converger. L'utilisation de la CTC en revanche permettait un apprentissage plus stable, mais était limitée par l'hypothèse d'indépendance conditionnelle. Le modèle utilisé dans la boîte à outils ESPnet (Watanabe *et al.*, 2018) que nous avons utilisé est un exemple d'une utilisation conjointe de ces deux

méthodes : ainsi l'attention et la CTC sont toutes les deux utilisées pour optimiser le modèle encodeur/décodeur et le décodage est également fait en utilisant une fusion de la CTC et d'un mécanisme d'attention.

Dans les approches E2E, la modélisation linguistique est assez faible. Si cela n'est pas nécessaire, il est d'usage d'utiliser à l'étape du décodage, des ressources externes dans le but d'augmenter le pouvoir prédictif du modèle. Le décodage est donc souvent fait en utilisant un modèle de langue externe (souvent sous forme de RNN), appris sur de très grands corpus de texte. Si les modèles E2E sont aujourd'hui état de l'art et ont permis un saut dans les performances, ils sont cependant particulièrement gourmands en données d'apprentissage et ont l'inconvénient d'être assez opaques dans leurs modélisations.

3.2 Métrique et campagnes d'évaluation

3.2.1 Les campagnes d'évaluation : tâches structurantes de la recherche

La recherche en ASR et en TAL a été largement conditionnée par de grandes campagnes d'évaluation. Les premiers progrès ont été réalisés lorsque, en plein contexte de guerre froide, la DARPA (*Defense Advanced Research Projects Agency*) a imposé un agenda de recherche en finançant largement la recherche nord-américaine en traitement automatique de la parole. Ces financements ont permis de grands développements dans la recherche tout en pérennisant en parallèle, l'organisation de campagnes d'évaluation, pour mesurer les progrès faits par les équipes de recherche et justifier des dépenses. Encore aujourd'hui NIST (*National Institute of Standards and Technology*) propose des campagnes d'évaluation (notamment la campagne OpenSAT). Ces campagnes régulièrement permettent de visualiser sur le long terme les évolutions dans le domaine (voir Figure 12). Ce fonctionnement de la recherche en TAL a donné lieu à ce que Gilles Adda *et al.* (1998) appelle *the Evaluation Paradigm* et qui structure la recherche au rythme des campagnes. Ces campagnes d'évaluation s'organisent selon 4 phases :

— la phase d'entraînement (*training phase*) : après la diffusion du corpus d'entraînement aux différentes équipes participantes, les équipes bénéficient d'un temps pour développer et calibrer leur système ;

— la phase d'essais (*dry-run phase*), pendant laquelle est réalisée une première évaluation, sur une petite quantité de données, souvent de très bonne qualité (*gold standard*). Suite à cette phase, des réajustements des systèmes peuvent encore être réalisés ;

— la phase de test (*test phase*) : un corpus de test est distribué pour l'évaluation finale de l'ensemble des systèmes ;

— la phase d'adjudication (*adjudication phase*), qui implique la validation des résultats de chacune des équipes. Cette phase donne souvent lieu à l'organisation d'un workshop durant lequel les différents systèmes seront présentés, et leurs avantages et limites discutées par la communauté.

Les discussions et échanges suscités autour de ces campagnes a permis de mutualiser les efforts pour développer des technologies performantes sur une courte durée. Mais au-delà de l'émulation scientifique créée autour de ces campagnes d'évaluations, celles-ci ont également permis de la mise en commun de ressources et de données : on citera par exemple Sphinx de l'Université Carnegie Mellon (CMU) (Lee *et al.*, 1990), Julius de l'Université de Kyoto University (Lee *et al.*, 2001) et Kaldi de l'Université Johns Hopkins (Povey *et al.*, 2011). La constitution de nombreux corpus dans le cadre de ces campagnes d'évaluation, dont le corpus TIMIT (résultant d'un partenariat entre Texas Instrument et le MIT) (Garofolo *et al.*, 1993), le corpus DARPA Resource Management (RM) (Price *et al.*, 1988), les données du Air Travel Information Service (ATIS) (Hemphill *et al.*, 1990), mais aussi les corpus téléphoniques Switchboard (Godfrey *et al.*, 1992) et Fisher (Cieri *et al.*, 2004), a mené à la création du *Linguistic Data Consortium* (LDC) en 1992 pour gérer la distribution des ressources langagières nécessaires pour le support et le développement de la recherche en TAL (Liberman et Cieri, 1998).

Comme le souligne Patrick Paroubek *et al.* (2007), la France et l'Europe ne disposent pas de structures similaires au partenariat DARPA/NIST. En 1995, l'association ELRA (*European Language Resources Association*) est créée (Choukri et Nilsson, 1998), dans le but de combler ce manque, puis celle-ci se transformera en ELDA (*Evaluation and Language Resources Distribution Agency*). ELRA/ELDA sera à l'initiative de l'organisation des conférences LREC (*Language Resources and Evaluation Conference*), avec une première édition en 1998 et qui continue aujourd'hui de rassembler les chercheurs et chercheuses autour de la question des données, ressources et évaluations des technologies du TAL. La publication du journal *Language Resources and Evaluation Journal* débuta également en 2005. Mais si des campagnes ont été organisées à l'échelle française et/ou européenne, elles n'ont pas le caractère permanent des campagnes NIST (Mariani, 2005). On peut néanmoins citer le programme TECHNOLANGUE [5] en France, qui a donné lieu aux campagnes ESTER décrites en section 5.1.1

Le paradigme d'évaluation a donc imposé un certain rythme dans la recherche, orientant les tâches traîtées mais a également contribué à une certaine harmonisation des pratiques. Pour citer Patrick Paroubek *et al.* (2007, p. 21) : « Since evaluation aims at providing a common ground to compare systems and approaches, it is by its nature an activity that is both a source and a user of standards ». Ainsi le corpus TIMIT sera à l'origine de la standardisation du taux d'échantillonnage de 16kHz et de la quantization 16 bits (Pallett, 2003). Les campagnes NIST ont également été à l'origine du script d'évaluation

5. `http://technolangue.net/`

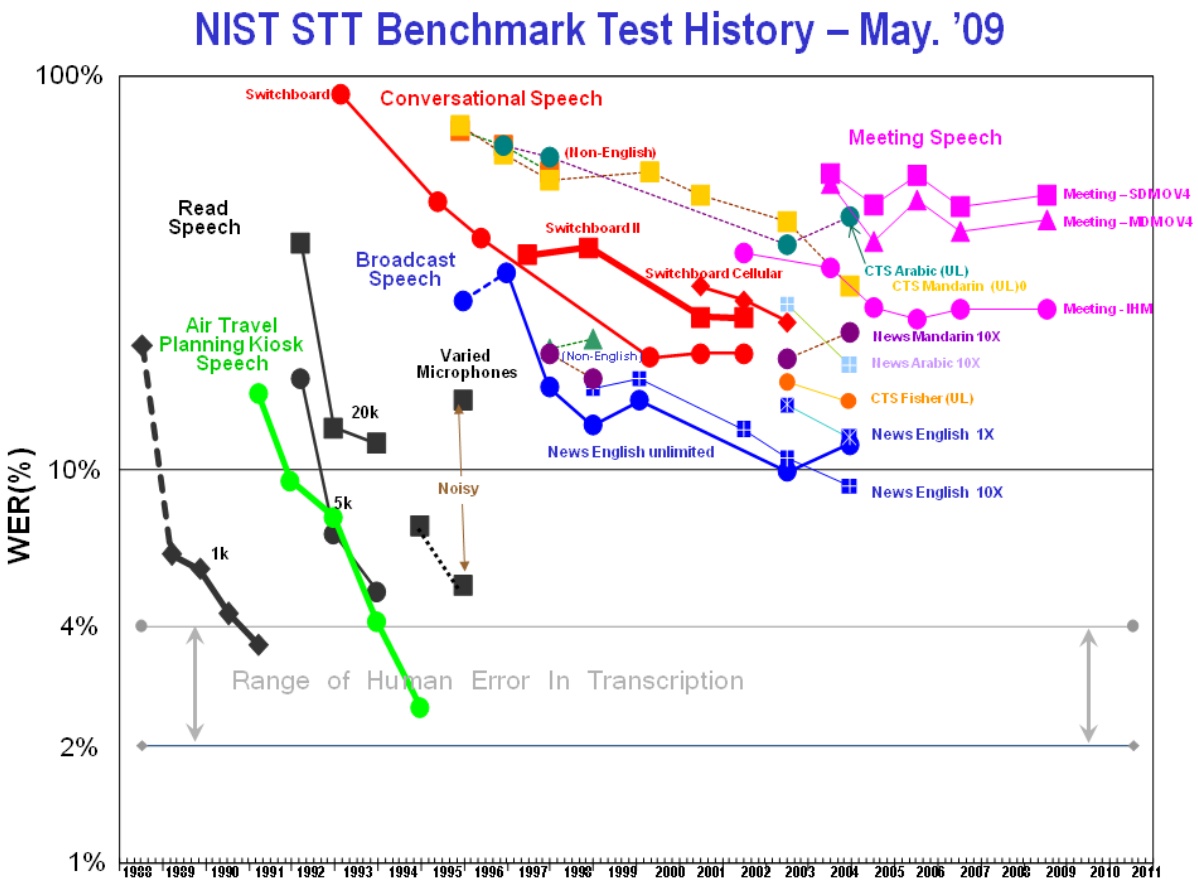

FIGURE 12 – Évolution des résultats obtenus par les campagnes de *Rich Transcription* de NIST jusqu'à mai 2009. La campagne Rich Transcription a depuis été intégrée dans le programme OpenSAT qui continue de travailler sur des tâches d'ASR.

SCLITE, aujourd'hui implémenté dans le package SCTK [6] (*Speech Recognition Scoring Toolkit*) librement distribué et largement utilisé par la communauté. Le paradigme d'évaluation a également trouvé toute sa place dans les grandes conférences du domaine avec l'organisation de *shared tasks* dans des sessions spéciales, chacune permettant d'adresser collectivement des complexités du domaine, avec par exemple la reconnaissance automatique de la parole d'enfants non-natifs et non-natives durant INTERSPEECH 2020 [7] ou la reconnaissance de parole "atypique" de l'édition 2021. [8]

3.2.2 Le WER

Lors des campagnes d'évaluation et plus généralement pour évaluer un système d'ASR, on utilise la métrique du WER (*Word Error Rate*) ou taux d'erreur-mot. Le WER est basé sur la distance de Levenshtein et se calcule comme la somme des erreurs (insertion, délétion et substitution) de l'hypothèse divisée par le nombre de mots total dans la référence. Il se calcule, après alignement dynamique de l'hypothèse de transcription et de sa référence, selon la formule suivante :

$$WER = 100 * \frac{I + D + S}{N} \tag{3.7}$$

où I représente le nombre d'insertions, D le nombre de suppressions, S le nombre de substitutions et N le nombre de mots total de la transcription de référence. Sa valeur peut être supérieure à 100% du fait de la présence d'insertions. D'autres mesures utilisées sont celles du CER (*character error rate*) et du PER (*phone error rate*). Lorsque l'on rapporte les performances d'un système, il est d'usage de calculer le WER à l'échelle du corpus de test.

3.3 Genèse de la tâche et variation individuelle

3.3.1 Bref historique de l'ASR [9]

Les premiers travaux reliés directement à la reconnaissance automatique de la parole sont ceux de Jean Dreyfus-Graf, inventeur du premier phonétographe et créateur du "Phonétographe III" en 1961, qui était capable d'écrire, sous dictée de Jean Dreyfus-Graf, toutes les lettres de l'alphabet (Dreyfus-Graf, 1961). Les années 1950 ont vu émerger les premiers systèmes analogiques : aux États-Unis, Ken Davis, R. Biddulph et Stephen Balashek au Bell Labs réalisent le premier reconnaisseur de chiffres, "Audrey" permettait

6. `https://github.com/usnistgov/SCTK`
7. `https://sites.google.com/view/wocci/home/interspeech-2020-special-session`
8. `https://sites.google.com/view/atypicalspeech-interspeech2021`
9. Cette sous-partie est largement inspirée du chapitre XVII de Calliope (1989) et du chapitre 1 de Kamath *et al.* (2019).

de reconnaître les dix chiffres isolément pour un locuteur donné à l'aide d'un dispositif câblé (Davis *et al.*, 1952). Parallèlement ,la première "machine à écrire phonétique" est présentée en 1956 par Harry Olson et Herbert Belar du RCA Laboratories et permettait de reconnaitre dix syllabes isolément, prononcées par un même locuteur (Olson et Belar, 1956). Durant les années 1960, les tâches adressées se sont donc concentrées sur la reconnaissance de petits vocabulaires ou de voyelles. L'utilisation de l'ordinateur a été introduite par James Forgie et Carma Forgie au Lincoln Laboratory, en 1959. L'échantillonnage des sorties d'un banc de trente-cinq filtres était analysé par leur programme pour permettre la reconnaissance de dix voyelles différentes, grâce à la valeur des deux premiers formants et sans adaptation au locuteur (Forgie et Forgie, 1959). Les années 1960 ont vu s'imposer l'utilisation de l'ordinateur pour la reconnaissance de la parole mais également le traitement du signal (Sholtz et Bakis, 1962). La taille du vocabulaire reconnu augmenta progressivement et on commença à intégrer des informations de plus haut niveau pour aider à la reconnaissance : en 1965, à l'université de Kyoto, Shuji Doshita présente un système de reconnaissance de parole continue pour le japonais, basé sur l'utilisation de tri-phones pour modéliser les phénomènes de coarticulation présents dans la parole continue (Doshita, 1965). Puis des informations syntaxiques ont également été prises en compte avec notamment les travaux de Jean-Pierre Tubach, qui a utilisé les niveaux syntaxiques et sémantiques pour lever des ambiguïtés et corriger des erreurs sur les mots reconnus. Cela était possible grâce à l'utilisation d'une grammaire hors contexte, qui prédisait les mots suivants possibles respectant la syntaxe définie, pour en favoriser la probabilité (Tubach, 1970), introduisant ainsi les modèles de langues.

Les années 1970 ont vu l'émergence des premiers systèmes de reconnaissance de parole continue, notamment aux États-Unis, où le département de la Défense finança un vaste projet, ARPA/SUR (*Advanced Research Projects Agency/Speech Understanding Research*), de 1971 à 1976. Il donnera naissance aux systèmes HARPY, HEARSAY I et II, et HWIM. Parallèlement en France, plusieurs systèmes virent également le jour : Myrtille I et II au CRIN de Nancy, Keal au CNET de Lannion, Esope au LIMSI à Orsay, le système de Groc et Tufelli à l'ICP Grenoble et Arial au CERFIA de Toulouse. Mais face à la complexité de la tâche, et motivés par l'objectif de produire des systèmes industriels, la majorité des travaux se sont focalisés sur la reconnaissance de mots isolés prononcés par un seul locuteur. Le premier système de reconnaissance de mots isolés est commercialisé en 1972 par Threshold ; le VP100 était capable de reconnaître 32 mots prononcés par un locuteur unique. Puis les travaux se sont intéressés à la reconnaissance de mots isolés pour plusieurs locuteurs et locutrices, notamment via des techniques d'adaptation des modèles (évoquées en section 3.3.2).

L'insertion des HMM pour traiter les aspects temporels du signal de parole vers la fin des années 1970 par Leonard Baum, a fait basculer le domaine dans des approches statistiques (comme décrites en section 3.1.2) à partir des années 1980. Les réseaux de neurones

ont également été introduits à cette période pour estimer les probabilités d'émissions des modèles acoustiques HMM, mais les modèles HMM-GMM sont restés le standard jusqu'au début des années 2000. Les financements de la DARPA et les campagnes d'évaluation NIST ont amené à la réalisation de divers outils pour soutenir l'ASR : comme Sphinx de la Carnegie Mellon University et Decipher du Standford Research Institute (aujourd'hui SRI), ou HTK de Cambridge. Les années 1990 ont été marquées par l'utilisation d'apprentissage automatique amenant à la commercialisation de systèmes de dictée vocale comme Dragon, qui était adapté à la voix de l'utilisateur ou utilisatrice. Après un retour en force des réseaux neuronaux dans les modèles acoustiques, les réseaux de neurones sont utilisés pour l'ensemble des systèmes depuis 2012.

Il est intéressant au regard de ce bref résumé de l'histoire de l'ASR de noter la place qu'ont joué les gouvernements dans le développement de la technologie. Jacqueline Léon (2015), dans son histoire de l'automatisation des sciences du langage (principalement focalisée sur la traduction automatique), insiste sur cette interpénétration entre sciences et ingénierie, amenée par les enjeux politiques sécuritaires derrière le développement des technologies du TAL. Si elle écrit que « [d]ans le cadre de la culture scientifique de guerre, la linguistique n'a pas de place » (Léon, 2015, p .12), il est intéressant de retirer de son propos que ce contexte a amené à une conception du langage comme donnée, et la parole comme contenant du message, entrant ainsi en tension avec une approche plus linguistique dans laquelle la parole est un objet d'étude et un système en soi.

3.3.2 La place du locuteur (et de la locutrice)

L'objectif final de la reconnaissance automatique de la parole est d'accéder au message et donc au contenu lexical et sémantique d'un signal de parole. De fait, l'idée a donc d'abord été de chercher les invariants de la parole. Tout ce qui relevait de la variation, qu'elle soit acoustique (changement de canal, d'environnement sonore) ou phonostylistique a été considéré comme du bruit. On a donc cherché à proposer des méthodes paramétriques ainsi que des techniques de normalisation pour gommer la variation due à l'environnement (téléphone, radio, variabilité des microphones, etc.) mais également celle due au locuteur ou à la locutrice (Ono *et al.*, 1993; Lee et Rose, 1996; Wegmann *et al.*, 1996; Garcia *et al.*, 2011). Les premières recherches s'étant intéressées à la parole mono-locuteur dans un contexte masculin (l'armée), l'évolution à de la reconnaissance de la parole multi-locuteur a donc cherché à retrouver une unicité de modèle face à la variation inter-individuelle. Plusieurs techniques de normalisation ont donc été proposées pour adapter les systèmes aux locuteurs et locutrices.

La technique de VTLN (*vocal tract length normalization*) se base sur le postulat que la variation inter-individuelle est en partie due à des variations de taille du tractus vocal, particulièrement pour les variations de genre, différence que l'on va chercher à gommer

via un coefficient de normalisation. En parallèle de ces opérations de normalisations, plusieurs méthodes ont été proposées pour adapter les modèles et passer d'une modélisation indépendante du locuteur ou de la locutrice à une modélisation adaptée ou dépendante de l'individu. Ces méthodes se classent en trois catégories : MAP (maximum a posteriori), MLLR (*maximum likelihood ratio*) et les méthodes basées sur le clustering ou les *speaker spaces* (Woodland, 2001).

Dans le cas des modèles HMM-GMM, il était courant d'entraîner un modèle acoustique différent pour les hommes et pour les femmes, comme écrit par Dan Jurafsky et Martin (2009) :

> « Since women and men have different vocal tracts and other acoustic and phonetic characteristics, we can split the training data by gender and train separate acoustic models for men and women. Then, when a test sentence comes in, we use a gender detector to decide which of the two acoustic models to use. Gender detectors can be built out of binary GMM classifiers based on cepstral features. Such gender-dependent acoustic modeling is used in most LVCSR systems. »(Jurafsky et Martin, 2009, p.387)

On constate donc que la variation de genre est donc conçue soit via un prisme biologique (VTLN), soit évacuée par la recherche de modèles indépendants du locuteur et de la locutrice. Si la conception du genre en ASR mérite d'être réfléchie à la lumière de travaux en phonétique et socio-phonétique, il est intéressant de souligner que l'absence de réflexion à ce sujet et la quête d'invariants et d'indépendance des systèmes vis-à-vis de la variation individuelle a amené à l'absence d'évaluation en fonction du genre, alors même qu'il existait en parallèle un discours sur la complexité des voix des femmes.

Dans les grandes campagnes d'évaluations, l'objet de l'évaluation étant le système, face à une tâche commune, il est de coutume de reporter les résultats à l'échelle du corpus de test. Dans le cas de certaines tâches, les résultats peuvent être présentés en fonction de certains types de variations considérées comme posant des problèmes techniques (environnement sonore, phonostyle) comme dans les campagnes CHiME [10] (Barker *et al.*, 2015) et MGB [11] (Bell *et al.*, 2015). Mais peu d'études se consacrent à la variation des performances en fonction des caractéristiques des individus. Le sexe, l'âge, l'appartenance ethnique, qui sont pourtant des variables protégées aux yeux de la loi française, ne sont pas prises en compte explicitement comme facteurs de variation des performances des systèmes, ce qui est justifié par l'aspect fortement international (voire universel) du paradigme d'évaluation. Pour autant les campagnes d'évaluation ayant amené à la cristallisation de pratiques dans le domaine, certains aspects concernant l'évaluation sont donc restés à l'échelle du système, dans une vision "ingénieriste" de la parole comme donnée, décontextualisée de l'individu et du contexte à l'origine de sa production.

10. `https://chimechallenge.github.io/chime6/`
11. `http://www.mgb-challenge.org/`

Conclusion

Après avoir défini la tâche de reconnaissance automatique de la parole à travers ses principes et ses approches, nous avons réinscrit les systèmes dans une histoire à travers la description des grandes campagnes d'évaluation francophones et anglophones, qui sont à l'origine de nombreuses pratiques en usage encore aujourd'hui. Parmi ces pratiques, on retrouve la métrique d'évaluation qu'est le *word-error rate* ou WER et les habitudes de report de performances globales, gommant la variabilité des contextes de production de la parole en situation écologique.

Si nous posons la question de l'existence de biais prédictif genré dans les systèmes, c'est au regard des impacts que peuvent avoir ces technologies sur les personnes. Or, en retraçant l'histoire du développement de la technologie, on observe que les variations individuelles dans la parole se sont souvent retrouvées décrites et comprises comme des écarts à normaliser, pour accéder au contenu stable, à savoir le message. Ce choix d'une tâche restreinte et contrôlée, du fait des limites techniques de la technologie de l'époque, contribue encore aujourd'hui à concevoir l'ASR via une configuration canonique de reconnaissance de la parole (lue) d'un homme et tout écart à ce standard constitue un ajout de difficulté. Cette perspective nous conforte dans l'idée que poser la question de l'existence de biais prédictifs en ASR est une question importante pour le domaine, car elle pose également la question de quelle vision du monde et de la parole ces systèmes modélisent ?

Chapitre 4

Problématique de recherche

La question à la base de ce travail était de savoir s'il existait un biais de genre dans les systèmes d'ASR. Cette question est partie de l'émulation autour des "biais de l'intelligence artificielle" et notamment du travail de Joy Buolamwini et Timnit Gebru dans leur étude *GenderShades* (2018). Si les systèmes d'IA reproduisent les déséquilibres de nos sociétés alors certainement que les disparités de genre devaient se retrouver en ASR. Mais de cette question a émergé une problématique de recherche plus large qui mérite d'être précisée. Qu'est-ce que nous entendons par biais ? Par genre ? Comment des disparités sociales trouvent leur place, s'immiscent dans des systèmes technologiques ?

La notion de biais, comme nous l'avons vu en chapitre 1 est complexe et polysémique. Le biais que nous adresserons dans ce travail est le biais prédictif, qui suppose une disparité de traitement par le système d'ASR, entre les catégories genrées binaires homme/femme [1].

Le concept de genre, étant le produit d'une évolution historique et ayant des acceptions différentes selon les approches théoriques, est adopté ici selon une définition hybride : on considère le système binaire basé sur les catégories sexuées homme/femme dont découlent des différentiels de pouvoir, de représentation ainsi que des attributs, valeurs et comportements considérés comme normaux, et donc normés, pour chaque catégorie. Cette conception du genre nous permet de nous intéresser à l'invisibilisation des femmes dans la technologie, à la construction du discours sur la difficulté représentée par la voix des femmes, et donc à la reproduction de ces rapports de pouvoir dans les systèmes d'ASR. Mais en questionnant les discours essentialisants, notamment en nous appuyant sur les travaux en sociophonétique oeuvrant à déconstruire cette conception binaire du genre dans la voix, et en montrant comment la technologie contribue également au maintien de ces catégories, nous cherchons également à penser la non-binarité de la voix genrée, du genre et comment cette dernière peut s'articuler avec les technologies d'ASR. En posant ces questions, nous retombons donc à la fois sur une vision performative du genre, par l'individu en contexte, ce qui nous ramène à la question de la prise en compte de la variation individuelle dans les systèmes d'ASR.

Dans leur article sur les biais dans le TAL, Su Lin Blodgett *et al.* (2020) soulignaient un ensemble de questions particulièrement pertinentes et qui entrent en écho avec la volonté derrière le présent travail. Cette volonté était celle de mettre en discussion plusieurs

1. Le choix de ces catégories fait l'objet d'une discussion dans le chapitre 5.

disciplines, à savoir la sociolinguistique, les études de genre et le TAL, pour dépasser la simple question du "existe-t-il des biais dans les systèmes d'ASR" et aller tenter de rendre compte des mécanismes à l'oeuvre dans les systèmes, dans une approche peut-être moins technique que socio-technique. Pour éviter de se perdre dans des débats flous aux termes polysémiques et aux objectifs vagues, Su Lin Blodgett *et al.* (2020) proposent aux chercheurs et chercheuses de garder à l'esprit un ensemble de questions sur l'interaction entre langages, langues et hiérarchies sociales. Nous reprenons ici ces différentes interrogations comme autant de points de départ à la réflexion derrière notre travail sur l'ASR et les biais prédictifs genrés.

La question générale interroge comment les hiérarchies sociales et les idéologies de langage influencent les décisions pendant le développement et le cycle de vie d'une technologie. Quels types de systèmes en résultent ? Comment la définition de la tâche rend compte de cette conception sociale ? Quelles normes linguistiques pérennisent et légitiment ces systèmes, quelles pratiques sont considérées comme standards ? Mais également pour qui les systèmes sont-ils développés ? Et comment l'implémentation d'un système discrétise le monde à travers des attributs démographiques pré-définis ? Comment sont conceptualisées les catégories de genre ? Quel impact cette conception a-t-elle sur les systèmes ? Comment sont constitués les corpus, comment sont-ils annotés, pré-traités et quelle est l'influence de ces annotations et pré-traitements ? Comment les systèmes sont-ils ensuite évalués ? Que représentent ces métriques ? Que permettent-elles de mesurer ? Que ne considèrent-elles pas ? Nous n'avons pas prétention à répondre à la totalité de ces questions, néanmoins, ces dernières ont guidé l'ensemble de ce travail.

Le travail de thèse ne pouvant se focaliser que sur un petit objet et un petit contexte, nous avons choisi d'essayer de tracer la génétique d'un système d'ASR à l'aune du genre. Il nous semble donc pertinent de se poser la question de la représentation du genre dans nos données. Si les biais prédictifs peuvent émerger du fait de biais de sélection, alors la première étape est d'essayer de décrire les pratiques actuelles concernant l'annotation du genre dans les données. La sous-représentation historique des femmes dans la majorité des espaces de parole publics et notamment des médias, nous laissent poser l'hypothèse que les femmes seront également sous-représentées dans les données disponibles pour l'ASR. Nous explorerons ces relations entre données et représentations des catégories genrées dans le chapitre 7.

Ce lien entre données et performance est ensuite creusé de manière plus technique à l'aide d'expériences mobilisant deux types de systèmes différents : un système hybride et un système E2E. À la question existe-t-il des biais prédictifs dans les systèmes d'IA, nous ajoutons celle de savoir si les systèmes E2E, en diminuant la part de modélisation et donc de contrôle, extrapolent ces disparités. Les apprentissages E2E accordent encore plus de poids aux données. La question se pose également de savoir si la variation de la représentation du genre amène un problème de robustesse dans les systèmes. Ou pour reformuler :

est-ce que les performances du système d'ASR sur les catégories de genre sont directement conditionnées par la représentation des catégories dans les données d'apprentissage ? (Chapitres 6 et 8)

Enfin, la dernière question que nous essayerons de traiter est celle d'une sortie de la binarité. Si la variable du genre a été utilisée pour développer les systèmes et si son utilisation est aujourd'hui peu questionnée, nous souhaitions explorer la possibilité de sortir d'une catégorisation genrée (chapitre 9). Est-il possible de décrire nos corpus d'une autre manière, pour garantir les performances du système, tout en sortant d'une vision binaire de la voix genrée, à l'instar des travaux en sociophonétique ?

Deuxième partie

Méthodologie

Chapitre 5

Cadre méthodologique

Ce chapitre présentera l'articulation faite entre les questions générales posées par notre approche théorique et les réalisations techniques de nos contributions. Dans une première partie nous présenterons l'ensemble des données sur lesquelles nous avons travaillé, à savoir les grands corpus médiatiques du français, le corpus Librispeech et la plateforme de dépôt OpenSLR. Dans un second temps, nous présenterons les architectures des systèmes que nous avons utilisés, ainsi que les métriques choisies pour l'évaluation. Enfin, nous reprendrons chaque question de recherche pour présenter nos différents plans d'expérience.

5.1 Données

Un des principes que nous posons dans notre travail est celui de l'importance des données. Si ce postulat n'a en soi rien de novateur, il s'inscrit dans une démarche de science ouverte et il nous semblait important de ne pas considérer comme anodins les différents facteurs nous ayant conduits à travailler sur nos corpus. Le présent chapitre consistera donc en une description détaillée des choix de données que nous avons faits au cours de la thèse.

Nous avons travaillé sur deux langues : le français et l'anglais, dans deux types d'interactions bien définies, à savoir la parole médiatique (enregistrements radiophoniques et télévisuels) et les livres audio (ou *audiobooks*). Le choix de ces données trouve une justification double : nous avons d'abord rencontré une limite de disponibilité, en effet, les systèmes d'apprentissage automatique de reconnaissance automatique de la parole sont gourmands et nécessitent de grandes collections de données d'apprentissage. Les corpus médiatiques francophones que sont ESTER1, ESTER2, ETAPE et REPERE ainsi que le corpus anglophone de Librispeech remplissaient donc cette contrainte quantitative. Plus récemment des corpus ouverts comme CommonVoice de la fondation Mozilla ont vu le jour et pourraient être utilisés pour étendre nos analyses à d'autres langues (Ardila *et al.*, 2020). Une seconde raison à ces choix de données réside dans le type de parole que sont la parole médiatique et les audiobooks. En effet, ces deux pratiques communicationnelles sont fortement normées et les prises de parole sont souvent faites par des professionnel·les ce qui nous permet de poser l'hypothèse que les principales sources de variations seront le fruit de différences liées aux individus.

5.1.1 Les grands corpus médiatiques du français

Comme expliqué dans le Chapitre 3, l'utilisation de l'apprentissage machine en traitement automatique de la parole (TAL) a donné une place centrale aux corpus, qui sont devenus indispensables pour la conception des systèmes. Cependant la production de ces corpus constitue un coût et un investissement que tous les laboratoires ne peuvent pas se permettre (Gravier *et al.*, 1998). L'organisation de campagnes d'évaluations mettant à disposition des corpus de données a permis aux équipes de recherche d'avoir accès à des données de qualité à moindre coût. De ces campagnes d'évaluations sont nés 4 grands corpus du français, ESTER1, ESTER2, ETAPE et REPERE utilisés dans le cadre de la thèse. Un résumé de la constitution de l'ensemble de ces corpus est présenté dans l'Annexe A. Ces corpus ont été utilisés dans nos travaux sur le français, décrits dans le Chapitre 6.

La campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques, appelée ESTER (1 et 2), a eu lieu dans le cadre du projet EVALDA du programme TECHNOLANGUE, entre 2003 et 2005. Soutenue par l'Association francophone de la communication parlée (AFCP), par le Centre d'Expertise Parisien de la Délégation Générale pour l'Armement (DGA/CEP) et par l'European Language Ressources Association (ELRA), elle avait pour but de permettre une évaluation commune des performances des systèmes du traitement de la parole en France. Deux tâches étaient proposées : une tâche de transcription orthographique de contenu radiophonique et une tâche de segmentation. En plus de ces deux tâches, une tâche supplémentaire de reconnaissance d'entités nommées a été rajoutée dans la deuxième phase de la campagne. L'ensemble des catégories et des métriques d'évaluations sont consultables sur le site de l'AFCP. [1] Le corpus ESTER a donc été constitué dans le cadre de cette campagne à partir de 4 sources radiophoniques francophones : *France Inter*, *France Info*, *Radio France International* (RFI), *Radio Télévision Marocaine* (RTM). Le corpus est organisé en deux grandes parties : la première est constituée d'enregistrements faits entre 1998 et 2003, pour un total de 100h de parole annotées manuellement (Galliano *et al.*, 2006). La seconde partie, non annotée, contient 1677h de parole, enregistrées sur les mêmes périodes. Un corpus de test d'une dizaine d'heures, provenant des mêmes sources, ainsi que de deux sources supplémentaires (*France Culture* et *Radio Classique*) a également été fourni. Les émissions de ce corpus de test ont été enregistrées en 2004. Dans l'optique d'une observation longitudinale des progrès réalisés par les systèmes, un deuxième corpus a été constitué : ESTER2. La campagne d'évaluation associée a eu lieu entre 2007 et 2009 et le corpus contient des émissions diffusées entre 2006 et 2008 (Galliano *et al.*, 2009). Les tâches adressées étaient les mêmes que celles d'ESTER1, auxquelles se sont ajoutées la transcription avec données contemporaines et la reponctuation. ESTER2 visait aussi à élargir le type de données pris en

1. `http://www.afcp-parole.org/category/actions-passees/`

compte avec l'ajout notamment de parole accentuée et de parole spontanée. Les ressources acoustiques disponibles pour la campagne étaient les ressources d'ESTER1, complétées par un corpus d'une centaine d'heures contenant des émissions de radio africaines transcrites (provenant d'Africa1 et TVME), dans le but d'étudier l'impact de l'accent. Une partie du corpus EPAC (Estève *et al.*, 2010), transcrite par le Laboratoire Informatique de l'Université du Mans (LIUM) a également été distribuée. Le corpus EPAC provient de la partie non-transcrite d'ESTER1 qui contient les transcriptions d'environ 100h de parole conversationnelle.

Suite aux deux campagnes ESTER, une nouvelle campagne francophone a été lancée entre 2011 et 2012, tout en continuant d'ajouter des difficultés aux tâches de reconnaissance. La campagne ETAPE (Évaluation en Traitement Automatique de la Parole) a donc intégré, en plus d'émissions radiophoniques, des émissions télévisées. Cette utilisation de données télévisuelles a permis de se rapprocher de contexte de parole spontanée, les phénomènes de parole superposée y étant particulièrement importants. Là où les campagnes ESTER s'intéressaient principalement aux émissions de type "nouvelles", ETAPE a choisi de proposer des données plus variées pour permettre une évaluation et une amélioration des systèmes sur un ensemble plus large de contenus médiatiques professionnels (Gravier *et al.*, 2012). Les tâches étudiées reprenaient celles des campagnes précédentes avec une attention particulière pour les tâches de segmentation impliquant de la détection de parole superposée. Le corpus d'ETAPE, de taille plus restreinte que les corpus ESTER, totalise une quarantaine d'heures de parole divisées en 13,5h de radio et 29h de télévision. Les émissions sont principalement des émissions d'informations et de débats mais également des programmes de divertissement, permettant de couvrir ainsi de nombreuses configurations communicationnelles. Les données radiophoniques privilégient les débats favorisant ainsi l'interaction et la parole superposée, avec parfois des conditions acoustiques difficiles, comme dans l'émission *Un Temps de Pauchon* sur France Inter.

La campagne REPERE (Reconnaissance des PERsonne dans des Émissions télévisuelles) s'est intéressée uniquement à la reconnaissance de personnes dans les émissions télévisées (Giraudel *et al.*, 2012). Elle a eu lieu entre 2011 et de 2014, a été financée par l'Association Nationale pour la Recherche (ANR) et la DGA et a été encadrée par le Laboratoire National de Métrologie et d'Essai (LNE). Le corpus, distribué par la société ELDA, est constitué de 60h de parole provenant d'émissions des chaînes BFM TV et LCP. Comme le corpus ETAPE, il regroupe des émissions de type news, débat mais également des programmes de divertissement dans lesquels la proportion de parole spontanée est plus forte. La constitution du premier corpus d'évaluation (corpus de développement et de test) est reportée dans l'Annexe A, mais la constitution précise du corpus complet n'est pas renseignée.

Partitions	Livres	Loc.			Partitions	Livres	Loc.		
		Tot.	F	H			Tot.	F	H
train-clean-100	585	251	125	126	train-other-500	2784	1166	564	602
train-clean-360	2097	921	439	482					
dev-clean	97	40	20	20	dev-other	91	33	16	17
test-clean	87	40	20	20	test-other	90	33	17	16

TABLE 5.1 – Récapitulatif du nombre de livres et nombre de locuteurs et locutrices pour les deux ensembles clean et other du corpus Librispeech.

5.1.2 Un corpus de référence : Librispeech

Pour nos expériences portant sur de l'anglais (présentées dans le Chapitre 8) nous avons utilisé le jeu de données *Librispeech* (Panayotov *et al.*, 2015). Diffusé pour la communauté en 2015, le corpus contient des livres audio en anglais, produits dans le cadre du projet LibriVox[2] débuté en 2005, qui propose à des bénévoles d'enregistrer des lectures de livres tombés dans le domaine public.

L'ensemble de données d'entraînement original contient un total de 5466 livres lus par 2338 locuteurs et locutrices anglophones résidant aux États-Unis. 2671 livres sont lus par des femmes et 2795 par des hommes. Le corpus total est divisé en plusieurs partitions d'apprentissage et de test. Les partitions, initialement réalisées pour des questions de tailles des archives, ont été créées en fonction de critères de "qualité". Les auteurs ont utilisé une procédure automatique pour sélectionner les enregistrements de meilleure qualité et contenant un anglais au plus proche de l'accent standard étatsunien : après avoir entrainé un modèle acoustique sur le Wall Street Journal corpus, ils ont transcrit les livres audio à l'aide dudit modèle et d'un modèle de langue à bigrammes. Après avoir trié leur résultat par WER décroissants, ils ont sélectionné la première moitié de leur données considérée par la suite comme "propre" ou "'clean". Parmi ces données, 20 locuteurs et locutrices ont été sélectionnées aléatoirement pour la partition de développement et l'opération a été répétée pour créer une partition de test. Le reste des données "propres" a été séparé en deux partitions de 100h et 360h respectivement et les 500h restantes constitue la partition d'entrainement "other", comme reporté dans le Tableau 5.1.

La séparation des sous-corpus de développement et de test a été faite pour que chaque locuteur et locutrice cumule environ 8 minutes de paroles. Le test-clean contient 87 livres lus par 40 locuteurs et locutrices différentes, parmi lesquels 49 livres sont lus par des femmes et 38 par des hommes. L'ensemble test-other, quant à lui contient 90 livres lus par 33 locuteurs et locutrices, parmi lesquels 44 livres sont lus par des femmes et 46 par des hommes.

2. `https://librivox.org`

5.1.3 OpenSLR

Open Speech Language Resources [3] (OpenSLR) est une plateforme créée par Daniel Povey, ayant pour objectif de centraliser des ressources langagières accessibles et téléchargeables gratuitement pour aider au développement de systèmes de parole. La plateforme héberge actuellement 114 ressources. [4] Ces ressources sont constituées d'enregistrements audio transcrits, de logiciels, ainsi que de lexiques et de données textuelles nécessaires à la création de modèles de langue. D'autres plateformes comme le LDC ou ELRA/ELDA proposent également ce type de ressources, mais celles-ci sont la plupart du temps payantes.

5.1.4 Corpus et méta-données de genre

Nous avons pu voir que le genre peut s'appréhender soit comme continuum, soit comme système de catégorisation sous-tendu par un rapport de pouvoir. Si nous questionnons une approche catégorielle homme/femme (voir Chapitre 9), il est cependant important de souligner que dans nos données, les informations, lorsque fournies, ne couvraient que ces deux catégories. Nous avons donc fait le choix dans nos expériences concernant l'équilibrage des données d'apprentissages de continuer à s'inscrire dans ces catégories binaires.

Pour questionner le rapport de pouvoir intégré dans la technologie, il nous semblait pertinent de rester dans les catégories utilisées par la communauté. Si la prise en compte de la non-binarité dans les systèmes fait partie des travaux menés actuellement, il n'en reste pas moins que la catégorisation homme/femme est historiquement prédominante car non questionnée et fait donc partie intégrante du contexte socio-culturel dans lequel ont été produites et récoltées ces données. De plus, le discours public sur la non-binarité étant relativement récent, nous postulons qu'il a peu de chances de se retrouver dans nos données, les média et les audio-books constituant, dans la majorité des cas, des lieux de représentation et de réaffirmation des normes sociales, et donc des normes de genre. Nous sommes néanmoins consciente des limites qui accompagnent ce choix et qui constituent donc une limite de notre travail.

5.2 Systèmes d'ASR utilisés

5.2.1 Système hybride : HMM-DNN

Kaldi est un projet open-source qui a débuté en 2009 à l'université John Hopkins aux États-Unis. Il a donné lieu à la publication et à la diffusion de la boîte à outils Kaldi, qui permet le développement de systèmes d'ASR (Povey *et al.*, 2011). Le premier système

3. `http://www.openslr.org`.
4. Dernière consultation au 10 octobre 2021

utilisé dans ce travail de thèse est un système hybride HMM-DNN développé par Elloumi *et al.* (2018) à l'aide de cette boîte à outils.

Les données d'apprentissage regroupent une centaine d'heures d'émissions radiophoniques et télévisuelles francophones, issues des corpus ESTER, ETAPE, REPERE et QUAERO. Une étape d'adaptation au locuteur est faite en appliquant une régression linéaire à maximum de vraisemblance (fMLLR) aux paramètres acoustiques.

Le modèle de langue utilisé est un modèle 5-grammes entrainé grâce à l'outil SRILM (Stolcke, 2002) sur plusieurs corpus de français écrits : les sous-corpus français des corpus EUbookshop, GlobalVoices, Europarl-v7, MultiUN, OpenSubtitles2016, DGT, News Commentary, WMT News Test Sets et Wikipedia distribués via OPUS [5] (Tiedemann, 2012) ainsi que les corpus Gigaword (Graff *et al.*, 2011), TED2013 (Rousseau *et al.*, 2014), Wit3 (Cettolo *et al.*, 2012), Le Monde (NA, 2016), Trames et les transcriptions du corpus d'apprentissage du modèle acoustique, pour un total de 3323 millions de mots. Le modèle étant volumineux, il a été filtré en ne gardant que les n-grammes ayant une probabilité supérieure à 10^{-9}.

Le modèle de prononciation utilise la ressource BDLEX (De Calmès et Pérennou, 1998) ainsi que la transcription graphème-phonème pour obtenir les différentes variantes du vocabulaire (limité à 80000 unités).

5.2.2 Modèle end2end : ESPNET

Nous avons entraîné notre système E2E avec l'outil ESPnet (Watanabe *et al.*, 2018), utilisant une recette déjà existante adaptée au corpus Librispeech : notre modèle est un modèle encodeur-décodeur avec attention, avec un encodeur VGG-BLSTM à 5 couches et un décodeur à 2 couches. Les couches de l'encodeur et du décodeur ont 1024 unités cachées et le vocabulaire de sortie est composé de 5 000 sous-mots générés par *byte pair encoding*. Nous avons utilisé le backend PyTorch pour l'entraînement ASR et le décodage a été effectué en utilisant à la fois un modèle de langage RNN entraîné sur le corpus de texte Librispeech et les scores attentionnels et CTC du modèle ASR (poids CTC=0,5, poids LM=0,7).

Avec cette configuration, nous avons obtenu (avec un modèle appris sur l'ensemble d'entraînement complet) un WER moyen de 4,2% sur l'ensemble de données test-clean et un WER moyen de 14,3% sur l'ensemble test-other. Les résultats rapportés dans le dépôt ESPnet [6] étaient de 4,0% sur l'ensemble test-clean et de 12,7% sur l'ensemble test-other, ce qui nous permet d'affirmer que notre système obtient un ordre de performance équivalent.

5. `https://opus.nlpl.eu`
6. `https://github.com/espnet/espnet/blob/master/egs/librispeech/asr1/RESULTS.md`

5.2.3 Métriques et évaluation

5.2.3.1 Approche critique du WER

Lorsque sont reportés des résultats de systèmes de reconnaissance automatique de la parole, la métrique utilisée est le taux d'erreur-mots ou WER (*word-error rate*). En pratique, le WER est calculé à l'échelle du corpus de test, lissant ainsi les variations dues à la longueur des énoncés. Le développement des systèmes d'ASR s'étant principalement fait à travers des campagnes d'évaluation (voir section 3.2.1), le report d'une mesure unique permettait la comparaison directe des systèmes entre eux, sur des données de test communes.

Mais Patrick Paroubek, en 2007, faisait déjà la différence entre *technology oriented evaluation* et *user oriented evaluation*. En évaluant des systèmes les uns par rapport aux autres, dans le but de faire avancer la recherche et le développement de ces derniers, on s'ancre dans une conception complètement désincarnée du langage ou l'évaluation des perfomances est décorrélée du fait que cette parole est produite de manière située, par un individu, en contexte. Cependant, si la variation individuelle est peu prise en compte, en revanche, certains types de variation sont considérés dans l'évaluation : le MGB Challenge[7] en est un exemple, les performances étant reportées en fonction des différentes émissions, chaque émission étant censée correspondre à un "style" différent, avec des variations de WER pouvant aller jusqu'à plus de 30 points (Bell *et al.*, 2015). De même, si les conditions sonores varient, il sera d'usage de reporter les résultats en fonction de celles-ci. À notre connaissance, en dehors des travaux portant explicitement sur les différences de performances en fonction du genre, Thomas Pellegrini *et al.* (2019) est une des seules études à reporter des WER pour les hommes et pour les femmes. L'absence de prise en compte de ces variations fait l'objet du travail de Piotr Szymański *et al.* (2020). Les auteurs et autrices soulignent que le paradigme d'évaluation et le peu de corpus disponibles en parole conversationnelle, nous laisse supposer que la technologie est aujourd'hui bien meilleure qu'elle ne l'est en réalité. Si les systèmes obtiennent de bonnes performances sur des corpus datant d'une vingtaine d'années, les performances dans des environnements plus écologiques de parole conversationnelle sont bien moins bonnes, avec des performances sur le sous-corpus "*dinner party*" du challenge CHiME5 WER variant entre 46% et 73%. Ils et elles écrivent donc :

> « We need to collect and annotate audio datasets that are much better aligned with contemporary application domains of ASR systems, work on extended and more inclusive acoustic models representing a much broader spectrum of dialects, account for technological advances which influence physical properties of processed audio signals, and develop language models for multi-domain conversations. The situation where most available spontaneous conversation

7. `http://www.mgb-challenge.org/`

> datasets are over 20 years old is both easy to overlook and hard to believe. » (Szymański *et al.*, 2020, p. 3293)

La prise en compte de la variation passe donc également par des procédures d'évaluations différenciées. Pour estimer correctement la capacité des systèmes actuels, les auteurs et autrices proposent notamment de construire de nouvelles mesures de qualité des transcriptions, mais également d'investir les liens entre académie et industrie pour proposer des corpus d'apprentissage et de test en adéquation avec les applications visées aujourd'hui et sortir des contextes très normés des conversations téléphoniques Fisher ou Switchboard ou de la lecture de Librispeech. Le recours au crowd-sourcing, comme dans le cas du corpus Common Voices de Mozilla, constitue des étapes dans la construction de jeux de données plus diversifiés, bien que restant de la lecture. [8]

Si dans notre travail, nous continuerons d'utiliser le WER, nous questionnons cependant le recours aux moyennes. En effet, si une comparaison de moyennes à l'échelle du corpus de test a du sens dans une évaluation orientée système, en faisant le choix de nous intéresser aux biais prédictifs et aux discriminations qui en découleraient, nous nous inscrivons dans un paradigme d'évaluation orientée utilisateurs et utilisatrices. Le recours à un WER moyen ne nous semble pas refléter suffisamment la manière dont la parole de différents individus est retranscrite par le système. Nous faisons donc le choix de reporter des WER médians, en plus des moyennes, ainsi que celui de représenter graphiquement nos distributions de performances à l'aide de boîtes à moustache (Chapitre 6) ou de graphiques en violon (Chapitre 8).

5.2.3.2 Tests statistiques

Nos échantillons de test étant petits et les distributions de WER ne suivant pas une loi normale, nous avons fait le choix de tests non-paramétriques pour évaluer la significativité des différences de performances observées. Nous utilisons le test des rangs signés de Wilcoxon ainsi que sa généralisation à un nombre d'échantillons supérieur à 2, le test de Kruskall-Wallis, pour comparer nos modèles (Wilcoxon *et al.*, 1963).

Nous utilisons également le test de la somme des rangs de Wilcoxon (aussi appelée test de Mann-Whitney), pour comparer nos distributions de WER en fonction des catégories de genre (Wilcoxon, 1945; Mann et Whitney, 1947). Ces tests estiment la probabilité des distributions de WER sur chaque sous-groupe étudié, de provenir de la même population. Nous fixons notre seuil de confiance à 99% ($\alpha = 0.01$).

L'ensemble des scripts d'analyse est disponible sur le dépôt GitHub suivant :

`https://github.com/mgarnerin/phd_scripts`.

8. Il est intéressant de noter que la représentativité est d'ailleurs le coeur de la nouvelle campagne publicitaire du projet comme en témoigne par exemple ce spot publicitaire : `https://www.youtube.com/watch?v=H_3FdvYJlHQ` (dernière consultation le 18/11/21).

Corpus	Émissions	Durée	Medium	Type
Apprentissage	BFM Story	25h 36min	TV	P
	France Info Infos	11h 23min	Radio	P
	France Inter Infos	42h 45min	Radio	P
	LCP Infos	10h 6min	TV	P
	RFI Infos	1h 49min	Radio	P
	Top Questions	7h 59min	TV	P
	Total	**99h 38min**	-	P
Test	Africa1	1h 21min	Radio	P
	Comme On Nous parle	2h 14min	Radio	S
	Culture et Vous	1h 16min	TV	S
	La Place du Village	1h 24min	TV	S
	Le Masque et la Plume	4h 12min	Radio	S
	Pile et Face	7h 52min	TV	P
	Planete Showbiz	1h 12min	TV	S
	RFI Infos	24h 14min	Radio	P
	RTM Infos	22h 0min	Radio	P
	Service Public	2h 30min	Radio	S
	TVME Infos	57min	Radio	P
	Un Temps de Pauchon	1h 31min	Radio	S
	Total	**70h43min**	-	-

TABLE 5.2 – Description des corpus d'apprentissage et de test. P correspond à de la parole préparée et S à de la parole spontanée.

5.3 Plans d'expérience

5.3.1 Évaluer le biais prédictif

Nous nous posons donc la question de l'impact d'une représentation déséquilibrée dans les données d'apprentissage sur les performances d'un système d'ASR. Pour vérifier l'existence de biais prédictif, nous avons travaillé sur un système d'ASR hybride HMM-DNN décrit en section 5.2.1 et entraîné sur les corpus médiatiques du français présentés en section 5.1.1.

Notre corpus d'apprentissage contient 27 085 énoncés produits par 2 506 locuteurs et locutrices, pour un total d'environ 100h de parole. Le corpus de test, quant à lui, contient 74 064 énoncés produits par 1 268 locuteurs et locutrices pour un total de 70 heures de parole. Les données par émissions, type de médium et type de discours sont décrites dans le Tableau 5.2 pour le corpus d'apprentissage et le corpus de test. Les données d'évaluation présentent une plus grande variété d'émissions avec des types de discours à la fois préparés (P) et spontanés (S) (le discours accentué d'une émission de radio africaine est également inclus dans l'ensemble d'évaluation).

Pour évaluer le lien entre représentations du genre dans les données d'apprentissage et performances du système, nous décrivons d'abord la représentation du genre dans les

données d'entraînement. La représentation du genre est mesurée en termes de nombre de locuteurs et de locutrices, mais également en termes de tours de parole et de longueur des tours.

Nous introduisons la notion de rôle du locuteur ou de la locutrice pour affiner notre exploration de la disparité entre les genres, suite à des études qui ont quantifié la présence des femmes en termes de rôle dans les média (Macharia *et al.*, 2015). Dans notre travail, nous définissons la notion de rôle par deux critères quantifiant l'expression du locuteur ou de la locutrice, à savoir le nombre de tours de parole et la durée cumulée de son temps de parole dans une émission. Ces calculs sont réalisés sur la base des transcriptions de discours et des méta-données disponibles. Ces deux critères servent à définir le rôle : un locuteur ou une locutrice est considérée comme parlant souvent (respectivement rarement) s'il ou elle accumule un nombre d'interventions supérieur (respectivement inférieur) à 1% du nombre total de tours de parole dans une émission donnée. Le même processus est appliqué pour identifier les locuteurs et locutrices parlant longtemps ou peu. Ces deux critères divisent nos locuteurs et locutrices en 4 groupes de rôles. Nous obtenons deux rôles saillants, appelés Ancres et Ponctuel·les [9] :

— les Ancres (A) sont au-dessus du seuil de 1% pour les deux critères (de durée et de nombre d'interventions), ce qui signifie qu'ils et elles interviennent souvent et longtemps et occupent donc une place importante dans l'interaction ;

— les Ponctuel·les (P), au contraire, sont en dessous de ce seuil à la fois pour le nombre d'interventions et pour le temps de parole total.

Ces rôles sont définis au niveau de l'émission, c'est-à-dire qu'une même personne peut théoriquement avoir un rôle d'Ancre dans une émission et de Ponctuel·le dans une autre. Cette personne sera alors considérée comme deux locutrices distinctes. Les rôles peuvent être assimilés à la catégorisation "hôte/invité·e" des émissions de radio et de télévision. Les Ancres pourraient être décrites comme des orateurs et oratrices professionnelles, produisant principalement des discours préparés, tandis que les Ponctuel·les sont plus susceptibles d'être des "gens ordinaires". Nous utilisons le terme de rôle pour faire écho à la notion théâtrale du rôle. Ainsi la fonction de présentateur ou de présentatrice sera assumée par les Ancres, qui en revêtiront également les caractéristiques, notamment discursives et acoustiques, la parole médiatique supposant une voix particulière. [10] On peut faire le lien entre notre notion de rôle et celle d'ethos (Maingueneau, 2002) ou de posture ou *stance* (Jaffe, 2009) utilisées en sociolinguistique et qui décrivent, avec leurs différences théoriques :

9. Nous n'avons pas analysé les résultats pour les deux autres groupes, car les effectifs étaient petits.

10. Si c'est le cas dans une moindre mesure pour les émissions radiophoniques et télévisuelles, on peut néanmoins penser à la construction de la voix de JT comme développé par Victoire Tuaillon, dans son documentaire *Et là c'est le drame* disponible sur Arte Radio. `https://www.arteradio.com/son/61658634/et_la_c_est_le_drame`. Si la voix de JT constitue un prototype marqué de voix médiatique, il nous semble pertinent de considéré que cette "voix" doit également se trouver, se travailler, pour d'autres types d'émissions.

> « [la] constitution d'un « soi » relativement stable dans et pour une collectivité, [...] un comportement verbal socialement évalué, qui ne peut être appréhendé hors d'une situation de communication historiquement déterminée. Chaque prise de parole engage une construction d'identité à travers les représentations que se font l'un de l'autre les partenaires de l'énonciation. » (Maingueneau, 2013, p. 2)

Ainsi, une Ancre est plus susceptible d'avoir un temps de parole long de par sa fonction, mais il ou elle aura aussi probablement une façon professionnelle de parler, doublée d'un nombre élevé d'énoncés, augmentant la quantité de données disponibles, ainsi que la qualité de l'élocution, peut-être plus adaptée aux capacités des systèmes. Les Ponctuel·les en revanche, ne bénéficiant pas de représentations prototypiques, sont plus difficiles à définir, mais on peut s'accorder sur le fait que la parole sera moins préparée, moins professionnalisée et peut-être enregistrée dans de moins bonnes conditions (téléphone, micro-trottoirs, etc.).

Comme nous étudions les liens entre genre et rôle du locuteur ou de la locutrice et performance du système, nous avons pris le parti de recalculer les WER pour chaque locuteur et locutrice au niveau de l'épisode (occurrence d'une émission), et non pas au niveau de l'énoncé. Ce choix de granularité nous permet d'éviter les grandes variations du WER qui pourraient être observées au niveau de l'énoncé (en particulier pour les tours de parole courts) mais implique également d'obtenir plusieurs valeurs de WER pour un locuteur ou une locutrice donnée, il ou elle pouvant intervenir dans plusieurs épisodes. Le genre du locuteur a été fourni par les méta-données et le rôle a été obtenu en utilisant les critères définis ci-dessus et calculés pour chaque émission. Nous avons enlevé de notre échantillon les locuteurs et locutrices ne rentrant pas dans nos deux catégories, ce qui a réduit la taille de notre échantillon de 2476 WER/loc/émission à 2384. Ces WER, à l'aide des méta-données, ont ensuite été analysés par catégories de genre et de rôle, à l'aide des tests de Mann-Whitney ou de la somme des rangs de Wilcoxon, comme défini plus haut. Les analyses correspondantes se trouvent dans la section 6.2 du Chapitre 6.

5.3.2 Penser le biais de sélection : analyse des meta-données disponibles

À la base d'un système d'ASR se trouvent donc des données. Si nous postulons que les biais prédictifs genrés proviennent en partie de biais de sélection, analyser la représentation de ces catégories genrées dans les données disponibles pour les systèmes nous semble être un point de départ pertinent pour tenter de comprendre les mécanismes de production de ces types biais. Nous avons donc fait une analyse des corpus de données disponibles pour l'apprentissage et l'évaluation des systèmes d'ASR. Nous avons fait le choix de travailler sur les corpus disponibles sur la plateforme OpenSLR en raison de leur libre accès. Notre

objectif était d'étudier à grande échelle la représentation du genre dans les corpus de parole, OpenSLR nous semblait constituer un échantillon intéressant car ne fournissant pas de format défini et n'ayant pas d'exigences explicites concernant les structures de données, les ressources présentes peuvent être envisagées comme le reflet des pratiques des créateurs et créatrices de ressources concernant les méta-données.

Au moment de notre étude (en novembre 2019), 83 ressources étaient disponibles sur la plateforme. Parmi ces 83 ressources, nous nous sommes focalisée uniquement sur les corpus de parole. Nous n'avons pris en compte ni les versions multiples d'une même ressource ni les sous-ensembles de ressources (par exemple LibriTTS, n'a pas été inclus, car ces données sont comprises dans LibriSpeech). Dans le cas des corpus ayant été publiés en plusieurs versions, seule la dernière version a été conservée (c'est le cas par exemple du corpus TED LIUM). Certains corpus étant multilingues ou pluridialectaux, nous avons, pour ces exemples, étudié chaque langue ou variation séparément. Au total nous avons donc obtenu 66 corpus, dans 33 langues différentes contenant 51 variantes dialectales/accentuelles. Les types de discours sont également variés : on retrouve de la parole élicitée et lue, des émissions radiophoniques, des TEDTalks, mais aussi des enregistrements de réunions, des appels téléphoniques, des livres audio, etc.

Pour évaluer la représentation des catégories genrées dans les corpus, nous nous sommes appuyée sur les notions de "taux de présence" (nombre d'individus) et de "taux d'expression" (temps de parole) utilisées par Doukhan *et al.* (2018).

Après téléchargement, ces informations ont donc été extraites manuellement des corpus, ainsi que l'ensemble des caractéristiques décrites ci-dessous :

- l'identifiant de la ressource (*id*) tel que défini sur OpenSLR
- la langue (*lang*)
- le dialecte ou l'accent s'il est spécifié (*dial*)
- le nombre total de locuteurs et locutrices ainsi que leur nombre dans chaque catégorie de genre (*#spk*, *#spk_m*, *#spk_f*)
- le nombre total d'énoncés ainsi que le nombre total d'énoncés par catégorie de genre (*#utt*, *#utt_m*, *#utt_f*)
- la durée totale, ou temps de parole (en heures), ainsi que la durée par catégorie de genre (*dur*, *dur_m*, *dur_f*)
- la taille de la ressource en gigaoctets (*sizeGB*) ainsi qu'un label qualitatif (*size*, prenant sa valeur entre "grand", "moyen", "petit")
- la fréquence d'échantillonnage des fichiers son (*sampling*)
- la tâche de discours ciblée pour la ressource (*task*)
- le caractère élicité ou non de la parole (*elicited*) [11]

11. Nous définissons comme données de parole non-élicitées, des données qui auraient existé sans la création des ressources (par exemple : TedTalks, livres audio, etc.). Les autres données de parole sont considérées comme élicitées.

— le statut de la langue (*lang_status*) : une langue est considérée comme ayant peu (low-resource) ou beaucoup (high-resource) de ressources. Le statut de la langue est défini d'un point de vue technologique (c'est-à-dire : y a-t-il des ressources ou des systèmes de TAL disponibles pour cette langue ?) Il est fixé à la granularité de la langue (d'où le nom), quel que soit le dialecte ou l'accent (si renseigné)
— l'année de la publication (*year*)
— les auteurs et autrices de la ressource (*producer*).

En plus des caractéristiques générales listées ci-dessus, et face à la difficulté que nous avons rencontrée pour récupérer l'ensemble de ces informations, nous avons ajouté une variable renseignant si des informations sur le genre étaient fournies en premier lieu (attribut *provided*) et le cas échéant de quelle manière (attribut *found_in*). La procédure de recherche était la suivante : nous avons d'abord examiné le fichier de métadonnées (si existant) et dans le cas contraire, nous avons cherché si le genre était indexé dans la structure des données. Si aucune information n'était trouvée, nous avons cherché s'il existait un article décrivant les données. Après recensement des données disponibles sans traitement supplémentaire, nous avons complété les données manuellement lorsque cela était possible, pour maximiser la taille de notre échantillon. Nous avons donc obtenu des données genrées sur les locuteurs et locutrices pour 47 corpus et des nombres d'énoncés pour 41 corpus.

La réalisation de ce travail de recensement nous a permis de constater que les informations sur le temps de parole ne sont pas standardisées. Lorsque des informations de durée sont fournies, la granularité utilisée varie selon les corpus. Certains auteurs indiquent les temps de parole en heures (e.g. Panayotov *et al.* (2015); Hernandez *et al.* (2018)), d'autres le nombre d'énoncés ou de phrases (e.g. Juan *et al.* (2015); Google (2019)), la définition de ces deux termes n'étant jamais explicite. Nous avons également constaté qu'il n'y avait pas de cohérence entre la durée de parole et le nombre d'énoncés, ce qui exclut la possibilité d'approximer l'une par l'autre.

Une version du tableau final est présentée dans l'Annexe C, la version numérique ainsi que le script R utilisé pour l'analyse sont disponibles sur le GitHub [12] et les résultats seront présentés dans le Chapitre 7.

5.3.3 Compenser le biais de sélection

Les précédentes expériences ont permis de penser les liens entre représentation des femmes dans les données et les médias et performances d'un système d'ASR. Nous formulons également l'hypothèse que l'absence de modélisation explicite des systèmes E2E risque d'extrapoler les disparités. Pour évaluer notre hypothèse, nous avons donc entraîné un système E2E en faisant varier la représentation homme/femme dans nos données d'ap-

12. `https://github.com/mgarnerin/phd_scripts`

prentissage. Les systèmes E2E étant particulièrement gourmands en termes de données d'apprentissage, cette expérience n'était pas conductible sur les grands corpus du français. Nous avons donc utilisé le corpus Librispeech décrit en section 5.1.2.

Nous décidons de travailler à la granularité du livre. Notre objectif étant d'évaluer les performances pour des utilisateurs et utilisatrices, le niveau de l'énoncé nous paraissait trop petit et décontextualisé. Une évaluation par locuteur ou locutrice en revanche aurait considérablement réduit la taille de notre échantillon, rendant peu robustes nos analyses statistiques. La granularité du livre nous semblait pertinente car correspondant à un individu, dans un contexte (l'enregistrement d'un livre particulier). Cela signifie que chaque point de mesure correspond au WER obtenu sur un livre particulier. Ce qui implique également que nous pouvons avoir plusieurs valeurs de WER par personne, pour des livres distincts.

Il n'y a pas de chevauchement de locuteurs et locutrices entre les corpus d'entraînement et de test. Pour des raisons de lisibilité, lorsque nous présentons les résultats du WER pour les locuteurs masculins et féminins, nous faisons en fait référence aux résultats du WER obtenus pour des livres lus par des locuteurs masculins ou féminins.

Librispeech ayant été conçu de manière équilibrée en termes de catégories de genre, nous avons recréé 3 ensembles de données d'entraînement en faisant varier le pourcentage de livres lus par des femmes et par des hommes. Nous avons construit 3 sous-corpus, dans lesquels 30%, 50% ou 70% des livres ont été lus par des femmes, afin d'observer l'impact de la représentation du genre sur les performances. Nous avons appelé les ensembles d'entraînement résultants : wper30, wper50 et wper70. Le Tableau 5.3 récapitule le contenu de chaque sous-corpus.

Pour assurer la comparabilité, le nombre total de livres (N=3816) est le même pour chaque ensemble d'entraînement. Ce nombre est choisi pour maximiser l'utilisation des données, c'est-à-dire que les 70% de femmes du wper70 correspondent à la totalité des données disponibles pour les femmes. Les données de chaque catégorie de genre sont sélectionnées à l'aide d'un script Python, qui après avoir défini une graine aléatoire et un partitionnement (dans notre cas 30/70, 50/50 ou 70/30), sélectionne aléatoirement des livres pour constituer les deux sous-corpus de voix d'hommes et de voix de femmes.

Corpus	**Femmes**	**Hommes**	**Total**
train original	2671	2795	5466
wper30	1145	2671	3816
wper50	1908	1908	3816
wper70	2671	1145	3816
test-clean	49	38	87
test-other	44	46	90

TABLE 5.3 – Composition des différents sous-corpus. Les effectifs reportés sont en nombre de livres lus par des hommes ou par des femmes.

Variation étudiée	**Modèle**	**m_seed**	**d_seed**	**Femmes**	**Hommes**	**Total**
Référence	m1-wper50	1	67	1908	1908	3816
Représentation du genre	m1-wper30	1	67	1145	2671	3816
	m1-wper70	1	67	2671	1145	3816
Aléatoire du modèle	m2	42	67	1908	1908	3816
	m3	13	67	1908	1908	3816
Aléatoire des données	d2	1	26	1908	1908	3816
	d3	1	42	1908	1908	3816
Cas limite (monogenre)	féminin	1	-	2671	0	2671
	masculin	1	67	0	2671	2671

TABLE 5.4 – Différentes configurations des systèmes E2E en fonction de la variable observée. m_seed représente la graine aléatoire du modèle et d_seed la graine aléatoire pour la sélection des données. Cette dernière n'a pas de valeur dans le modèle féminin, car toutes les données disponibles sont utilisées.

Réunis, ces deux échantillons mono-genre constituent notre corpus d'apprentissage. La graine aléatoire n'étant pas modifiée entre les différentes partitions, il y a recouvrement entre les données des différents systèmes : les 30% de livres lus par des femmes dans wper30 sont également présents dans les corpus wper50 et wper70, et les 50% du wper50 sont contenus dans le wper70. Il en va de même pour les livres lus par les hommes. Après avoir obtenu nos 3 corpus d'apprentissage, nous avons ensuite entraîné un système avec chacun d'entre eux. Nous nous attendons à de meilleures performances sur la catégorie représentée à 70% et une absence de différence significative pour le système entraîné sur le corpus wper50.

En plus de ces systèmes, nous avons entraîné des modèles mono-genre, pour observer les performances du système dans ce cas limite. Comme expliqué plus haut, les corpus wper30, wper50 et wper70 ayant été constitués pour maximiser la taille des données, les corpus d'apprentissage mono-genre ne sont pas directement comparables, car de taille plus petite. En effet le corpus mono-genre de voix de femmes contient l'ensemble de 2671 livres contenu dans le wper70, soit toutes les données disponibles. Le corpus mono-genre masculin contient également la même quantité de données, soit 2671 livres lus par des hommes, qui sont les 2671 livres contenus dans le corpus wper30.

Pour nous assurer que les différences observées sont bien uniquement dues à la variation introduite par nos différents jeux de données d'apprentissage, nous avons également entraîné deux systèmes avec notre corpus équilibré (wper), en faisant varier la graine d'apprentissage du modèle, qui définit l'initialisation des poids du réseau de neurones.

Enfin, nous avons entraîné 2 systèmes en faisant varier cette fois la graine aléatoire de sélection des données, tout en gardant un équilibre de 50-50 dans nos corpus d'apprentissage, pour essayer de mesurer l'impact de la variabilité individuelle. Les graines aléatoires pour le modèle et pour la sélection des données ont été choisies arbitrairement. Le Tableau 5.4 récapitule ces différentes configurations.

5.3.4 Sortir du genre ?

Enfin, le dernier volet de ce travail de thèse s'est intéressé à penser une manière de sortir des étiquettes du genre. Pour cela, nous avons essayé de regarder s'il existait des mesures acoustiques particulières qui pourraient permettre de s'affranchir de la catégorisation du genre dans les données.

À l'aide d'un script Praat (Boersma et Weenink, 2018), nous avons extrait pour nos corpus d'apprentissage et de test, un ensemble de caractéristiques acoustiques pour chaque signal. Les caractéristiques extraites sont :

— le nombre de syllabes
— le nombre de pauses
— la durée (en secondes)
— la durée de parole (en secondes)
— le débit (en nombre de syllabes par seconde)
— le débit articulatoire (en nombre de syllabes par seconde)
— la durée moyenne d'une syllabe (en secondes)
— la f0 médiane (en Hertz)
— l'écart interquartile de la f0
— le delta f0 calculé comme le quotient de l'écart interquartile par la médiane
— la f0 médiane en semi-tons (100Hz)
— l'écart interquartile de la f0 en semi-tons
— le delta f0 en semi-tons

Le choix de ces mesures se justifie par le fait que nous considérons la fréquence fondamentale comme paramètre saillant dans l'expression de la voix genrée, comme expliqué dans la section 2.2. Les caractéristiques extraites (débit, nombre de pauses, etc.) sont des paramètres acoustiques relevant plus largement du type de parole produite et peuvent être corrélées à des perceptions d'accents et pourraient potentiellement expliquer les variations de WER observées.

Troisième partie

Contributions

Chapitre 6

Évaluer le biais prédictif : étude d'un système d'ASR développé sur de la parole médiatique française [1]

Nous avons montré, dans le Chapitre 3, le rôle que jouent les campagnes d'évaluation et les corpus dans le développement des systèmes d'ASR. Pour ce qui est du français, les grands corpus de la parole disponibles sont des corpus médiatiques, constitués dans le cadre du projet TECHNOLANGUE, à savoir ESTER, ETAPE et REPERE. Mais l'espace médiatique est caractérisé notamment, par un déséquilibre de représentation entre les femmes et les hommes, ce qui a donné lieu à un ensemble de mesures gouvernementales visant à promouvoir une représentation plus égalitaire à l'antenne. Après une reproblématisation de la présence et de la représentation des femmes dans les médias, nous regarderons comment ces déséquilibres s'expriment dans les données issues d'émissions radiophoniques et télévisuelles francophones. Nous mesurerons dans un second temps, l'impact de ces écarts de représentation sur les performances d'un système d'ASR entraîné sur ce type de données. Notre analyse se fera également à la lumière de la notion de "rôle", qui montre selon nous, une fois de plus, comment le monde social façonne les données.

6.1 Présentes. Média, données, systèmes... Quelle place pour les femmes ? [2]

6.1.1 Femmes et média

Nous l'avons déjà écrit, mais les données sont à la base du développement des systèmes. Or celles-ci sont coûteuses, à récupérer ou à produire, et c'est d'ailleurs ce qui a contribué à rendre les campagnes d'évaluation centrales, dans la mesure où elles permettent un accès à des corpus de données de bonne qualité. Une manière d'alléger le coût

1. Ce travail a donné lieu à un article (Garnerin *et al.*, 2019), dont le présent chapitre est partiellement inspiré.

2. Ce titre est largement inspiré de celui de l'essai de la journaliste féministe Lauren Bastide, *Présentes. Villes, média, politique... Quelle place pour les femmes ?*

de constitution des corpus réside dans l'utilisation de données produites par ailleurs : émissions médiatiques, livres audio, etc. L'avantage supplémentaire de ce type de données tient à leurs conditions d'enregistrement : en effet, des enregistrements de parole visant à être diffusés publiquement sont souvent de bonne voire très bonne qualité. Mais une fois les enregistrements récupérés, il faut également transcrire ces données, ce qui rajoute au coût général : on compte environ 8h de travail pour transcrire une heure de radio (Bell *et al.*, 2015). Ce coût des données explique l'utilisation des corpus médiatiques, mais également leur réutilisation : on comprend alors pourquoi les grands corpus du français datent tous d'il y a plus de 10 ans, date des grandes campagnes d'évaluation financées par TECHNOLANGUE. À un moment où de nouvelles architectures systèmes sont publiées tous les ans, les données, elles, ne suivent pas les mêmes évolutions.

Si nous arguons que les données encapsulent des représentations sociales, ce postulat se trouve largement vérifié dans le cas des données médiatiques. Comme l'écrit Éric Macé : « Comme tout objet social, la télévision est une forme particulière de traduction des rapports sociaux en représentations culturelles. » (Macé, 2000, p. 248). Ainsi, si la télévision (et nous considérons, par extension de l'analyse, la radio) n'est pas "le reflet" de la réalité, mais bien « une forme spécifique de construction des représentations de la réalité sociale », il n'en reste pas moins que les inégalités sociales se retrouvent souvent reproduites dans ces réprésentations, notamment en termes de genre. En 1979, l'étude *Image, rôle et condition sociale de la femme dans les médias*, conduite par l'UNESCO questionnait déjà la représentation et l'image des femmes dans les médias. On y lit que « [par] rapport aux autres types de programmes, c'est dans les informations que les femmes apparaissent le moins, à la fois comme journalistes chargés de couvrir les nouvelles et comme sujet d'actualité » (Ceulemans et Fauconnier, 1979, p. 29). Ce constat était également fait par le rapport *Who Makes the News* du Global Monitoring Media Project (GMMP) de 2015, qui montrait à l'échelle mondiale que si l'on comptait 57% de présentatrices télé, elles n'étaient que 41% à la radio, et uniquement 37% parmi les reporters (Macharia *et al.*, 2015). En France, force est de constater que la situation est similaire. La loi 2014-873 du 4 août 2014 pour l'égalité réelle entre les femmes et les hommes a donc ajouté aux missions du Conseil Supérieur de l'Audiovisuel (CSA) de garantir les respects des droits des femmes dans le domaine de la communication audiovisuelle : veiller "à une juste représentation des hommes et des femmes dans les programmes des services de communication audiovisuelle" [3] devient une des fonctions du Conseil et des rapports annuels sont publiés depuis, réalisés en partenariat avec l'Institut National de l'Audiovisuel. Le rapport d'exercice 2017 (dernier en date lors de la réalisation de cette étude) rapportait que les femmes repré-

3. Article 3-1 de la loi 86-1067 du 30 septembre 1986 relative à la liberté de communication, modifié par l'article 56 de la loi 2014-873 du 4 août 2014 pour l'égalité réelle entre les femmes et les hommes. `https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000029330832`

sentent 40% des personnes intervenant dans les média français, ce pourcentage tombant à 29% pendant les heures à forte audience (créneau 18h-20h).

Pour faciliter la réalisation de ces rapports, David Doukhan et Jean Carrive proposaient en 2018 un système permettant de suivre la présence dans les flux télévisuels français des hommes et des femmes (Doukhan et Carrive, 2018). En utilisant un système neuronal à base de CNN, les auteurs segmentent le flux audiovisuel entre musique et parole pour ensuite attribuer automatiquement les extraits de parole à des hommes ou à des femmes. Ce système permet donc d'avoir accès automatiquement à des mesures de leurs temps de parole respectifs. En 2021, le dernier rapport portant sur l'exercice 2020 annonce une présence des femmes en hausse à la télévision entre 18h et 20h (ayant atteint 40%) et une stabilisation à 41% pour la part des femmes présentes à l'antenne, télévision et radio confondues (CSA, 2021). Pour autant, le corpus *RadioTalk*, publié par le MIT et contenant des enregistrements ayant été diffusés entre octobre 2018 et mars 2019 aux États-Unis, reportait que les femmes ne représentaient qu'un tiers de leur corpus (27,8%), que celles-ci étaient plus présentes dans les appels au studio que dans les personnes présentes sur les plateaux (32,6% contre 27,3%) et que les tours de paroles des hommes étaient 21,9% plus longs pour les hommes que pour les femmes (Beeferman *et al.*, 2019). On ne peut donc s'attendre à une amélioration systématique de la représentation des femmes dans l'ensemble des corpus issus de la radio.

Si les représentations médiatiques évoluent au rythme de la société, les données, elles, une fois constituées en corpus, restent fixées au temps de leur création. Ainsi, les corpus ESTER, ETAPE et REPERE cristallisent une représentation de la scène médiatique dans laquelle les femmes sont sous-représentées. Par conséquent, nous supposons que c'est également le cas dans le corpus qui nous a servi de corpus d'apprentissage. Nous avons donc analysé la représentation des femmes dans ces données, pour mettre en lumière les liens entre cette représentation des catégories de genre dans nos données et les performances de notre système.

6.1.2 Femmes et données

Dans leur étude, David Doukhan et Jean Carrive mobilisaient les notions de "taux de présence" et de "taux d'expression" (Doukhan et Carrive, 2018). Nous avons donc repris leur terminologie pour décrire la représentation des femmes dans nos données. Le Tableau 6.1 récapitule cette répartition dans nos données d'apprentissage. Sans surprise, on observe une représentation majoritaire des hommes qui composent plus de 65% des personnes présentes (résultat en accord avec les chiffres donnés par le rapport du GMMP 2015) et occupent 75% du temps de parole. Nos observations, bien que portant sur des données antérieures, sont concordantes avec celles de David Doukhan et Jean Carrive qui observent un taux d'expression des femmes variant entre 7,73% et 47,44% selon les chaînes

	Femmes	Hommes	NA
Taux de présence	33,16% (831)	65,32% (1637)	1,52% (38)
Taux d'expression	22,57% (22h 30min)	75,75% (75h 30min)	1,68% (1h 40min)

TABLE 6.1 – Représentation du genre dans les données d'apprentissage.

(2018, p. 499) et pas si éloignées de celles du dernier rapport du CSA qui reporte un taux de présence de 41% et un taux d'expression de 35% pour les femmes.

6.1.3 Femmes et rôles

Le temps de parole, qui conditionne la quantité de données disponibles pour la constitution de corpus, est largement dépendant du statut de la personne dans l'interaction médiatique. Ainsi, une présentatrice, un chroniqueur ou une experte invitée cumuleront plus de temps de parole, qu'un ou une invitée, venue faire une promotion ponctuelle ou interrogée pour un reportage. Ce point est d'ailleurs important car dans le rapport de 2020, le CSA notait que :

> « [l]e temps de parole des femmes à l'antenne - télévision et radio confondues -, mesuré automatiquement par l'INA, est inférieur au taux de présence et relativement stable par rapport à 2019 (35%), ce qui laisse supposer qu'à présence égale, les femmes s'expriment toujours moins que les hommes. » (CSA, 2021, p.5)

La question se pose alors de savoir si les femmes s'expriment moins, ou si ces dernières ont moins accès à des statuts leur permettant de parler plus longuement. Nous avons donc tenté d'introduire une notion de rôle dans nos analyses, en considérant que cette information éclairait à la fois des caractéristiques du type de parole (professionnelle ou non) mais également une implication vis-à-vis des quantités de données disponibles. Les rôles sont définis en fonction du nombre d'interventions et du temps de parole total (voir Chapitre 5). Nous appelons Ancres les locuteurs et locutrices qui parlent longtemps et souvent (présentatrice, chroniqueur, etc.) et Ponctuel·les, celles et ceux qui interviennent peu et peu souvent (interviewée de micro-trottoir, auditeur via le standard téléphonique,

	Ancres (A)	Ponctuel·les (P)	Autres (O)
Taux de présence	3,79% (95)	92,78% (2325)	3,43% (86)
Taux d'expression	35,71% (35h 36min)	49,59% (49h 25min)	14,70% (14h 39min)

TABLE 6.2 – Représentation des rôles dans les données d'apprentissage.

etc.). Le Tableau 6.2 présente la distribution des rôles dans les données d'entraînement et montre que malgré le faible nombre d'Ancres dans nos données (3,79%), elles concentrent néanmoins 35,71% du temps de parole total, validant ainsi l'approximation faite par nos rôles définis automatiquement.

En croisant les deux paramètres, nous pouvons observer que la distribution des rôles n'est pas homogène parmi les catégories de genre (voir Tableau 6.3). Les femmes représentent 29,47% des Ancres, moins que chez les Ponctuel·les, où elles représentent 33,68%. On peut donc observer que ces dernières sont déjà peu représentées dans notre échantillon global, mais qu'elles le sont encore moins, dans des rôles plus exposés médiatiquement. La distribution homme/femme globale est équivalente à la distribution homme/femme chez les Ponctuel·les, ce qui est assez logique, lorsqu'on voit la disparité de taille d'échantillon, les Ancres représentant moins de 4% de notre échantillon total. En revanche, il est intéressant de noter, que sur nos données, même en tant qu'Ancres, les femmes parlent moins, avec un taux d'expression à 20,89% contre 74,86% pour les hommes. En calculant le temps de parole moyen pour les femmes Ancres, on obtient une valeur de 15,9 min contre 25,2 min pour un homme de la même catégorie. La différence de taux d'expression relevée par David Doukhan et Jean Carrive ne peut donc pas être expliquée uniquement par un différentiel de rôle, mais bien une tendance des hommes à parler davantage à rôle équivalent. Ces disparités dans les données d'apprentissage, au-delà de la description sociologique qu'elles contiennent, nous laissent supposer que des différences vont être observables dans les performances de notre système.

	Rôle	Femmes	Hommes	NA
Taux de présence	A	29,47% (28)	66,32% (63)	4,21% (4)
	P	33,68% (783)	64,95% (1510)	1,37% (32)
Taux d'expression	A	20,89% (7h26min)	74,86% (26h 29min)	4,25% (1h 40min)
	P	25,00% (12h 21min)	74,86% (37h 00min)	0,14%, (4min)

TABLE 6.3 – Représentation des rôles en fonction des catégories de genre dans les données d'apprentissage. Les NA correspondent à de la parole superposée ou des locuteurs et locutrices pour lesquels les méta-données de genre n'étaient pas disponibles.

6.2 Analyse des performances

Après entraînement, nous avons donc évalué notre système sur un ensemble de données de test. Au sein de ces données, la répartition des rôles en fonction des catégories de genre est sensiblement la même que celle observée dans nos données d'apprentissage (voir

	Rôle	Femmes	Hommes
Taux de présence	A	31,82% (70)	68,18% (150)
	P	33,27% (337)	66,73% (676)

TABLE 6.4 – Représentation des rôles en fonction des catégories de genre dans les données de test.

Tableau 6.4). Nous avons reporté nos résultats à la fois par WER moyens (les WER moyens étant habituellement reportés dans les articles) mais également en termes de WER médians (Tableau 6.6). Il est intéressant de noter l'écart qui peut exister entre ces valeurs. Les tests statistiques que nous avons utilisés étant des tests non-paramétriques se basant sur les médianes, reporter les valeurs médianes est également un bon indicateur de la significativité statistique des différences de performances observées.

6.2.1 Performances globales

Notre système présente un WER moyen de 38,10% et un WER médian de 26%. On observe une grande variabilité dans les performances obtenues en fonction des émissions. En effet, comme décrit dans le Chapitre 5, le corpus d'apprentissage du système utilisé contenait uniquement de la parole préparée, alors que le corpus de test également de la parole spontanée, dans le cadre d'émissions de divertissement notamment, comme *Planète Showbiz*. Ainsi, on observe de variations allant jusqu'à 36,51 points de pourcentage pour les WER moyens et 39,50 points pour les WER médians (voir Tableau 6.5).

Émission	Effectif	WER moy.	WER med.
Africa1 Infos	57	33,79%	23,00%
Comme On Nous Parle	80	51,45%	31,50%
Culture et Vous	276	55,58%	52,00%
La Place du Village	30	45,10%	44,50%
Le Masque et la Plume	68	42,57%	30,50%
Pile et Face	123	20,02%	19,00%
Planète Showbiz	312	56,53%	57,50%
RFI Infos	804	31,09%	21,00%
RTM Infos	394	22,04%	18,00%
Service Public	88	50,68%	29,50%
TVME Infos	52	25,44%	22,50%
Un Temps de Pauchon	100	53,28%	48,00%
Total	2384	38,10%	26,00%

TABLE 6.5 – Effectifs, WER moyen et médian et effectifs par émissions.

6.2.2 Impact du genre et du rôle sur les performances

Lorsqu'on s'intéresse aux valeurs par catégories de genre, on observe que les valeurs de WER sont plus élevées pour les femmes avec un WER médian de 30% contre 25% pour les hommes. Cet écart de performance est significatif avec un W = 731854 et une p-valeur = $2{,}34e^{-6}$. Les performances sont également bien meilleures sur les Ancres (22%) que sur les Ponctuel·les (31%) avec un écart de presque 10 points de pourcentage de WER. Cet écart est également significatif (W = 588683 ; p-valeur = $5{,}32e^{-11}$)

Lorsqu'on regarde l'interaction de ces deux facteurs, la différence de rôle se retrouve dans la valeur des WER obtenus, qui sont bien meilleures pour les Ancres que pour les Ponctuel·les. Le WER médian se situe à 22% pour les hommes Ancres et 24% pour les femmes Ancres, alors qu'il monte à 29% pour les locuteurs ponctuels et atteint la valeur maximale de 38% pour les locutrices ponctuelles. Il est intéressant de noter que malgré une quantité de données plus importantes et une parole plus professionnelle pour les Ancres, une différence de genre est toujours observée avec un écart de WER moyen de 10 points de pourcentage (W = 143286 ; p-valeur = 0.0022). D'un point de vue graphique, on observe une queue de distribution plus longue pour les femmes (voir Figure 13), marquant cette différence de performance avec des WER relativement élevés pour certaines locutrices. Concernant les Ponctuel·les, l'écart entre catégories de genre est également significatif (W = 228508 ; p-valeur = $3{,}43e^{-5}$) et va dans le sens attendu, à savoir de meilleures performances pour les hommes, largement plus représentés dans nos données d'apprentissage.

Rôles	WER	Femmes	Hommes	Tous·tes
Ancres	médian	**24,00 %**	**22,00%**	22,00%
	moyen	38,02 %	28,20%	32,18%
Ponctuelles	médian	**38,00 %**	**29,00%**	31,00%
	moyen	48,86 %	39,60%	42,63%
Tous·tes	médian	**30,00%**	**25,00%**	26,00%
	moyen	43,58%	35,00%	38,10%

TABLE 6.6 – WER médian et moyen par rôle et par catégorie de genre.

6.2.3 Effet de l'émission et du type de parole

Pour autant, les variations de genre ne sont pas les plus importantes en termes d'effet. Le système utilisé pour cette expérience est un système développé par Zied Elloumi (2018) dont l'objectif était, à terme, de construire un système de prédiction du WER. Comme expliqué plus haut, les données d'apprentissage contiennent donc uniquement de la parole préparée, alors que le corpus de test contient presque autant de parole préparée que de parole spontanée, permettant ainsi de tester la précision de la prédiction du WER (voir

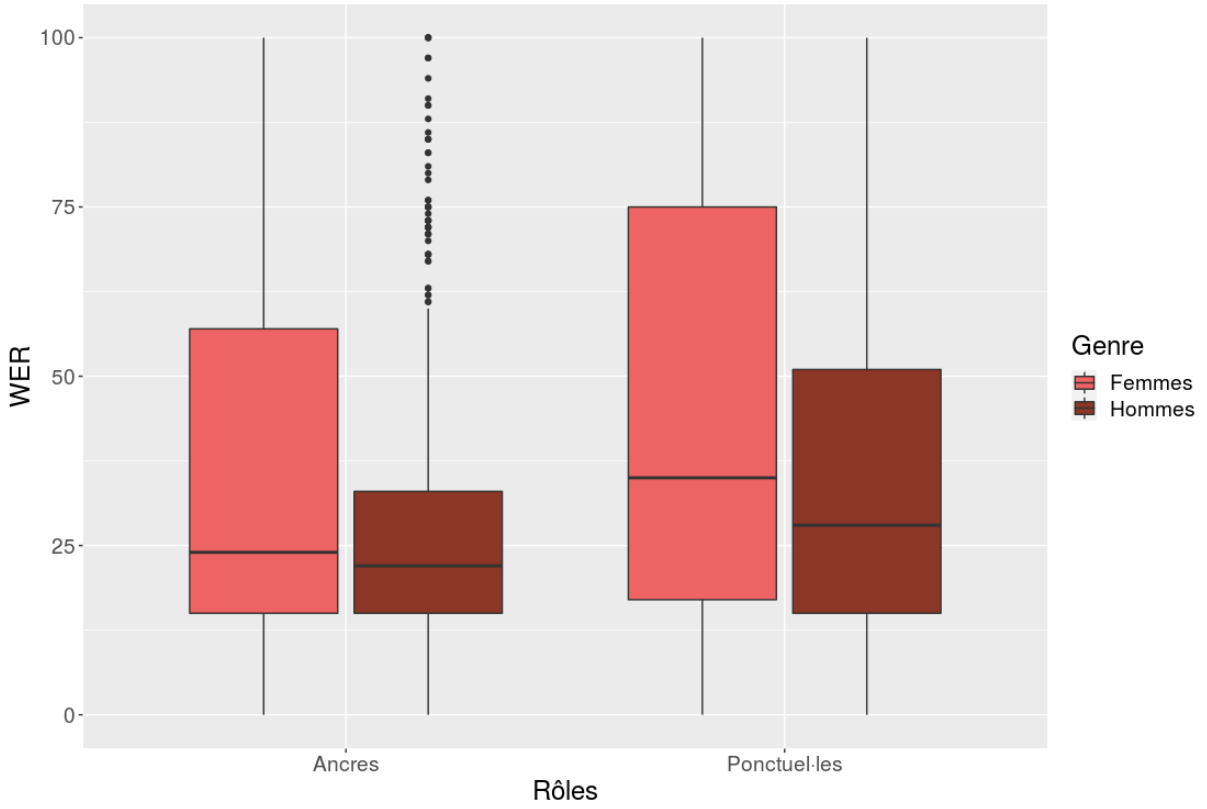

FIGURE 13 – Distribution des WER en fonction du rôle par catégorie de genre.

section 5.3.1). On s'attend effectivement à un WER plus important sur de la parole spontanée, où les disfluences sont nombreuses, que sur de la parole préparée sur laquelle la majorité des travaux a porté jusqu'alors. Au sein des émissions, les distributions de WER varient grandement comme en témoigne la Figure 14, dans laquelle les émissions sont classées par WER moyens croissants. Il est intéressant de souligner que l'on observe également que plus les performances du système se dégradent, plus les écarts en fonction du genre se creusent.

Lorsqu'on s'intéresse aux distributions de WER par catégorie de genre, par émission, les tendances ne sont pas nettes : les WER médians sont plus bas pour les femmes pour les 7 premières émissions à gauche de la figure, à savoir *Pile et Face*, *RTM Infos*, *TVME Infos*, *RFI Infos*, *Africa 1 Infos*, *Le Masque et la Plume* et *La Place du Village*. Or ces émissions, à l'exception de *Le Masque et la Plume* et de *La Place du Village*, ne contiennent que de la parole préparée, soit dans le cadre de programmes d'informations radiophoniques, soit, dans le cas de *Pile ou Face* de débats politiques télévisés. On peut donc questionner l'existence d'un lien entre la différence de performance homme/femme et le type de parole. De fait, on observe que, dès lors que les performances sont particulièrement mauvaises,

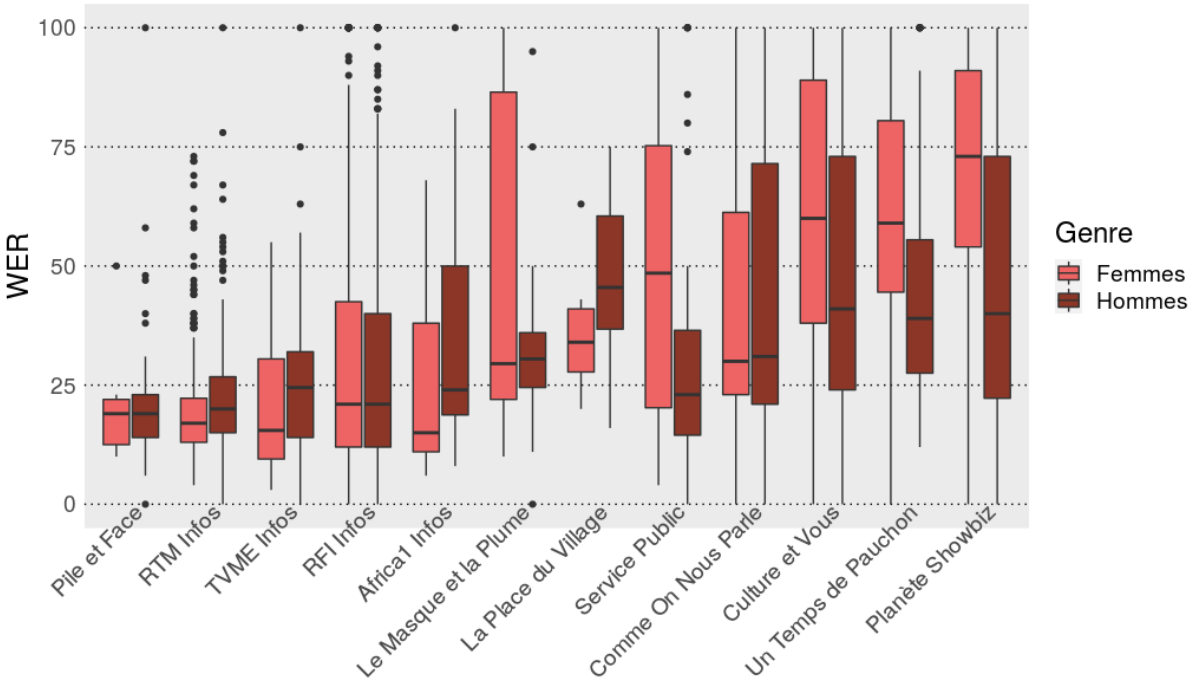

FIGURE 14 – Distribution des WER par émissions et par catégorie de genre.

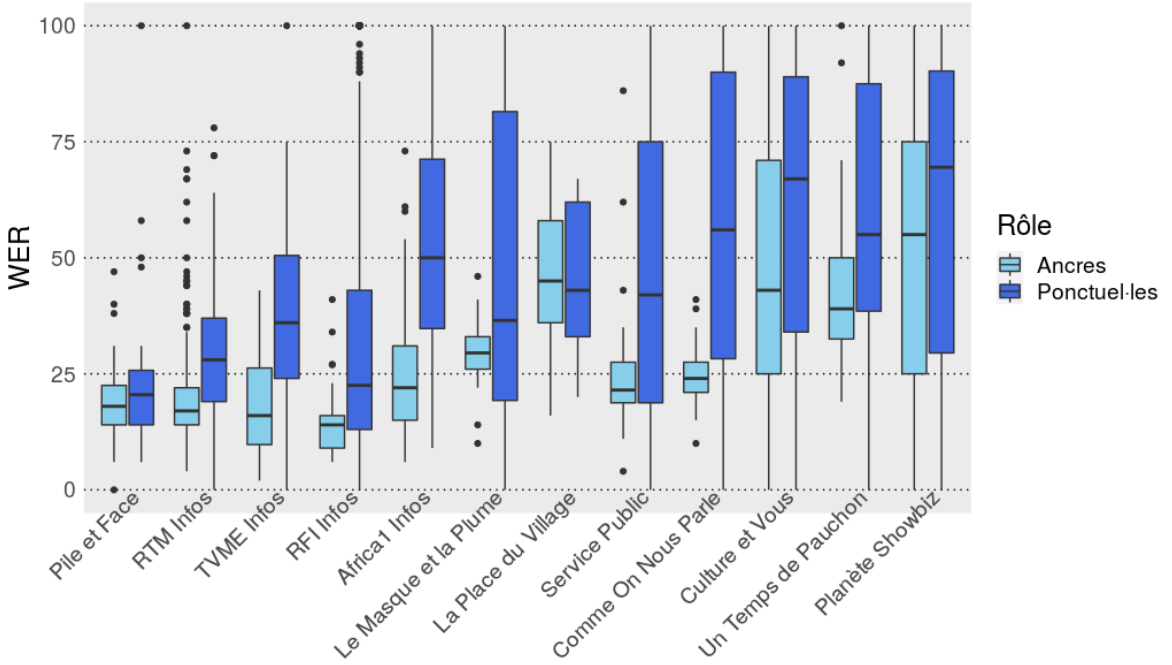

FIGURE 15 – Distribution des WER par émissions et par rôles.

	WER	Femmes	Hommes	Tous·tes
Parole préparée	médian	18%	21%	20%
	moyen	28,53%	27,06%	27,55%
Parole spontanée	médian	61%	38%	48%
	moyen	62,11%	48,33%	53,91%

TABLE 6.7 – Représentation des rôles en fonction des catégories de genre dans les données de test.

les WER médians des hommes deviennent meilleurs alors que les WER médians obtenus pour les femmes explosent et dépassent pour la majorité le seuil de 50%. [4]

La Figure 15 représente également nos distributions de WER mais cette fois-ci en fonction du rôle des locuteurs et locutrices. Les écarts entre Ancres et Ponctuelles sont bien plus importants que les écarts entre catégories de genre et ce, même sur de la parole spontanée. Ce que ces résultats nous disent c'est qu'un système est entraîné sur un type de données et un type de parole. Le corpus d'apprentissage du système ne contenant que de la parole préparée et les Ancres représentant une grande partie de ces données, le système modélise finement ce type de parole. La qualité de leur énonciation est peut-être également un facteur influant sur la reconnaissance. Nous soulignons également qu'en séparant nos WER/loc/émission en catégorie de genre et en rôle, nous finissons par obtenir des effectifs très petits, qui n'assurent pas la robustesse de nos résultats statistiques (voir Annexe B pour une description détaillée des WER) .

On observe donc un WER moyen plus bas pour les femmes sur les émissions *Africa 1 Infos*, *Comme On Nous Parle*, *La Place du Village*, *Pile et Face*, *RTM Infos* et *TVME Infos*, et des WER plus bas pour les hommes pour *Culture et Vous*, *Le Masque et la Plume*, *Planète Showbiz*, *RFI Infos*, *Service Public* et *Un Temps de Pauchon*. À partir de ces résultats, il nous semble délicat de conclure sur l'existence ou l'absence d'un biais prédictif genré. Ce que l'on peut affirmer en revanche, c'est que notre système est globalement peu performant sur la parole spontanée avec des WER bien supérieurs à 25% (voir Tableau 6.7). La Figure 16 nous montre bien que les différences de genre se lissent sur de la parole préparée (W = 212209 ; p-valeur = 0,047), alors que l'écart explose, avec une détérioration importante des performances sur les voix de femmes dans le cas de parole spontanée (W = 140053 ; p-valeur = $3{,}148e^{-13}$). On peut donc supposer que nous sommes face à un phénomène d'amplification des difficultés. Le modèle canonique contient une parole préparée, professionnelle (d'Ancres) et masculine, et s'en éloigner rend explicite l'existence d'un canon acoustique masculin en détériorant drastiquement les performances sur les locutrices.

4. L'ensemble des valeurs de WER par émission et catégorie de genre sont présentées dans l'Annexe B.

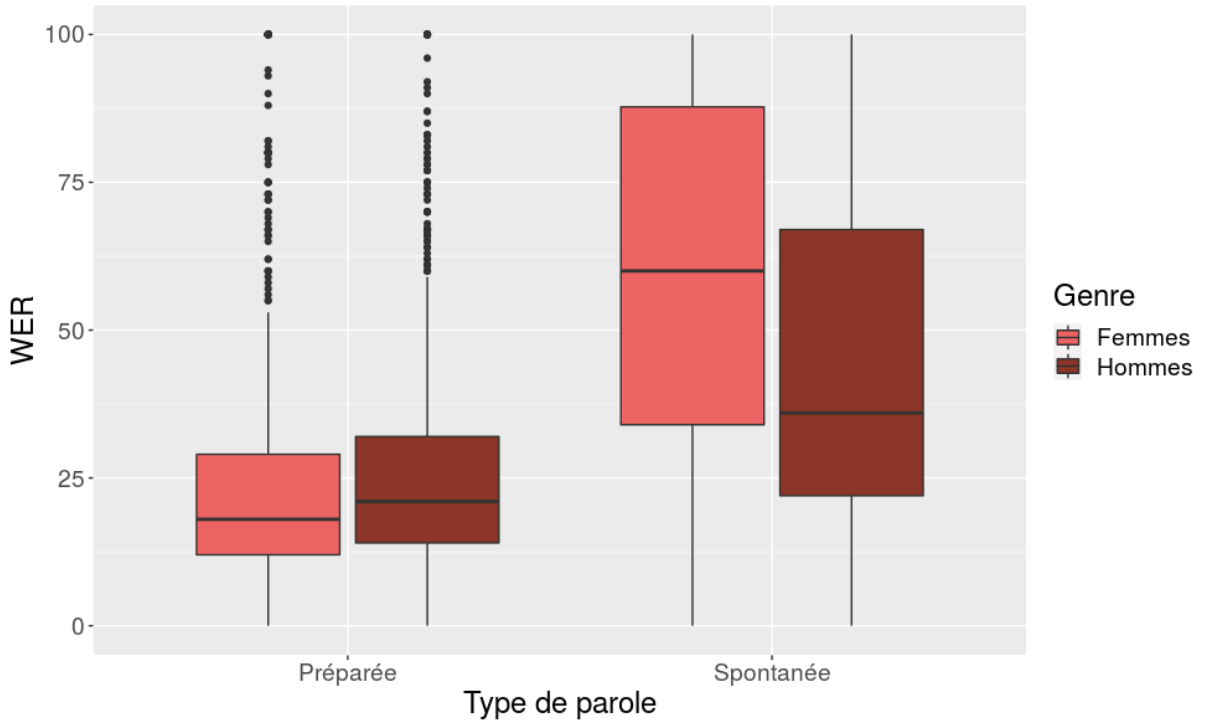

FIGURE 16 – Distribution des WER en fonction du type de parole par catégorie de genre.

6.3 Conclusion

Nous constatons dans nos données issues des média français une nette disparité en termes de présence des femmes et de quantité de parole. Nos données contenant des émissions diffusées entre 1998 et 2013, on peut s'attendre à ce que cette disparité soit moins importante sur des enregistrements plus récents, d'autant plus que le gouvernement français affiche des efforts vers la parité dans la représentation médiatique. On peut également noter que même si notre analyse a été menée sur un grand nombre de données, elle n'atteint pas l'exhaustivité d'études à grande échelle comme celle de Doukhan *et al.* (2018). Néanmoins, nos résultats posent question, car si la représentation des genres dans le monde réel est peut-être plus équilibrée aujourd'hui, ces corpus sont toujours utilisés comme données d'entraînement pour les systèmes d'ASR.

Dans cette première étude, nous observons de nombreuses variations de performances, en fonction du genre des locuteurs et locutrices, mais également de leur rôle ainsi que du type d'émission et de parole. Il est impossible d'isoler clairement l'impact de la représentation du genre sur ces données. Dans le cas de la parole préparée, on observe des WER médians plus bas pour les femmes (voir Figure 14, en accord avec les résultats de Martine Adda-Decker et Lori Lamel (2005)). Cependant, à l'échelle globale, cet écart de performances en fonction du genre dans la configuration de parole canonique, à savoir de la parole préparée, n'est pas significatif ($W = 212209$; p-valeur $= 0{,}047$). Des disparités apparaissent dès que l'on s'éloigne de ce standard. Les femmes obtiennent des WER

moins bons chez les Ponctuel·les et dans le cas de la parole spontanée, rendant visible la modélisation du masculin comme valeur par défaut.

L'inter-relation du rôle, du type de parole et du genre doit également nous conduire à questionner la manière dont nous tirons des conclusions sur les performances de nos systèmes. Dans leur étude sur l'ASR pour le trafic aérien, Thomas Pellegrini *et al.* (2019) notent que la parole de femmes est mieux reconnue que celle des hommes. Les auteurs et autrices mettent cela en relation avec les différents rôles occupés par les femmes : en effet celles-ci sont majoritairement présentes en tour de contrôle, où les conditions acoustiques sont meilleures, alors que les pilotes, dont la parole est difficilement reconnue, sont majoritairement des hommes. Dès lors, lorsque l'on s'intéresse à des données non-élicitées, le genre peut constituer une grille d'analyse, mais les interactions sont souvent multiples et se cantonner à une analyse genrée, peut conduire à une interprétation partielle voir faussée des résultats.

La deuxième question soulevée par ce travail est celle de la capacité de généralisation des systèmes. On peut supposer qu'un écart à l'usage (ici appliqué à de la parole préparée) risque de faire apparaitre des disparités de traitement sur d'autres facteurs sous-représentés à l'apprentissage (ici le rôle et la catégorie de genre). Une faible capacité de généralisation d'un système entraîné sur de la parole médiatique peut conduire à une invisibilisation de certaines paroles. Dans le cas de notre exemple, si ce système était utilisé pour du sous-titrage ou de l'indexation d'archives, on peut se questionner sur l'impact de mauvaises performances du système sur l'indexation de la parole des femmes par exemple ou des personnes moins représentées dans la sphère médiatique (on peut penser à la parole accentuée, par exemple). Le système présenté dans le cadre de cette étude n'était pas destiné à un tel usage, mais il semble que dans le cas de systèmes hybrides, un écart au modèle canonique entraîne des phénomènes d'amplification des difficultés.

Notre étude met également en avant le fait que les femmes sont sous-représentées dans les données issues des médias, ces données étant le reflet des inégalités sociales actuelles, comme le souligne l'ensemble des rapports du CSA relatifs à la représentation des femmes à la télévision et à la radio. Par conséquent, afin de créer des systèmes équitables (selon un conception de l'équité à définir), il est nécessaire de prendre en compte les problèmes de représentation dans la société qui vont être encapsulés dans les données. Ceci est en accord avec le concept de *Fairness by Design* proposé par Ahmed Abbasi *et al.* (2018).

Pour évaluer l'impact du genre sur un système d'ASR, il est donc nécessaire i) de décrire exhaustivement les données, et de comprendre les processus sociaux à la base de leur production, ii) de contrôler au maximum les autres variables pouvant influencer les performances du système. C'est ce à quoi s'intéresseront les chapitres 7 et 8 de ce manuscrit.

Chapitre 7

Représentation du genre dans les données de parole [1]

> « A data set may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a data set, we need to know where data is coming from ; it is similarly important to know and account for the weaknesses in that data. » (Boyd et Crawford, 2012, p. 668)

Cette citation est extraite de l'article de danah Boyd et Kate Crawford intitulé *Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon*. Parmi ces questions essentielles, se pose celle de la représentativité des données. Si nous postulons qu'une des sources des biais prédictifs dans les systèmes d'ASR se trouve dans le biais de sélection, alors les données doivent être considérées avec attention. Nous avons vu dans le chapitre précédent que les femmes sont majoritairement sous-représentées dans les corpus issus des médias. Si l'impact de la représentation du genre dans les données sur le WER reste complexe à analyser, nos résultats posent néanmoins la nécessité d'une description exhaustive des données. Cette description est nécessaire pour être en mesure d'évaluer nos systèmes et l'existence de biais prédictif genré, mais également dans une démarche de compréhension de ce que les systèmes modélisent réellement. Nous nous sommes donc intéressée à décrire les corpus de parole disponibles sur une plateforme de diffusion de ressources pour aider au développement de technologie du TAL : la plateforme OpenSLR (*Open Speech Language Resources*).

La question à la base de cette étude était multiple : dans le cadre de notre travail de thèse, nous souhaitons questionner l'existence de biais prédictifs de genre dans les systèmes d'ASR, mais pour attester de l'existence d'un tel biais prédictif, encore faut-il pouvoir être capable de connaître la représentation de genre dans les données utilisées par les systèmes et de les évaluer en fonction de ces catégories. Nous nous posons donc la question de savoir si ces informations sont disponibles. En suivant la méthodologie décrite en section 5.3.2, nous avons récupéré des informations sur 66 corpus de parole, disponibles sur la plateforme au moment de la réalisation de notre travail. À partir de ces informations, nous avons à la fois dressé un aperçu des pratiques, et mené une réflexion

1. Ce travail a donné lieu à deux articles (Garnerin *et al.*, 2020b) et (Garnerin *et al.*, 2020a), dont le présent chapitre est partiellement inspiré.

sur ce que nos observations viennent dire de la manière dont est envisagé le genre dans nos données.

Comme expliqué plus haut, la méthodologie de récolte des informations est décrite en section 5.3.2. Dans le présent chapitre, nous nous intéresserons à la présentation et à l'analyse des résultats. [2] Dans un premier temps, nous discuterons de la faible quantité d'informations fournies avec les corpus et des impacts de ce manque de transparence. Puis nous nous intéresserons à la représentation des catégories de genre dans nos données, en questionnant les multiples facteurs à l'origine des différences observées. Enfin, nous conclurons sur un ensemble de bonnes pratiques, réactualisant ainsi une partie des recommandations que de nombreux auteurs et autrices (Wilkinson *et al.*, 2016; Bender et Friedman, 2018; Gebru *et al.*, 2018; Mitchell *et al.*, 2019) préconisent.

7.1 Disponibilités des méta-données

En commençant ce travail de recensement, nous ne pensions pas rencontrer de difficultés outre mesure. Cependant, force a été de constater que les informations sur les locuteurs et locutrices sont rares. Sur les 66 corpus étudiés, 36,4% ne fournissaient aucune information concernant le genre des locuteurs et locutrices. Pour les corpus fournissant ces informations, seulement 9 les indiquaient dans un fichier de méta-données, 28 l'avaient indexé de manière explicite dans les données elles-mêmes et 5 mentionnaient ces informations dans l'article publié à l'occasion de la diffusion publique du corpus (voir Tableau 7.1). On peut se demander si la prise en compte tardive du genre dans la technologie est à l'origine de l'absence de mise à disposition de données démographiques sur le genre par les créateurs et créatrices de ressources.

Lorsque des informations sont fournies, ce sont majoritairement le nombre de personnes appartenant à chaque catégorie. Dans le sous-ensemble des 42 corpus contenant des informations sur le genre des locuteurs et locutrices, seuls cinq corpus fournissent le temps de parole pour chaque catégorie (voir Tableau 7.2).

La deuxième difficulté rencontrée concerne l'absence de standard dans le report d'informations sur le temps de parole, rendant impossible la comparaison directe de temps de parole entre individus ou entre catégories de genre. Lorsque des informations de durée sont fournies, la granularité utilisée varie selon les corpus. Certains auteurs indiquent les temps de parole en heures (e.g. Panayotov *et al.* (2015); Hernandez *et al.* (2018)), d'autres le nombre d'énoncés ou de phrases (e.g. Juan *et al.* (2015); Google (2019)), la définition de ces deux termes n'étant jamais explicite : le découpage est-il syntagmatique ? basé sur les groupes de souffles ou du aux limites techniques du système ? Il n'existe pas de cohérence entre nombre d'énoncés et temps de parole, excluant ainsi la possibilité d'approximer l'une par l'autre.

2. L'ensemble des données recueillies sont présentées dans l'Annexe C

Informations disponibles		#corpus
Non		24 (36.4%)
Oui	metadata	9 (13.6%)
	indexed	28 (42.4%)
	paper	5 (7.6%)
Total	-	66

TABLE 7.1 – Disponibilité des informations concernant le genre dans les corpus OpenSLR.

Informations disponibles	#corpus
Nombre de loc.	40
Nombre d'énoncés	32
Durée de parole	5
Nombre total de corpus	42

TABLE 7.2 – Type d'information disponible en fonction du genre dans les 42 corpus contenant des informations genrées.

Notre approche se voulait inspirée de celle du travail de David Doukhan et Jean Carrive (2018), en mobilisant les notions de "taux de présence" et de "taux d'expression". Mais la difficulté à obtenir des informations de durée de parole comparables nous a mené à nous focaliser principalement sur le taux de présence. Cependant, au vu des résultats présentés dans le précédent chapitre, nous encourageons le lecteur ou la lectrice à garder à l'esprit que des taux de présence égaux n'impliquent pas nécessairement des quantités de parole égales, même si cette mesure en reste un premier indicateur.

En plus des 42 corpus pour lesquels nous avons réussi à trouver des informations sur le genre des locuteurs et locutrices, nous avons recueilli manuellement des informations sur les locuteurs et locutrices pour 7 autres corpus, atteignant une taille d'échantillon finale de 47 corpus. [3]

7.2 Genre et taux de présence

À l'échelle générale de notre corpus d'étude composé de 47 corpus de parole pour lesquels des données démographiques sur le genre ont été récupérées, nous observons la répartition suivante : sur les 6 072 locuteurs et locutrices, 3050 sont des femmes et 3022 des hommes, on peut donc considérer la parité comme atteinte.

Nous avons ensuite croisé la répartition homme/femme avec le caractère élicité ou non des données, une parole non élicitée décrivant toute parole qui aurait existé sans la

3. Les données sur le corpus TEDLIUM renseignées dans l'article de ne portant que sur la durée de parole, le corpus ne fait pas partie de notre analyse sur la représentation des catégories genrées basées sur le nombre de locuteurs et locutrices. De plus le corpus The Free ST Chinese Mandarin Corpus, de SurfingTech, fournissait des informations de genre, mais nous avons rencontré des problèmes avec les fichiers et ne sommes pas parvenue à l'intégrer à notre analyse, d'où un total de 47 et non 49 corpus.

création du corpus, comme les TEDTalks, les interviews, les émissions de radio, etc. Nous supposons que si les données n'ont pas été élicitées, un déséquilibre entre les catégories de genre pourrait apparaître, comme ce que nous avons pu observer dans le chapitre précédent. Ces attentes sont renforcées par des exemples tels que la ressource des TED Talks espagnols, qui indique dans sa description concernant les locuteurs et locutrices que "most of them are men" (Hernandez-Mena, 2019). Les données élicitées à l'inverse, ayant été produites et récoltées dans le cadre d'un protocole de création de corpus sont beaucoup moins susceptibles de présenter une représentation déséquilibrée puisque celle-ci aura été réfléchie en amont par les auteurs et autrices de la ressource, lors du protocole de recueil des enregistrements.

Les autres facteurs que nous avons analysés sont la tâche adressée (à savoir l'ASR ou la synthèse vocale) et le statut de la langue. Le statut de la langue (bien ou peu dotée) est défini ici de manière approximative et ne prend pas en compte les variations dialectales. Ainsi un dialecte de l'anglais, même si peu représenté dans les technologies, sera quand même considéré comme bien doté. Ces deux facteurs nous semblaient pertinents car ils réinscrivent les données et les technologies dans des espaces sociaux et culturels.

7.2.1 Parole élicitée vs non-élicitée

Nous présentons les résultats dans le Tableau 7.3. Que ce soit dans le cas de la parole élicitée ou non, la différence de nombre de locuteurs et locutrices est relativement faible (respectivement 5,6 et 5,8 points de pourcentage), loin de la différence de 30 points de pourcentage observée dans le chapitre précédent. On peut néanmoins noter que dans le cas de parole élicitée, les femmes sont plus représentées et que cette tendance s'inverse pour les données non-élicitées. Ces résultats laissent supposer qu'élicités ou non, les corpus sont le résultat d'un processus contrôlé de création de données, et donc la disparité de genre est réduite autant que possible par les auteurs et autrices du corpus. Nous remarquons cependant qu'à l'exception de Librispeech (Panayotov *et al.*, 2015), tous les corpus non-élicités sont de petits corpus. En retirant Librispeech de l'analyse, nous retrouvons un ratio

Type de parole	#corpus	#F	#H
Élicitée	41	1782 52,8%	1596 47,2%
Non-élicitée	5	1268 47,1%	1426 52,9%
Non-élicitée (sans Librispeech)	4	67 31.9%	143 68.1%

TABLE 7.3 – Distribution des catégories de genre en fonction du type de parole. (N.B. : Les deux dernières lignes reportent les résultats obtenus pour la parole élicitée sans prendre en compte le corpus Librispeech).

femmes/hommes de 1/3-2/3, cohérent avec nos résultats précédents. Ce qui ressort de ces résultats est que lorsque les données ne sont pas élicitées ou soigneusement équilibrées, il existe un risque que la disparité de représentation s'insinue. Ce déséquilibre n'est pas observé à l'échelle de l'ensemble de la plateforme OpenSLR, en raison du fait que la plupart des corpus sont élicités (89,1%). Par conséquent, l'existence d'un tel écart entre les genres est évitée par un contrôle minutieux pendant le processus de création des ensembles de données. Les créateurs du corpus Librispeech, par exemple se sont assurés de maintenir une « gender balance at the speaker level and in terms of the amount of data available for each gender » (Panayotov *et al.*, 2015, p. 5208).

7.2.2 "How can I help ?" : impact de la tâche

Les données disponibles sur OpenSLR sont des données de paroles pour l'entraînement des systèmes. Si notre travail porte uniquement sur l'ASR, les corpus de parole peuvent également servir au développement de systèmes de synthèse vocale (*text-to-speech* ou TTS). Nous nous sommes intéressée à l'impact de la tâche sur la composition des corpus, les résultats sont reportés dans le Tableau 7.4. Nous observons que si les taux de présence sont presque équilibrés au sein des corpus d'ASR, les femmes sont plus représentées dans les ensembles de données pour la synthèse, avec 63,9% de locutrices contre 36.1% de locuteurs. Il y a donc retournement de la répartition 1-3/2-3 avec cette fois-ci une sous-représentation des hommes. Cette observation entre en résonance avec un ensemble de travaux démontrant que nous serions censées préférer les voix de femmes. Si ces hypothèses sortent du champ de notre compétence, nous pouvons néanmoins montrer que celles-ci permettent aux systèmes de rester en adéquation avec les stéréotypes de genre. En effet, dans *2001 : l'Odyssée de l'Espace*, HAL qui se veut figure d'autorité et de pouvoir (et qui incarne au final une persona violente) est représenté avec une voix d'homme, alors que les assistants vocaux, pensés comme des secrétaires personnelles, appellent à des valeurs majoritairement associées à la féminité. Ce constat fait écho au rapport de recommandation de l'ONU pour une éducation numérique égalitaire, qui indique qu'aujourd'hui la plupart des assistants vocaux ont une voix de femme :

> « Today and with rare exception, most leading voice assistants are exclusively female or female by default, both in name and in sound of voice. Amazon has Alexa (named for the ancient library in Alexandria), Microsoft has Cortana (named for a synthetic intelligence in the video game Halo that projects itself as a sensuous unclothed woman), and Apple has Siri (coined by the Norwegian co-creator of the iPhone 4S and meaning 'beautiful woman who leads you to victory' in Norse). While Google's voice assistant is simply Google Assistant and sometimes referred to as Google Home, its voice is unmistakably female. » (West *et al.*, 2019, p. 94)

Tâche	#corpus	#F	#H
ASR	12	2523 49,1%	2615 50,9%
TTS	10	124 63,9%	70 36,1%
NA	25	403 54,5%	337 45,5%

TABLE 7.4 – Distribution des catégories de genre en fonction de la tâche adressée. ASR correspond à la reconnaissance automatique de la parole, TTS à la synthèse vocale à partir de texte et NA correspond aux corpus pour lesquels aucune tâche particulière n'était renseignée.

On peut néanmoins soulever que, par défaut, la voix française de Siri est masculine, rare exception dans le tableau général. Les stéréotypes tels que "les voix féminines sont perçues comme plus serviables, plus sympathiques ou plus agréables" justifient de leur utilisation dans les systèmes. [4] Les systèmes de synthèse vocale étant souvent utilisés pour créer des assistants vocaux, on peut supposer que l'utilisation de voix féminines est devenue pratique courante pour garantir l'adhésion du public au système, ce qui expliquerait leur sur-représentation dans les corpus de données destinés à la synthèse. Cette pratique, d'utilisation de voix féminines dans les technologies reposant sur une utilisation de la parole et justifiée par des phénomènes d'identification sociale ou de stéréotypes culturels liés au genre est discutée dans les travaux de Clifford Nass et Scott Brave (2005).

7.2.3 Statut de la langue

Dans les corpus élicités mis à disposition sur OpenSLR, certains contiennent des données de langues peu dotées et d'autres de langues largement dotées (principalement des variations régionales de langues largement dotées, l'idée étant de permettre une amélioration de la couverture dialectale des systèmes). Les stéréotypes de genre et les disparités de représentation étant le résultat de processus sociaux et culturels, il nous semblait intéressant de questionner si les écarts de représentation observés jusqu'alors étaient un phénomène constant à l'échelle globale, où si des différences par langue émergeaient. Lorsque nous examinons la répartition des catégories de genre dans ces corpus élicités,

4. Même si ces associations tendent à être mises à distance par les créateurs et créatrices de ces artefacts. On peut citer Deborah Harisson, ayant travaillé sur le développement de Cortana et citée par François Perea (2018) : « We always get the question why is she a she and not a he and there are two genders and you can start with one so there are reasons we went with a female but at the end it's sort of fifty fifty. So she's a woman, there is a legacy of what women are expected to be like in an assistant role and we wanted to be really careful that Cortana got a female voice but she's not a woman, she's an AI assistant, a digital assistant with a voice set as a female voice ». Il est intéressant de souligner qu'en tant qu'artefact, Cortana est pensée comme ne relevant d'aucune catégorie de genre, mais que dans la relation utilisateur (ou utilisatrice), elle incarne une persona féminine (avec ce que cela invoque en termes de représentations).

Statut de la langue	#corpus	#F	#H	Total
Peu dotée	23	677 55,7%	539 44,3%	1216 100%
Fortement dotée	19	1105 51,1%	1057 48,9%	2162 100%

TABLE 7.5 – Répartition des catégories de genre en fonction du statut de la langue.

	F	H
Nombre de loc.	591 51,8%	551 48,2%
Nombre d'énoncés	72280 33,5%	143342 66,5%

TABLE 7.6 – Nombre de locuteurs et locutrices et nombres d'énoncés par catégories de genre, dans le sous-échantillon de 41 corpus pour lesquels nous avons pu récupérer des fréquences d'énoncés. (N.B. : l'ensemble de ces corpus ne renseigne pas le nombre de locuteurs et locutrices, ces chiffres sont donc simplement donnés à titre indicatif).

nous n'observons pas de différence en fonction du statut de la langue, comme reporté dans le Tableau 7.5. On pourrait conclure que les pratiques sont relativement uniformisées à travers la communauté de TAL. Nous remarquons également que les corpus de langues largement dotées contiennent deux fois plus de locuteurs et locutrices, les corpus de langues peu dotées étant tous de petits corpus.

7.3 Genre et taux d'expression

En raison d'un manque global d'informations sur le temps de parole, nous n'avons pas analysé le taux d'expression par catégorie. Cependant, le nombre d'énoncés est souvent renseigné, ou facilement retrouvable dans les corpus, et nous avons pu récupérer des valeurs par catégorie de genre pour 41 corpus, soit les 32 corpus mentionnés dans le Tableau 7.2 et 9 corpus pour lesquels nous avons récupéré les informations manuellement. Si l'équilibre entre hommes/femmes est presque atteint d'un point de vue du taux de présence, les hommes sont plus représentés lorsqu'on s'intéresse au nombre d'énoncés (voir Tableau 7.6). Cependant, cette disparité n'est en réalité que l'effet de trois corpus mono-genre, contenant respectivement 51 463, 26 567 (Korvas *et al.*, 2014) et 8376 (Hernandez-Mena, 2019) énoncés de locuteurs, alors que le nombre moyen d'énoncés par corpus sur l'ensemble de notre sous-échantillon est respectivement de 1942 pour les hommes et 1983 pour les femmes. Après avoir retiré ces trois valeurs extrêmes, la quantité de parole est équilibrée entre les catégories de genre. Le nombre élevé d'énoncés des trois valeurs extrêmes est cependant surprenant, ces trois corpus étant "petits" (2,1 Go, 2,8 Go) et "moyens" (5,2

Go).[5] Cela met une fois de plus en évidence le flou autour de la notion d'énoncé (*sentence* ou *utterances*) qui n'est jamais explicitement définie. Une telle différence de granularité rend difficile la comparaison entre les corpus, ces résultats sont donc uniquement donnés à titre indicatif.

7.4 Évolution dans le temps

Lors de la collecte des données, nous avons remarqué que plus les ressources étaient récentes, plus il était facile de trouver des informations sur le genre, attestant de l'émergence du genre dans la technologie en tant que sujet pertinent. Comme l'a souligné Kate Crawford (2017) dans son discours d'ouverture de NeurIPS, les questions d'éthique et d'équité en IA ont été investies par la communauté de recherche en IA et en apprentissage automatique. La Figure 17 montre l'évolution de la disponibilité des informations sur le genre au cours des 10 dernières années. Nous pouvons voir que ce pic d'intérêt est également présent dans nos données, avec plus de ressources contenant des informations sur le genre après 2017.

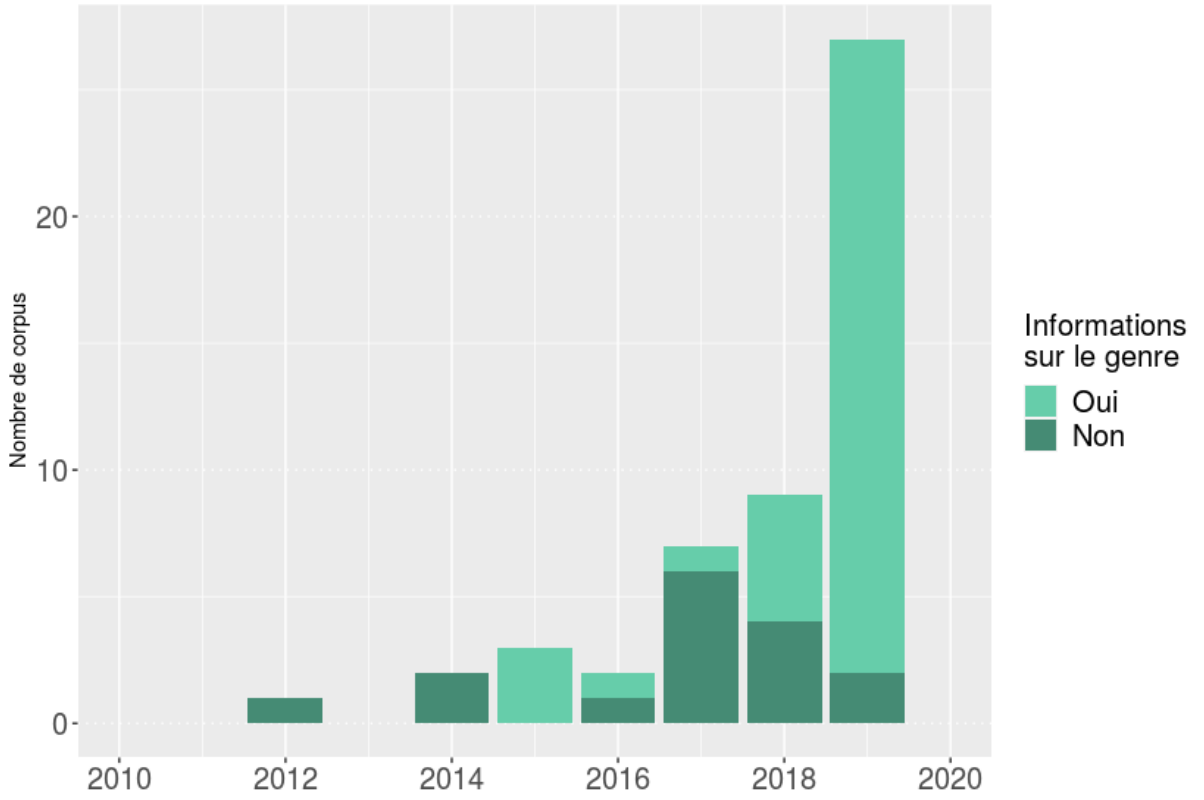

FIGURE 17 – Évolution de la disponibilité des informations sur le genre dans les ressources OpenSLR de 2010 à 2019

5. D'après notre classification en taille définie en section 5.3.2 du Chapitre 5.

7.5 Sur l'importance des méta-données

Cet effort de recherche a donné lieu à des articles méthodologiques, parmi lesquels on peut citer le travail d'Emily Bender et Batya Friedman (2018), qui ont proposé la notion de *data statements*. Les autrices supposent que :

> « [...] data statements will help alleviate issues related to exclusion and bias in language technology, lead to better precision in claims about how natural language processing research can generalize and thus better engineering results, protect companies from public embarrassment, and ultimately lead to language technology that meets its users in their own preferred linguistic style and furthermore does not misrepresent them to others. » (Bender et Friedman, 2018, p. 587)

Décrire ses données de manière explicite devient ainsi un moyen de contrôle de la capacité de généralisation des modèles. Si la question se pose aujourd'hui de la transparence des modèles et de l'explicabilité du fonctionnement des systèmes, il reste cependant en amont et en aval de ceux-ci, des lieux qui se doivent d'être investis : les données et l'évaluation.

Dans les principes FAIR de Mark Wilkinson *et al.* (2016) pour la gestion des données scientifiques, les deux premiers concernent la repérabilité (*foundability*) et l'accessibilité (*accessibility*). C'est seulement une fois que ces deux aspects sont pris en compte que les questions de l'interopérabilité (*interoperability*) et de la réutilisabilité (*reusability*) peuvent se poser. Si dans notre cas, la repérabilité et l'accessibilité sont en partie prises en compte par la mise à disposition des ressources via la plateforme OpenSLR, l'absence de description systématique de leur contenu est un frein à l'interopérabilité et à leur réutilisation. Une autre discussion sur la description des données au sein de la communauté du TAL a été initiée par Alain Couillault *et al.* (2014), qui ont proposé une Charte sur l'éthique et les *big data* (*Ethics and Big Data Charter*), pour aider les créateurs et créatrices de ressources à décrire leurs données d'un point de vue juridique et éthique. En dehors du domaine du TAL, cette nécessité d'une description exhaustive des données, tant d'un point de vue juridique qu'éthique est également la proposition de Timnit Gebru *et al.* (2018). Avec leur notion de *datasheets for datasets*, ils et elles proposent une méthodologie de standardisation des corpus, à l'instar du travail d'Emily Bender et Batya Friedman.

Dans notre enquête, 13 des 66 corpus étaient accompagnés d'un article décrivant les ressources. Lorsque les performances des systèmes d'ASR étaient reportées, aucune évaluation en termes de genre n'était faite, même si des informations sur la représentation du genre dans les données étaient renseignées. Si le report d'informations démographiques est une pratique relativement commune dans la description des ressources, la variabilité qu'elle introduit dans le système ne fait pour autant pas l'objet d'une évaluation. Pourtant, la communication des résultats pour les différentes catégories de genre constitue un

moyen simple de vérifier l'existence de biais prédictif dans les performances du système. Décrire ses données constitue un premier pas inévitable, mais l'étape suivante devrait être de prendre également en compte ces informations dans les processus d'évaluation. C'est également le point de vue défendu par Margaret Mitchell *et al.* (2019) avec leur notion de *model cards* : une fois la variabilité des données décrites, il est important de rendre compte des performances des modèles sur ces différents groupes, pour faciliter une interopérabilité des modèles, sans introduire des risques de biais prédictifs. Montrer la robustesse d'un système à différentes catégories amène obligatoirement à penser et remettre en question sa capacité de généralisation. Dans cette perspective, Josh Meyer *et al.* (2020) ont proposé le corpus *Artie Bias* pour évaluer les performances des systèmes en fonction du genre, de l'âge et de l'accent.

7.6 Conclusion

La première conclusion de notre enquête concerne la difficulté à obtenir une description exhaustive des locuteurs et locutrices présentes dans les ressources de parole. Ce manque de méta-données est problématique d'un point de vue scientifique, car il empêche de garantir la généralisation des systèmes ou des résultats linguistiques basés sur ces corpus, comme le soulignent Emily Bender et Batya Friedman (2018), mais également d'un point de vue éthique rendant impossible tout contrôle quant à l'existence d'une disparité de représentation pouvant conduire à des biais de performance genrés. Cette absence d'informations contextuelles sur la parole traduit aussi une conception du langage comme entité abstraite, plutôt que comme production située, qui mérite d'être questionnée (Hovy et Spruit, 2016).

Lorsque des informations sur la représentation du genre dans les données étaient fournies, celles-ci se portaient majoritairement sur le nombre de locuteurs et locutrices. Mais comme montré dans le chapitre précédent, le taux de présence n'est pas nécessairement relié au taux d'expression. Il serait intéressant d'avoir accès à la durée des ensembles de données en heures ou minutes, à l'échelle du corpus mais également par individu et/ou catégorie de genre. Des informations de durée standardisées pourraient permettre de vérifier rapidement l'équilibre entre les catégories de genre, sans s'appuyer sur une notion d'énoncé peu fiable. Lors de la collecte des données, nous avons remarqué que plus les ressources étaient récentes, plus il était facile de trouver des informations sur le genre, attestant de la visibilité croissante des thématiques de genre dans la technologie. Mais si ce travail descriptif est important pour les futurs corpus, il doit également être effectué pour les ensembles de données déjà publiés, car ils sont susceptibles d'être utilisés à nouveau par la communauté.

Comme écrit dans le rapport de l'UNESCO :

> « Even if far from a panacea, establishing balance between men and women in the technology sector will help lay foundations for the creation of technology products that better reflect and ultimately accommodate the rich diversity of human societies. » (West *et al.*, 2019, p. 89)

Si la parité n'est pas une fin en soi, elle constitue un premier pas vers une représentation de la diversité de nos sociétés pour créer des technologies adaptées à ces dernières. Dès lors, cette représentation des femmes doit se faire à l'échelle des entreprises et des laboratoires, mais également à l'échelle des données. Dans le cas de la plateforme OpenSLR, nous avons pu observer que cette parité est globalement atteinte lorsque les données sont élicitées. En revanche, lorsque celles-ci sont récupérées, à partir d'émissions radiophoniques par exemple, nous pouvons observer des disparités de représentations. Cependant ces pratiques ont tendance à évoluer, le corpus TED LIUM (Rousseau *et al.*, 2014) par exemple, dans sa dernière version, reporte une quantité de données équivalente pour les hommes et les femmes.

Nous avons observé également que les voix féminines sont majoritaires dans la plupart des corpus visant à développer les systèmes de synthèse vocale. Cela pourrait être expliqué par les stéréotypes associant la voix féminine aux activités de *care*, la majorité des voix de synthèse étant utilisées dans des produits de services. Ce type de résultat permet de réaffirmer le caractère hautement social des données, et la mesure dans laquelle celles-ci représentent une certaine vision et organisation du monde. Notre étude ne peut cependant pas prétendre à l'exhaustivité et il serait intéressant de mener un recensement comparable à plus grande échelle sur l'ensemble des ressources de la plateforme LDC ou ELDA-ELRA. Nous pensons néanmoins que notre travail constitue un bon aperçu de l'état actuel des pratiques.

Enfin, la disponibilité croissante de données sur le genre est de bon augure pour une réflexion commune sur sa conception dans les systèmes actuels d'ASR et leurs données. Pour autant, si les informations de genre sont fournies à propos des données, premier maillon de la chaîne du développement des systèmes d'ASR, celles-ci restent largement absentes lorsque l'on s'intéresse à la question de l'évaluation. Or pour évaluer la capacité de généralisation d'un système, il semble pertinent de pouvoir évaluer les performances de celui-ci sur différents groupes de la population, d'autant plus quand ces groupes s'inscrivent dans des rapports de pouvoir, comme nous avons pu le soulever avec l'invisibilisation des femmes en tant que locutrices dans les technologies de la parole.

Chapitre 8

Maîtriser la répartition des genres : une étude sur Librispeech [1]

Penser l'existence de biais prédictif en fonction des catégories de genre commence donc avec les données. Mais quel est l'impact de la représentation des catégories de genre sur les performances d'un système ? Nous avons vu dans le Chapitre 6 que l'interaction de facteurs multiples (canal, type de parole, rôle du locuteur ou de la locutrice) rendait impossible l'interprétation de l'impact du facteur genre de manière isolée. Le présent chapitre s'intéressera donc à questionner cet impact dans un contexte beaucoup plus contraint, puisque nous travaillons sur le corpus Librispeech contenant des livres audio, contrôlant ainsi les variations dues au rôle et au type de parole, ne laissant place qu'à la variation individuelle dans la tâche de lecture.

8.1 Variation de la proportion homme/femme

Pour ce faire, nous avons donc entraîné trois systèmes E2E avec trois jeux de données différents (voir section 5.3.3) : un premier corpus d'apprentissage contenant une majorité de livres lus par des hommes (70%), un corpus avec une répartition équilibrée à 50-50% et un corpus contenant une majorité de livres lus par des femmes (70% également). L'idée ici étant de venir vérifier si la variation de la représentation des catégories de genre influence les performances du système. Nos systèmes étant des systèmes E2E, c'est-à-dire sans modélisation ni adaptation au locuteur ou à la locutrice explicites, nous nous attendions à un impact fort de ces représentations sur les résultats.

8.1.1 Performances globales

Les résultats pour ces trois modèles sont présentés dans le Tableau 8.1 et les distributions sont présentées dans l'Annexe D. Les WER moyens sont de 9,7 %, 10,2% et 9,0% pour nos 3 conditions sur le corpus test-clean, c'est-à-dire sur le corpus contenant la parole la plus "simple" à reconnaitre (puisque les corpus clean sont ceux ayant obtenus les

1. Ce travail a donné lieu à un article (Garnerin *et al.*, 2021), dont le présent chapitre est partiellement inspiré.

Modèles	WER	test-clean			test-other		
		F	H	Tous·tes	F	H	Tous·tes
wper30	médian	**10,3%**	**8,0%**	9,5%	21,2%	23,4%	22,5%
	moyen	10,9%	8,3%	9,7%	23,3%	26,7%	25,0%
wper50	médian	11,1%	8,9%	9,7%	21,9%	25,0%	24,0%
	moyen	11,0%	9,1%	10,2%	25,1%	30,2%	27,7%
wper70	médian	9,4%	8,1%	8,9%	19,5%	22,8%	21,6%
	moyen	9,6%	8,3%	9,0%	20,9%	25,9%	23,4%

TABLE 8.1 – WER par genre obtenus sur les corpus d'évaluation test-clean et test-other pour nos 3 modèles.

meilleurs WER avec un premier modèle acoustique appris sur le WSJ, d'après les auteurs du corpus (Panayotov *et al.*, 2015)). Sur le corpus test-other, plus complexe pour le système donc, les WER moyens sont de 25%, 27,7% et 23,4%. Malgré les écarts de WER médians et moyens observés entre les trois systèmes, aucune de ces différences n'est statistiquement significative [2] (p-valeur = 0,14 pour le test-clean et p-valeur = 0,11 pour le test-other). On ne peut donc pas conclure que la performance globale du système est affectée par la variation de la représentation du genre dans les données d'entraînement, ce qui nous laisse supposer une bonne robustesse des systèmes face à ces variations.

8.1.2 Performances par catégorie de genre

En revanche, le constat n'est pas tout à fait le même dès que l'on s'intéresse aux performances par catégorie de genre. Un rapide coup d'œil à nos distributions [3] de WER par catégorie montre que les performances obtenues pour les femmes sont généralement moins bonnes que celles obtenues pour les hommes (voir Figures 18). Sur le test-clean, cette différence est statistiquement significative (p-valeur = 0,003) lorsque notre ensemble d'entraînement ne contient que 30% des livres lus par des femmes. Plus la proportion de livres lus par des femmes dans le corpus d'apprentissage augmente, plus la p-valeur augmente, jusqu'à dépasser notre risque alpha (p-valeur = 0,04 pour wper50 et p-valeur = 0,10 pour wper70). Notre hypothèse selon laquelle la représentation des catégories de genre dans le corpus d'apprentissage influence les performances obtenues se vérifie dans le cas des femmes, avec de meilleurs WER au fur et à mesure que l'on augmente leur proportion dans les données d'apprentissage, mais ce phénomène n'est pas symétrique, ce qui nous empêche d'affirmer que la sous-représentation d'une catégorie de genre amène une dégradation du WER sur cette catégorie. De manière surprenante, lorsque l'ensemble d'entraînement contient 70% de livres lus par des femmes, il n'y a pas de différence

2. L'ensemble des test statistiques discutés dans ce chapitre est reporté dans l'Annexe E.

3. Le choix de représentations graphiques à l'aide de diagrammes en violons permet d'associer les avantages de la boîte à moustache et de l'histogramme, pour donner un aperçu visuel de la répartition des performances comme ensemble continu, plutôt qu'à l'aide d'une moyenne unique, montrant ainsi la variabilité des performances en fonction des types de données et d'individus.

significative entre les WER obtenus pour les hommes et ceux obtenus pour les femmes. Les WER des femmes diminuent de presque 1 point de pourcentage par rapport au modèle wper30 alors que le WER des hommes n'augmentent que de 0,1 point de pourcentage. On obtient également les meilleures performances globales de nos 3 systèmes avec un WER médian de 8,9% et un WER moyen de 9,0%. Pour autant, peu importe le pourcentage de femmes représentées à l'apprentissage, les WER sont toujours meilleurs pour les hommes même si la significativité statistique n'est pas toujours atteinte.

Sur le test-other, les tendances des distributions de WER par genre ne vont pas dans le même sens (voir Figure 19). Outre une dégradation générale des performances avec un WER pratiquement doublé, les WER moyens et médians obtenus sont toujours plus faibles pour les femmes : on obtient un WER moyen de 23,30% (respectivement 25,1% et 20,9%) pour le modèle appris sur le corpus wper30 (respectivement wper50 et wper70) et ces valeurs sont toujours inférieures aux WER moyens obtenus pour les locuteurs. En revanche ces différences ne sont jamais significatives avec une p-valeur de 0,21 pour le wper30, 0,15 pour le wper50 et 0,03 pour le wper70. On peut néanmoins observer sur la Figure 19 que l'écart des distributions entre catégories de genre semble se creuser avec l'augmentation du pourcentage de livres lus par des femmes dans le corpus d'apprentissage.

8.2 Cas limite : les modèles mono-genre

Face à de tels résultats, on peut se demander si la représentation du genre dans les données d'apprentissage joue un rôle dans la distribution de nos performances, ou si celles-ci ne sont que le fruit de l'aléatoire contenu dans nos modèles. En effet, la performance globale de notre système semble être robuste à la variation de la représentation du genre dans les données d'apprentissage. Pour autant dans le modèle wper30, nous observons une différence de WER statistiquement significative entre livres lus par des hommes et par des femmes. Nous avons donc choisi d'explorer le cas limite de systèmes mono-genre (par souci de lisibilité nous nous référerons à ces modèles comme modèle féminin et modèle masculin par la suite, même si ces appellations relèvent d'un abus de langage), pour vérifier la pertinence de notre hypothèse. Nous avons maximisé la taille de notre ensemble d'entraînement, atteignant un nombre de livres dans ces systèmes mono-genre de 2671. [4]

Concernant les performances globales, le WER moyen est de 12,3% pour le modèle masculin et de 11,7% pour le modèle féminin sur le test-clean, sans différence statistiquement significative entre les modèles (p-valeur = 0,48) et de 32,3% et 30,9% sur le test-other toujours sans différence statistique (p-valeur = 0,23). On observe, de manière cohérente avec les résultats obtenus sur notre modèle wper70, qu'une plus grande représentation de femmes, et dans le cas présent, une représentation uniquement de femmes, conduit à

4. Il convient donc de noter que la taille des données d'entraînement de ces systèmes est inférieure à celle des données d'entraînement de nos modèles précédents.

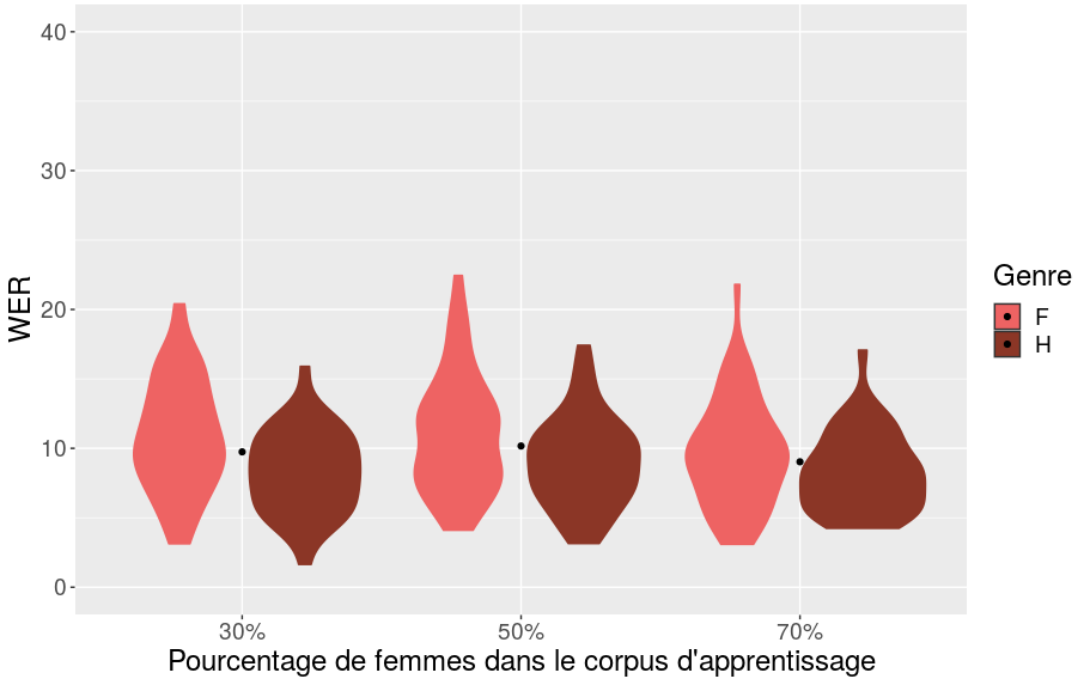

FIGURE 18 – Variabilité due à la représentation du genre : distribution des performances pour nos 3 modèles sur le corpus test-clean. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 40%).

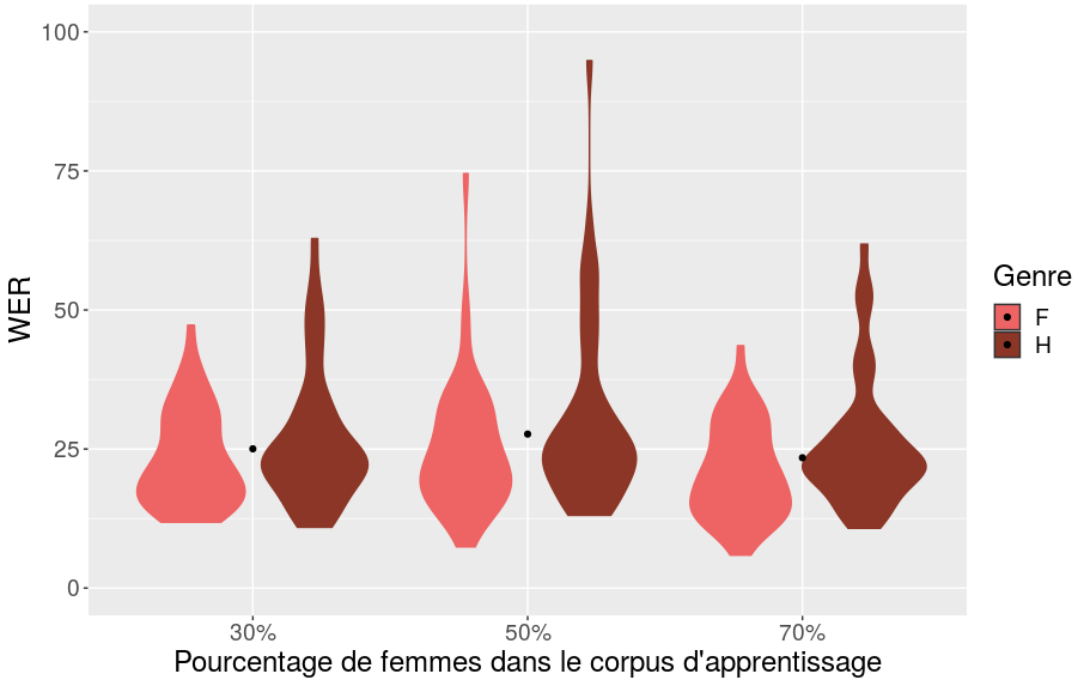

FIGURE 19 – Variabilité due à la représentation du genre : distribution des performances pour nos 3 modèles sur le corpus test-other. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 100%).

Modèle	WER	test-clean			test-other		
		F	H	Tous·tes	F	H	Tous·tes
Masculin	médian	**14,3%**	**8,7%**	12,0%	**36,8%**	**24,9%**	29,2%
	moyen	14,7%	9,11%	12,3%	35,8%	28,8%	32,2%
Féminin	médian	10,6%	11,9%	11,0%	**21,1%**	**36,0%**	26,9%
	moyen	10,9%	12,7%	11,7%	22,7%	38,8%	30,9%

TABLE 8.2 – WER moyens par genre obtenu sur les corpus d'évaluation test-clean et test-other pour nos modèles mono-genre.

de meilleures performances. On observe que les performances globales sont un peu moins bonnes dans les configurations mono-genre que dans nos systèmes précédents, mais on peut supposer que cet écart vient de la quantité de données, qui est plus faible dans le cas des systèmes mono-genre (N=2671) que dans nos précédents modèles (N=3816).

Lorsqu'on analyse le comportement de nos modèles sur les catégories de genre, les performances ne semblent pas suivre des distributions similaires (voir Figures 20 et 21). L'ensemble de valeurs est présenté dans le Tableau 8.2.

Sur le test-clean, on observe une différence statistique dans le modèle masculin (p-valeur $< 10^{-6}$), avec un WER moyen de 9,11% sur les livres lus par les hommes et de 14,7% sur les livres lus par les femmes. Cette significativité statistique est à mettre en lien de la significativité statistique également observée sur le modèle wper30 contenant 70% d'hommes, et dont le modèle masculin peut être envisagé comme un prolongement. Cependant, ce n'est pas le cas pour le modèle féminin. En effet, malgré une différence de WER moyen de 2,2 points de pourcentage, cette différence n'est pas significative (p-valeur = 0,11). Nous retombons sur le résultat obtenu avec notre modèle wper70, où nous observions qu'une sur-représentation des femmes augmentait les performances générales du système, en augmentant les performances pour les livres lus par des femmes sans diminuer celles obtenues sur les livres lus par des hommes. Cette différence de comportement des modèles se retrouve également graphiquement (voir Figure 20). Il est également intéressant de noter que c'est uniquement dans le cas d'un modèle mono-genre féminin que l'on finit par inverser la tendance observée jusqu'à présent d'obtenir de meilleurs résultats de WER pour les voix d'hommes sur le test-clean (sans atteindre la significativité statistique cependant).

Sur le test-other, on observe également de meilleures performances sur la catégorie de genre présente à l'apprentissage. En revanche, les écarts sont cette fois significatifs pour les deux modèles (p-valeur $< 10^{-6}$ pour le modèle féminin et p-valeur = 0,002 pour le modèle masculin), ce qu'on retrouve graphiquement sur les distributions (voir Figure 21).

FIGURE 20 – Cas limite. Distribution des performances par catégorie de genre pour nos 2 modèles monogenre sur le corpus test-clean. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 40%).

FIGURE 21 – Cas limite. Distribution des performances par catégorie de genre pour nos 2 modèles monogenre sur le corpus test-other. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 100%).

8.3 Variabilité du modèle

Nous nous posons également la question de l'impact du modèle lui-même sur les performances du système. En effet, il se pourrait tout à fait que l'aléatoire introduit dans la phase d'initiation des poids du modèle soit responsable des écarts observés, les performances globales n'étant pas significativement différentes d'un point de vue statistique. Afin de contrôler que les comportements que nous observons sont uniquement dus à la variation des données, nous avons réalisé une expérience pour tester la robustesse des modèles à la variabilité introduite par la graine aléatoire lors de l'entraînement. Pour ce faire, nous avons entraîné deux nouveaux modèles (m2 et m3) avec l'ensemble d'entraînement wper50 (équilibré entre les catégories de genre) en changeant uniquement la graine du modèle. Avec le modèle wper50 déjà présenté plus haut, nous avons donc 3 modèles entrainés sur les mêmes données et nous avons comparé leurs performances. Les distributions WER obtenues sont représentées dans les Figures 22 et 23 pour les performances obtenues sur le test-clean et dans l'Annexe D pour les performances sur le test-other.

En effectuant le test de Kruskall-Wallis, aucune différence statistiquement significative n'est observée entre les 3 distributions (p-valeur = 0,17 pour le test-clean et p-valeur = 0,06 pour le test-other). La même observation est faite en comparant les modèles deux à deux. Nous avons, à tort, conclu que notre modèle était robuste à l'aléa introduit lors de l'initialisation dans notre article (Garnerin *et al.*, 2021). Pourtant, lorsqu'on explore les performances en fonction du genre, une fois encore, on observe des variations dans la significativité des écarts entre hommes et femmes : on observe une différence statistiquement significative pour les performances sur le test clean dans le cas du modèle m2 (p-valeur = 0,002) et dans le cas du modèle m3 (p-valeur = 0,005). L'absence de significativité statistique observée dans notre modèle m1 entre WER obtenus sur des livres lus par des hommes et par des femmes pourrait donc s'expliquer uniquement par l'aléatoire du modèle n'ayant dans ce cas précis pas convergé de la même manière. On observe d'ailleurs que des trois modèles, le modèle m1 est celui présentant les plus mauvaises performances même si cette variation n'est pas significative. On retrouve également la tendance selon laquelle les WER sont toujours plus bas pour les hommes, que les écarts soient significatifs ou non. Il nous semble impossible d'exclure l'hypothèse selon laquelle la variabilité intrinsèque des modèles est responsable d'une partie des variations de performances observées entre catégorie de genre, puisque les résultats présentés sont le fruit de modèles appris sur le même corpus d'apprentissage équilibré.

8.4 Variabilité individuelle

Nous pensons que le genre en tant qu'attribut du locuteur ou de la locutrice ne peut pas s'envisager comme donnée homogène et que la variabilité du genre n'est pas capturée

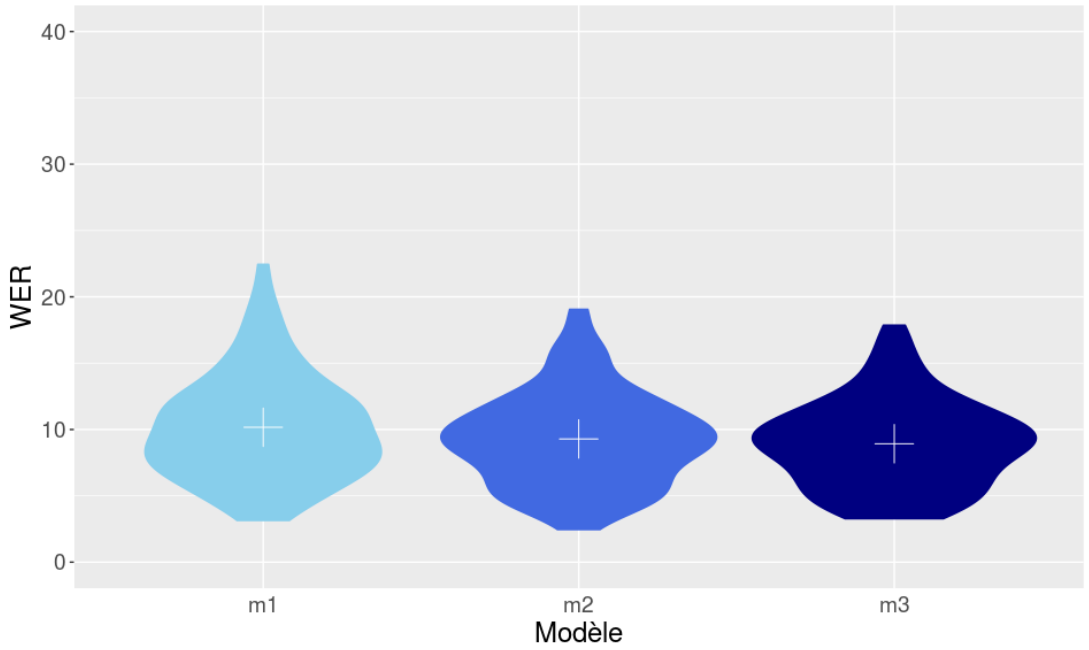

FIGURE 22 – Variabilité due au modèle. Distribution des performances pour nos 3 modèles sur le corpus test-clean. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

FIGURE 23 – Variabilité due au modèle. Distribution des performances pour chaque catégorie de genre pour nos 3 modèles sur le corpus test-other. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 40%).

Modèles	WER	test-clean			test-other		
		F	H	Tous·tes	F	H	Tous·tes
m1	médian	11,1%	8,9%	9,7%	21,9%	25,0%	24,0%
	moyen	11,0%	9,1%	10,2%	25,1%	30,2%	27,7%
m2	médian	**10,2%**	**8,6%**	9,0%	19,6%	22,3%	20,9%
	moyen	10,4%	7,9%	9,3%	21,4%	25,7%	23,6%
m3	médian	**9,7%**	**7,4%**	8,9%	18,7%	22,4%	21,8%
	moyen	9,8%	7,8%	8,9%	21,3%	26,3%	23,9%

TABLE 8.3 – WER par genre obtenus sur les corpus d'évaluation test-clean et test-other pour nos 3 modèles.

Modèles	WER	test-clean			test-other		
		F	H	Tous·tes	F	H	Tous·tes
m1	médian	11,1%	8,9%	9,7%	21,9%	25,0%	24,0%
	moyen	11,0%	9,1%	10,2%	25,1%	30,2%	27,7%
d2	médian	9,1%	7,5%	8,3%	19,7%	21,7%	20,7%
	moyen	9,5%	7,7%	8,7%	20,9%	25,3%	23,2%
d3	médian	9,1%	7,7%	8,6%	20,3%	21,8%	21,3%
	moyen	10,0%	8,0%	9,1%	21,8%	25,5%	23,7%

TABLE 8.4 – WER par genre obtenu sur les corpus d'évaluation test-clean et test-other pour nos 3 modèles.

par des statistiques démographiques. La variabilité intrinsèque de l'indexation du genre (Ochs, 1992) nous amène à affirmer que deux personnes partageant la même "étiquette" de genre ne seront pas interchangeables dans un ensemble de données. Cette hypothèse nous a amené à creuser l'impact de la variabilité individuelle au sein de nos modèles.

Afin de tester cette hypothèse, nous avons créé deux autres ensembles d'entraînement avec le même équilibre 50-50 mais ne contenant pas exactement les mêmes données. Pour ce faire, nous avons fait varier la graine aléatoire du processus de brassage et de sélection des livres pour ces corpus d'entraînement. [5] Nous appelons cet élément aléatoire la "graine de données". En résulte deux modèles (d2 et d3) que nous comparons à notre modèle m1, entrainé sur wper50 et déjà présenté plus haut. Nous avons obtenu les distributions présentées dans les Figures 24 et 25. Lorsqu'on compare les distributions générales, le test de Kruskal-Wallis n'est pas statistiquement significatif pour nos deux ensembles de test, même si les faibles p-valeur et les différences graphiques peuvent nous faire supposer la présence d'une tendance (p-valeur = 0,052 pour le test-clean et 0,043 pour le test-other). On observe également sur le test-clean l'absence de queue de distribution pour le modèle d2, qui obtient de fait les meilleures performances parmi nos trois modèles. L'ensemble des performances sont reportées dans le Tableau 8.4.

5. Les valeurs des graines de données ont été choisies arbitrairement.

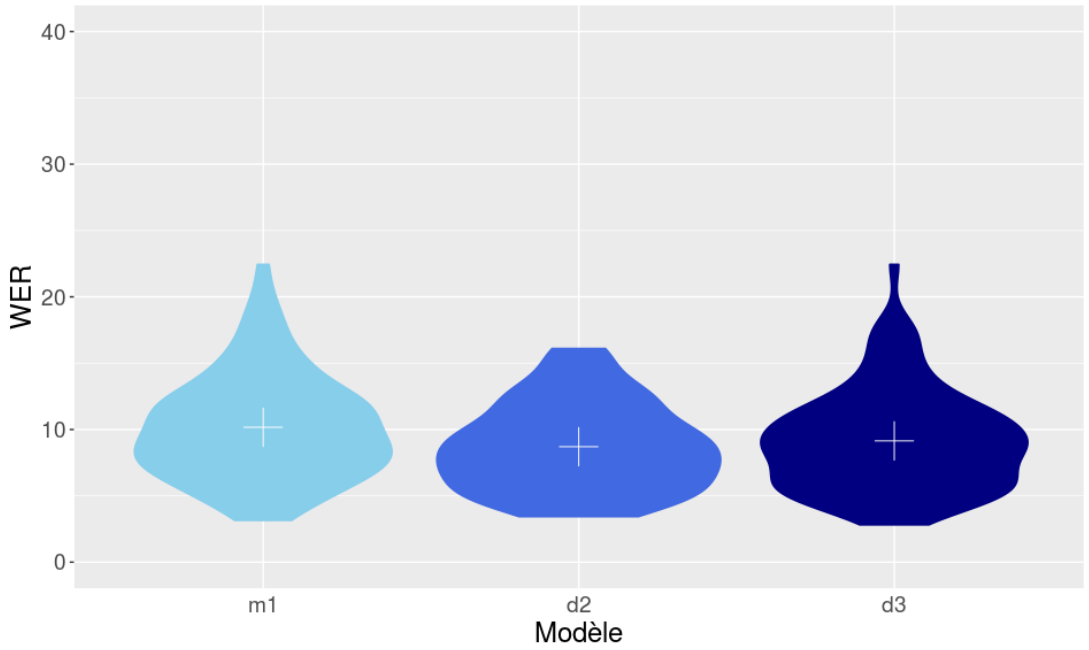

FIGURE 24 – Variabilité due aux données. Distribution des performances pour nos 3 modèles sur le corpus test-clean. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

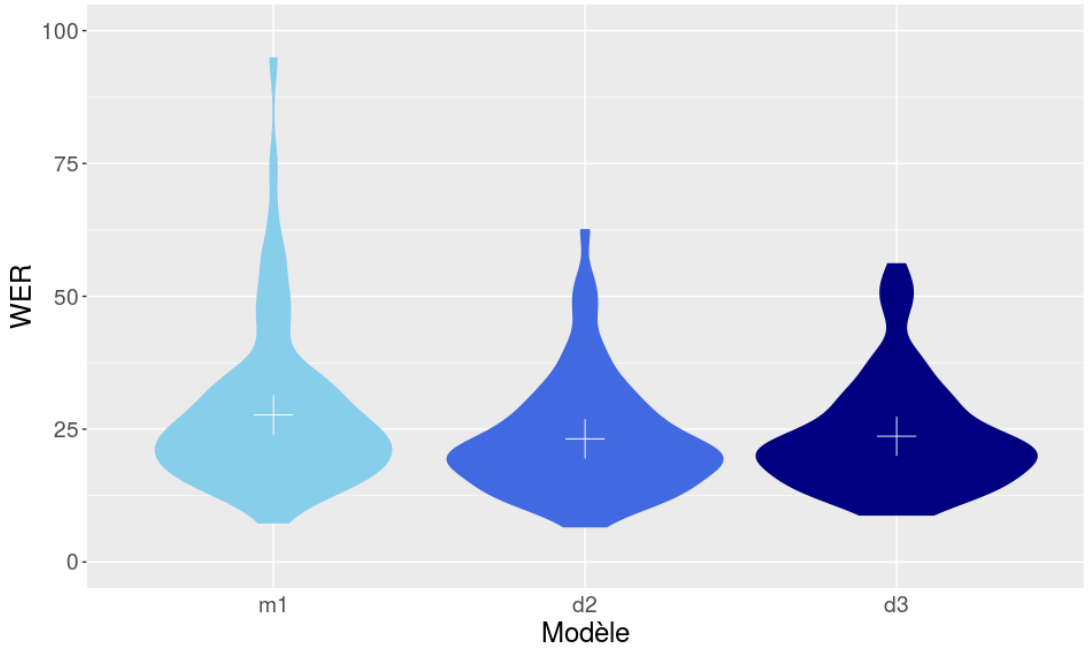

FIGURE 25 – Variabilité due aux données. Distribution des performances pour nos 3 modèles sur le corpus test-other. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 100%).

En examinant les performances obtenues par catégories de genre (voir Figures 26 et 27), nous observons également des comportements distincts entre les modèles. Comme observé jusqu'à maintenant, les WER obtenus sur des livres lus par les hommes sont toujours plus bas que les WER obtenus sur des livres lus par des femmes. Si aucun modèle n'atteint notre seuil de significativité statistique attestant d'une différence de performances entre catégories de genre, la tendance reste quand même forte sur le test clean (p-valeur = 0,04 pour le modèle m1, p-valeur = 0,02 pour le modèle d2 et p-valeur = 0,01 pour d3). Nous ne pouvons donc conclure à un impact de la variation individuelle à partir de ces résultats. Dans notre cadre expérimental, il semblerait que nos modèles soient plus sensibles à l'aléatoire introduit dans le modèle à l'étape d'initialisation des poids qu'à l'aléatoire de sélection des données.

8.4.1 Clarification expérimentale

Lors de la première réalisation de cette expérience, nous avions commis une erreur qui nous avait conduit à constituer deux ensembles d'apprentissage équilibrés mais ne maintenant pas l'effectif total de 3816 livres assurant la comparabilité de nos modèles. Nous avions donc obtenu deux modèles équilibrés à 50-50 mais contenant un total de 5342 livres, dont la totalité de nos 2671 livres lus par des femmes disponibles dans le corpus original et 2671 livres lus par les hommes. Par la suite, nous nous référerons à ces modèles par d2-5342 et d3-5342. Le nombre de livres lus par les hommes étant de 2795 dans le corpus original, notre tirage aléatoire n'a modifié que 118 livres sur les 2671 livres lus par des hommes entre les deux modèles, soit 2,2% de l'effectif total des corpus. Pourtant, nous avons observé des différences de performances significatives. En effet, avec le test de Kruskall-Wallis nous obtenons une p-valeur de 0,003 sur le test-clean dépassant largement notre risque alpha, et nous observons une tendance forte avec une p-valeur de 0,016 sur le test-other.

En analysant les distributions de performances entre catégories de genre, les deux modèles présentent des comportements différents : pour notre modèle d2-5342 nous obtenons un WER médian de 10,7% pour les femmes et 7,7% pour les hommes sur le test-clean, avec une différence significative (p-valeur de 0,008) mais pas sur le test-other (p-valeur = 0,09) malgré un écart de WER du même ordre avec des valeurs de 20,1% pour les femmes et 22,7% pour les hommes. Pour notre modèle d3-5342 nous obtenons des WER de 7,7% pour les femmes et 6,9% pour les hommes sur le test-clean et de 17,69% pour les femmes et 22,9% pour les hommes sur le test-other. Aucune différence significative n'était observée, ni sur le test-clean (p-valeur = 0,038), ni sur le test-other (p-valeur = 0,146). Les distributions sont reportées dans l'Annexe D.

Ces résultats sont particulièrement surprenants, car une variation minime de notre ensemble d'apprentissage (2,2% des données) nous conduit à des différences statistique-

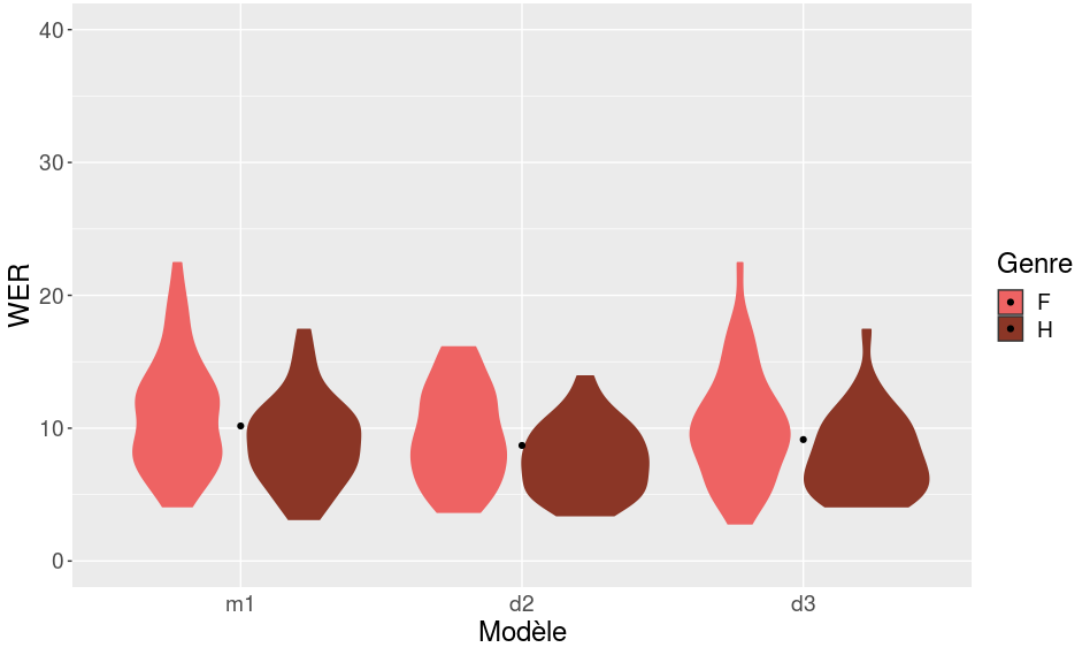

FIGURE 26 – Variabilité due aux données. Distribution des performances pour chaque catégorie de genre pour nos 3 modèles sur le corpus test-clean. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 40%).

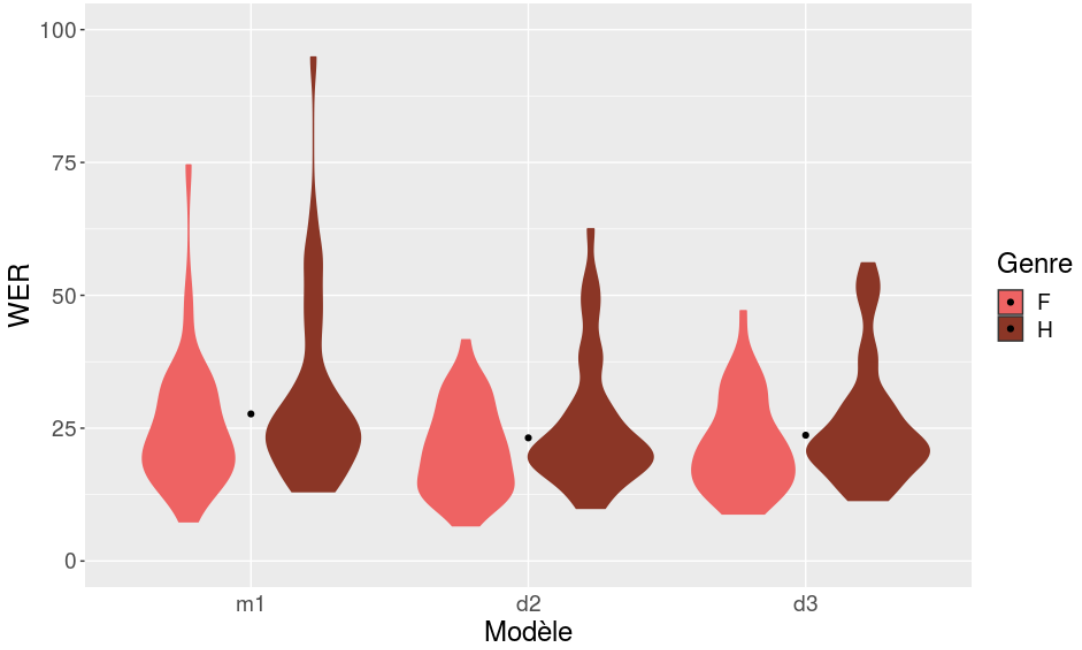

FIGURE 27 – Variabilité due aux données. Distribution des performances pour pour chaque catégorie de genre nos 3 modèles sur le corpus test-other. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 100%).

ment significatives. Nous nous posons la question de savoir comment interpréter de tels résultats, une meilleure compréhension du fonctionnement interne des systèmes pourrait nous aider à éclairer les sources de telles variations.

8.5 Impact du contenu textuel

Notre travail s'est principalement articulé sur la représentation du genre dans les enregistrements audio servant de base au système. Mais nous nous sommes également posée la question de savoir si les contenus des livres lus par des hommes et par des femmes pouvaient différer, et si ceux-ci pouvaient être à l'origine de certaines variations de performances. Nous avons donc tenté de voir si nos valeurs de WER pouvaient être corrélées avec des mesures de perplexité, puisque pour qu'un modèle de langue soit efficace, il faut que sa perplexité soit minimisée. Mais comme d'autres résultats de la littérature (Iyer *et al.*, 1997; Chen *et al.*, 1998), aucune corrélation n'a été décelée entre nos performances et les valeurs de perplexité par livres.

8.6 Discussion & Conclusion

Il pourrait relever du bon sens d'affirmer que toutes les catégories de genre doivent être représentées dans le corpus d'entraînement d'un système ASR afin d'être capable de transcrire la parole indépendamment du genre de l'utilisateur ou de l'utilisatrice. Contrairement à ce que nous pensions, les résultats obtenus dans nos expériences semblent questionner l'évidence d'une telle proposition. Nous nous attendions à ce que les performances obtenues sur chaque catégorie de genre reflètent leur représentation dans les données d'apprentissage. Or lorsqu'on s'intéresse aux performances générales des modèles, il semble que ceux-ci soient relativement robustes à la représentation du genre dans les données d'apprentissage.

En revanche, la distribution des WER n'est pas toujours également répartie entre catégories de genre, venant réactualiser notre discours sur la nécessité d'évaluation en fonction du genre des locuteurs et locutrices. On observe une différence significative, entraînant un taux d'erreur plus important chez les femmes dans le cas de notre modèle entraîné sur une majorité de livres lus par des hommes, et évalué sur notre corpus de test-clean. Pour autant, dans le cas de systèmes entraînés sur des corpus équilibrés, on observe tantôt des écarts de performances entre hommes et femmes, tantôt pas de différence significative. Il nous semble donc qu'une partie de cette variabilité ne s'explique pas par les données présentées au système, mais par l'aléatoire intrinsèque du modèle lors de l'initialisation des poids. Ce résultat nous laisse à penser que les démarches actuelles visant à comprendre le fonctionnement interne des systèmes sont nécessaires pour expliquer ce type d'observations.

Si les écarts de WER entre nos catégories ne sont pas réguliers, on observe quand même une tendance générale sur l'ensemble de nos modèles, tendance selon laquelle les WER sont plus bas sur les livres lus par des hommes, à l'exception du modèle mono-genre féminin (voir le Tableau 8.5). De manière surprenante, les modèles entraînés sur des données majoritairement féminines (les modèles wper70 et mono-genre féminin) réussissent très bien à généraliser sur des voix d'hommes et permettent de minimiser à la fois le WER et les différences de WER médians entre hommes et femmes. Cela pourrait nous amener à conclure, de manière légèrement provocative que "l'homme est une femme comme les autres". On peut alors se poser la question de savoir quelles sont les caractéristiques acoustiques des voix de femmes, impliquant qu'un système ayant appris à modéliser correctement ces voix soit à même de reconnaître une voix d'homme, là où l'inverse est moins observé dans nos expériences.

Expérience	Modèles	test-clean			
		F	H	W	p-val.
Référence	Original	4,26	3,36	1108	0,131
Var. représ.	wper30	**10,3%**	**8,0%**	**1279**	**0,003**
	wper50	11,1%	8,9%	1176	0,036
	wper70	9,4%	8,3%	1123	0,101
Var. modèle	m2	**10,2%**	**8,6%**	**1296**	**0,002**
	m3	**9,7%**	**7,4%**	**1259,5**	**0,005**
Var. données	d2	9,1%	7,5%	1199	0,022
	d3	9,1%	7,7%	1216	0,014
Sys. mono.	Féminin	10,6%	11,9%	746	0,114
	Masculin	**14,3%**	**8,7%**	**1540,5**	$\mathbf{1{,}871}e^{-7}$
	d2-5342	**10,7%**	**7,7%**	**1240**	**0,008**
	d3-5342	7,7%	6,9%	1173	0,038

TABLE 8.5 – WER médians, statistique W et p-valeur du test de la somme des rangs de Wilcoxon comparant la différence entre les distributions de WER entre hommes et femmes pour nos différents modèles sur le test-clean. Risque α fixé à 0,01.

Lorsqu'on s'intéresse aux résultats obtenus sur le test-other (voir Tableau 8.6), on observe une tendance inverse, avec de meilleures performances sur les voix de femmes. Pour autant, la significativité statistique n'est jamais atteinte (sauf dans le cas des modèles mono-genre) et nos variations de genre étant assez fines, nous supposons qu'elles se perdent dans la variabilité générale. Le test-other contenant des livres que les systèmes ont du mal à transcrire (car ayant obtenu les plus mauvais WER avec un système appris sur le WSJ) nous pensons que les facteurs de variabilité et de difficulté sont trop nombreux. La représentation du genre ne permet donc pas d'expliquer les variations observées et il faudrait creuser les données dans des travaux ultérieurs pour comprendre les mécanismes à l'oeuvre, notamment peut-être en termes de différences d'accent puisque l'article de Librispeech soulignait que les deux partitions de clean contenaient en moyenne, des

enregistrements de meilleure qualité avec des accents plus près de l'accent standard de l'anglais des États-Unis (Panayotov *et al.*, 2015, p.5208).

Expérience	Modèles	test-other			
		F	H	W	p-val.
Référence	Original	10,4	13,0	771	0,052
Var. représ.	wper30	21,2%	23,4%	856	0,211
	wper50	21,9%	25,0%	834	0,153
	wper70	19,5%	22,8%	749	0,034
Var. modèle	m2	19,6%	22,3%	797	0,083
	m3	18,7%	22,4%	750	0,035
Var. données	d2	19,7%	21,7%	827	0,137
	d3	20,3%	21,8%	826	0,135
Sys. mono.	Féminin	**21,1%**	**36,0%**	**387**	**1,375**e^{-7}
	Masculin	**36,8%**	**24,9%**	**1391**	**0,002**
	d2-5342	20,1%	22,7%	799	0,086
	d3-5342	17,7%	20,4%	831	0,146

TABLE 8.6 – WER médians, statistique W et p-valeur du test de la somme des rangs de Wilcoxon comparant la différence entre les distributions de WER entre hommes et femmes pour nos différents modèles sur le test-other. Risque α fixé à 0,01.

Comme écrit plus haut, les meilleures performances sont obtenues pour le modèle wper70, dont le corpus d'apprentissage contenait 70% de livres lus par des femmes et 30% de livres lus par des hommes. On pourrait creuser si une sur-représentation des femmes conduit toujours à de meilleures performances globales et par catégorie de genre et si une quantité de données "suffisante" de voix d'hommes est nécessaire (30% ou moins) ou si à taille de corpus d'apprentissage égale, on n'observe pas de différence entre système mono-genre féminin et système majoritairement féminin.

Dans l'ensemble, il nous est difficile de conclure que la représentation du genre dans les données d'entraînement a une forte influence sur les performances du système. Celle-ci ne joue pas au niveau des performances globales, et dans le cas des distributions de WER par catégorie de genre cet impact n'est pas symétrique ni systématique. La question n'est donc pas aussi simple qu'il nous avait semblé de prime abord. Ce que ces résultats nous montrent, c'est que nous ne sommes aujourd'hui pas capables de définir quels sont les paramètres acoustiques qui jouent dans nos systèmes. Les étiquettes binaires de genre, si elles recouvrent peut-être une partie de ces variations acoustiques ne sont pas suffisamment robustes dans leur capacité descriptive pour expliquer d'où proviennent les différences de performances observées. Nous montrons également que la vérification de l'existence de biais prédictif en fonction du genre ne peut pas être évacuée avec un simple contrôle de la parité démographique dans les corpus d'apprentissage. Dans notre cas, s'il apparaît que les voix d'hommes sont généralement mieux reconnues (sur le test-clean), il semble que l'augmentation de la proportion de voix de femmes dans le corpus d'entraînement contribue à réduire les différences de performances différenciées entre catégories, tout en

assurant le même niveau de performance globale. Ce type de résultat nous empêche de conclure simplement que nos systèmes sont suffisamment robustes au genre, pour évacuer la question dans sa totalité. Si le genre a constitué pour nous une entrée dans la variabilité de la production vocale qui doit être prise en compte pour que le système soit capable de généraliser, il nous est aujourd'hui impossible de séparer cette variation genrée de la variation individuelle, la première étant contenue dans la seconde. Si la prise en compte de l'équilibre des genres dans les données d'entraînement est un point de départ pour des systèmes plus équitables, essayer de quantifier l'intra-variabilité de nos ensembles d'entraînement pour estimer une mesure d'adéquation avec nos données de test apparaît comme une piste forte pour comprendre les mécanismes à l'oeuvre dans nos modèles. De plus, revenir à une conception de la variabilité individuelle via des mesures acoustiques telles que la fréquence fondamentale et le débit de parole pour évaluer cette adéquation pourrait nous permettre de sortir de la matrice binaire du genre.

Chapitre 9

Sortir de la binarité ?

Jusqu'à présent nous avons utilisé dans nos expériences la terminologie du genre pour décrire une réalité binaire composée d'hommes et de femmes. Nous avons justifié notre utilisation de la terminologie du genre et non celle du sexe car celle-ci permet selon nous de faire un pas de côté face aux discours biologisants de la différence des sexes, nous permettant ainsi de mettre en lumière le système de pouvoir qui la constitue. Nous avons également justifié notre choix en référence aux travaux théoriques d'inspiration butlerienne qui conçoivent le genre comme une performance. Parler de genre ouvre donc une possibilité de sortir du binarisme et de reconsidérer le genre et la voix comme des continuum, des ensembles de variations possibles, et non plus comme des catégories fixes et anhistoriques. Si nous avons questionné cette catégorisation discrète pour envisager le genre et la voix comme des objets relevant du continu, il n'en reste pas moins que nos expériences restent dans cette matrice binaire des étiquettes homme/femme, au risque de venir réactualiser des discours sur la différence des sexes. Nous avons justifié ce choix par la qualité même de nos données et par la nature de notre questionnement, s'inscrivant dans une histoire de l'invisibilisation et de la problématisation de la voix des femmes. Pour autant, il existe aujourd'hui des personnes qui s'identifient en dehors de ces catégories et remettent en question le système hégémonique du genre. Notre expérience sur Librispeech nous laisse également à penser que les catégories binaires du genre, si elles recouvrent une réalité sociale aujourd'hui, ne parviennent pas à expliquer les variations observées.

Le présent chapitre s'intéressera donc à discuter de la possibilité de sortir de la binarité dans le cadre des systèmes d'ASR. Sortir de la binarité permettrait selon nous, à la fois d'articuler les systèmes avec les évolutions de nos réalités sociales, mais aussi d'en comprendre mieux le fonctionnement. Quelles en seraient les conséquences et comment articuler une nouvelle typologie ou l'absence de typologie avec nos systèmes techniques ? Ce sont les questions auxquelles nous essayerons de répondre dans les prochaines sections, en essayant de nous affranchir des étiquettes du genre via un retour aux mesures acoustiques.

9.1 Catégorisation & binarité

Le fait de compter, dénombrer, classifier façonne une réalité sociale. Nous avons déjà évoqué dans le Chapitre 2, les travaux de Geoffrey Bowker et Susan Star (2000) qui ques-

tionnent les processus de création des systèmes de classification et leurs conséquences. Choisir nos catégories d'analyse, ce que l'on compte, constitue donc déjà une prise de position et une manière de voir le monde. Les mouvements queer et plus récemment non-binaires sont l'expression collective du refus du système de catégorisation du genre hégémonique. En multipliant les catégories du genre, on en montre les variabilités intrinsèques ainsi que l'absence d'homogénéité dans les réalités que les catégories binaires veulent décrire. La création de nouvelles catégories souligne également toutes les réalités que les catégories homme/femme ne parviennent pas à couvrir. Il s'agit alors de dépasser la binarité, voire de montrer les insuffisances d'un système catégoriel du genre dans son ensemble. Dans ce travail, on entend la non-binarité comme désignant aussi bien l'expérience d'une personne ayant une identité de genre en dehors des catégories hégémoniques "homme" et "femme" (*genderqueer*, *pangender*, *androgyne*), mais aussi celle d'une personne se sentant "homme" ou "femme" voire d'un troisième genre selon les contextes (*genderfluid*, *bigender*), voire rejetant toute entière l'identité de genre comme caractéristique identitaire (*agender*, *genderless*, *neuter*). [1]

Compter et raconter les expériences des minorités (pas uniquement celles de genre) permet de les rendre visibles pour adapter les politiques publiques (et/ou les systèmes), mais rend également ces populations identifiables et donc vulnérables. Dans leur ouvrage *Data Feminism*, Catherine d'Ignazio et Lauren Klein exposent les dangers qui peuvent survenir avec l'accès à la visibilité via la notion de *paradox of exposure* : « the double bind that places those who stand to significantly gain from being counted in the most danger from the same counting (or classifying) act. » (2020, p. 105). Elles prennent pour exemple le cas des personnes en situation irrégulière, mais le *paradox of exposure* peut aussi s'appliquer aux personnes existant en dehors des représentations binaires du genre, majoritaires aujourd'hui dans notre société. [2] Si ce travail se pense comme un premier questionnement face à la possibilité de sortir du binarisme du genre dans le cadre des systèmes d'ASR, il n'en reste pas moins nécessaire de rester vigilants et vigilantes quant aux possibles implications de ces choix techniques.

1. On peut également souligner que la non-binarité qui nous semble être un phénomène relativement récent dans le monde occidental existe déjà dans d'autres cultures. Parmi les exemples les plus cités se trouvent, par exemple, les *hijras* en Inde (Nanda, 1986), les *two-spirit* chez les peuples Premiers (Robinson, 2020), et les *muxes* dans la communauté zapotèque au Mexique (Mirandé, 2016).

2. Selon les rapports annuels sur les LGBTIphobies (l'ensemble des actes de violence envers les personnes LGBTI) de 2021, l'année 2020 est la première depuis 2015 pour laquelle on note une baisse du nombre de témoignages recueillis par l'association. L'objectif de l'association SOS Homophobies étant principalement l'écoute, le genre des personnes n'est pas toujours une information prise en compte dans leur statistique, mais il n'existe pas à notre connaissance d'organisme qui ne recense que les violences dues au genre.

9.2 Coder la non-binarité : quelles limites techniques ?

Changer nos catégories implique également de penser comment ces nouvelles classifications vont interagir avec nos systèmes. Le genre étant une catégorie particulièrement saillante dans nos vies [3], il est à prévoir qu'un changement social de typologie du genre, soit amené à être pris en compte par les systèmes. C'est d'ailleurs le choix qu'a fait Facebook (et d'autres réseaux sociaux [4]) : en 2014, le réseau social a changé les options proposées aux utilisateurs et utilisatrices souhaitant créer un compte : en plus des catégories binaires "homme" "femme", une troisième option "personnalisée" donnait accès à une liste déroulante de plus de 52 catégories, parmi laquelle se trouvaient des catégories comme *gender questionning* ou encore *agender*. Cette solution a ensuite été remplacée par un champ de saisie, laissant à la personne la possibilité de s'auto-définir (ou de laisser le champ vide), tout en lui proposant de choisir ses pronoms, parmi *she*, *he* ou *them*. [5] Comme le notait Sam Bourcier :

> « Qu'il s'agisse de revendiquer une identité existante comme celle d'"homosexuel", de "transsexuel" ou de "gay" ou d'identités sexuelle et de genre nouvelles directement issue des subcultures LGBTQI comme "genderqueer" ou "genderfluid", les autodéterminations de genre offertes sur le réseau social s'inscrivent dans ce mouvement de visibilisation et d'irruption dans l'espace public des minorités sexuelles et de genre et des identités qu'elles génèrent. C'est important, parce que même ceux qui ne vivent pas dans ces subcultures comme tous ceux et toutes celles qui étouffent dans des formes de masculinité ou de féminités imposées vont voir qu'il existe non pas 2 sexes/2 genres mais n genres et n sexes et pouvoir l'afficher et le vivre. » [6]

Rena Bivens, dans un article de 2017, s'est intéressée à l'implémentation de ce changement de paradigme (Bivens, 2017). Elle a montré, à l'aide de requêtes via l'API Facebook, comment était encodé le genre , dans la base de données du réseau social. Elle observe que la multiplication des catégories proposées n'est au final pas une réalité dans la structure des données : le choix de pronom (rendu obligatoire par Facebook) supplantait l'auto-description du genre, et une personne ayant décidé de s'auto-catégoriser comme non-binaire par exemple, sera codée comme "femme" si le choix de son pronom s'est porté sur *she*. Le choix de Facebook quant à ce codage du genre suppose que pour les publicitaires, hommes et femmes constituent des groupes sociaux aux comportements identifiés

3. Lorsqu'il nous est demandé de nous décrire, on commence souvent par "je suis un.e X", l'existence même d'un enjeu politique autour des catégories de genre démontre l'importance de celles-ci dans notre organisation sociale.

4. Voir Bivens et Haimson (2016)

5. Si les interfaces Facebook peuvent varier en fonction des pays, cette interface et le choix obligatoire des pronoms est également implémentée dans la version française. En revanche l'option "neutre" correspond à l'utilisation du pronom "lui", ce qui est largement discutable.

6. http ://www.slate.fr/culture/83605/52-genre-facebook-definition

et de bonnes “cibles” marketing. Dans leurs objectifs de vente, l'existence de personnes en dehors de ces catégories ne constitue pas une part de marché suffisante pour remettre en cause leur système de catégorisation. Dans un monde où les normes comportementales sont construites autour d'un système de classification du genre binaire, sortir de cette binarité dans les systèmes nous demande peut-être un peu de volontarisme.

Les modèles d'ASR E2E ne reposent plus sur des modèles acoustiques du genre contrairement à certains systèmes HMM-GMM ou hybrides, et ils constituent de ce point de vue, une possibilité de prise de distance par rapport à la construction de la voix comme intrinsèquement genrée. Mais qu'est-ce que la non-binarité modifie dans la perception et dans la production de la voix genrée ? C'est la question du projet de recherche NoBiPho, dont les objectifs sont, entre autres :

> « [...] de mettre à l'épreuve les paradigmes émergents qui envisagent le genre comme un système non-binaire, affranchi du postulat de dimorphisme sexuel. Le second objectif est de proposer des modélisations nouvelles susceptibles de rendre compte de la diversité humaine ainsi que des capacités à s'adapter à de nouveaux paradigmes. Dans le protocole proposé nous évaluons la possibilité de décrire le genre selon des modèles avec deux, trois, quatre ou plus de catégories, ou même d'éliminer le genre comme critère pertinent et opérationnel dans les descriptions. Cette démarche méthodologique sera utilisée pour proposer de revisiter les nombreuses études précédentes qui ont mis en avant une bi-partition genrée de différents marqueurs ou paramètres phonétiques, afin de réinterpréter les résultats obtenus dans un paradigme binaire. » [7]

Si nous ne sommes pas en mesure de proposer un nouveau modèle du genre qui serait pertinent dans le cas de nos systèmes, s'affranchir d'un paradigme catégoriel en revenant à des mesures acoustiques pourrait être une piste intéressante pour penser la sortie de la binarité pour les systèmes d'ASR.

9.3 Des catégories au continuum : retour à l'acoustique

Nous avons tenté jusqu'à présent de questionner l'existence de biais prédictifs genrés en fonction de la représentation des catégories de genre dans les données d'apprentissage. Nos observations sur les systèmes E2E, nous laissent à penser que le genre ne constitue peut-être pas une grille d'analyse pertinente pour questionner les variabilités à l'origine des disparités de performances observées dans ces systèmes.

Nous avions d'abord imaginé que ce qui se jouait dans nos systèmes relevait plus d'une question d'adéquation de données d'apprentissage et de test que d'une biais de

7. `https://nobipho.hypotheses.org/a-propos`

sélection basé sur des étiquettes démographiques recouvrant des réalisations vocales parfois très différentes. Pour autant, nos résultats obtenus dans le Chapitre 8 selon lesquels une sur-représentation de voix de femmes ne conduisait pas à une baisse significative des performances obtenues sur les voix d'hommes tout en augmentant celles obtenues pour les femmes, ne nous permettent pas d'affirmer que les différences observées se résument à une question de couverture d'un ensemble de variabilités. Cela nous amène à questionner ce qui se joue acoustiquement dans les voix, pour nous amener à de tels résultats. Nous avons donc souhaité explorer nos WER en fonction, non plus d'étiquettes de genre définies à priori, mais à travers des mesures acoustiques effectuées sur nos données.

9.3.1 Analyses préliminaires

Une manière d'essayer de comprendre nos systèmes tout en sortant des catégories de genre serait de creuser quels sont les facteurs acoustiques qui jouent dans le WER, lorsqu'une différence significative est observée : à savoir, le modèle wper30 et les modèles mono-genre. [8] Passer par des paramètres acoustiques permet de sortir de la perception catégorielle, qui est, elle, construite socialement, et de revenir sur un continuum physique.

9.3.1.1 Rôle de la fréquence fondamentale

Nous avons discuté du rôle que joue la fréquence fondamentale dans la perception du genre dans la voix. Une hypothèse sur l'existence de différences de performances entre nos catégories serait que des différences de fréquence fondamentale jouent sur les performances du système et que, notre société ayant construit un lien entre cette fréquence et le genre, des sensibilités aux variations de fréquence fondamentale peuvent être interprétées comme des différences genrées. Nous avons donc regardé si dans nos résultats, une corrélation (qui n'est pas pour autant une relation de causalité) existait entre nos mesures de F0 médianes et nos WER (voir Figure 28).

Tout d'abord, on observe que dans notre échantillon de test, la F0 permet de discriminer de manière presque parfaite nos locuteurs et locutrices (hormis un locuteur avec une F0 médiane élevée), dès lors une sensibilité aux F0 élevées pourrait être interprétée comme une différence de genre. On observe une tendance dans le cas de notre modèle wper30 corrélant le WER avec la F0 médiane (r = 0,22, p-valeur = 0,04) celui-ci augmente lorsque la F0 augmente. Cette corrélation est faible et décroit encore dans nos modèles appris sur les corpus wper50 et wper70. On remarque que la pente diminue au fur et à mesure que le pourcentage de femmes dans notre corpus d'apprentissage augmente. On peut mettre en parallèle ces absences de corrélation entre F0 et WER avec la diminution de l'écart entre WER médians obtenus sur des livres lus par des femmes et lus par des hommes,

8. Nous n'avons pas travaillé sur nos modèles m2 et m3 car les données d'apprentissage restaient les mêmes et nous avons conclu que les différences de performances étaient dûes à la variabilité du modèle.

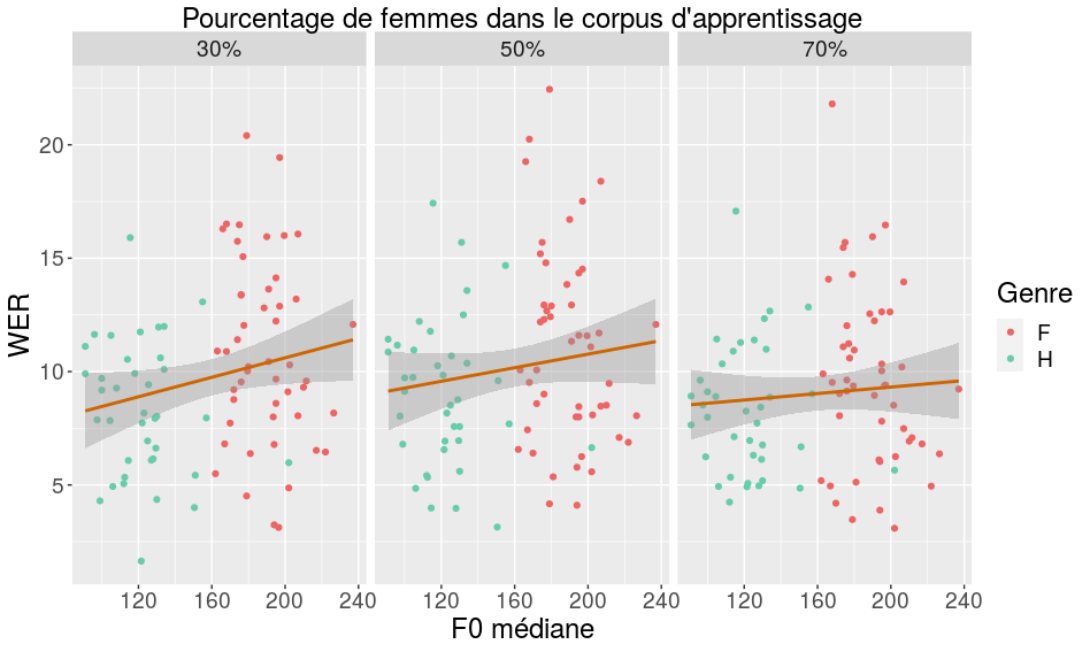

FIGURE 28 – WER en fonction de la F0 médiane obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

puisque la fréquence fondamentale a été érigée en caractère discriminant pour définir les catégories de genre dans la voix.

Nous avons également exploré l'existence d'une corrélation entre WER et écart interquartile de F0, mais aucune corrélation n'a été trouvée. Les mêmes expériences ont été conduites avec des mesures de F0 en semi-tons et des résultats similaires ont été obtenus, que ce soit pour la valeur médiane de la F0 ou pour l'écart interquartile. L'ensemble des graphiques correspondants est reporté dans l'Annexe F.

Dans le cas des modèles mono-genre, on observe aussi une corrélation entre WER et F0 médiane comme le montre la Figure 29. Cette association est plus forte dans le cas du modèle masculin, comme le montre la pente, et on observe bien des changements de direction selon le modèle mono-genre utilisé (corrélation positive dans le cas du modèle masculin et négative dans le cas du modèle féminin). On peut donc conclure que dans nos systèmes sont construites des modélisations acoustiques dépendantes de la F0 et que plus on s'éloigne de ces valeurs canoniques, plus mauvaises seront les performances. Dans le cas de notre modèle féminin, la corrélation entre F0 médiane et WER n'est qu'une tendance ($r = -0{,}22$, p-valeur $= 0{,}04$), là où elle dépasse largement le seuil de significativité dans notre modèle masculin ($r = 0{,}48$, p-valeur $< 10^{-5}$). Ces différences de significativité sont encore une fois à mettre en lien avec nos écarts de WER médians par catégorie de genre, statistiquement significatif dans le cas du modèle masculin et pas dans le cas du modèle féminin. On peut donc se poser la question de savoir si ces résultats sont l'expression

d'un lien entre F0 et WER, ou si un modèle appris sur des fréquences fondamentales plus élevées est plus performant pour généraliser sur des F0 plus basses, que dans le scénario contraire.

Sur le test-other où les différences de WER sont significatives entre catégories de genre, on retrouve également des corrélations statistiquement significatives entre WER et F0 médiane, comme le montre la Figure 30. On observe également que sur le test-other, la catégorisation homme/femme via la mesure de F0 est moins discriminante que dans le cas de notre test-clean.

9.3.1.2 Rôle du débit

Dans le cadre de notre première expérience sur des données médiatiques francophones, nous avions observé l'impact du rôle des locuteurs et locutrices sur les performances. Les rôles étant associés à un temps de parole, mais également ce que nous avons appelé une parole "professionnelle", nous avons effectué des mesures de débit articulatoire. Nous posons donc qu'une partie de la réalité physique de nos catégories de rôles réside dans une certaine pratique articulatoire, dont le débit serait un facteur.

Nous avons fait des mesures de moyenne et d'écart-type et si des corrélations sont parfois trouvées, leur significativité ou absence de significativité ne correspond pas aux systèmes pour lesquels des écarts de performances entre les catégories de genre sont observées. Cela n'est pas surprenant dans l'absolu, car à notre connaissance, il n'existe pas d'association dans nos représentations, entre débit et catégorie de genre. (Les graphiques sont reportés dans l'Annexe F).

9.3.2 Autres paramètres

Nous n'avons pas observé de différences de performances en fonction de la durée qui contribuerait à expliquer nos résultats. Pour autant, il nous semble que d'autres facteurs identifiés comme permettant de percevoir le genre dans la voix, comme les fréquences de résonances, pourraient constituer une piste supplémentaire à creuser pour expliquer nos différences de performances.

9.4 Sortir de la binarité ou sortir du genre ?

Notre travail questionnait l'existence de biais prédictif en fonction des catégories genrées dans les systèmes d'ASR. L'existence de biais prédictif était abordée à la fois comme problème technique, mais également comme problème social, avec les discriminations qui peuvent en découler. Sortir des catégories binaires du genre permet de donner de la visibilité aux minorités de genre et d'appuyer la lutte contre les discriminations de genre. Comme l'écrivent Rena Bivens et Oliver Haimson (2016, p. 8) : « The values they bake

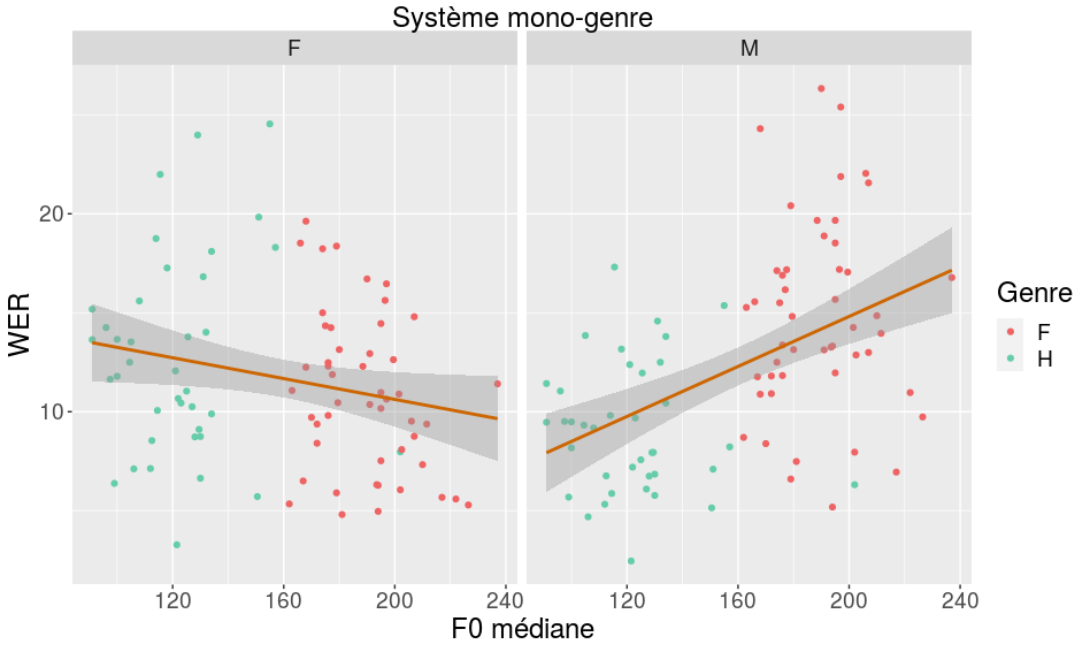

FIGURE 29 – WER en fonction de la F0 médiane obtenus sur le test-clean pour nos 2 systèmes mono-genre.

FIGURE 30 – WER en fonction de la F0 médiane obtenus sur le test-other pour nos 2 systèmes mono-genre.

into their software have the capacity to influence the next generation of platforms that will go on to play an intermediary role in shaping society's construction of itself. ». Pourtant, si sortir du genre peut nous amener à mieux comprendre certains phénomènes que nous n'envisageons aujourd'hui qu'à travers des causalités d'un binarisme homme/femme construit, se débarrasser des catégories hégémoniques n'est pas non plus sans conséquences. Les auteurs et autrices écrivent également :

> « While a genderless Internet would force the advertising industry to focus on behavioral and taste-based data, or at least other demographic characteristics, this move would not eliminate gender-based discrimination and marginalization (Kendall, 1998). In the context of race and ethnicity, for instance, online spaces that do not explicitly allow race to be presented or specified do not become raceless utopias ; instead, these spaces tend to erase race while positioning whiteness as the default (Kolko, Nakamura, & Rodman, 2013 ; Nakamura, 2002). » (Bivens et Haimson, 2016, p. 8)

Le risque en voulant sortir de nos catégorisations, et donc de celle du genre, serait de revenir sur un impensé des différences que cette catégorisation a contribué à construire au cours de l'histoire et qui est aujourd'hui une réalité sociale. C'est d'ailleurs le propos de Betsy Anne Williams *et al.* (2018), même si leur étude porte principalement sur la race aux États-Unis. Nous soutenons donc qu'il serait important de réussir à proposer une mesure de couverture vocale (en creusant les mesures mathématiques possibles et les facteurs acoustiques pouvant être responsables de variations de WER) pour comprendre le fonctionnement de nos systèmes et en réduire les erreurs, tout en continuant conjointement à les évaluer au regard des catégories de genre existantes dans nos sociétés. Si les étiquettes de genre n'ont peut-être pas de capacité descriptive suffisante pour expliquer les variations des systèmes d'ASR, elles restent cependant une réalité sociale, et perçue, sur laquelle se basent les processus de discriminations. Cela vaut pour les étiquettes du genre binaire, mais n'entre pas en opposition avec l'utilisation d'autres étiquettes, les personnes non-binaires étant également victimes de discriminations et invisibilisées dans nos sociétés. Il nous semble donc pertinent de continuer d'affirmer que des étiquettes de genre sont nécessaires pour contrôler la représentation des femmes dans les données, mais également pour évaluer les performances des systèmes en fonction des catégories, ces catégories ne se limitant pas nécessairement aux catégories binaires. Le choix du système de catégorisation en revanche, devra faire l'objet d'autres travaux par des experts et expertes du domaine, à l'instar du projet de recherche NoBiPho.

Conclusion

Dans ce travail, nous nous sommes posée la question de l'existence de biais en fonction du genre dans les systèmes d'ASR. Nous avons ainsi, au fil de notre travail théorique, délimité notre question aux biais prédictifs, notamment comme résultat de biais de sélection des données. Nous nous sommes appuyée sur une conception hybride du genre compris à la fois comme un rapport social et une performance individuelle, et circonscrit dans un premier temps à la matrice binaire hégémonique, soulevant le risque de renforcer cette catégorisation binaire. Nous avons présenté deux systèmes d'ASR sur lesquels nous avons travaillé, à savoir un système hybride HMM-DNN entraîné sur des données issues d'émissions radiophoniques et télévisuelles francophones ainsi qu'un système E2E entraîné sur des livres audio en anglais.

Nous avons exploré l'existence de ce biais via nos expériences faites sur les grands corpus médiatiques du français dans lesquels les femmes étaient sous-représentées (Chapitre 7), puis sur le grand corpus LibriSpeech plus homogène pour lequel nous avons fait varier la représentation des hommes et des femmes (Chapitre 8). Ainsi, nous avons observé l'existence de biais prédictifs genrés dans certaines configurations : les performances étaient moins bonnes sur les voix de femmes dans le cas des locutrices ponctuelles. En revanche, nous n'observons pas de différence significative entre les Ancres – les locuteurs et locutrices largement représentées dans nos données, et que nous assimilons à des professionnelles de la parole (présentatrice, expert, chroniqueuse, etc.).

Cette première étude montre que le genre est un facteur de variation dont l'impact sur les distributions de performances varie en fonction d'autres facteurs.

On constate également qu'il n'est d'ailleurs pas la source de variations la plus importante, le rôle ou le type de parole conduisant à des disparités plus grandes. En revanche, nous remarquons un phénomène de cumul des difficultés, et s'éloigner du rôle ou du style de parole canonique vient accentuer les disparités hommes/femmes. Dans l'étude sur le corpus LibriSpeech, plus homogène en termes de type de parole, nous observons que la performance globale du système n'est pas affectée par la variation de la représentation du genre dans les données d'entraînement, ce qui nous laisse supposer une bonne robustesse globale des systèmes face à ces variations. Cependant, les performances du système sur les voix d'hommes sont meilleures que celles de femmes lorsque le corpus d'apprentissage contient seulement 30% de livres lus par des femmes. L'étude des cas limites, avec des apprentissages mono-genre, montre également des performances globales équivalentes avec des performances significativement meilleures pour les livres lus par des hommes par rapport à ceux lus par des femmes, mais ce uniquement pour le modèle masculin. Ainsi, un

biais de sélection du genre important dans les données d'apprentissage contribue donc de façon assez partielle au biais prédictif du système d'ASR comme observé sur Librispeech, pour autant un biais prédictif genré peut néanmoins émerger lorsque les données de parole regroupent des situations d'énonciation et des rôles de locuteurs et locutrices différents.

Dès lors, une connaissance globale des caractéristiques des corpus utilisés pour l'apprentissage des systèmes automatiques de traitement de la parole apparaît comme un élément important pour évaluer les possibles biais prédictifs d'un système, non seulement la répartition en fonction du genre mais également les types de parole recueillis. Dans le Chapitre 6, nous avons réalisé une enquête sur la représentation du genre au sein des ressources disponibles sur la plateforme OpenSLR. Nous avons observé que la parité entre hommes et femmes était contrôlée dans le cas de données construites (pour la plupart élicitées) mais que dans le cas de corpus issus de données produites par ailleurs (via les média par exemple) nous retrouvions une sous-représentation des femmes. Celles-ci, en revanche, tendent à être sur-représentées dans les corpus construits pour des systèmes de synthèse, constituant souvent le premier bloc de développement d'assistantes vocales, reposant largement sur les stéréotypes associant le service à la féminité. Alors que le souci d'une parité des locuteurs et locutrices montre une réelle progression depuis plusieurs années, nous avons vu également les difficultés rencontrées pour obtenir des informations fiables sur les corpus, notamment en termes de temps de parole. Étant donné les résultats précédents mettant en lumière les processus d'interférence possible entre le genre des personnes enregistrées, les conditions d'élocutions des locuteurs et locutrices, nous soulignons que l'absence de telles méta-données est un frein au contrôle de l'existence de biais en fonction du genre.

L'existence de biais prédictifs dans les systèmes d'ASR constitue une question scientifique mais également une question sociale. En effet, si les travaux sur l'explicabilité des systèmes d'IA se multiplient, c'est à la fois pour la connaissance scientifique qu'ils produisent, mais également car expliquer les fonctionnements des systèmes permet d'ouvrir le dialogue sur les impacts sociaux et les problèmes éthiques que peuvent soulever leur utilisation. Dans le cas de l'ASR, et notamment car la voix se veut être "la nouvelle interface entre l'homme et la machine" [9], l'existence de biais prédictifs d'ASR peut constituer un frein à l'accessibilité à tout un ensemble de services. De plus, l'indexation des contenus médiatiques reposant de plus en plus sur des outils de sous-titrage, une moins bonne transcription des voix de femmes contribuerait à reconduire l'invisibilisation des femmes dans le discours public, alors même que des politiques publiques et des initiatives collectives oeuvrent à une meilleure représentation de ces dernières, comme en témoigne la création de l'organisme Les Expertes. [10]

9. `https://www.lenouveleconomiste.fr/lesdossiers/la-voix-nouvelle-interface-entre-lhomme-et-la-machine-69317/`
10. https ://expertes.fr/

Ce travail nous a également donné des éclairages sur les fonctionnements des systèmes E2E utilisés en ASR. En nous intéressant aux variations des performances en fonction du genre, nous avons été amenée à questionner les variations de performances de ces systèmes dues aux données d'apprentissage et à la graine du modèle. Nous observons une part d'aléatoire dans ces modèles qui n'en est pas moins importante à prendre en compte. Ainsi que nous l'avons montré dans la section 8.3, l'aléatoire de la graine du modèle peut, lors d'une répartition équilibrée entre livres lus par des hommes et lus par des femmes, conduire ou non à des différences de performances significatives entre hommes et femmes. Nous n'avons pas observé de différences significatives en faisant varier l'aléatoire de la sélection des données. Cependant, pour ces différents modèles construits avec une même architecture et un même nombre de livres du corpus LibriSpeech pour les hommes et pour les femmes, nous obtenons un WER moyen de 8,7% à 10,2% sur le test-clean (entre 7,7% et 9,1% pour les hommes, et entre 9,5% et 11% pour les femmes). La contrepartie à cette part d'aléatoire est la grande robustesse de ces modèles. En effet, ils montrent une forte capacité de généralisation, avec des performances équivalentes le plus souvent entre les hommes et les femmes, pour des répartitions variant de 30%, 50% et 70% de locuteurs (inversement, 70%, 50% et 30% de locutrices) dans les données d'apprentissage. Les écarts de performances, même lorsqu'ils sont significativement différents entre hommes et femmes, restent faibles, avec par exemple pour wper30, un WER moyen de 8,3% pour les hommes et de 10,9% pour les femmes. Les systèmes E2E apparaissent ainsi relativement robustes à des biais de sélection. Il faut cependant nuancer ce résultat car nous n'avons testé pour ce modèle qu'un type de parole très largement étudié, celui de la parole lue. Le système hybride HMM-DNN n'a pas montré non plus de différences importantes pour le type de parole préparée (cf. section 6.2), se rapprochant le plus du type de parole lue tandis que le biais prédictif est apparu plus nettement sur de la parole spontanée et/ou pour les locutrices ponctuelles. Nous pouvons donc conclure à une certaine robustesse uniquement pour le type de parole lue de LibriSpeech, et nous ne nous risquerons pas à extrapoler cette stabilité des performances face à des biais de sélection important à d'autres types de production langagière.

Enfin, ce travail nous a conduit à élaborer une réflexion sur le genre et sur son utilisation dans des approches que l'on peut qualifiées de "quantitatives". Si une évaluation en fonction du genre revêt un sens au regard de la fonction sociale du système du genre, notamment dans une perspective de visibilisation des femmes, l'ensemble de nos résultats semble nous montrer que les étiquettes genrées, conçues comme des catégories démographiques binaires posées a priori, échouent à décrire la variabilité de nos données, et permettent donc peu de contrôler et de prédire l'existence de biais prédictifs dans nos systèmes. Ces résultats nous confortent dans l'idée que la voix et les données de parole doivent être reconsidérées comme profondément incarnées et largement dépendantes des locuteurs et locutrices ainsi que des contextes interactionnels de leur production, mais

nous questionnent également sur l'utilisation de catégories descriptives binaires du genre dans la voix. Nous avons donc tenté de réfléchir à une manière de sortir de ces étiquettes, notamment via des mesures acoustiques pour permettre d'évaluer la capacité d'un système à reconnaître un ensemble de variabilités vocales et donc des individus, de genre mais aussi d'âge, de sexe, et d'accents différents. S'affranchir des étiquettes binaires permet de concevoir la variabilité du genre comme une variabilité individuelle, mais également contextuelle, car la voix, en tant que pratique sociale, permet à chacun·e de performer des rôles, statuts et postures dépendantes des situations et des enjeux de communication. Il nous semble pertinent de poursuivre des efforts de recherche en ce sens, pour permettre une meilleure compréhension de nos systèmes mais également questionner et réaffirmer le caractère construit de la matrice binaire du genre.

Table des figures

Liste des tableaux

Bibliographie

Abbasi, A., Li, J., Clifford, G. et Taylor, H. (2018). Make “Fairness by Design" part of machine learning. *Harvard Business Review*.

Adda, G., Mariani, J., Lecomte, J., Paroubek, P. et Rajman, M. (1998). The GRACE French part-of-speech tagging evaluation task. *In Proceedings of the 1st International Conference on Language Resources and Evaluation*, LREC 1998, pages 433–441, Grenada, Spain. ELRA.

Adda-Decker, M. et Lamel, L. (2005). Do speech recognizers prefer female speakers ? *In Proceedings of the 6th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2005, pages 2205–2208, Lisbon, Portugal. ISCA.

Aebischer, V. (1985). *Les femmes et le langage : représentations sociales d'une différence*. Presses Universitaires de France.

Amir, O., Engel, M., Shabtai, E. et Amir, N. (2012). Identification of children's gender and age by listeners. *Journal of Voice*, 26(3):313–321.

Angwin, J., Larson, J., Mattu, S. et Kirchner, L. (2016). Machine bias. *ProPublica*, (23).

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F. et Weber, G. (2020). Common Voice : A massively-multilingual speech corpus. *In Proceedings of the 12th International Conference on Language Resources and Evaluation*, LREC 2020, pages 4218–4222, Marseille, France. ELRA.

Arnold, A. (2012). Le rôle de la fréquence fondamentale et des fréquences de résonance dans la perception du genre. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (28).

Arnold, A. (2015). *La voix genrée, entre idéologies et pratiques – Une étude sociophonétique*. Thèse de doctorat, Sorbonne Paris Cité.

Austin, J. (1975). *How To Do Things With Words*. Oxford University Press.

Auzépy, M.-F. et Cornette, J. (2017). *Histoire du poil*. Belin Éditeur.

Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. et Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. *In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2016, pages 4945–4949, Shangai, China. IEEE.

Bahl, L. R., Jelinek, F. et Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

Bamman, D., Dyer, C. et Smith, N. A. (2014a). Distributed representations of geographically situated language. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, ACL 2014, pages 828–834, Baltimore, Maryland, USA. ACL.

Bamman, D., Eisenstein, J. et Schnoebelen, T. (2014b). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Barker, J., Marxer, R., Vincent, E. et Watanabe, S. (2015). The third 'CHiME' speech separation and recognition challenge : Dataset, task and baselines. *In Proceedings of the 2015 Workshop on Automatic Speech Recognition and Understanding*, ASRU 2015, pages 504–511, Scottsdale, Arizona - USA. IEEE.

Barlow, A. L. (2005). Globalization, racism, and the expansion of the American penal system. *In* Conrad, C., éditeur : *African American in the US Economy*, pages 223–230. Rowman and Littlefield Publishers, New York.

Barocas, S. et Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671.

Barthes, R. (1973). Théorie du texte. *Encyclopaedia universalis*, 15:1013–1017.

Beeferman, D., Brannon, W. et Roy, D. (2019). RadioTalk : A Large-Scale Corpus of Talk Radio Transcripts. *In Proceedings of the 21th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2019, pages 564–568, Graz, Autriche. ISCA.

Belin, P., Fecteau, S. et Bédard, C. (2004). Thinking the voice : neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135.

Bell, P., Gales, M. J., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M. et Woodland, P. C. (2015). The mgb challenge : Evaluating multi-genre broadcast media recognition. *In Proceedings of the 2015 Workshop on Automatic Speech Recognition and Understanding*, ASRU 2015, pages 687–693, Scottsdale, Arizona - USA. IEEE.

Bender, E. M. et Friedman, B. (2018). Data statements for natural language processing : Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bennett, S. et Montero-Diaz, L. (1982). Children's perception of speaker sex. *Journal of Phonetics*, 10(1):113–121.

Berendt, B. et Preibusch, S. (2012). Exploring discrimination : A user-centric evaluation of discrimination-aware data mining. *In 2012 IEEE 12th International Conference on Data Mining Workshops*, pages 344–351. IEEE.

Berendt, B. et Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining : foundations and empirical evidence. *Artificial Intelligence and Law*, 22(2):175–209.

Berger, P. L. et Luckmann, T. (1966). *The social construction of reality : A treatise in the sociology of knowledge*. Doubleday & Company Inc., NY.

Bertail, P., Bounie, D., Clémençon, S. et Waelbroeck, P. (2019). Algorithmes : biais, discrimination et équité. Telecom ParisTech & Fondation Abeona.

Bertau, M.-C. (2008). Voice : A pathway to consciousness as "social contact to oneself". *Integrative Psychological and Behavioral Science*, 42(1):92–113.

Bivens, R. (2017). The gender binary will not be deprogrammed : Ten years of coding gender on facebook. *New Media & Society*, 19(6):880–898.

Bivens, R. et Haimson, O. L. (2016). Baking gender into social media design : How platforms shape categories for users and advertisers. *Social Media + Society*, 2(4).

Blanche-Benveniste, C. et Jeanjean, C. (1986). *Le français parlé : transcription et édition*. Institut National de la Langue Française.

Blodgett, S. L., Barocas, S., Daumé III, H. et Wallach, H. (2020). Language (technology) is power : A critical survey of "bias" in NLP. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, pages 5454–5476, Online. ACL.

Boersma, P. et Weenink, D. (2018). Praat : doing phonetics by computer [computer program]. version 6.0.37. Web download on 14 March 2018 from `http://www.praat.org/`.

Bogen, M. et Rieke, A. (2018). Help wanted : An examination of hiring algorithms, equity, and bias. Upturn.

Bollier, D., Firestone, C. M. *et al.* (2010). *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC.

Bolt, R. H., Cooper, F. S., David, E. E., Denes, P. B., Pickett, J. M. et Stevens, K. N. (1973). Speaker identification by speech spectrograms : some further observations. *The Journal of the Acoustical Society of America*, 54(2):531–534.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. et 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 2016, pages 4349–4357, Barcelona, Spain. ACM.

Bonastre, J.-F. (2020). 1990-2020 : retours sur 30 ans d'échanges autour de l'identification de voix en milieu judiciaire. *In* Adda, G., Amblard, M. et Fort, K., éditeurs : *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 38–47, Nancy, France. ATALA.

Bonastre, J.-F., Bimbot, F., Boe, L.-J., Campbell, J. P., Reynolds, D. A. et Magrin-Chagnolleau, I. (2003). Person authentication by voice : a need for caution. *In Proceedings of the 8th European Conference on Speech Communication and Technology*, EUROSPEECH 2003 (INTERSPEECH), pages 33–36, Geneva, Switzerland. ISCA.

Bower, A., Niss, L., Sun, Y. et Vargo, A. (2018). Debiasing representations by removing unwanted variation due to protected attributes. *In Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Bowker, G. C. et Star, S. L. (2000). *Sorting things out : Classification and its consequences*. MIT Press.

Boyd, D. et Crawford, K. (2012). Critical questions for big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

Boë, L.-J., Contini, M. et Rakotofiringa, H. (1975). Étude statistique de la fréquence laryngienne. *Phonetica*, 32(1):1–23.

Bucholtz, M. et Hall, K. (2009). Locating identity in language. *In* Llamas, C. et Watt, D., éditeurs : *Language and identities*, chapitre 2, pages 18–28. Edinburgh University Press.

Buolamwini, J. et Gebru, T. (2018). Gender shades : Intersectional accuracy disparities in commercial gender classification. *In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, ACM FAccT 2018, pages 77–91, New-York City, USA. ACM.

Butler, J. (2004[1997]). *Le pouvoir des mots : politique du performatif*. Editions Amsterdam.

Butler, J. (2005[1990]). *Trouble dans le genre : pour un féminisme de la subversion*. Editions La Découverte, Paris.

Caliskan, A., Bryson, J. J. et Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Calliope (1989). Ergonomie et évaluation du traitement de la parole par ordinateur. *In* Tubach, J., éditeur : *La parole et son traitement automatique*, chapitre 26, pages 689–705. Paris : Masson.

Cettolo, M., Girardi, C. et Federico, M. (2012). Wit3 : Web inventory of transcribed and translated talks. *In Proceedings of the 16th Annual Conference of European Association for Machine Translation*, EAMT 2012, pages 261–268, Trento, Italy. ACL.

Ceulemans, M. et Fauconnier, G. (1979). *Image, rôle et condition sociale de la femmes dans les médias : recueil et analyse des documents de recherche*. UNESCO.

Chen, G.-T. (1974). The pitch range of English and Chinese speakers. *Journal of Chinese Linguistics*, pages 159–171.

Chen, L., Ma, R., Hannák, A. et Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. *In Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems*, CHI 2018, pages 1–14, Montréal, QC, Canada. ACM.

Chen, S., Beeferman, D. et Rosenfeld, R. (1998). Evaluation metrics for language models. *In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 275–280.

Chevrie-Muller, C. et Gremy, F. (1967). Contribution à l'établissement de quelques constantes physiologiques de la voix parlée de l'adulte. *Journal Français d'Oto-Rhino-Laryngologie XV1*, pages 149–154.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K. et Bengio, Y. (2015). Attention-based models for speech recognition. *In Proceedings of the 29th International Conference on Neural Information Processing Systems*, volume 28 de *NIPS 2016*, Montreal, Canada. ACM.

Choukri, K. et Nilsson, M. (1998). The European Language Resources Association. *In Proceedings of the 1st International Conference on Language Resources and Evaluation*, LREC 1998, pages 153–159, Grenada, Spain. ELRA.

Chouldechova, A. (2017). Fair prediction with disparate impact : A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Cieri, C., Miller, D. et Walker, K. (2004). The Fisher corpus : a resource for the next generations of speech-to-text. *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 4 de *LREC 2004*, pages 69–71, Lisbon, Portugal. ELRA.

Collins, P. H. (2002). *Black feminist thought : Knowledge, consciousness, and the politics of empowerment*. Routledge.

Costanza-Chock, S. (2018). Design justice : Towards an intersectional feminist framework for design theory and practice. *Proceedings of the Design Research Society*.

Cotton, N. (2016). Du performatif à la performance : la « performativité » dans tous ses états. *Sens public*.

Couillault, A., Fort, K., Adda, G. et Mazancourt, H. (2014). Evaluating corpora documentation with regards to the ethics and big data charter. *In Proceedings of the 9st International Conference on Language Resources and Evaluation*, LREC 2014, pages 4225–4229, Reykjavik, Islande. ELRA.

Crawford, K. (2017). The trouble with bias. NIPS 2017 Keynote.

Crenshaw, K. W. (2005). Cartographies des marges : intersectionnalité, politique de l'identité et violences contre les femmes de couleur. *Cahiers du Genre, 39*, 39(2):51–82.

CSA (2021). La Représentation des Femmes à la Télévision et à la Radio. Rapport d'Exercice 2020.

Davis, K. H., Biddulph, R. et Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642.

De Calmès, M. et Pérennou, G. (1998). BDLEX : a lexicon for spoken and written french. *In Proceedings of the 1st International Conference on Language Resources and Evaluation*, LREC 1998, pages 1129–1136, Grenada, Spain. ELRA.

Derrida, J. (1972). *Signature événement contexte*, pages 365–393. Editions de Minuit.

Descola, P. (2005). *Par-delà nature et culture*. Gallimard Paris.

Devlin, J., Chang, M.-W., Lee, K. et Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL 2019, pages 4171–4186, Minneapolis, Minnesota, USA. ACL.

D'Ignazio, C. et Klein, L. F. (2020). *Data feminism.* MIT Press.

DiPasquale, S. L. (2019). Women in radio : a (Her)Story. *In Communications : Student Scholarship & Creative Works*. 3.

Dong, L., Xu, S. et Xu, B. (2018). Speech-transformer : A no-recurrence sequence-to-sequence model for speech recognition. *In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2018, pages 5884–5888, Calgary, Alberta, Canada. IEEE.

Doshita, S. (1965). *Studies on the analysis and recognition of Japanese speech sounds.* Thèse de doctorat, Université de Kyoto, Japon.

Doukhan, D. et Carrive, J. (2018). Description automatique du taux d'expression des femmes dans les flux télévisuels français. *In Actes des 32e Journées d'Études sur la Parole*, JEP 2018, pages 496–504, Aix-en-Provence, France. AFCP.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A. et Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. *In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2018, pages 5214–5218, Calgary, Alberta, Canada. IEEE.

Dreyfus-Graf, J. (1961). Phonétographe, présent, futur. *Bulletin techniques des PTT suisses*, 5:363–379.

Dubet, F. (2014). *Les Places et les Chances. Repenser la justice sociale : Repenser la justice sociale.* Média Diffusion.

Duchêne, A. et Moïse, C. (2011). *Langage, genre et sexualité.* Nota bene.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. et Zemel, R. (2012). Fairness through awareness. *In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Ehrick, C. (2010). "Savage dissonance". Gender, voice, and women's radio speech in Argentina, 1930–1945. *In* Suisman, D. et Strasser, S., éditeurs : *Sound in the age of mechanical reproduction*, pages 69–92. University of Pennsylvania Press.

Elloumi, Z., Besacier, L., Galibert, O., Kahn, J. et Lecouteux, B. (2018). ASR performance prediction on unseen broadcast programs using convolutional neural networks. *In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2018, pages 5894–5898, Calgary, Alberta, Canada. IEEE.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F. et Farinas, J. (2010). The EPAC corpus : Manual and automatic annotations of conversational speech in French broadcast news. *In Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC 2010, Valetta, Malta. ELRA.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton & Co. N.V., Publishers, The Hague.

Fassin, É. (2008). *L'empire du genre. L'histoire politique ambiguë d'un outil conceptuel*. Numéro 187-188. Editions de l'EHESS.

Fausto-Sterling, A. (2000). *Sexing the body : Gender politics and the construction of sexuality*. Basic Books.

Feng, S., Kudina, O., Halpern, B. M. et Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. (Submitted to INTERSPEECH 2021).

Fitch, W. T. et Giedd, J. (1999). Morphology and development of the human vocal tract : A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522.

Fleck, L. (2005[1934]). Genèse et développement d'un fait scientifique (n. jas, trad.). *Paris : Les belles lettres)*.

Fletcher, H. et Munson, W. A. (1933). Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4):377–430.

Flores, A. W., Bechtel, K. et Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses : A rejoinder to Machine bias : There's software used across the country to predict future criminals. And it's biased against blacks. *Fed. Probation*, 80:38.

Forgie, J. W. et Forgie, C. D. (1959). Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, 31(11):1480–1489.

Fox, R. A. et Nissen, S. L. (2005). Sex-related acoustic changes in voiceless english fricatives. *Journal of Speech, Language, and Hearing Research*, 48(4):753–765.

Fracchiolla, B. et Sini, L. (2020). La haine, c'est les autres ! *In La haine en discours*. Editions Le bord de l'eau.

Fraisse, G., Dauphin, S. et Sénac-Slawinski, R. (2008). Le gender mainstreaming, vrai en théorie, faux en pratique ? *Cahiers du Genre*, 44(1):17.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651–661.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D. et Choukri, K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC 2006, Genova, Italy. ELRA.

Galliano, S., Gravier, G. et Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Proceedings of the 10th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2009, pages 2583–2586, Brighton, United-Kingdom. ISCA.

Garcia, L., Benitez, C., Segura, J. C. et Umesh, S. (2011). Combining speaker and noise feature normalization techniques for Automatic Speech Recognition. *In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2011, pages 5496–5499, Prague, Czech Republic. IEEE.

Garfinkel, H. (1967). Passing and the managed achievement of sex status in an "intersexed" person. *In Studies in Ethnomethodology*, pages 116–185. Englewood Cliffs, NJ : Prentice Hall.

Garg, N., Schiebinger, L., Jurafsky, D. et Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Garnerin, M., Rossato, S. et Besacier, L. (2019). Gender representation in French broadcast corpora and its impact on ASR performance. *In Proceedings of the 1st ACM Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV 2019, pages 3–9, Nice, France. ACM.

Garnerin, M., Rossato, S. et Besacier, L. (2020a). Gender representation in open source speech resources. *In Proceedings of the 12th International Conference on Language Resources and Evaluation*, LREC 2020, pages 6599–6605, Marseille, France. ELRA.

Garnerin, M., Rossato, S. et Besacier, L. (2020b). Représentation du genre dans des données open source de parole. *In Actes de la 6e conférence conjointe Journées*

d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole, pages 244–252, Nancy, France. ATALA et AFCP.

Garnerin, M., Rossato, S. et Besacier, L. (2021). Investigating the impact of gender representation in ASR training data : a case study on librispeech. *In Proceedings of the 3rdrd Workshop on Gender Bias in Natural Language Processing*, GeBNLP 2021, pages 86–92, Online. ACL.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G. et Pallett, D. S. (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D. et Crawford, K. (2018). Datasheets for datasets. *In Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, Stockholm, Sweden.

Gingras, Y. (2020). *Sociologie des sciences*. Presses Universitaires de France.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O. et Quintard, L. (2012). The REPERE corpus : a multimodal corpus for person recognition. *In Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey. ELRA.

Godfrey, J. J., Holliman, E. C. et McDaniel, J. (1992). Switchboard : Telephone speech corpus for research and development. *In Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 de *ICASSP92*, pages 517–520. IEEE.

Goldwater, S., Jurafsky, D. et Manning, C. D. (2010). Which words are hard to recognize ? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200.

Gonen, H. et Goldberg, Y. (2019). Lipstick on a pig : Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL 2019, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Google (2019). Crowdsourced high-quality UK and Ireland English Dialect speech data set. Web download at `http://www.openslr.org/83/`.

Graddol, D. (1986). Discourse specific pitch behaviour. *In* Johns-Lewis, C. M., éditeur : *Intonation in discourse*, pages 221–237. Croom Helm, London and Sidney.

Graff, D., Mendonça, A. et DiPersio, D. (2011). French gigaword third edition ldc2011t10. Web download at `http://www.openslr.org/38/`.

Graves, A., Fernández, S., Gomez, F. et Schmidhuber, J. (2006). Connectionist temporal classification : Labelling unsegmented sequence data with recurrent neural networks. *In Proceedings of the 23rd International Conference on Machine Learning*, ICML 2006, page 369–376, New York, NY, USA. ACM.

Gravier, G., Adda, G., Paulsson, N., Carré, M., Giraudel, A. et Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based tv content processing in the French language. *In Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey. ELRA.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K. et Choukri, K. (1998). The ESTER evaluation campaign for the rich transcription of French broadcast news. *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, LREC 2004, Lisbon, Portugal. ELRA.

Greenwald, A. G., McGhee, D. E. et Schwartz, J. L. (1998). Measuring individual differences in implicit cognition : the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Grey, C. et Kopp, G. (1944). Voiceprint identification. *Bell Telephone Laboratories Report*, pages 1–14.

Guillaumin, C. (1992). *Sexe, race et pratique du pouvoir*. Les Éditions iXe.

Guiraud, P. (1966). Le système du relatif en français populaire. *Langages*, (3):40–48.

Hannun, A. (2017). Sequence modeling with CTC. *Distill*, 2(11).

Haraway, D. (1988). Situated knowledges : The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599.

Harding, S. (1992). Rethinking standpoint epistemology : What is "strong objectivity" ? *The Centennial Review*, 36(3):437–470.

Hardt, M., Price, E., Price, E. et Srebro, N. (2016). Equality of opportunity in supervised learning. *In* Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. et Garnett, R., éditeurs : *Proceedings of the 30th International Conference on Neural Information Processing Systems*, volume 29 de *NIPS 2016*, pages 3315–3323, Barcelona, Spain. ACM.

Haton, J.-P., Cerisara, C., Fohr, D., Laprie, Y. et Smaïli, K. (2006). *Introduction à la reconnaissance automatique de la parole*, chapitre 1, pages 1–15. Paris : Dunod.

Hemphill, C. T., Godfrey, J. J. et Doddington, G. R. (1990). The ATIS spoken language systems pilot corpus. *In Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

Henton, C. G. (1989). Fact and fiction in the description of female and male pitch. *Language & Communication*, 9(4):299–311.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N. et Estève, Y. (2018). TED-LIUM 3 : Twice as much data and corpus repartition for experiments on speaker adaptation. *In Proceedings of the 20th International Conference on Speech and Computer*, SPECOM 2018, pages 198–208, Leipzig, Germany. Springer.

Hernandez-Mena, C. D. (2019). TEDx Spanish corpus. audio and transcripts in Spanish taken from the TEDx Talks ; shared under the CC BY-NC-ND 4.0 license. Web Download at http ://www.openslr.org/67/.

Hey, T., Tansley, S. et Tolle, K. (2009). *Jim Gray on eScience : A transformed scientific method*. Microsoft Research.

Hillenbrand, J. M. et Clark, M. J. (2009). The role of F0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, 71(5):1150–1166.

Hirano, M., Kurita, S. et Nakashima, T. (1983). Growth, development and aging of human vocal folds. *In* Bless, D. et Abbs, J., éditeurs : *Vocal Fold Physiology : Contemporary Research and Clinical Issues*, pages 22–43. San Diego, CA : College Hill Press.

Hirata, H., Laborie, F., Le Doaré, H. et Senotier, D., éditeurs (2000). *Dictionnaire critique du féminisme*. Presses Universitaires de France.

Hollien, H. et Paul, P. (1969). A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls. *Language and Speech*, 12(2):119–124. PMID : 5792372.

Honda, K., Hirai, H., Masaki, S. et Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in f0 control. *Language and Speech*, 42(4):401–411. PMID : 10845244.

Horii, Y. (1975). Some statistical characteristics of voice fundamental frequency. *Journal of Speech and Hearing Research*, 18(1):192–201.

Hovy, D. et Spruit, S. L. (2016). The social impact of Natural Language Processing. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL 2016, pages 591–598, Berlin, Germany. ACL.

Ingemann, F. (1968). Identification of the speaker's sex from voiceless fricatives. *The Journal of the Acoustical Society of America*, 44(4):1142–1144.

Iyer, R., Ostendorf, M. et Meteer, M. W. (1997). Analyzing and predicting language model improvements. *In Proceedings of the 1997 Workshop on Automatic Speech Recognition and Understanding*, ASRU 1997, pages 254–261, Santa Barbara, CA, USA. IEEE.

Jaffe, A., éditeur (2009). *Stance : sociolinguistic perspectives*. Oxford University Press.

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Jelinek, F. et Mercer, R. (1980). Interpolated estimation of Markov source parameters from sparse data. *In* Gelsema, E. S. et Kanal, L. N., éditeurs : *Pattern recognition in practice*, pages 381–397. Amsterdam, Holland.

Jelinek, F., Mercer, R. L., Bahl, L. R. et Baker, J. K. (1977). Perplexity — a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Johns-Lewis, C. M. (1986). Prosodic differentiation of discourse modes. *In* Johns-Lewis, C. M., éditeur : *Intonation in discourse*, pages 199–219. Croom Helm, London and Sidney.

Johnson, K. (2006). Resonance in an exemplar-based lexicon : The emergence of social identity and phonology. *Journal of Phonetics*, 34(4):485–499.

Johnson, K. et Azara, M. (2000). The perception of personal identity in speech : Evidence from the perception of twins' speech. *Unpublished manuscript*.

Joseph, J. E. (2009). Identity. *In* Llamas, C. et Watt, D., éditeurs : *Language and identities*, chapitre 1, pages 9–17. Edinburgh University Press.

Juan, S. S., Besacier, L., Lecouteux, B. et Dyab, M. (2015). Using resources from a closely-related language to develop ASR for a very under-resourced language : a case study for Iban. *In Proceedings of the 16th Annual Conference of the International*

Speech Communication Association, INTERSPEECH 2015, pages 1270–1274, Dresden, Germany. ISCA.

Jurafsky, D. et Martin, J. H. (2009). *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Kahane, J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *American Journal of Anatomy*, 151(1):11–19.

Kamath, U., Liu, J. et Whitaker, J. (2019). *Deep Learning for NLP and Speech Recognition*. Springer.

Kergoat, D. (2005). Rapports sociaux et division du travail entre les sexes. *In Femmes, genre et sociétés*, chapitre 12, pages 94–101. La Découverte.

Kersta, L. G. (1962). Voiceprint identification. *Nature*, 196(4861):1253–1257.

Kessler, S. J. et McKenna, W. (1978). *Gender : An ethnomethodological approach*. University of Chicago Press.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D. et Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *In Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS 2017, pages 656–666, Long Beach, CA, USA. ACM.

Kitzing, P. (1979). *Glottografisk frekvensindikering : En undersökningsmetod för mätning av röstläge och röstomfang samt framställning av röstfrekvensdistributionen*. Oronkliniken i Malmö, Luds universitet.

Kneser, R. et Ney, H. (1995). Improved backing-off for M-gram language modeling. *In Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1995, pages 181–184, Detroit, Michigan, USA. IEEE.

Koolen, C. et van Cranenburgh, A. (2017). These are not the stereotypes you are looking for : Bias and fairness in authorial gender attribution. *In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Korvas, M., Plátek, O., Dušek, O., Žilka, L. et Jurčíček, F. (2014). Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. *In Proceedings of the 9st International Conference on Language Resources and Evaluation*, LREC 2014, pages 4423–4428, Reykjavik, Islande. ELRA.

Krauss, R. M., Freyberg, R. et Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, 38(6):618–625.

Kreiman, J. et Sidtis, D. (2011). Physical characteristics and the voice : Can we hear what a speaker looks like ? *In Foundations of Voice Studies*, chapitre 4. Wiley-Blackwell.

Krook, M. I. P. (1988). Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis. *Folia Phoniatrica et Logopaedica*, 40(2):82–90.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kuhn, T. S. (1963). The function of dogma in scientific research. ac crombie. *In* Crombie, A. A., éditeur : *Scientific Change. Historical Studies in the Intellectual, Social, and Technical Conditions for Scientific Discovery and Technical Invention, From Antiquity to the Present*, pages 347–69. Heinemann Educational Books, Ltd.

Kurita, K., Vyas, N., Pareek, A., Black, A. W. et Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. *In Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, GeBNLP 2019, pages 166–172, Florence, Italy. ACL.

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight ? *Phonetica*, 46(1-3):117–125.

Labov, W. (1966). *The social stratification of English in New York city*. Washington DC, Center for Applied Linguistics.

Labov, W. (1976). *Sociolinguistique*. Les Editions de Minuit.

Lacey, K. (2013). *Listening Publics : The Politics and Experience of Listening in the Media Age*. Malden, MA : Polity Press.

Lalande, A. (1950[1926]). *Vocabulaire technique et critique de la philosophie*, volume 2. Presses Universitaires de France.

Larson, B. (2017). Gender as a variable in natural-language processing : Ethical considerations. *In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Lass, N. J., Almerino, C. A., Jordan, L. F. et Walsh, J. M. (1980). The effect of filtered speech on speaker race and sex identifications. *Journal of Phonetics*, 8(1):101–112.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T. et Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, 59(3):675–678.

Lass, N. J., Tecca, J. E., Mancuso, R. A. et Black, W. I. (1979). The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics*, 7(2):105–118.

Laver, J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H. *et al.* (2009). Social science. computational social science. *Science*, 323(5915):721–723.

Le, H., Pino, J., Wang, C., Gu, J., Schwab, D. et Besacier, L. (2020). Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. *In Proceedings of the 28th International Conference on Computational Linguistics*, COLING 2020, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

LeCun, Y., Bengio, Y. *et al.* (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Lecun, Y., Bottou, L., Bengio, Y. et Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, A., Kawahara, T. et Shikano, K. (2001). Julius — an open source real-time large vocabulary recognition engine. *In Proceedings of the 7th European Conference on Speech Communication and Technology*, EUROSPEECH 2001 (INTERSPEECH), pages 1691–1694, Aalborg, Denmark. ISCA.

Lee, K.-F., Hon, H.-W. et Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(1):35–45.

Lee, L. et Rose, R. (1996). Speaker normalization using efficient frequency warping procedures. *In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1996, pages 353–356, Atlanta, Georgia, USA. IEEE.

Léon, J. (2015). La traduction automatique comme technologie de guerre. *In Histoire de l'automatisation des sciences du langage*. ENS Éditions.

Levi-Strauss, C. (1967). *Les structures élémentaires de la parenté*. De Gruyter Mouton.

Lewis, S. (2019). The racial bias built into photography. 25/04/19. *The New York Times*.

Li, S., Raj, D., Lu, X., Shen, P., Kawahara, T. et Kawai, H. (2019). Improving Transformer-Based Speech Recognition Systems with Compressed Structure and Speech Attributes Augmentation. *In Proceedings of the 21th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2019, pages 4400–4404, Graz, Autriche. ISCA.

Liberman, M. et Cieri, C. (1998). The creation, distribution and use of linguistic data : the case of the Linguistic Data Consortium. *In Proceedings of the 1st International Conference on Language Resources and Evaluation*, LREC 1998, pages 159–164, Grenada, Spain. ELRA.

Llamas, C. et Watt, D. (2009). *Language and identities*. Edinburgh University Press.

Macharia, S., Ndangam, L., Saboor, M., Franke, E., Parr, S. et Opoku, E. (2015). Who makes the news. Global Media Monitoring Project (GMMP).

Macé, E. (2000). Qu'est-ce qu'une sociologie de la télévision ? Esquisse d'une théorie des rapports sociaux médiatisés. 1. La configuration médiatique de la réalité. *Réseaux*, 18(104).

Maingueneau, D. (2002). Problèmes d'ethos. *Pratiques : linguistique, littérature, didactique*, (113-114).

Maingueneau, D. (2013). L'èthos : un articulateur. *COnTEXTES*, (13).

Makhoul, J. (1975). Linear prediction : A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

Mann, H. B. et Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

Mariani, J. (2005). Developing language technologies with the support of language resources and evaluation programs. *Language Resources and Evaluation*, 39(1):35–44.

Marignier, N. (2016). *Les matérialités discursives du sexe : la construction et la déstabilisation des évidences du genre dans les discours sur les sexes atypiques*. Theses, Université Sorbonne Paris Cité.

Mathieu, N.-C. (1991). *L'Anatomie politique du sexe. Catégorisations et idéologies du sexe*. Éditions Côté-femmes, Paris.

Mbiatong, J. (2019). Ethnométhodologie. *In* Delory-Momberger, C., éditeur : *Vocabulaire des histoires de vie et de la recherche biographique*, pages 219—222. Erès, Toulouse, France.

McGurk, H. et MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Meditch, A. (1975). The development of sex-specific speech patterns in young children. *Anthropological Linguistics*, pages 421–433.

Meyer, J., Rauchenstein, L., Eisenberg, J. D. et Howell, N. (2020). Artie bias corpus : An open dataset for detecting demographic bias in speech applications. *In Proceedings of the 12th International Conference on Language Resources and Evaluation*, LREC 2020, pages 6462–6468, Marseille, France. ELRA.

Miller, C. L., Younger, B. A. et Morse, P. A. (1982). The categorization of male and female voices in infancy. *Infant Behavior and Development*, 5(2-4):143–159.

Mirandé, A. (2016). Hombres mujeres : An indigenous third gender. *Men and Masculinities*, 19(4):384–409.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. et Gebru, T. (2019). Model cards for model reporting. *In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, ACM FAccT 2019, pages 220–229, Atlanta, GA, USA. ACM.

Money, J., Hampson, J. G. et Hampson, J. L. (1955). An examination of some basic sexual concepts : The evidence of human hermaphroditism. *Bulletin of the Johns Hopkins Hospital*, 97(4):301–319.

Mullany, L. (2009). Gendered identities in the professional workplace : Negotiating the glass ceiling. *In* Llamas, C. et Watt, D., éditeurs : *Language and identities*, chapitre 16, pages 179–190. Edinburgh University Press.

Mulligan, D. K., Kroll, J. A., Kohli, N. et Wong, R. Y. (2019). This thing called fairness : disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–36.

NA (2016). Corpus de textes du journal "Le Monde". ELRA-W0015. Web download at `http://catalog.elra.info/en-us/repository/browse/ELRA-W0015/`.

Nanda, S. (1986). The hijras of india : Cultural and individual dimensions of an institutionalized third gender role. *Journal of Homosexuality*, 11(3-4):35–54.

Nass, C. I. et Brave, S. (2005). *Wired for Speech : How Voice Activates and Advances the Human-computer Relationship*. MIT Press.

Neiman, G. S. et Applegate, J. (1990). Accuracy of listener judgments of perceived age relative to chronological age in adults. *Folia Phoniatrica et Logopaedica*, 42(6):327–330.

Nolan, F. et Oh, T. (1996). Identical twins, different voices. *International Journal of Speech Language and the Law*, 3(1):39–49.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T. et Staab, S. (2020). Bias in Data-driven AI Systems - An Introductory Survey. *arXiv preprint 2001.09762*.

Oakley, A. (1972). *Sex, Gender and Society*. London, Temple Smith.

Ochs, E. (1992). Indexing gender. *In* Duranti, A. et Goodwin, C., éditeurs : *Rethinking Context : Language as an interactive phenomenon*, pages 335—350. Cambridge University Press.

Olson, H. F. et Belar, H. (1956). Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6):1072–1081.

O'Neil, C. (2016). *Weapons of math destruction : How big data increases inequality and threatens democracy*. Crown.

Ono, Y., Wakita, H. et Zhao, Y. (1993). Speaker normalization using constrained spectra shifts in auditory filter domain. *In Proceedings of the 3rd European Conference on Speech Communication and Technology*, EUROSPEECH 1993, pages 355–358, Berlin, Germany. ISCA.

Ormezzano, Y. (2000). *Le guide de la voix*. Odile Jacob.

Pallett, D. S. (2003). A look at NIST'S benchmark ASR tests : past, present, and future. *In Proceedings of the 2003 Workshop on Automatic Speech Recognition and Understanding*, ASRU 2003, pages 483–488. IEEE.

Panayotov, V., Chen, G., Povey, D. et Khudanpur, S. (2015). Librispeech : an ASR corpus based on public domain audio books. *In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2015, pages 5206–5210, Brisbane, Australie. IEEE.

Parini, L. (2010). Le concept de genre : constitution d'un champ d'analyse, controverses épistémologiques, linguistiques et politiques. *Socio-Logos. Revue de l'association française de sociologie*.

Paroubek, P., Chaudiron, S. et Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *Revue TAL*, 48(1):7–31.

Parsheera, S. (2018). A gendered perspective on artificial intelligence. *In 2018 ITU Kaleidoscope : Machine Learning for a 5G Future (ITU K)*, pages 1–7. IEEE.

Pastoureau, M. et Simonnet, D. (2007). *Le petit livre des couleurs*, volume 377. Points.

Pavard, B., Rochefort, F. et Zancarini-Fournel, M. (2020). *Ne nous libérez pas, on s' en charge*. La Découverte.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. et Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pedreshi, D., Ruggieri, S. et Turini, F. (2008). Discrimination-aware data mining. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2008, page 560–568, Las Vegas, Nevada, USA. ACM.

Pellegrini, T., Farinas, J., Delpech, E. et Lancelot, F. (2019). The Airbus Air Traffic Control Speech Recognition 2018 Challenge : Towards ATC Automatic Transcription and Call Sign Detection. *In Proceedings of the 21th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2019, pages 2993–2997, Graz, Autriche. ISCA.

Pépiot, E. (2013). *Voix de femmes, voix d'hommes : différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones*. Theses, Université Paris VIII Vincennes-Saint Denis.

Perea, F. (2018). Cortana est-elle une humaine comme les autres ? *Semen*, (44).

Perry, T. L., Ohde, R. N. et Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America*, 109(6):2988–2998.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. et Zettlemoyer, L. (2018). Deep contextualized word representations. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peterson, G. E. et Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2):175–184.

Pingflow (2019). *Développer une cuture "Data driven" : mode d'emploi*. Pingflow.

Popper, K. R. (1973). *La Logique de la Découverte Scientifique*. Editions Payot.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. *et al.* (2011). The Kaldi speech recognition toolkit. *In Proceedings of the 2011 Workshop on Automatic Speech Recognition and Understanding*, ASRU 2011, Hawaii, USA. IEEE.

Prates, M. O. R., Avelar, P. H. C. et Lamb, L. C. (2020). Assessing gender bias in machine translation - A case study with Google Translate. *Neural Computing and Applications*, 32(10):6363–6381.

Price, P., Fisher, W. M., Bernstein, J. et Pallett, D. S. (1988). The darpa 1000-word resource management database for continuous speech recognition. *In Proceedings of the 1988 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1988, pages 651–654, Shangai, China. IEEE.

Prost, F., Thain, N. et Bolukbasi, T. (2019). Debiasing embeddings for fairer text classification. *In Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, GeBNLP 2019, pages 69–75, Florence, Italy. ACL.

Qian, Y., Bi, M., Tan, T., Yu, K., Qian, Y., Bi, M., Tan, T. et Yu, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 24(12):2263–2276.

Rabiner, L. et Juang, B. (1986). An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rappaport, W. (1958). über messungen der tonhöhenverteilung in der deutschen sprache. *Acta Acustica united with Acustica*, 8(4):220–225.

Rendall, D., Owren, M. J., Weerts, E. et Hienz, R. D. (2004). Sex differences in the acoustic structure of vowel-like grunt vocalizations in baboons and their perceptual discrimination by baboon listeners. *The Journal of the Acoustical Society of America*, 115(1):411–421.

Resch, B. (2003). Automatic speech recognition with HTK. Signal Processing and Speech Communication Laboratory. Infelldgase, Austria.

Rider, J. F. (1928). Why is a radio soprano unpopular ? *Scientific American*, 139(4):334–337.

Robinson, M. (2020). Two-spirit identity in a time of gender fluidity. *Journal of Homosexuality*, 67(12):1675–1690.

Rosaldo, M. Z., Lamphere, L. et Bamberger, J. (1974). *Woman, culture, and society*, volume 133. Stanford University Press.

Rose, P. (1991). How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency ? *Speech Communication*, 10(3):229–247.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory.

Rousseau, A., Deléglise, P. et Estève, Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED Talks. *In* Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J. et Piperidis, S., éditeurs : *Proceedings of the 9st International Conference on Language Resources and Evaluation*, LREC 2014, Reykjavik, Islande. ELRA, European Language Resources Association (ELRA).

Rubin, G. (1975). The traffic in women : Notes on the "political economy" of sex. *In* Reiter, R. R., éditeur : *Toward an Anthropology of Women*, pages 157–210. Monthly Review Press New York.

Rudinger, R., Naradowsky, J., Leonard, B. et Van Durme, B. (2018). Gender bias in coreference resolution. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Rumelhart, D. E., Hinton, G. E. et Williams, R. J. (1986). *Learning Internal Representations by Error Propagation*, page 318–362. MIT Press, Cambridge, MA, USA.

Ryan, E. B. et Capadano, H. L. (1978). Age perceptions and evaluative reactions toward adult speakers. *Journal of Gerontology*, 33(1):98–102.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y. et Ramabhadran, B. (2013a). Improvements to deep convolutional neural networks for LVCSR. *In Proceedings of the 2013 Workshop on Automatic Speech Recognition and Understanding*, ASRU 2013, pages 315–320, Olomouc, Czech Republic. IEEE.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B. et Ramabhadran, B. (2013b). Deep convolutional neural networks for lvcsr. *In Proceedings of the 2013 IEEE International*

Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, pages 8614–8618, Vancouver, Canada. IEEE.

Sánchez-Monedero, J., Dencik, L. et Edwards, L. (2020). What does it mean to solve the problem of discrimination in hiring ? social, technical and legal perspectives from the uk on automated hiring systems. *In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, ACM FAccT 2020, Barcelona, Spain. ACM.

Santolaria, N. (2016). *"Dis Siri". Enquête sur le génie à l'intérieur du smartphone.* Anamosa.

Sap, M., Card, D., Gabriel, S., Choi, Y. et Smith, N. A. (2019). The risk of racial bias in hate speech detection. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL 2019, pages 1668–1678, Florence, Italy. ACL.

Scheler, M. (1993). *Problèmes de sociologie de la connaissance.* Presses Universitaires de France.

Schuster, M., Lohscheller, J., Kummer, P., Eysholdt, U. et Hoppe, U. (2005). Laser projection in high-speed glottography for high-precision measurements of laryngeal dimensions and dynamics. *European Archives of Oto-Rhino-Laryngology*, 262:477–481.

Schwartz, M. F. (1968). Identification of speaker sex from isolated, voiceless fricatives. *The Journal of the Acoustical Society of America*, 43(5):1178–1179.

Scott, J. W. (1986). Gender : A useful category of historical analysis. *The American Historical Review*, 91(5):1053–1075.

Shah, D. S., Schwartz, H. A. et Hovy, D. (2020). Predictive biases in natural language processing models : A conceptual framework and overview. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL 2020, pages 5248–5264, Online. ACL.

Shipp, T. et Hollien, H. (1969). Perception of the aging male voice. *Journal of Speech and Hearing Research*, 12(4):703–710.

Sholtz, P. N. et Bakis, R. (1962). Spoken digit recognition using vowel-consonant segmentation. *The Journal of the Acoustical Society of America*, 34(1):1–5.

Simpson, A. P. (2002). Gender-specific articulatory–acoustic relations in vowel sequences. *Journal of Phonetics*, 30(3):417–435.

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Linguistics Compass*, 3(2):621–640.

Stacey, J. et Thorne, B. (1985). The missing feminist revolution in sociology. *Social problems*, 32(4):301–316.

Stevens, S. S. et Volkmann, J. (1940). The relation of pitch to frequency : A revised scale. *The American Journal of Psychology*, 53(3):329–353.

Stolcke, A. (2002). SRILM - An extensible language modeling toolkit. *In Proceedings of the 7th International Conference on Spoken Language Processing*, ICSLP 2002 - INTERSPEECH 2002, pages 901–904, Denver, Colorado, USA. ISCA.

Stoller, R. J. (1964). A contribution to the study of gender identity. *International Journal of Psycho-Analysis*, 45:220–226.

Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J. et Carmiel, Y. (2020). WER we are and WER we think we are. *In Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 3290–3295, Online. ACL.

Takefuta, Y., Jancosek, E. G. et Brunt, M. (1972). A statistical analysis of melody curves in the intonation of american english. *In Proceedings of the seventh International Congress of Phonetic Sciences / Actes du Septième Congrès international des sciences phonétiques*, pages 1035–1039. De Gruyter.

Tatman, R. (2017). Gender and dialect bias in YouTube's Automatic Captions. *In Actes de ACL Workshop on Ethics in Natural Language Processing*, pages 53–59.

Tatman, R. et Kasten, C. (2017). Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube Automatic Captions. *In Proceedings of the 19th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2017, pages 934–938, Stockholm, Sweden. ISCA.

Thébaud, F. (2005). Sexe et genre. *In* Maruani, M., éditeur : *Femmes, genre et sociétés*, pages 57–66. La Découverte.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. *In Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC 2012, Istanbul, Turkey. ELRA.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707.

Tolonen, H., Kuulasmaa, K. et Ruokokoski, E. (2000). MONICA population survey data book. `http://www.ktl.fi/publications/monica/surveydb/title.htm`. World Health Organization.

Traunmüller, H. et Eriksson, A. (1994). The frequency range of the voice fundamental in the speech of male and female adults. Rapport technique, Department of Linguistics, University of Stockholm.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in society*, 1(2):179–195.

Tubach, J.-P. (1970). *Reconnaissance automatique de la parole : étude et réalisation fondées sur les niveaux acoustique, morphologique et syntaxique*. Thèse de doctorat, Université Joseph-Fourier-Grenoble I, France.

Vaissière, J. (2006). *La phonétique*. Presses Universitaires de France.

Vanmassenhove, E., Harmeier, C. et Way, A. (2018). Getting gender right in neural machine translation. *In Proceedings of 2018 the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2018, pages 3003–3008, Bruxelles, Belgique. ACL.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. et Polosukhin, I. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems*, pages 5998–6008.

Villani, C., Bonnet, Y., Berthet, C., Levin, F., Schoenauer, M., Cornut, A. C. et Rondepierre, B. (2018). *Donner un sens à l'intelligence artificielle : pour une stratégie nationale et européenne*. Conseil National du Numérique.

Visweswaran, K. (1994). *Fictions of Feminist Ethnography*. University of Minnesota Press.

Vorperian, H. et Kent, R. (2007). Vowel acoustic space development in children : a synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, 50(6):1510–1545.

Wang, F.-Y. (2011). AI's hall of fame. *IEEE Intelligent Systems*, 26(4):5–15.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Enrique Yalta Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. et Ochiai, T. (2018). Espnet : End-to-end speech processing toolkit. *In Proceedings of the 20th Annual Conference of the International Speech Communication Association*, INTERSPEECH 2018, pages 2207–2211, Hyderabad, India. ISCA.

Wegmann, S., McAllaster, D., Orloff, J. et Peskin, B. (1996). Speaker normalization on conversational telephone speech. *In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 1996, pages 339–341, Atlanta, Georgia, USA. IEEE.

West, M., Kraut, R. et Ei Chew, H. (2019). I'd blush if I could : closing gender divides in digital skills through education. UNESCO.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J. et Schwartz, O. (2018). *AI now report 2018*. AI Now Institute at New York University New York.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wilcoxon, F., Katti, S. et Wilcox, R. A. (1963). *Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test*. American Cyanamid Company, Pearl River, NY, USA.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E. *et al.* (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3.

Williams, B. A., Brooks, C. F. et Shmargad, Y. (2018). How algorithms discriminate based on data they lack : Challenges, solutions, and policy implications. *Journal of Information Policy*, 8:78–115.

Witten, I. H. et Bell, T. C. (1991). The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression. *Ieee transactions on information theory*, 37(4):1085–1094.

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs : A review. *In Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, pages 11–19, Sophia Antipolis, France. ISCA.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V. et Chang, K.-W. (2018). Gender bias in coreference resolution : Evaluation and debiasing methods. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. ACL.

Zhou, Y. et Chellappa, R. (1988). Computation of optical flow using a neural network. *IEEE 1988 International Conference on Neural Networks*, 2:71–78.

Annexes

Annexe A

Description des corpus originaux

Source	**Phase 1**		**Phase 2**		
	train/dev	test	train/dev	non-trans	test
France Info	–	–	8h/2h	643h	2h
France Inter	19h40/2h40	2h40	8h/2h	337h	2h
RFI	11h/2h	2h	8h/2h	445h	2h
RTM	–	–	18h/2h	–	2h
France Culture	–	–	–	–	1h
Radio Classique	–	–	–	–	1h
Total	40h		42h/8h	2000h	10h
Période	1998-2000		2003	2004	2004

Table A.1 – Composition du corpus ESTER1. Tiré de Gravier *et al.* (1998) et Galliano *et al.* (2006).

Partition	**Source**		**Durée**	**Période**
Train	ESTER1		100h	1998-2004
	Contenu additionnel	trans. riche	100h	1999-2003
		trans. rapide	50h	1999-2003
	EPAC		45h	2003-2004
Dev.	–		6h	2007
Test	–		7h	2008
Total	–		308h	–

Table A.2 – Composition du corpus ESTER2. Tiré de Galliano *et al.* (2009).

Genre	**Train**	**Dev**	**Test**	**Sources**
TV news	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
TV debates	10h30	2h40	2h40	Pile et Face, Ça vous regarde, Entre les lignes (LCP)
TV entertainment	–	1h05	1h05	La Place du Village (TV8)
Radio shows	7h50	3h00	3h00	Un Temps de Pauchon, Service Public, Le Masque et la Plume, Comme on nous parle, Le Fou du Roi
Total	25h30	8h20	8h20	42h10

TABLE A.3 – Composition du corpus ETAPE. Tiré de (Gravier *et al.*, 2012).

Émission	**Chaînes**	**Dev/test set (min)**
BFM Story	BFM	60
Planète Showbiz	BFM	15
Ça vous regarde	LCP	15
Entre les lignes	LCP	15
Pile et Face	LCP	15
LCP Info	LCP	30
Top Question	LCP	30
Total	–	180

TABLE A.4 – Composition du corpus REPERE. Tiré de (Giraudel *et al.*, 2012).

Partition	**#heures**	**μ(min par loc.)**	**#femmes**	**#hommes**	**#loc.**
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
Total	982.1	–	–	–	–

TABLE A.5 – Composition du corpus Librispeech. Tiré de (Panayotov *et al.*, 2015).

Annexe B

Performances par émissions

Émission	Rôle	Effectif		WER moy.		Effectif	WER moy.
		F	H	F	H		
Africa1 Infos	P	3	13	40,33%	55,69%	16	52,81%
	A	6	35	17,50%	27,89%	41	26,37%
Comme On Nous Parle	P	20	33	51,21%	56,31%	53	65,13%
	A	10	17	26,40%	23,59%	27	24,59%
Culture et Vous	P	45	79	73,04%	56,09%	124	62,26%
	A	75	77	56,71%	42,82%	152	50,13%
La Place du Village	P	5	4	38,00%	55,20%	21	43,89%
	A	1	20	26,00%	45,94%	9	45,62%
Le Masque et la Plume	P	22	18	48,20%	26,67%	40	51,93%
	A	9	19	23,33%	32,37%	28	29,21%
Pile et Face	P	7	21	19,83%	24,53%	28	24,79%
	A	4	91	13,00%	18,89%	95	18,62%
Planète Showbiz	P	24	89	72,28%	62,88%	113	62,61%
	A	91	108	69,69%	38,77%	199	53,10%
RFI Infos	P	224	501	37,41%	30,98%	725	32,97%
	A	23	56	15,91%	13,07%	79	13,90%
RTM Infos	P	24	73	31,58%	28,97%	97	29,62%
	A	164	133	19,37%	19,82%	297	19,57%
Service Public	P	29	35	61,59%	48,00%	64	59,91%
	A	11	13	19,30%	28,62%	24	26,08%
TVME Infos	P	8	12	33,50%	37,21%	20	37,75%
	A	12	20	13,17%	20,89%	32	17,75%
Un Temps de Pauchon	P	31	30	65,96%	48,12%	61	64,41%
	A	13	26	49,82%	38,96%	39	43,56%
Total		**861**	**1523**	**43,58%**	**35,00%**	**2384**	**38,10%**

Table B.1 – WER moyens et effectifs par émissions en fonction du rôle par genre, et pour l'ensemble des locuteurs et locutrices.

Émission	**Rôle**	**Effectif**		**WER med.**		**Effectif**	**WER med.**
		F	H	F	H		
Africa1 Infos	P	3	13	38,00%	50,00%	16	50,00%
	A	6	35	12,00%	22,00%	41	22,00%
Comme On Nous Parle	P	20	33	47,00%	69,00%	53	57,00%
	A	10	17	25,50%	23,00%	27	24,00%
Culture et Vous	P	45	79	78,00%	57,00%	124	67,50%
	A	75	77	52,00%	35,00%	152	44,50%
La Place du Village	P	5	4	35,00%	53,00%	21	43,00%
	A	1	20	26,00%	45,50%	9	45,00%
Le Masque et la Plume	P	22	18	77,50%	24,50%	40	39,50%
	A	9	19	25,00%	32,00%	28	29,50%
Pile et Face	P	7	21	21,00%	19,00%	28	20,50%
	A	4	91	11,50%	19,00%	95	18,00%
Planète Showbiz	P	24	89	78,00%	67,00%	113	71,00%
	A	91	108	72,00%	27,00%	199	56,00%
RFI Infos	P	224	501	22,00%	23,00%	725	23,00%
	A	23	56	16,00%	18,00%	79	14,00%
RTM Infos	P	24	73	28,50%	28,00%	97	28,00%
	A	164	133	16,00%	18,00%	297	17,00%
Service Public	P	29	35	61,00%	25,00%	64	46,00%
	A	11	13	19,00%	23,00%	24	21,50%
TVME Infos	P	8	12	36,00%	34,50%	20	36,00%
	A	12	20	11,50%	20,00%	32	16,00%
Un Temps de Pauchon	P	31	30	63,00%	52,50%	61	56,00%
	A	13	26	47,00%	37,00%	39	39,00%
Total		**861**	**1523**	**30,00%**	**25,00%**	**2384**	**26,00%**

TABLE B.2 – WER médians et effectifs par émissions en fonction du rôle par genre, et pour l'ensemble des locuteurs et locutrices.

Émission	**Effectif**		**WER moy.**		**WER med.**	
	F	H	F	H	F	H
Africa1 Infos	9	48	25,11%	35,42%	15,00%	24,00%
Comme On Nous Parle	30	50	44,53%	55,60%	30,00%	38,50%
Culture et Vous	120	156	62,83%	49,99%	61,50%	41,50%
La Place du Village	6	24	36,67%	47,21%	34,00%	45,50%
Le Masque et la Plume	31	37	52,19%	34,51%	30,00%	31,00%
Pile et Face	11	112	19,90%	20,04%	19,00%	19,00%
Planète Showbiz	115	197	70,27%	48,51%	73,00%	41,00%
RFI Infos	247	557	35,41%	29,17%	21,00%	21,00%
RTM Infos	188	206	20,93%	23,06%	17,00%	20,00%
Service Public	40	48	57,65%	44,88%	49,50%	23,50%
TVME Infos	20	32	21,30%	28,03%	15,50%	24,50%
Un Temps de Pauchon	44	56	65,32%	49,18%	60,50%	39,50%
Total	**861**	**1523**	**43,58%**	**35,00%**	**30,00%**	**25,00%**

TABLE B.3 – WER médians et effectifs par émissions en fonction du rôle par genre, et pour l'ensemble des locuteurs et locutrices.

Annexe C

Données OpenSLR

id	lang	dial	l_status			#spk	#spk_m	#spk_f	#utt	#utt_m	#utt_f	dur	dur_m	dur_f
SLR1	hebrew		Low-res.	ASR	yes	1	1	0	60	60	0		506,4	475,9
SLR12	english		High-res.	ASR	no	2484	1283	1201						
SLR16	english		High-res.	ASR	yes	189	123	66						
SLR18	chinese		High-res.	ASR	yes	50	9	31	11142			35		
SLR22	uyghur		Low-res.	ASR	yes	595	289	315	9468			23,6		
SLR24	iban		Low-res.	ASR	no	23	9	14	3132	1750	1382	8	4,3	3,7
SLR25	wolof		Low-res.	ASR	yes	18	10	8	17998			2,4		
SLR25	swahili		Low-res.	ASR	no				12169					
SLR25	ahmaric		Low-res.	ASR	yes									
SLR30	sinhala		Low-res.	TTS	yes	12	0	12	1251	0	1251			
SLR32	Afrikaans		Low-res.	TTS	yes	9	0	9	2927	0	2927			
SLR32	sesotho		Low-res.	TTS	yes	19	0	19	2096	0	2096			
SLR32	setswana		Low-res.	TTS	yes	26	0	26	2378	0	2378			
SLR32	isiXhosa		Low-res.	TTS	yes									
SLR33	chinese	mandarin	High-res.	ASR	yes	400	186	214						
SLR35	javanese		Low-res.	ASR	yes									
SLR36	sundanese		Low-res.	ASR	yes									
SLR37	bengali	bengladesh	Low-res.	TTS	yes	5	5	0	1890	1890	0			
SLR37	bengali	indian	Low-res.	TTS	yes	9	8	1	1365	1272	93			
SLR38	chinese	mandarin	High-res.	NA	yes	855			102600					
SLR39	spanish		High-res.	ASR	yes	114						52,8		
SLR40	korean		High-res.	ASR	yes	115	45	70	22720					
SLR41	javanese		Low-res.	TTS	yes	39	20	19	5820	2957	2863			
SLR42	khmer		Low-res.	TTS	yes	16	16	0	2905	2905	0			
SLR43	nepali		Low-res.	TTS	yes	18	0	18	2063	0	2963			
SLR44	sundanese		Low-res.	TTS	yes	41	21	20	4211	1811	2400			
SLR45	english	american	High-res.	NA	yes	10	5	5	3842	1656	2186			
SLR46	arabic	tunisian	High-res.	ASR	yes	122								
SLR47	chinese	mandarin	High-res.	ASR	yes	296						100		
SLR51	english		High-res.	ASR	no	2028			268231			452	316	134
SLR52	sinhala		Low-res.	ASR	NA									
SLR53	bengali		Low-res.	ASR	NA									
SLR54	nepali		Low-res.	ASR	NA									
SLR57	french	african	High-res.	ASR	yes									
SLR58	korean		High-res.	ASR	no	41	32	10	3091	2128	963	2,8	1,9	0,9
SLR59	catalan		Low-res.	NA	no							320		
SLR6	czech		High-res.	ASR	yes					26567	0	18,3		
SLR6	english		High-res.	ASR	yes					51463	0	45		
SLR61	spanish	argentinian	High-res.	NA	yes	47	13	34	5826	1817	4009			
SLR61	spanish	peninsular	High-res.	NA	no	3	0	3	89	0	89			
SLR62	chinese	mandarin	High-res.	ASR	yes	600						200		
SLR63	malayalam		Low-res.	NA	yes	43	18	25	4124	2022	2102			
SLR64	marathi		Low-res.	NA	yes	9	0	9	1568	0	1568			
SLR65	tamil		Low-res.	NA	yes	50	25	25	4283	1955	2334			
SLR66	telugu		Low-res.	NA	yes	47	23	27	4446	2153	2293			
SLR67	spanish		High-res.	ASR	no	142	102	40	11243	8376	2867	24,5	18,2	6,3
SLR68	chinese	mandarin	High-res.	ASR	yes	1080	526	554				755		
SLR69	catalan		Low-res.	NA	yes	36	16	20	4328	1918	2320			
SLR70	english	nigerian	High-res.	NA	yes	31	12	19	3357	1313	2044			
SLR71	spanish	chilean	High-res.	NA	yes	31	18	13	4372	2635	1737			
SLR72	spanish	columbian	High-res.	NA	yes	33	17	16	4901	2533	2368			
SLR73	spanish	peruvian	High-res.	NA	yes	38	20	18	5445	2917	2528			
SLR74	spanish	puerto rico	High-res.	NA	yes	5	0	5	616	0	616			
SLR75	spanish	venezuelan	High-res.	NA	yes	23	12	11	3355	1753	1602			
SLR76	basque		Low-res.	NA	yes	52	23	29	7134	3277	3857			
SLR77	galician		Low-res.	NA	yes	44	10	34	5585	1322	4263			
SLR78	gujarati		Low-res.	NA	yes	36	18	18	4270	2052	2218			
SLR79	kannada		Low-res.	NA	yes	59	36	23	4398	2213	2185			
SLR80	burmese		Low-res.	NA	yes	20	0	20	2529	0	2529			
SLR82	chinese		High-res.	ASR	no	1000			130109			274		
SLR83	english	irish	High-res.	NA	yes	3	3	0	450	450	0			
SLR83	english	midlands	High-res.	NA	yes	5	3	2	696	450	246			
SLR83	english	northern	High-res.	NA	yes	19	14	5	2847	2097	750			
SLR83	english	scottish	High-res.	NA	yes	17	11	6	2543	1649	894			
SLR83	english	southern	High-res.	NA	yes	57	29	28	8492	4331	4161			
SLR83	english	welsh	High-res.	NA	yes	19	11	8	2848	1650	1198			

id	lang	dial	lang_status	sizeGB	size	sampling	task	elicited	provided	found_in	year	producer
SLR1	hebrew		Low-res.	0,006	small	8000	ASR	yes	yes	metadata	NA	Anonymous
SLR12	english		High-res.	63,9	large	16000	ASR	no	yes	metadata	2015	Center for Lang. and Speech Proc. & Human Lang. Techn. Center
SLR16	english		High-res.	91,6	large	16000	ASR	yes	yes	metadata	2007	AMI Project
SLR18	chinese		High-res.	10,2	medium	16000	ASR	yes	yes	paper	2015	Center for Speech and Language Technology (CSLT)
SLR22	uyghur		Low-res.	3	small	16000	ASR	yes	yes	metadata	2015	Center for Speech and Language Technology (CSLT)
SLR24	iban		Low-res.	0,996	small	16000	ASR	no	yes	paper	NA	LIG
SLR25	wolof		Low-res.	1,5	small	16000	ASR	yes	yes	paper	2016	ALFFA Project
SLR25	swahili		Low-res.	2,6	small	16000	ASR	no	no	NA	2012	ALFFA Project
SLR25	ahmaric		Low-res.	2,5	small	16000	ASR	yes	no	NA	2005	ALFFA Project
SLR30	sinhala		Low-res.	1,2	small	48000	TTS	yes	no	manually	2016	Google
SLR32	Afrikaans		Low-res.	1,1	small	48000	TTS	yes	no	manually	2017	Google
SLR32	sesotho		Low-res.	1,1	small	44100	TTS	yes	no	manually	2017	Google
SLR32	setswana		Low-res.	1,2	small	48000	TTS	yes	no	manually	2017	Google
SLR32	isiXhosa		Low-res.	1,1	small	48000	TTS	yes	no	NA	2017	Google
SLR33	chinese	mandarin	High-res.	15,6	medium	16000	ASR	yes	yes	metadata	2017	Beijing Shell Shell Technology Co.,Ltd.
SLR35	javanese		Low-res.	18,9	medium	16000	ASR	yes	no	NA	2017	Google
SLR36	sundanese		Low-res.	23	medium	16000	ASR	yes	no	NA	2017	Google
SLR37	bengali	bengladesh	Low-res.	1	small	48000	TTS	yes	no	manually	NA	Google
SLR37	bengali	indian	Low-res.	0,7	small	48000	TTS	yes	no	manually	NA	Google
SLR38	chinese	mandarin	High-res.	12,7	medium	16000	NA	yes	yes	metadata	NA	Surfingtech
SLR39	spanish		High-res.	2,6	small	22050	ASR	yes	no	NA	NA	DFL & CTELL
SLR40	korean		High-res.	2,9	small	16000	ASR	yes	yes	metadata	NA	Lucas Jo(@Atlas Guide Inc.) and Wonkyum Lee(@Gridspace Inc.).
SLR41	javanese		Low-res.	2,4	small	48000	TTS	yes	yes	indexed	2018	Google
SLR42	khmer		Low-res.	1,4	small	48000	TTS	yes	yes	indexed	2018	Google
SLR43	nepali		Low-res.	0,967	small	48000	TTS	yes	yes	indexed	2018	Google
SLR44	sundanese		Low-res.	1,9	small	48000	TTS	yes	yes	indexed	2018	Google
SLR45	english	american	High-res.	0,547	small	16000	NA	yes	yes	indexed	NA	Surfingtech
SLR46	arabic	tunisian	High-res.	1,5	small	16000	ASR	yes	no	NA	NA	NA
SLR47	chinese	mandarin	High-res.	11,4	medium	16000	ASR	yes	no	NA	2018	Primewords Information Technology Co., Ltd.
SLR51	english		High-res.	63,4	large	16000	ASR	no	yes	paper	NA	LIUM
SLR52	sinhala		Low-res.	14,7	medium	16000	ASR	NA	no	NA	2018	Google
SLR53	bengali		Low-res.	14,6	medium	16000	ASR	NA	no	NA	2018	Google
SLR54	nepali		Low-res.	9,3	medium	16000	ASR	NA	no	NA	2018	Google
SLR57	french	african	High-res.	2,3	small	16000	ASR	yes	no	NA	NA	CTELL & RDECOM (army)
SLR58	korean		High-res.	0,175	small	16000	ASR	no	yes	paper	2018	Electronics Engineering, Inha University
SLR59	catalan		Low-res.	18,2	medium	16000	NA	no	no	NA	NA	Col·lectivaT
SLR6	czech		High-res.	2,1	small	16000	ASR	yes	no	NA	2014	Charles University in Prague
SLR6	english		High-res.	5,2	medium	16000	ASR	yes	no	NA	2014	Charles University in Prague
SLR61	spanish	argentinian	High-res.	2,8	small	48000	NA	yes	yes	indexed	2019	Google
SLR61	spanish	peninsular	High-res.	0,03	small	48000	NA	no	no	manually	2019	Google
SLR62	chinese	mandarin	High-res.	18,8	medium	16000	ASR	yes	no	NA	NA	Beijing DataTang Technology Co., Ltd.
SLR63	malayalam		Low-res.	1,9	small	48000	NA	yes	yes	indexed	2019	Google
SLR64	marathi		Low-res.	1	small	48000	NA	yes	yes	indexed	2019	Google
SLR65	tamil		Low-res.	2,4	small	48000	NA	yes	yes	indexed	2019	Google
SLR66	telugu		Low-res.	2	small	48000	NA	yes	yes	indexed	2019	Google
SLR67	spanish		High-res.	2,8	small	16000	ASR	no	yes	metadata	2019	Univ. Nacional Autonoma de Mexico + CIEMPIESS-UNAM project
SLR68	chinese	mandarin	High-res.	131	large	16000	ASR	yes	yes	metadata	2019	Magic Data Technology Co., Ltd.
SLR69	catalan		Low-res.	3,3	small	48000	NA	yes	yes	indexed	2019	Google
SLR70	english	nigerian	High-res.	2	small	48000	NA	yes	yes	indexed	2019	Google
SLR71	spanish	chilean	High-res.	2,5	small	48000	NA	yes	yes	indexed	2019	Google
SLR72	spanish	columbian	High-res.	2,6	small	48000	NA	yes	yes	indexed	2019	Google
SLR73	spanish	peruvian	High-res.	3,2	small	48000	NA	yes	yes	indexed	2019	Google
SLR74	spanish	puerto rico	High-res.	0,347	small	48000	NA	yes	yes	indexed	2019	Google
SLR75	spanish	venezuelan	High-res.	1,7	small	48000	NA	yes	yes	indexed	2019	Google
SLR76	basque		Low-res.	4,8	small	48000	NA	yes	yes	indexed	2019	Google
SLR77	galician		Low-res.	3,6	small	48000	NA	yes	yes	indexed	2019	Google
SLR78	gujarati		Low-res.	2,7	small	48000	NA	yes	yes	indexed	2019	Google
SLR79	kannada		Low-res.	2,9	small	48000	NA	yes	yes	indexed	2019	Google
SLR80	burmese		Low-res.	1,4	small	48000	NA	yes	yes	indexed	2019	Google
SLR82	chinese		High-res.	36,9	medium	16000	ASR	no	no	NA	2019	Center for Speech and Language Technologies, Tsinghua University
SLR83	english	irish	High-res.	0,247	small	48000	NA	yes	yes	indexed	2019	Google
SLR83	english	midlands	High-res.	0,416	small	48000	NA	yes	yes	indexed	2019	Google
SLR83	english	northern	High-res.	1,7	small	48000	NA	yes	yes	indexed	2019	Google
SLR83	english	scottish	High-res.	1,5	small	48000	NA	yes	yes	indexed	2019	Google
SLR83	english	southern	High-res.	5	medium	48000	NA	yes	yes	indexed	2019	Google
SLR83	english	welsh	High-res.	1,9	small	48000	NA	yes	yes	indexed	2019	Google

Annexe D

Distribution des performances

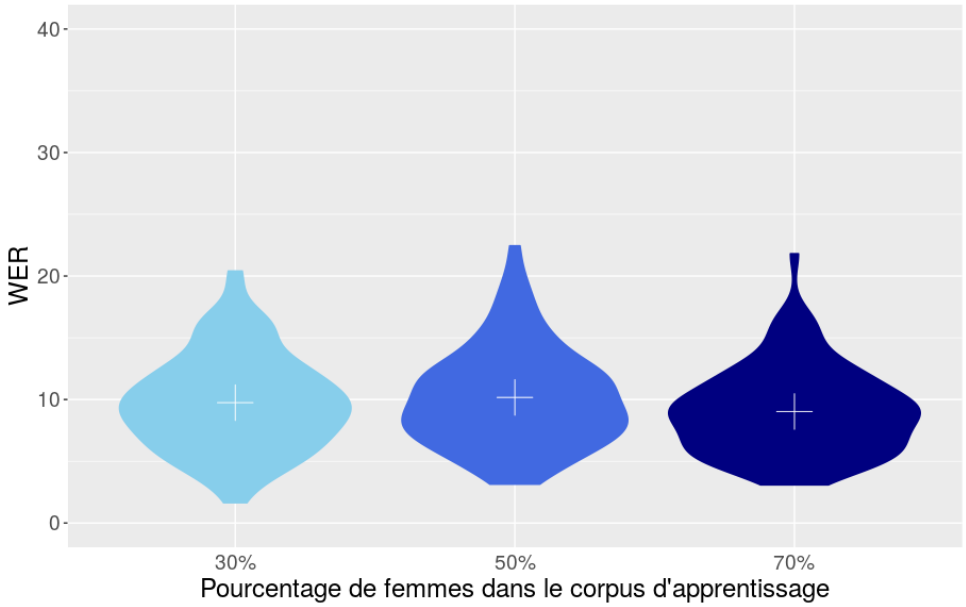

Figure 31 – Variabilité due à la représentation du genre. Distribution des performances pour nos 3 modèles sur le corpus test-clean. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

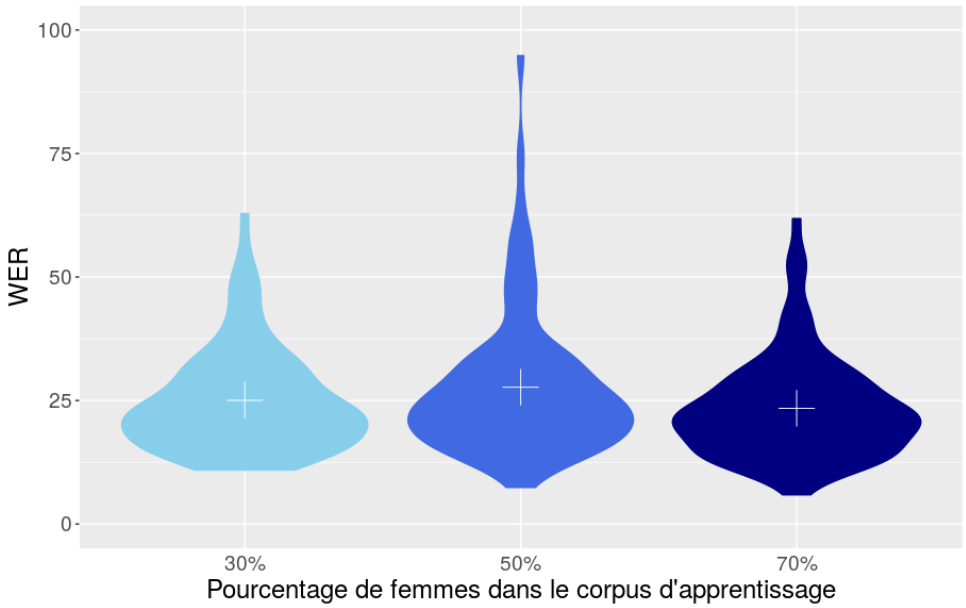

Figure 32 – Variabilité due à la représentation du genre. Distribution des performances pour nos 3 modèles sur le corpus test-other. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 100%).

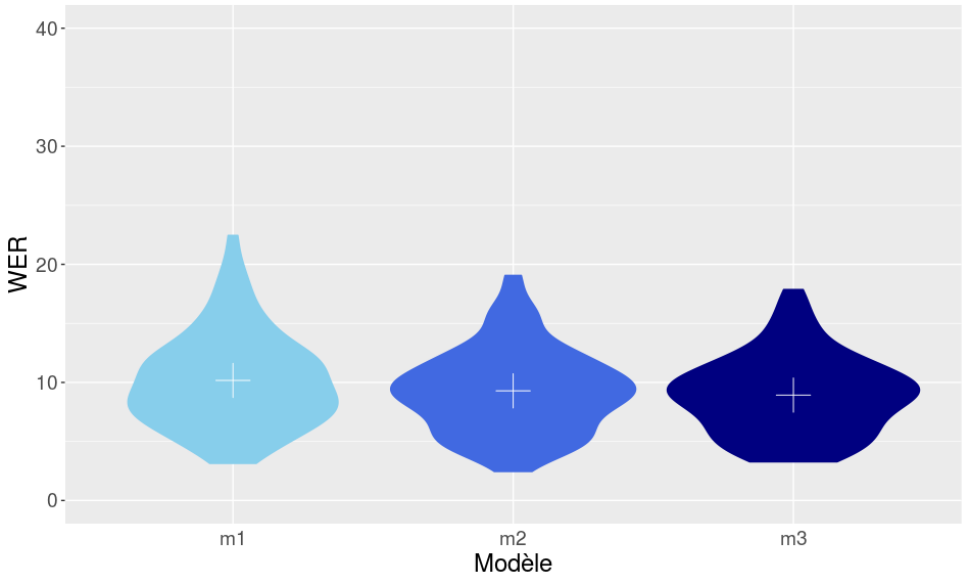

FIGURE 33 – Variabilité due au modèle. Distribution des performances sur le corpus test-clean pour nos 3 modèles entraînés sur le corpus wper50 avec des graines différentes. Les valeurs des graines ont été choisies arbitrairement. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

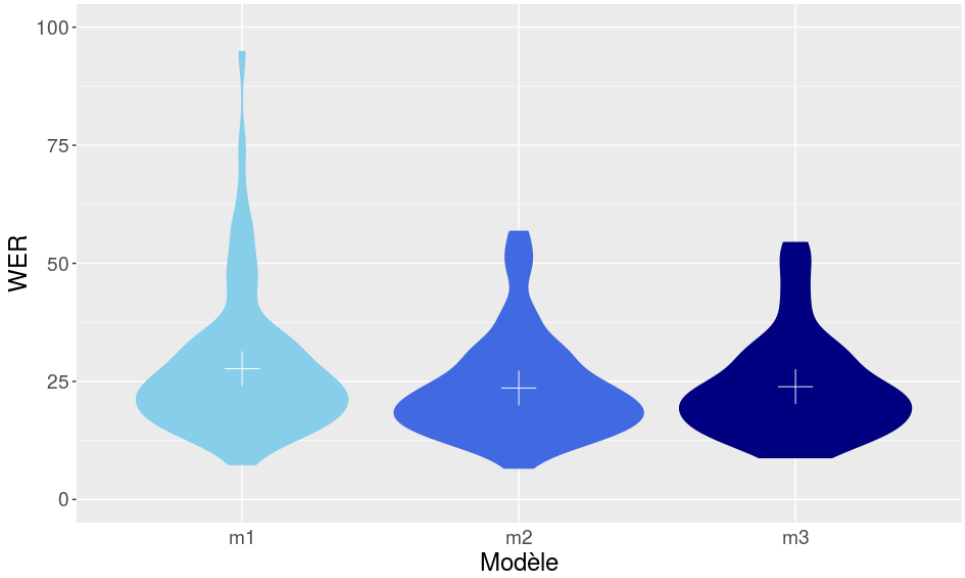

FIGURE 34 – Variabilité due au modèle. Distribution des performances sur le corpus test-other pour nos 3 modèles entraînés sur le corpus wper50 avec des graines différentes. Les valeurs des graines ont été choisies arbitrairement. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 100%).

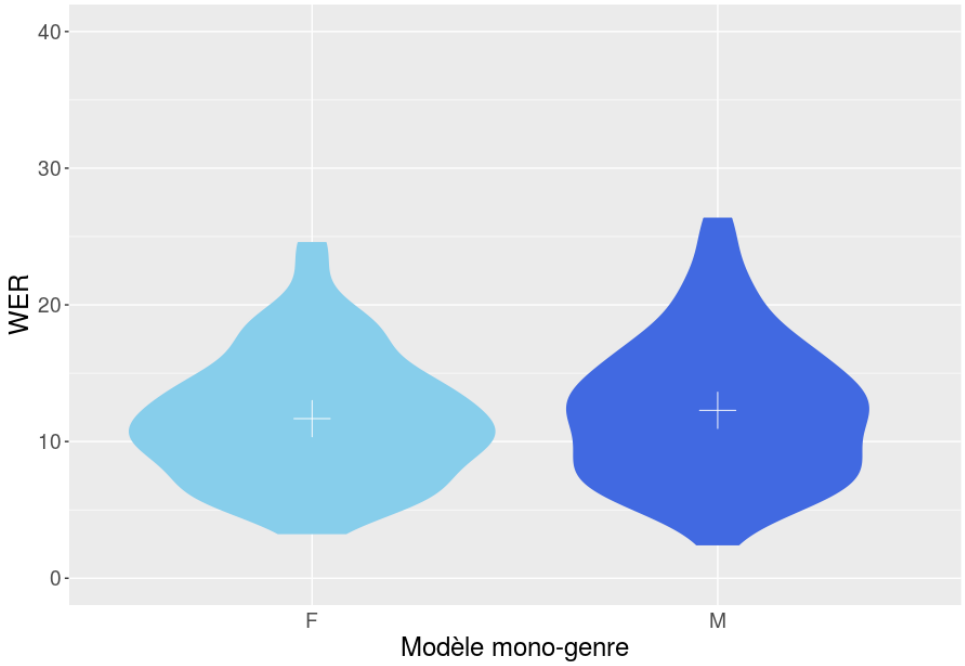

FIGURE 35 – Cas limite. Distribution des performances pour nos 2 modèles mono-genre sur le corpus test-clean. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

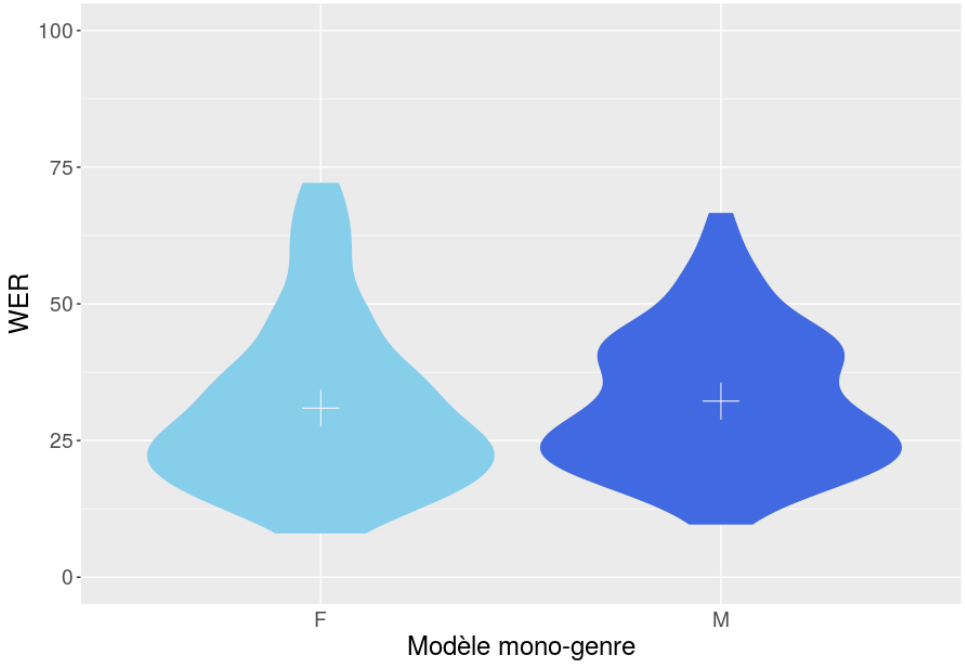

FIGURE 36 – Cas limite. Distribution des performances pour nos 2 modèles mono-genre sur le corpus test-other. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 100%).

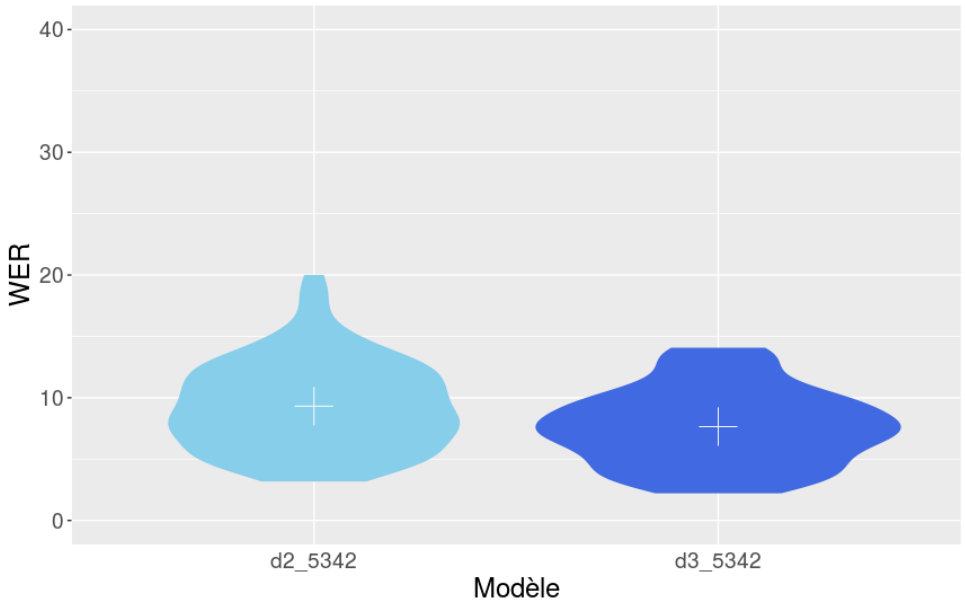

FIGURE 37 – Erreur expérimentale. Distribution des performances pour nos 2 modèles d2-5342 et d3-5342 sur le corpus test-clean. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 40%).

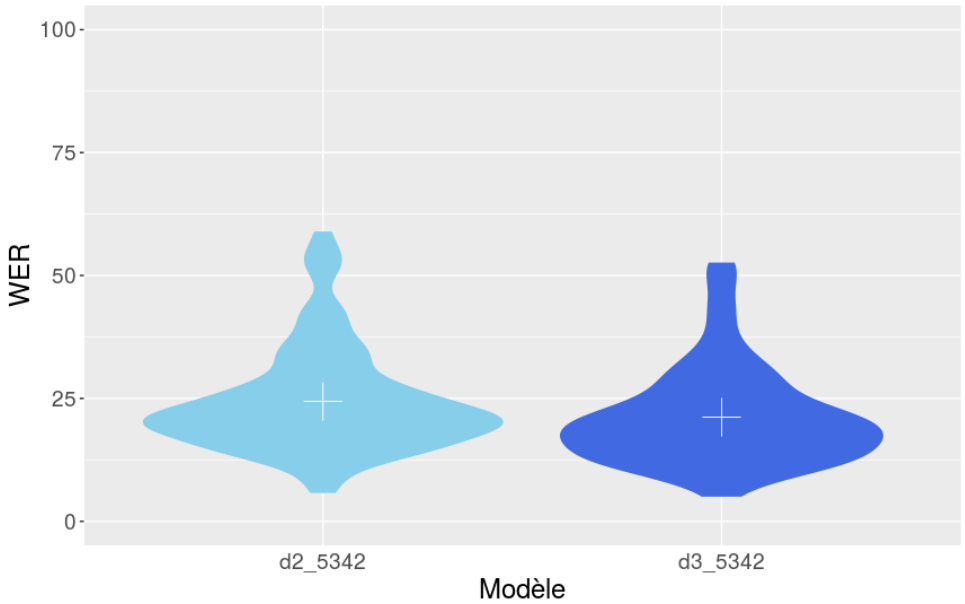

FIGURE 38 – Erreur expérimentale. Distribution des performances pour nos 2 modèles d2-5342 et d3-5342 sur le corpus test-other. La croix blanche représente la valeur moyenne. (N.B. : l'échelle de WER va de 0 à 100%).

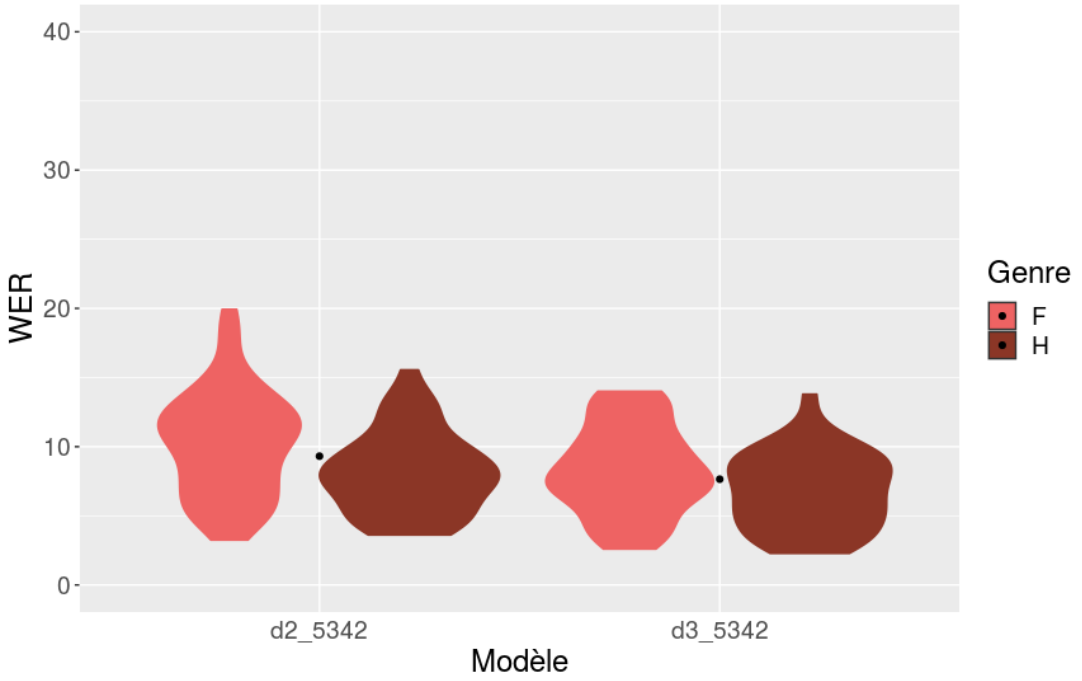

FIGURE 39 – Erreur expérimentale. Distribution des performances pour chaque catégorie de genre pour nos 2 modèles d2-5342 et d3-5342 sur le corpus test-clean. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 40%).

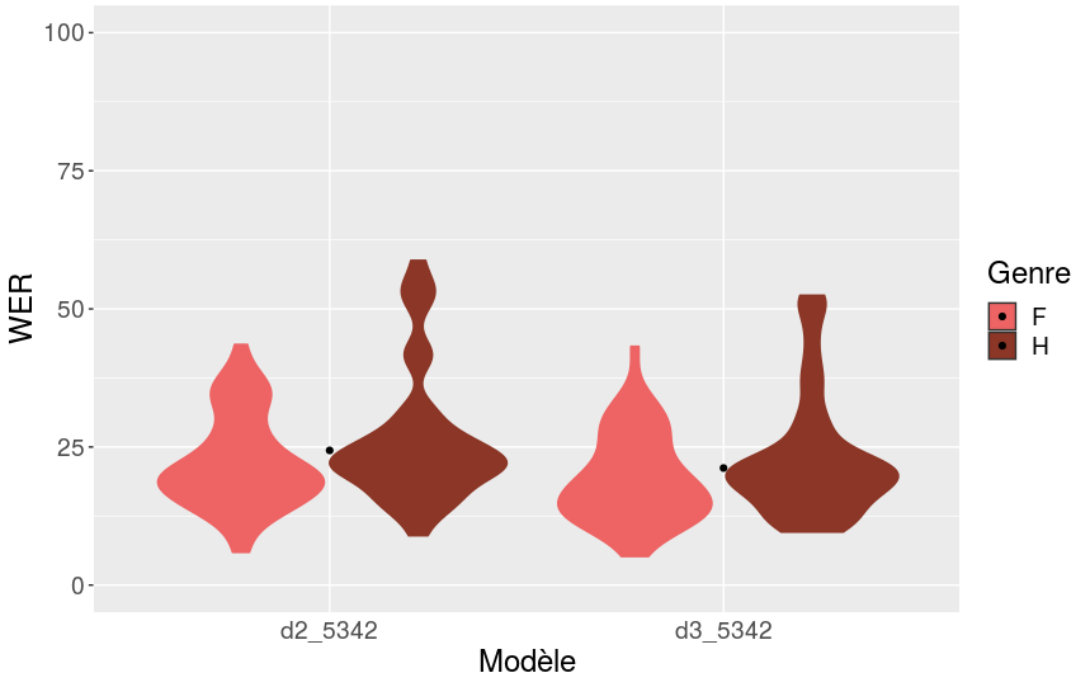

FIGURE 40 – Erreur expérimentale. Distribution des performances pour chaque catégorie de genre pour nos 2 modèles d2-5342 et d3-5342 sur le corpus test-other. Le point noir représente la valeur moyenne indépendamment des catégories. (N.B. : l'échelle de WER va de 0 à 100%).

Annexe E

Analyses statistiques

	wper30			wper50			wper70			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
wper30	-	-	-	0,327	1	0,568	1,851	1	0,174			
wper50	0,327	1	0,568	-	-	-	3,676	1	0,055	3,910	2	0,142
wper70	1,851	1	0,174	3,676	1	0,055	-	-	-			

Table E.1 – Variabilité due à la représentation des catégories de genre. Kruskall-Wallis. Test-clean. Risque α fixé à 0,01.

	wper30			wper50			wper70			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
wper30	-	-	-	1,029	1	0,311	1,245	1	0,265			
wper50	1,029	1	0,311	-	-	-	4,362	1	0,037	4,431	2	0,109
wper70	1,245	1	0,265	4,362	1	0,037	-	-	-			

Table E.2 – Variabilité due à la représentation des catégories de genre. Kruskall-Wallis. Test-other. Risque α fixé à 0,01.

	m1			m2			m3			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
m1	-	-	-	1,493	1	0,222	3,460	1	0,063			
m2	1,493	1	0,222	-	-	-	0,390	1	0,532	3,569	2	0,168
m3	3,460	1	0,063	0,390	1	0,532	-	-	-			

Table E.3 – Variabilité du modèle. Kruskall-Wallis. Test-clean. Risque α fixé à 0,01.

	m1			m2			m3			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
m1	-	-	-	4,428	1	0,035	3,718	1	0,054			
m2	4,428	1	0,035	-	-	-	0,065	1	0,799	5,483	2	0,064
m3	3,718	1	0,054	0,065	1	0,799	-	-	-			

TABLE E.4 – Variabilité du modèle. Kruskall-Wallis. Test-other. Risque α fixé à 0,01.

	m1			d2			d3			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
m1	-	-	-	5,420	1	0,012	3,069	1	0,080			
d2	5,420	1	0,012	-	-	-	0,0348	1	0,555	5,903	2	0,052
d3	3,069	1	0,080	0,0348	1	0,555	-	-	-			

TABLE E.5 – Variabilité due aux données. Kruskall-Wallis. Test-clean. Risque α fixé à 0,01.

	m1			d2			d3			Tous		
	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.	χ^2	df	p-val.
m1	-	-	-	5,225	1	0,022	4,005	1	0,045			
d2	5,225	1	0,022	-	-	-	0,198	1	0,656	6,291	2	0,043
d3	4,005	1	0,045	0,198	1	0,656	-	-	-			

TABLE E.6 – Variabilité due aux données. Kruskall-Wallis. Test-other. Risque α fixé à 0,01.

	test-clean			test-other		
	χ^2	df	p-val.	χ^2	df	p-val.
d2-5342/d3-5342	**8,826**	**1**	**0,003**	6,103	1	0,014

TABLE E.7 – Variabilité due aux données. Kruskall-Wallis. Modèles d2-5342 et d3-5342. Risque α fixé à 0,01.

Expérience	Modèles	test-clean				test-other			
		F	H	W	p-val.	F	H	W	p-val.
Var. représ.	wper30	**10,3%**	**8,0%**	**1279**	**0,003**	21,2%	23,4%	856	0,211
	wper50	11,1%	8,9%	1176	0,036	21,9%	25,0%	834	0,153
	wper70	9,4%	8,3%	1123	0,101	19,5%	22,8%	749	0,034
Var. modèle	m2	**10,2%**	**8,6%**	**1296**	**0,002**	19,6%	22,3%	797	0,083
	m3	**9,7%**	**7,4%**	**1259,5**	**0,005**	18,7%	22,4%	750	0,035
Var. données	d2	9,1%	7,5%	1199	0,022	19,7%	21,7%	827	0,137
	d3	9,1%	7,7%	1216	0,014	20,3%	21,8%	826	0,135
	d2-5342	**10,7%**	**7,7%**	**1240**	**0,008**	20,1%	22,7%	799	0,086
	d3-5342	7,7%	6,9%	1173	0,038	17,7%	20,4%	831	0,146
Sys. mono.	Féminin	10,6%	11,9%	746	0,114	**21,1%**	**36,0%**	**387**	**1,375**e^{-7}
	Masculin	**14,3%**	**8,7%**	**1540,5**	**1,871**e^{-7}	**36,8%**	**24,9%**	**1391**	**0,002**

TABLE E.8 – WER médians, statistique W et p-valeur du test de la somme des rangs de Wilcoxon comparant la différence entre les distributions de WER entre hommes et femmes pour nos différents modèles. Risque α fixé à 0,01.

Annexe F

Étude des corrélations entre WER et paramètres acoustiques

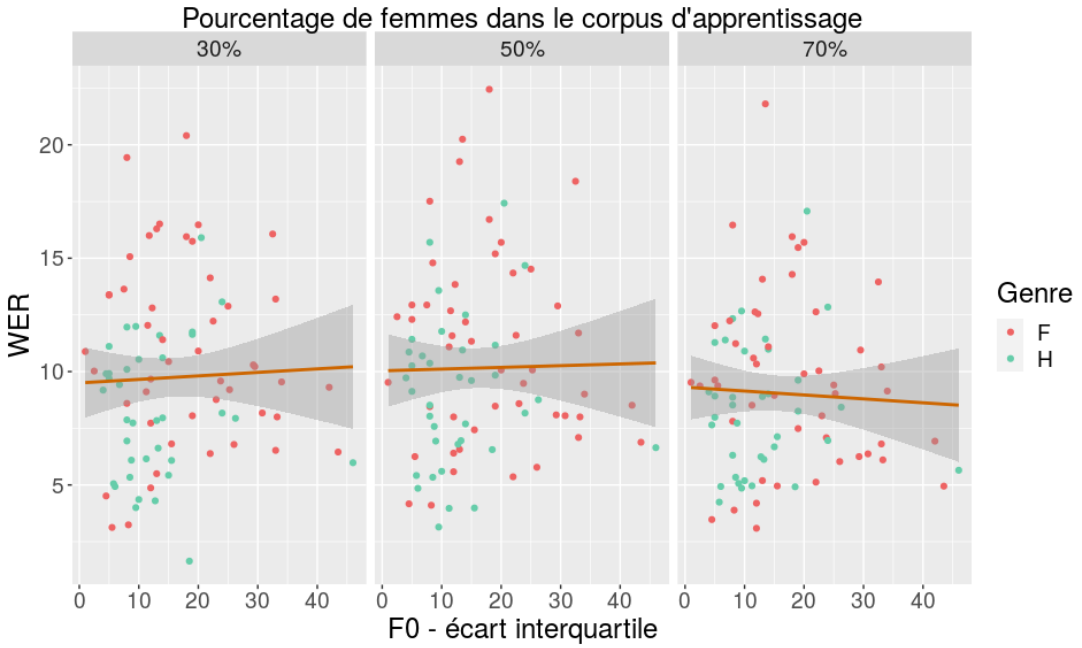

Figure 41 – WER en fonction de l'écart interquartile de F0 (en Hertz) obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

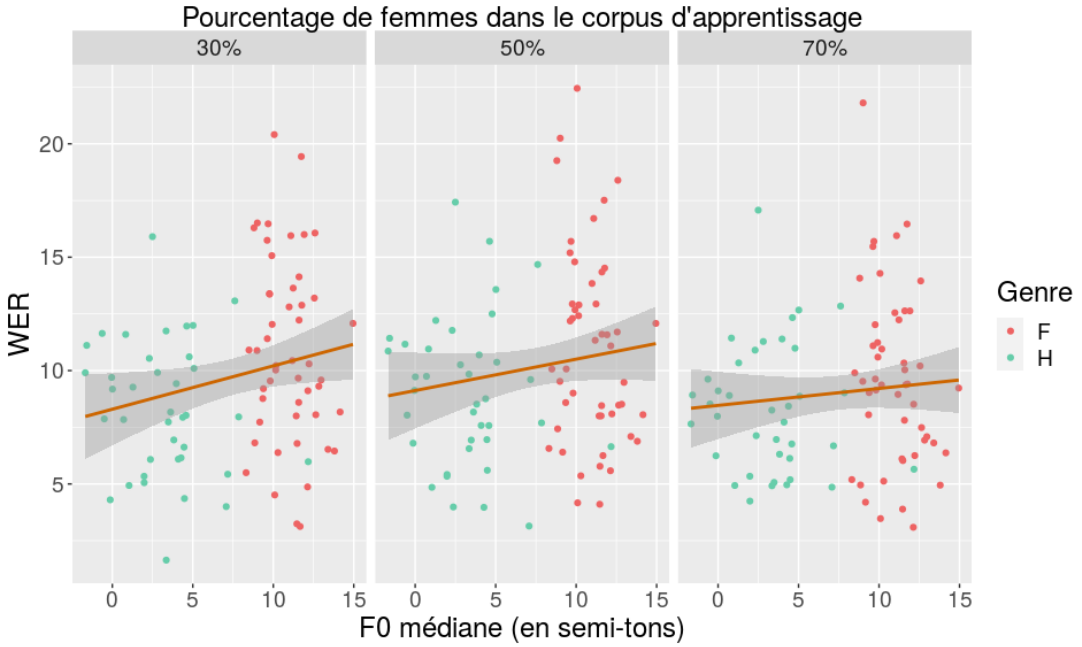

FIGURE 42 – WER en fonction de la F0 médiane (en semi-tons) obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

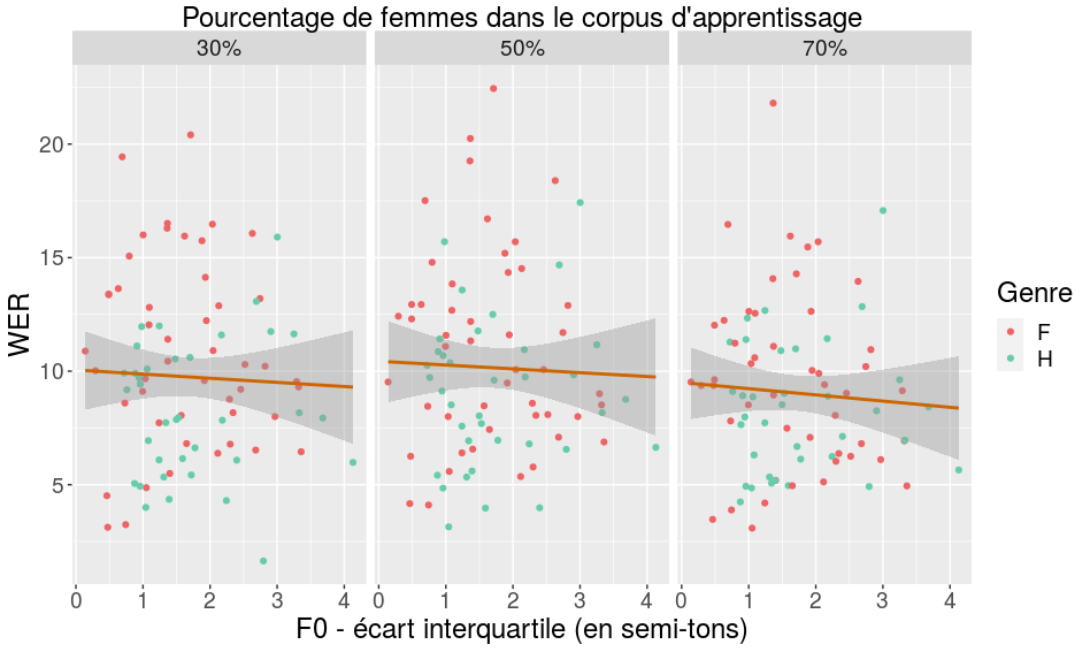

FIGURE 43 – WER en fonction de l'écart interquartile de F0 (en semi-tons) obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

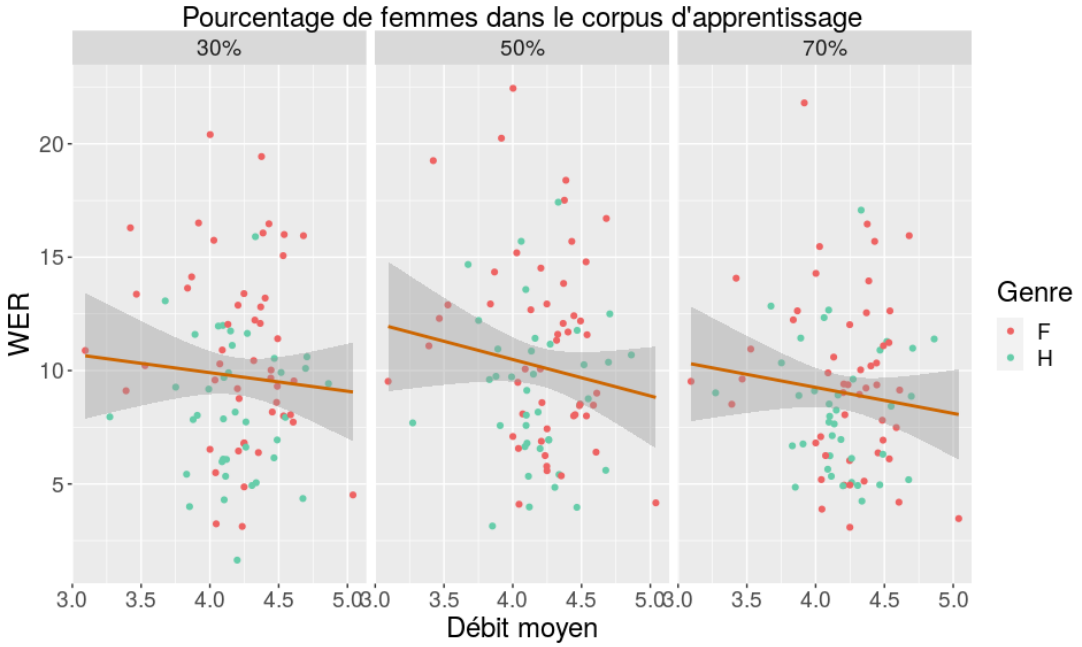

FIGURE 44 – WER en fonction du débit moyen obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

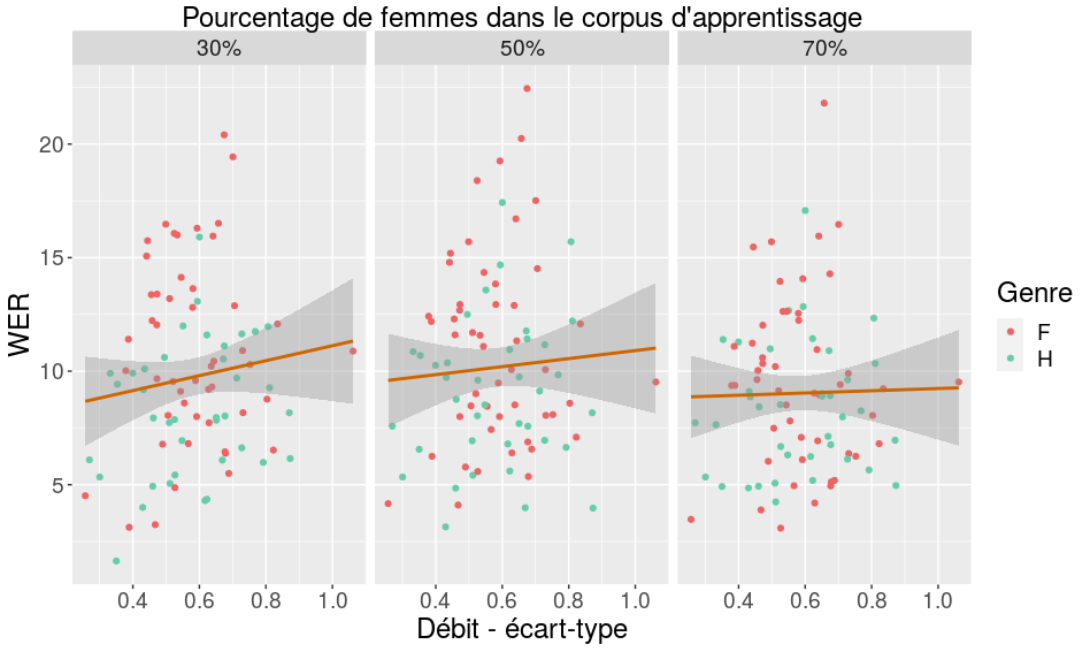

FIGURE 45 – WER en fonction de l'écart-type de débit obtenus sur le test-clean pour nos 3 systèmes faisant varier la représentation des catégories de genre dans les données d'apprentissage.

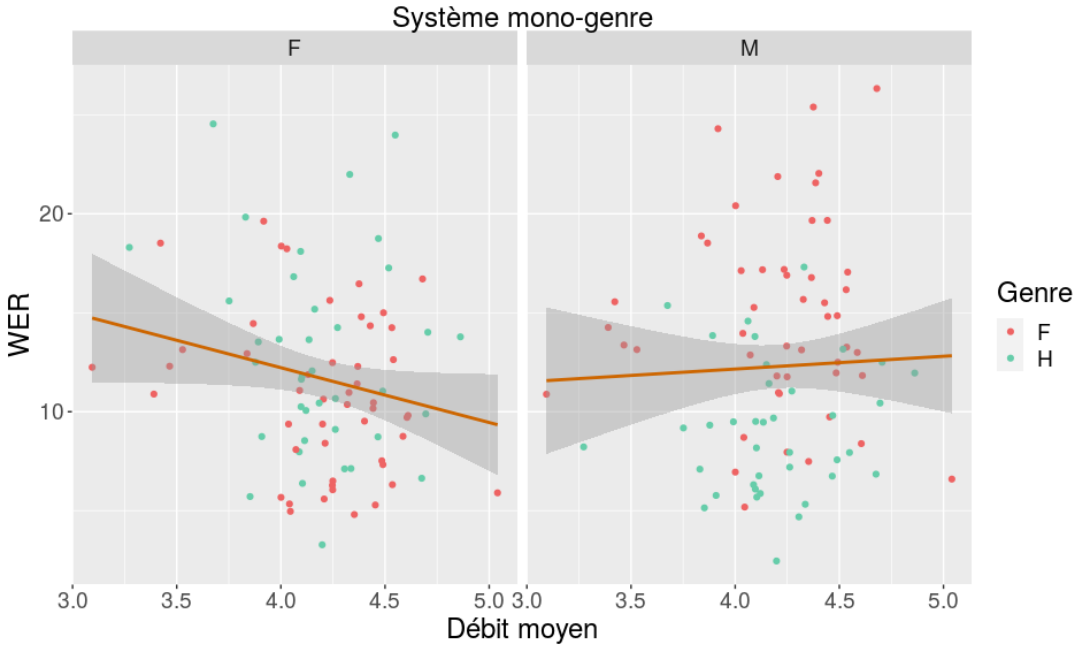

FIGURE 46 – WER en fonction du débit moyen obtenus sur le test-clean pour nos 2 systèmes mono-genre.

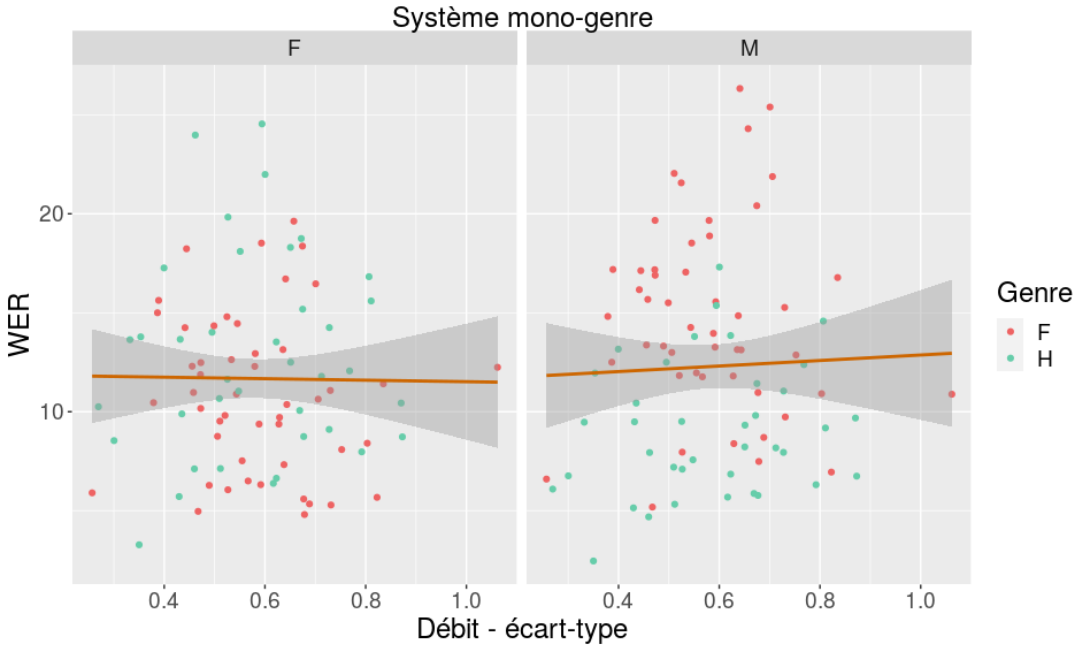

FIGURE 47 – WER en fonction de l'écart-type de débit obtenus sur le test-clean pour nos 2 systèmes mono-genre.

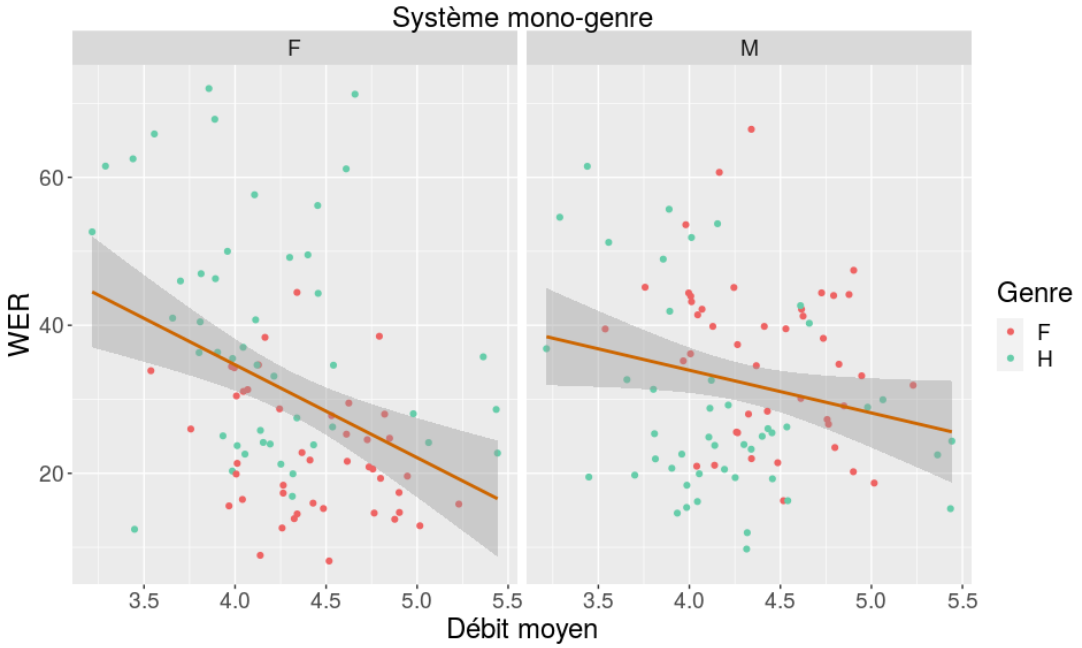

FIGURE 48 – WER en fonction du débit moyen obtenus sur le test-other pour nos 2 systèmes mono-genre.

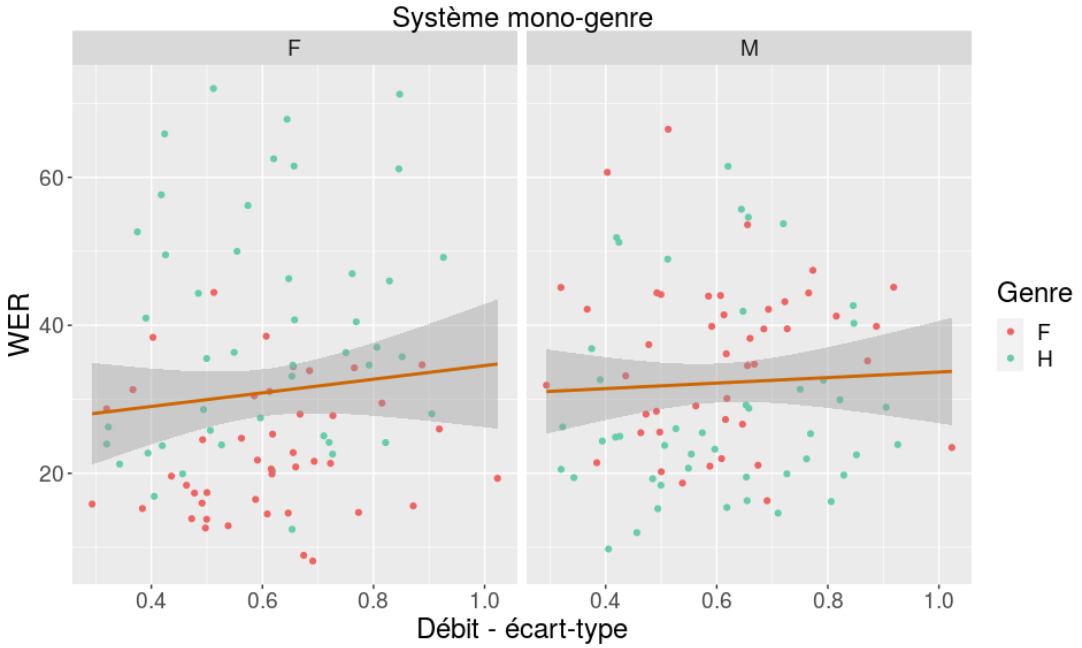

FIGURE 49 – WER en fonction de l'écart-type de débit obtenus sur le test-other pour nos 2 systèmes mono-genre.