Répartition hommes/femmes dans les systèmes d'IA : une étude pilote sur les grands corpus pour la transcription automatique de la parole

Mahault Garnerin

GARNERIN
Mahault

Sous la direction de Laurent Besacier, Claudine Moïse et Solange Rossato

Laboratoire : LIG-GETALP

UFR LLASIC
Département Sciences du Langage et FLE

Mémoire de master 2 Sciences du Langage - orientation Recherche - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2017-2018

Répartition hommes/femmes dans les systèmes d'IA : une étude pilote sur les grands corpus pour la transcription automatique de la parole

GARNERIN

Mahault

Sous la direction de Laurent Besacier, Claude Moïse et Solange Rossato

Laboratoire : LIG-GETALP

UFR LLASIC

Département Sciences du Langage et FLE

Mémoire de master 2 Sciences du Langage - orientation Recherche - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2017-2018

Remerciements

Je tiens à remercier tout d'abord Solange Rossato pour m'avoir soutenue et guidée dans ce projet ainsi que Laurent Besacier pour ses critiques, remarques et conseils précieux.

Je remercie également, et fort chaleureusement, Frédérique Letué pour le temps qu'elle m'a accordé ainsi que pour sa patience face à mon ignorance statistique.

Et enfin un grand merci à Céline (pour ses nombreuses relectures), William, Eric et Lucie pour m'avoir encouragée, motivée et supportée (dans toute la polysémie du mot) durant toute la durée de ce mémoire.

Ce travail a bénéficié du programme «Investissements d'avenir» portant la référence ANR-15-IDEX-02 et du Pole Grenoble Cognition.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : GARNERIN

PRENOM : Mahault

DATE : 28/05/18 SIGNATURE : [redacted]

Sommaire

Introduction

Le 1er juin 2011 un article[1] intitulé « *It's Not You, It's It: Voice Recognition Doesn't Recognize Women* » a été publié dans la rubrique technologique du Time. Cet article expliquait que les chances étaient fortes que les systèmes de reconnaissance automatique contenus dans les voitures fonctionnent moins bien si l'on était une femme. Le directeur du département client, David Champion expliquait que ce n'était pas que le genre, mais également les accents, sa voix n'étant jamais reconnue par les voitures de marque anglaise comme les Land Rover.

Que ce soit pour régler le GPS de votre voiture, pour dicter des SMS à votre téléphone ou interagir avec un service client téléphonique automatique, les technologies de la parole investissent de plus en plus notre quotidien. Ces nouvelles technologies sont de plus en plus utilisées mais il semblerait que les performances puissent varier en fonction de différents facteurs tels que le genre ou l'accent de l'utilisateur. Dans le cas de l'article précédemment cité, la solution proposée par les constructeurs automobiles était de changer la manière dont les gens parlent. D'après le vice-président des technologies de la parole du groupe ATX, Tom Schalk, *"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"*. Cette solution nous semblant quelque peu insultante et difficile à mettre en œuvre, nous avons souhaité aborder le problème sous un angle différent : l'entraînement non pas des utilisateurs mais des systèmes. Les systèmes automatiques sont développés sur un ensemble de données, et nous supposons que le choix de ces données à un impact non-négligeable sur les performances du système.

Le présent travail se propose donc d'étudier, dans le cadre restreint de la parole médiatique, les données utilisées par les systèmes de reconnaissance automatique de la parole et d'en évaluer l'influence sur les performances. Plus généralement, ce travail peut également constituer une première contribution concernant l'étude de l'impact des données d'apprentissage sur les performances des systèmes d'intelligence artificielle utilisés pour le traitement automatique de la parole en français.

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

Ce mémoire s'articulera en 3 parties : la première consistant en un état de l'art et une présentation de notre problématique de recherche, la deuxième développant notre méthodologie et la troisième présentant l'ensemble de nos résultats obtenus.

Partie 1 : État de l’art et positionnement du problème

1. Des systèmes développés sur corpus

Dernièrement le *machine learning* ou apprentissage automatique a investi de nombreux champs de recherche. L'explosion de la quantité de données disponibles notamment avec le World Wide Web, les réseaux de neurones artificiels, la puissance de calcul des machines actuelles sont autant de facteurs qui ont installé l'apprentissage automatique comme méthode incontournable. La norme est ainsi devenue de développer les systèmes sur des grands ensembles de données, grâce à des algorithmes d'apprentissage profond. Les données sont la matière première de ces nouveaux systèmes et la constitution de corpus d'apprentissage est maintenant une part importante du développement de ces nouveaux outils.

1.1. Les corpus dédiés

En traitement automatique de la parole (TAL) comme dans de nombreux domaines, les corpus sont donc 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., 2004). Différents corpus ont donc été créés, dans le cadre de campagnes d'évaluation, pour permettre aux équipes de recherche d'avoir accès à des données de qualité à moindre coût. Et ces campagnes ont également permis d'évaluer les performances des systèmes sur différentes tâches. Dans le cadre de ce travail nous nous sommes intéressée à 4 grands corpus du français, ESTER1, ESTER2, ETAPE et REPERE décrits ci-dessous. Un résumé de la constitution de l'ensemble des corpus est présenté dans l'Annexe 1.

ESTER 1

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. Le but de cette campagne[2] était de permettre une évaluation commune des performances des systèmes du traitement de la parole en France. Elle a été principalement 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).

[2] http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/plan-phase1-1.1.pdf [consulté le 23/05/18]

La campagne ESTER1 a eu lieu entre 2003 et 2005. Les tâches sur lesquelles se focalisait cette campagne s'organisaient selon deux axes : la transcription orthographique et la segmentation. Une tâche d'extraction d'information, la reconnaissance d'entités nommées, a été rajoutée dans la deuxième phase de la campagne. Une description des différentes catégories de tâches et des métriques d'évaluation est disponible dans les plans d'évaluation consultables sur le site de l'AFCP. Pour évaluer les systèmes sur chacune de ces tâches, un corpus a été constitué, contenant des enregistrements provenant de 4 sources différentes : 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.

ESTER2

ESTER2 a débuté fin janvier 2008 et avait pour but de mesurer les progrès effectués par les systèmes de transcription automatique depuis la publication des résultats d'ESTER1. Elle reprend donc les tâches étudiées dans la campagne d'ESTER1 et en introduit de nouvelles, comme la transcription avec données contemporaines et la reponctuation. ESTER2 visait aussi à élargir le type de données pris en compte comme la parole accentuée et la 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'Africa n°1 et TVME), dans le but d'étudier l'impact de l'accent.[3] Une partie du corpus EPAC (Estève et al., 2010), annoté par le Laboratoire Informatique de l'Université du Mans (LIUM) a également été distribuée. Le corpus EPAC provient de la partie non-annotée d'ESTER1 qui contient les transcriptions d'environ 100h de parole « conversationnelle ».

ETAPE

La campagne ETAPE (Évaluation en Traitement Automatique de la Parole) s'est déroulée entre 2011 et 2012 et se situe dans la continuité des deux campagnes ESTER. La

[3] http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/plan-ester2-1.0.pdf [consulté le 23/05/18]

campagne ETAPE avait pour but de diversifier les sources étudiées en évaluant notamment les systèmes sur des émissions télévisées. En plus de permettre l'observation de l'évolution des performances des systèmes sur les tâches précédemment étudiées dans le cadre des campagnes ESTER, un focus a été fait sur la parole spontanée, avec l'utilisation de données télévisées dans lesquelles les phénomènes de parole superposée sont 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 étaient une fois de plus des tâches de segmentation (avec une attention particulière pour la détection de parole superposée), de transcription et d'extraction d'information (entités nommées). Le corpus contient une quarantaine d'heures de parole, divisées en 13,5h de radio et 29h de télévision et contient des émissions d'informations, de débats mais aussi des programmes de divertissement, recouvrant ainsi de nombreuses configurations communicationnelles. Sur les enregistrements de radio, la plupart consistent en des débats, favorisant ainsi l'interaction et la parole superposée, avec parfois des conditions acoustiques difficiles (ex : Un Temps de Pauchon).

REPERE

REPERE (Reconnaissance des PERsonne dans des Émissions télévisuelles) s'intéresse à la reconnaissance de personnes dans les émissions télévisées (Giraudel et al., 2012). La campagne a eu lieu entre 2011 et de 2014, a été financée par la DGA et 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.

1.2. Un rôle prépondérant des grands corpus du français

Comme expliqué précédemment, les corpus se situent maintenant à la base du développement de la majorité, voire de la totalité, des systèmes. Mais créer ces corpus a un coût non négligeable. Les campagnes précédemment décrites avaient également pour but,

au-delà des tâches d'évaluation, de constituer des ressources de qualité pour le développement des systèmes. Pour cette raison, la plupart des systèmes développés se sont donc basés sur les corpus décrits précédemment. (Brun et al., 2004), (Meignier & Merlin, 2010), (Scheffer et al., 2005) sont autant d'articles publiés sur des systèmes développés sur ces corpus.

ESTER 1 et 2, ETAPE et REPERE sont donc des ressources inévitables pour tous les chercheurs ou compagnies souhaitant développer des systèmes de traitement automatique du français oral et constituent, réunis, un ensemble qu'on pourrait appeler « les grands corpus du français ».

Il est intéressant de souligner néanmoins, que par rapport aux données disponibles sur l'écrit, les corpus oraux restent relativement petits. « Même les « grands » corpus oraux restent désespérément petits par rapport à leurs homologues écrits. » écrivait Veronis (2004). En 2004, le moteur Google indexait 8 milliards d'occurrences pour le français, en comparaison avec le corpus GARS-DELIC qui contient environ 3,5 millions d'occurrences. Ces différences de tailles sont toujours d'actualité, le corpus REPERE par exemple contient environ 66 452 mots ; ces grands corpus du français restent donc de taille modeste.

Ce constat peut cependant se nuancer lorsque l'on s'intéresse aux corpus utilisés dans l'industrie, notamment avec le développement de nouveaux assistants personnels incarnés tels que Google Home, dont le corpus d'apprentissage contient 18000h de parole ou le corpus de Baidu contenant 12000h de parole. Mais ces chiffres se rapportent souvent à de l'anglais et ces corpus ne sont pas accessibles.

2. Corpus et performances des systèmes d'IA

2.1. Le cas médiatique de la reconnaissance faciale

L'utilisation grandissante de ces masses de données a également posé la question de l'éthique et des biais existants dans les corpus et les systèmes ainsi constitués. Dans son étude intitulée GenderShades (Buolamwini & Gebru, 2018), Buolamwini a exploré les performances en terme de genre et de pigmentation de la peau de différents systèmes de classification en genre basés sur la reconnaissance faciale. Les systèmes évalués (Microsoft, IBM et Face++) réalisaient des performances de l'ordre de 87.9 % à 93.7 % de

classification sur le corpus Pilot Parliament Benchmark, mais en y regardant de plus près, ces performances étaient inégales selon les sous-groupes étudiés. Les différences en terme de taux d'erreur étaient de l'ordre de 8,1 % à 20,6 % entre visages d'hommes et visages de femmes et de l'ordre de 11,8 % à 19,2 % entre les visages à la peau claire et ceux à la peau foncée. En croisant les résultats en terme de genre et de couleur de peau, les différences de performances vont jusqu'à 33,8 % pour le système Face++ entre les hommes blancs et les femmes de couleur. 93,6 % des visages mal reconnus par le système de Microsoft étaient ceux de personnes à la peau noire et 95,9 % des erreurs de Face++ concernait des visages de femmes. Ces écarts de performance s'expliquent très vraisemblablement par une représentation inégale des genres et des couleurs de peau dans les corpus d'apprentissage, créant ainsi des systèmes aux performances discriminantes. Cette discrimination algorithmique se doit d'être étudiée pour comprendre dans quelle mesure les systèmes reproduisent des biais contenus dans leur corpus d'apprentissage et ainsi dégager des perspectives pour améliorer les systèmes.

2.2. Les questions soulevées en TAL

Le cas de la reconnaissance faciale et l'étude de Buolamwini ont permis de questionner l'idée selon laquelle les algorithmes étaient « objectifs ». Il est facile de justifier l'utilisation de systèmes informatiques pour éviter les biais humains. Un système automatique n'est pas censé modifier son résultat en fonction du sexe de la personne, de son origine sociale ou raciale ou encore de son orientation sexuelle. Mais si on peut considérer qu'un système automatique sera moins influencé par des éléments contextuels, la réalité est plus complexe que cela et tenter de réduire le nombre d'interventions humaines n'empêche pas forcément la présence de biais. De nombreuses études comme celles de (Kilbertus et al., 2017) et (Hardt, Price, & Srebro, 2016) ont d'ailleurs essayé de réduire au maximum la présence des biais dans les algorithmes, mais cela est rarement fait dans le cadre des systèmes de TAL.

Les systèmes automatiques cherchent à extraire des motifs récurrents dans de grands ensembles de données et sont donc complètement dépendants des données sur lesquelles ils effectuent leur apprentissage. Le *chat-bot* Tay a été un parfait exemple de ce qu'il pouvait se passer lorsque l'on fournissait des données biaisées à un système d'apprentissage automatique : le 23 mars 2016, Microsoft met en ligne sur Twitter un *chat-*

bot, Tay, censé incarner une adolescente. Le robot conversationnel de Microsoft visait comme public les Américains de 18 à 24 ans. L'intelligence artificielle était basée sur un fonctionnement simple : le vocabulaire et le raisonnement du chat-bot était censé s'étoffer au fur et à mesure que les utilisateurs interagissaient avec lui. Tay apprenait donc de ses interactions, des données qui lui étaient fournies. Mais la communauté 4Chan a décidé de pousser le système dans ses retranchements et par force de discours biaisés a réussi à faire tenir des propos misanthropes, sexistes, racistes et xénophobes au *chat-bot*[4].

Un exemple un peu moins médiatisé est celui des *word embeddings* ou plongements de mots. Les plongements de mots se basent sur l'adage de Firth (1957) selon lequel « *you shall know a word by the company it keeps* ». La similarité syntaxique et sémantique des mots est capturée en étudiant la mesure dans laquelle ils se produisent dans des contextes similaires. Ainsi, les mots sont représentés dans un espace vectoriel et des mots au comportement similaire se retrouveront proches dans cet espace. Cette représentation vectorielle des mots permet d'utiliser des opérations algébriques pour inférer sur leur sens : par exemple, si l'on soustrait au vecteur représentant « roi » celui de « homme » et qu'on y ajoute le vecteur pour « femme » on s'attend à être près (en terme de distance euclidienne) du vecteur correspondant à « reine » (Mikolov et al., 2013).

Mais des études ont récemment mis en avant le fait que ces plongements de mots capturaient également des stéréotypes sexistes ou racistes. Les travaux de Bolukbasi et al. (2016) ont mis en avant l'existence de ces stéréotypes dans les plongements de mots entraînés sur le corpus Google News et préviennent du biais que cela peut induire dans les systèmes qui utiliseront de telles données. Dans leur article ils démontrent que les plongements de mots encodent des analogies du type *(man:computer programmer :: woman:homemaker)* qui pourrait se traduire comme « le programmeur est à l'homme ce que la ménagère est à la femme». Caliskan et al. (2017) ont trouvé des résultats similaires et montré l'existence de biais en terme de genre et de race.

Ce genre d'études permet de nous rappeler le fait que les données ne sont pas plus brutes qu'elles ne sont neutres. Elles existent et sont récoltées dans le cadre d'une société construite par sa vision du monde et sa structure, et ces dernières se retrouvent dans les systèmes. Dans son étude de l'évolution des plongements de mots au cours du temps, Garg

[4] http://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html [consulté le 27/03/18]
et https://motherboard.vice.com/fr/article/vv3m3j/comment-lia-adolescente-de-microsoft-est-devenue-nazie-en-moins-de-24h [consulté le 27/03/18]

et al. (2018) ont mis en avant l'évolution des associations avant et après les mouvements féministes montrant ainsi que les événements sociétaux se retrouvent effectivement dans les données textuelles de l'époque. Ainsi par exemple, si l'on s'intéresse aux adjectifs les plus associés aux femmes pour chaque décade entre 1910 et 1990, on observe un changement entre les années 1960-1970, point culminant des mouvements féministes aux États-Unis. En effet après 1960, les adjectifs associés à l'intelligence fortement biaisés en faveur des hommes, le sont moins : le nombre d'associations de ces adjectifs avec des termes représentant les femmes augmente, ce qui peut être interprété comme une conséquence des mouvements visant à mettre fin aux obstacles sociaux et légaux empêchant l'accès aux femmes à la sphère universitaire et professionnelle.

Il est donc possible que les systèmes automatiques contiennent des biais. Partant de ce constat il nous semble plus que nécessaire d'étudier ces biais à l'heure ou l'intelligence artificielle infiltre notre quotidien. Et bien que les études citées portaient sur l'anglais, il semble raisonnable de faire l'hypothèse que des résultats similaires pourraient être observés concernant le français.

2.3. Et les études en parole ?

Les études présentées précédemment témoignent donc de l'existence de biais dans les performances des systèmes automatiques, principalement dus aux caractéristique des données d'apprentissage. Si les outils d'IA, tels que la reconnaissance faciale et les plongements de mots, capturent les spécificités culturelles contenues dans les données, alors nous pouvons nous demander ce qu'il est en des systèmes de traitement automatique de la parole ? Les corpus de l'oral contiennent-ils également des biais ? Le corpus FABIOLE (Ajili et al., 2016) qui a été constitué pour des tâches de reconnaissance du locuteur dans un cadre criminalistique ne contient que des locuteurs masculins et cela pour la raison suivante : *« female are not selected because we does not find 30 women who have enough excerpts with the desired characteristics »*, ces critères étant une durée de minimum 30s de parole dans un contexte de radio ou de télévision. Au regard de ce constat, on peut s'attendre à retrouver des spécificités dans nos corpus oraux et on peut donc se poser la question de la mesure dans laquelle celles-ci se retrouvent dans les systèmes ?

Adda-Decker et Lamel (2005) ont montré que les systèmes de transcription automatique présentaient de meilleures performances sur la parole des femmes que sur celle des hommes. Ce constat peut sembler surprenant dans le sens où les femmes étant sous-représentées par rapport aux hommes dans les corpus servant à développer les systèmes, on aurait pu s'attendre à ce que le système fonctionne mieux sur le type de données qu'il a rencontré majoritairement, à savoir de la parole d'homme. Selon les autrices, ces différences de WER[5] sont principalement dues à un nombre de délétions plus important chez les hommes. Elles avancent l'hypothèse que les hommes parlent plus vite ou ne produisent pas des prononciations complètes des mots.

D'autres études ont trouvé des résultats similaires : Goldwater et al. (2010) ont comparé les systèmes SRI et Cambridge et là où les femmes avaient un WER de respectivement 16,7 % et 15,3 %, celui des hommes s'élevait à respectivement 19,8 % et 18 ,1 %. Sawalha & Abu Shariah (2013) ont également trouvé que dans leur système de reconnaissance automatique de l'arabe, les performances étaient meilleures chez les femmes, mais également chez les locuteurs de moins de 30 ans. Le genre ne serait donc pas le seul biais observé en parole, mais l'âge pourrait également jouer sur les performances.

D'autres études ont cependant trouvé des résultats contraires : Rodger & Pendharkar (2004) ont montré que le système d'ASR contenu dans une application de suivi médical reconnaissait moins bien la parole des femmes, ce qui est corroboré par le docteur Syed Ali dans un article[6] de l'American Roentgen Ray Society. Cela semble également être le cas pour les GPS, qui reconnaîtrait moins bien la parole des femmes, comme l'explique cet article[7] paru dans la rubrique technologique du Time. Dans son étude Tatman (2017) a étudié le système de génération de sous-titres sur la plateforme YouTube et a démontré qu'il semblait exister des biais en terme de genre et de dialecte, avec des performances moins bonnes pour les femmes et pour l'anglais d'Ecosse. En revanche dans son étude avec Kasten (Tatman & Kasten, 2017), où ces biais ont été étudiés plus avant dans le système de YouTube mais également dans le système BingSpeech de Microsoft, le genre ne semblait plus être significatif dans les performances. En revanche, l'appartenance

[5]Le WER (Word Error Rate) ou taux d'erreur mot est une métrique permettant l'évaluation des systèmes de reconnaissance de la parole. Il est calculé comme la somme des suppressions, ajouts et substitutions divisée par le nombre de mot de la référence.
[6]http://cf.arrs.org/Pressroom/Archives/info.cfm?prID=202 [consulté le 22/05/18]
[7]http://techland.time.com/2011/06/01/its-not-you-its-it-voice-recognition-doesnt-recognize-women/ [consulté le 23/05/18]

ethnique semblait être un facteur important. Les WER étant significativement plus bas pour les locuteurs blancs.

Les variations de performances existantes dans les systèmes de traitement automatique de la parole sont rarement étudiés bien qu'ils semblent exister à différents niveaux (genre, âge, dialecte, appartenance ethnique) dans les systèmes de reconnaissance automatique de la parole. Mais ces systèmes ne sont pas les seuls contenant des biais, en effet comme expliqué par Jean-François Bonastre lors de l'Atelier Sciences et Voix du 08/03/18 , les performances des systèmes de reconnaissance automatique du locuteur sont aussi variables en fonction du genre : « On perd beaucoup d'informations sur les voix féminines, car les évaluations sont majoritairement sur des conversations téléphoniques (à bandes étroites), ce qui supprime les fréquences hautes et entraîne de plus mauvais résultats sur les voix de femmes ».

Les variations de performances en fonction du genre dans les systèmes de traitement automatique de la parole n'ont pas été souvent étudiées et les différents travaux sur le sujet ne semblent pas aboutir à un consensus. Nous proposons donc avec le présent travail d'explorer dans quelle mesure le genre a un impact sur les performances des systèmes de reconnaissance automatique de la parole en français. Dans notre approche, et pour réduire la complexité d'une telle étude nous ne nous concentrerons que sur un type de variation, celle du genre, sur un type de discours, la parole médiatique et sur une tâche, celle de reconnaissance automatique de la parole.

3. Les corpus médiatiques

Les grands corpus du français précédemment décrits que sont ESTER1, ESTER2, ETAPE et REPERE sont tous des corpus médiatiques. Or la parole médiatique est contrainte par un ensemble de normes qui en fait un type de discours particulier. Dans le but d'étudier les biais présents dans les grands corpus du français, il nous est nécessaire de caractériser en amont la parole médiatique et les biais qu'elle encode. Nous avons décidé de nous intéresser à la représentation des genres dans les médias et c'est ce que nous nous proposons d'étudier dans cette partie.

Comme l'avait déjà montré l'étude *Image, rôle et condition sociale de la femme dans les médias*[8] publiée en 1977 par l'Unesco, les femmes sont sous-représentées à la radio. Bien entendu, les répartitions ont évolué depuis, mais le constat reste le même : la

[8]http://unesdoc.unesco.org/images/0013/001343/134357fo.pdf

représentation des genres reste inégale dans le paysage médiatique. Depuis 3 ans, les chaînes de télévision et de radio sont priées de remettre au CSA (Conseil Supérieur de l'Audiovisuel) un rapport sur la représentation des femmes et des hommes dans leurs programmes. La synthèse[9] réalisée sur l'année 2017 mettait en avant le fait que les femmes représentaient 40 % de la présence à l'antenne. Dans le cadre de la télévision uniquement, il a été observé que les femmes étaient beaucoup moins présentes sur les heures à forte audience (29% de femmes sur la tranche 18h-20h) que sur le temps d'antenne général (42%). De même, les expertes sont beaucoup moins présentes sur les plateaux, avec 35 % de femmes pour 65 % d'hommes. Ces constats ont motivé la création de l'annuaire électronique Les Expertes[10], qui a pour but de réunir les contacts de femmes expertes dans une vaste palette de domaines, pour que l'absence de femmes compétentes ne soient plus une excuse à leur sous-représentation dans les médias. Dans la même mouvance l'association Prenons La Une[11], qui réunit des femmes journalistes, milite pour une représentation plus juste des femmes dans la médias et l'égalité professionnelle dans les rédactions.

Les femmes sont donc moins présentes que les hommes dans les médias et ces tendances ne semblent pas être typiquement françaises. Le Projet Mondial du Monitorat des Médias (GMMP) qui a également étudié la représentation des hommes et des femmes dans les médias à travers le monde arrive à des résultats similaires[12]. En 2015, 41 % des émissions de radios européennes étaient présentées par des femmes, ce qui constitue une représentation stable depuis les années 2000. Cette proportion est plus importante à la télévision où 57 % des présentatrices sont des femmes. D'après leurs résultats la répartition des genres dans la fonction de présentation semble relativement paritaire, (49 % de présentatrices et 51 % de présentateurs). Lorsque l'on s'intéresse aux reporters, l'écart se creuse : en 2015, les femmes ne présentent que 38 % des reportages à la télévision et 41 % à la radio.

En conclusion, on observe une sous-représentation globale des femmes dans les médias, avec une grande disparité en fonction des rôles médiatiques. Partant de ce constat, il semble donc logique de supposer que les femmes sont également sous-représentées dans

[9]http://www.csa.fr/Etudes-et-publications/Les-autres-rapports/Rapport-relatif-a-la-representation-des-femmes-dans-les-programmes-des-services-de-television-et-de-radio-Exercice-2016 [consulté le 23/05/18]
[10] https://expertes.fr/ [consulté le 22/05/18]
[11]https://prenons-la-une.tumblr.com/ [consulté le 22/05/18]
[12]http://cdn.agilitycms.com/who-makes-the-news/Imported/reports_2015/global/gmmp_global_report_en.pdf [consulté le 23/05/18]

les corpus médiatiques qui sont à la base du développement des systèmes de traitement automatique de la parole. Nous avons donc choisi, avec le présent travail, de vérifier si les femmes étaient effectivement sous-représentées dans les grands corpus médiatiques du français et d'en étudier les répercussions sur les performances des systèmes.

Pour vérifier cette hypothèse, nous avons choisi de répondre aux questions suivantes :

- les femmes sont-elles sous représentées dans les grands corpus du français ?

- quelle est la place des femmes dans les différents rôles médiatiques ?

- puisque les systèmes de traitement automatiques de la parole sont basés sur des corpus médiatiques, quel est l'impact de la place des femmes sur les performances des systèmes ?

La première partie de ce mémoire sera consacrée à la méthodologie suivie pendant la réalisation de ce travail et une seconde partie en présentera les résultats.

Partie 2 : Méthodologie du travail de mémoire

Chapitre 1. Les ressources

1. Les données

Dans le cadre de notre travail nous allons étudier les grands corpus du français que sont ESTER1, ESTER2, ETAPE et REPERE. Chaque corpus est constitué de fichiers sons qui ne seront pas utilisés dans ce travail et de leur transcriptions en fichier .trs. Les transcriptions ont été faites à l'aide du logiciel Transcriber (Barras et al., 1998) et un exemple de fichier .trs est consultable dans l'Annexe 1. Les fichiers .trs sont encodés en XML et respectent les normes prescrites dans le guide d'annotation[13]. Toutes les métadonnées concernant les locuteurs sont regroupées entre des balises <Speakers> </Speakers> au début des fichiers et chaque locuteur se voit attribuer un identifiant local, qui sert de référence par la suite. Les tours de parole seront définis comme les segments se trouvant entre des balises <Turn> </Turn>.

Au total nous travaillons sur les quantités de données suivantes :

Tableau 1. Récapitulatif des données

	Durée	Nombre de tours de parole	Nombre de fichiers
ESTER1	95h50min	13269	194
ESTER2	12h33min	2946	46
ETAPE	34h24min	16616	74
REPERE	57h48min	27639	370

2. Structure de la base de données

Dans le but de faciliter l'exploration des données disponibles dans les grands corpus du français, nous avons fait le choix de constituer une base de données. Cette base s'organise autour de 5 tables nommées respectivement Speaker, Turn, Episode, Show et Corpus. Un diagramme UML représentant son organisation est présenté dans la Figure 1.

[13]http://trans.sourceforge.net/en/transguidFR.php

La table Speaker contient l'ensemble des locuteurs, représentés par un identifiant unique, un nom, un genre et une propriété nommée « native » censée représenté le fait que la langue parlée, en l'occurrence le français, soit la langue maternelle du locuteur ou non.

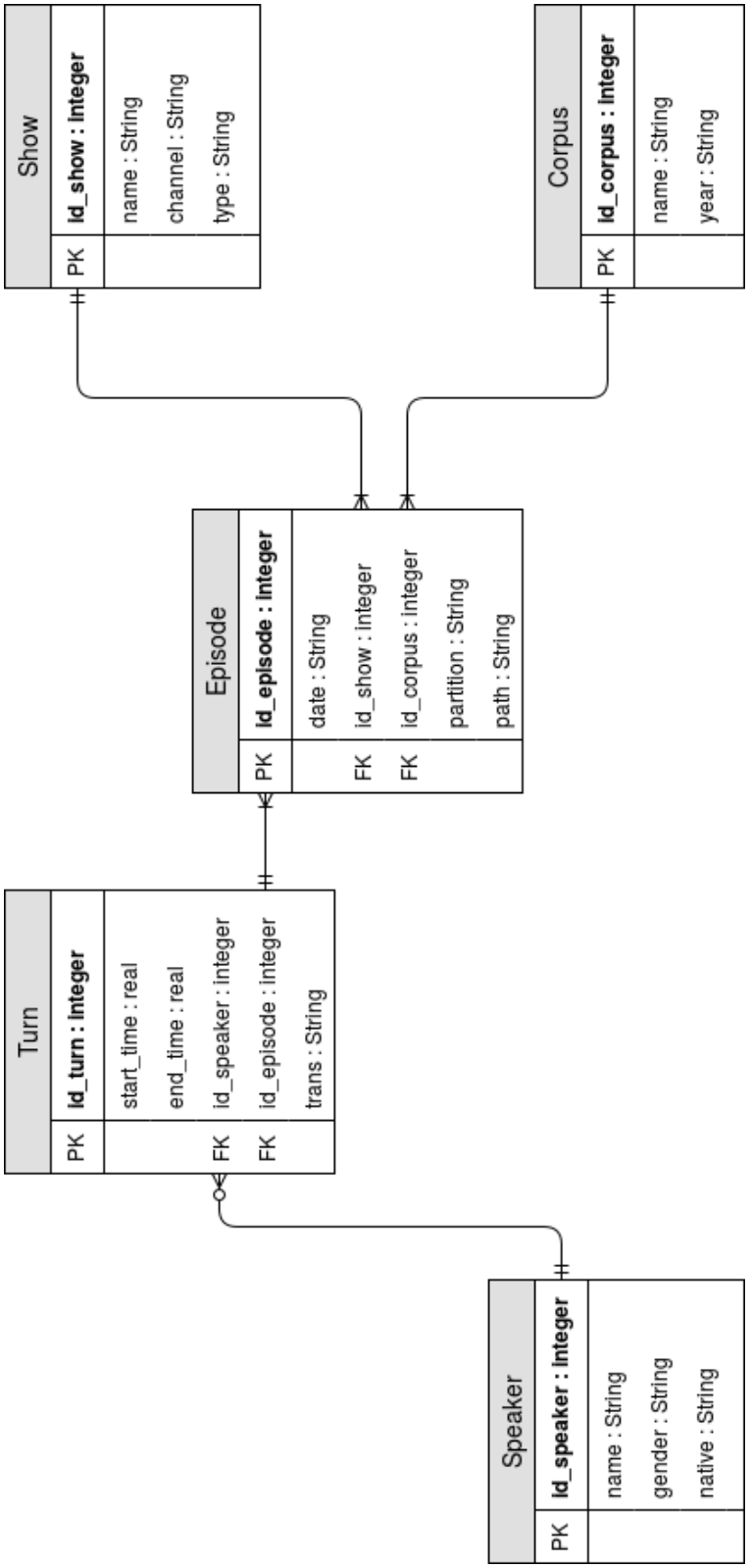
La table Turn contient l'ensemble des tours de parole, identifiés de manière unique par un numéro, et contenant le temps de début et de fin du tour, l'identifiant du locuteur qui l'a produit ainsi que l'identifiant de l'épisode dans lequel il apparaît. La transcription du tour est également un attribut de chaque instance de la table.

La table Episode contient la liste des différentes instances d'une émission. Chaque épisode a un identifiant unique, une date, l'identifiant de l'émission dont il est une occurrence, l'identifiant du corpus dans lequel il apparaît, ainsi que la partition (train, dev, test) et le chemin absolu dans lequel le fichier se trouvait. Dans l'optique d'une diffusion de cette base, ce dernier attribut pourrait être remplacé par le nom du fichier uniquement.

La table Show contient la liste des différentes émissions présentes dans le corpus, leur titre, la chaîne sur laquelle l'émission est diffusée ainsi qu'un attribut « type », représentant la catégorie de l'émission. Cet attribut est une simple chaîne de caractère et peut donc prendre n'importe quelle valeur, car celle-ci n'était pas directement présente dans les données. Une liste de valeur prédéfinies aurait aussi pu être considérée, mais ce n'est pas l'option que nous avons choisi de retenir.

Enfin la table Corpus contient les 4 corpus du français présentés précédemment ESTER1, ESTER2, ETAPE et REPERE, chacun représenté par un identifiant unique, leur nom, ainsi que leur date de publication.

Figure 1. Diagramme UML de la base de données



3. Remplissage de la base de données

3.1. Méthode générale

La base de données a été créée manuellement. Une version vide de la base est disponible sur GitHub[14]. Les tables Show et Corpus ont également été créées manuellement, car les informations n'étaient pas extractibles directement des fichiers de transcription. En revanche, les tables Turn, Speaker et Episode ont été remplies de manière automatique à l'aide d'un script[15] Python à partir des fichiers de transcription : `extract_data.py`, contenu dans le projet `trs_parsing`.

Le principe du script était de parcourir les différents dossiers contenant les fichiers de transcriptions et pour chaque fichier (chaque fichier représentant un épisode) et de parser le XML pour en extraire les informations sur le locuteur et sur les tours de parole. Les locuteurs se voyant attribuer un identifiant local dans chaque fichier, le script consulte également la base de données à chaque nouveau fichier, pour ne pas dédoubler les locuteurs.

Une partie des fichiers de REPERE contenaient des caractères empêchant le traitement (notamment le '&'), ceux-ci ont donc été enlevés des transcriptions pour permettre l'extraction automatique des données et ont ensuite été traités de manière séparée. La liste des fichiers problématiques se trouve dans le dossier `not-well-formed-trs`.

3.2. Choix de normalisation

En plus de ces fichiers mal-formés, nous avons rencontré deux problèmes majeurs dans la constitution de la base de données : la gestion des noms et la gestion de la parole superposée.

Concernant la gestion des noms, les normes différant selon les campagnes, certains locuteurs se sont retrouvés dédoublés, comme par exemple Arnaud Ardouin qui existait comme « Arnaud Ardouin » et « Arnaud_ARDOUIN » dans notre base de données. Pour éviter de fausser nos analyses, les noms ont donc été normalisés sous la forme Prénom_NOM.

[14]https://github.com/mgarnerin/mathesis_genderbias
[15]L'ensemble des scripts dont il sera question par la suite sont également disponibles sur le GitHub

Certains locuteurs n'étant pas nommés, ceux-ci sont notés comme « inconnu » ou « unknown » dans les transcriptions[16]. Mais ce type de noms peut se retrouver dans différents fichiers et ne correspond pas forcément aux mêmes locuteurs inconnus. Nous avons donc normalisé les noms de la manière suivante : spk-id_episode-compteur_local. Pour un certain nombre de locuteur, le genre était également indisponible (valeur égale à « unknown » ou « NA »), dans un souci de normalisation, celui-ci a été codé comme « NA » pour tous les locuteurs pour lesquels l'information n'avait pas été renseignée.

Concernant la parole superposée, les normes d'annotation varient selon les campagnes : dans ESTER1, des nouveaux locuteurs étaient créés pour les tours de parole superposée, contenant le nom de chaque locuteur individuel. En revanche pour ESTER2, ETAPE et REPERE, où l'étude de la parole superposée était un des axes des campagnes, la parole superposée était encodée au niveau de l'attribut speaker des balises <Turn>, qui contenait les identifiants locaux de chaque locuteur impliqué. Pour éviter de décupler le nombre locuteurs du à des tours de parole superposée dans ESTER1, nous avons fait le choix de créer un identifiant unique (id_speaker = 0) correspondant au « multiloc » et dont le genre est indéfini (valeur égale à « NA »). On notera le fait que les tours de parole superposée n'ont pas été transcrits et ne sont donc pas pris en compte dans les performances des systèmes présentées par la suite.

L'attribut *native* a été conservé comme annoté dans les données, cependant, il correspond plus à un accent qu'au statut de la langue pour le locuteur. Nous ne l'avons donc pas considéré comme pertinent pour nos analyses.

[16]Certains locuteurs ne sont pas directement nommés, mais certaines informations sont néanmoins disponibles, comme uniquement le prénom, la profession, ou plus généralement un statut quelconque. Il est également possible que certains locuteurs aient été fusionnés s'ils avaient le même prénom ou fonction et que cette information ait été la seule disponible mais nous avons décidé que ce cas était marginal et ne l'avons donc pas traité.

Chapitre 2. Les analyses : méthodes et outils

1. Extraction des données : requêtes SQL

La constitution de la base de données présentée précédemment a facilité l'exploration des données et a permis l'extraction des statistiques présentées dans la suite de ce travail.

Pour extraire les données dans le but de réaliser nos analyses statistiques nous avons eu recours à des requêtes SQL. L'ensemble de ces requêtes est consultable dans l'Annexe 2. Ont été extraits : un fichier contenant le nombre de tours par locuteur (`turn_by_speaker.csv`) ainsi que leur genre, et un fichier contenant l'ensemble des tours de parole (`turns_all_info.csv`). Un autre fichier ne contenant que les tours de parole de débats télévisés a également été généré.

2. Préparation des données : scripts Python

Si le fichier `turn_by_speaker.csv` a été directement analysé sous R avec le script `speaker_analysis.R`, l'analyse des tours a nécessité un prétraitement, réalisé à l'aide de scripts Python. Le script `turn_count_length.py` a permis de calculer la longueur totale des interventions des locuteurs et d'attribuer les classes médiatiques (script R `class_analysis.R`). Enfin le script `turn_id_class.py` a permis de réinjecter les classes médiatiques dans les tours de paroles dont la longueur a ensuite été analysée à l'aide du script `turn_length_analysis.R`

3. Les modèles statistiques

L'ensemble des analyses statistiques que nous avons faites ont été réalisées sous R, avec l'utilisation du logiciel Rstudio[17]. L'ensemble des scripts R est disponible sur le GitHub. Nous verrons dans la suite de notre travail que pour analyser la distribution des

[17]https://www.rstudio.com/

tours de parole par locuteurs ainsi que la longueur de ces tours de parole en fonction du genre, nous allons avoir besoin des modèles statistiques présentés ci-dessous.

3.1. Modèles linéaires généralisés

Pour étudier la distribution du nombre de tours de parole par locuteurs, nous aurions pu nous utiliser un test de Student ou t-test. Le t-test est utilisé lorsque l'on souhaite vérifier l'existence d'une différence statistiquement significative entre les moyennes de deux conditions. L'idée est de voir si la variance entre les conditions est plus grande que les variances intra-conditions, si c'est le cas, on peut alors conclure qu'il existe une différence entre nos deux conditions. Cependant, si le t-test est souvent utilisé, du fait de sa facilité de compréhension, il nécessite certaines conditions pour être applicable : les données doivent suivre une loi normale. Or lorsque l'on regarde la distribution du nombre de tours de parole par locuteur on se rend compte que ça n'est absolument pas le cas et qu'on est plus proche d'une loi de Poisson que d'une gaussienne. Pour cette raison, nous allons utiliser des modèles linéaires généralisés[18]. La régression de Poisson fait partie de la famille des modèles linéaires généralisés et permet d'étudier le lien entre une variable dépendante et un ensemble de variables explicatives. Dans notre cas la variable dépendante sera le nombre de tours de parole par locuteur et la variable explicative sera le genre.

Un exemple de résultat pourrait être :

```
Coefficients:

                          Estimate      Std. Error  z value     Pr(>|z|)
(Intercept)           2.697859      0.004864       554.68     <2e-16 ***
spk_all$genderfemale   -0.684333    0.010352       -66.11            <2e-16 ***
```

Ce qui s'interprète comme suit : la p-valeur étant très faible, les distributions sont significativement différentes entre les hommes et les femmes concernant le nombre de tours de parole. Le coefficient permet de quantifier cette différence, on obtient ici un coefficient de -0,68, ce qui peut s'interpréter de la manière suivante : une locutrice aura en moyenne $e^{-0.68} \approx 0.5$ fois le nombre de tours d'un locuteur masculin.

[18]Une présentation approfondie des modèles est disponible à l'adresse suivante : http://maths.cnam.fr/IMG/pdf/Presentation_MODGEN_02_2007.pdf

3.2. Modèles de survie

Nous avons également souhaité étudier la longueur des tours de parole. Contrairement au nombre de tours de parole par locuteur, cas dans lequel chaque locuteur pouvait être considéré comme une instance indépendante des autres, l'étude de la longueur des tours de parole ne peut pas utiliser des tests considérant les variables comme étant indépendantes, car la longueur des tours est fortement liée à tous les tours de parole ayant été produits précédemment. Pour étudier ces durées, nous avons donc eu recours à des modèles de Cox. Les modèles de Cox sont des modèles de survie, une présentation en est faite dans l'article de Letué et al. (2018). Principalement utilisés en médecine, ils permettent de quantifier le risque qu'un événement survienne à un instant t, avec la possibilité d'intégrer des effets fixes ainsi que des effets aléatoires. Cette prise en compte des effets est particulièrement intéressante dans notre cas, car la longueur des tours de parole est le résultat de multiples facteurs : le genre du locuteur peut-être, mais également le type d'émission et/ou le locuteur lui-même.

Dans notre cas, l'événement t que nous cherchons à prédire est la fin du tour de parole. Le genre est considéré comme un effet fixe et le locuteur comme un effet aléatoire. Augmenter le risque revient donc à augmenter la probabilité que la fin du tour de parole survienne plus tôt. Le modèle que nous avons utilisé définit comme facteur fixe le genre du locuteur, le corpus (qui peut se voir comme une opposition de la radio, pour ESTER1 et ESTER2, à la télévision, pour ETAPE et REPERE) et nous avons inséré un effet aléatoire dû au locuteur.

Les résultats que nous pouvons interpréter dans la sortie du test sont surtout les coefficients et les p-valeurs. Comme sur l'ensemble des tests statistiques, la p-valeur nous indique la significativité du résultat. Le coefficient, quant à lui, nous permet de quantifier la liaison entre la variable dépendante observée et les effets fixes. Soit une variable X représentant la longueur d'un tour de parole étudiée selon un effet fixe de genre pouvant prendre la valeur F (femme) ou H (homme) , on pourrait obtenir une sortie de R comme suit :

```
         coef        se(coef)      se2           Chisq  DF     p
X$H   -0.04854              0.01936       0.01935       6.29     1    1.2e-02
```

que l'on interprétera comme ceci : par rapport à la condition F considérée comme référence, le risque relatif par rapport à un homme est $e^{-0.04854} \approx 0{,}95$. Or réduire le risque,

revient à augmenter la probabilité que l'événement « fin de tour de parole » arrive plus tard. On peut donc en conclure que les tours de parole des hommes sont plus longs que ceux des femmes.

4. Le système de reconnaissance automatique de la parole

Le système d'ASR que nous avons choisi a été développé par Elloumi et al. (2018) dans le cadre d'une tâche de prédiction de performance des systèmes de transcription de la parole. Les performances du système réalisé au LIG sont comparables à celles obtenues par d'autres systèmes sur les mêmes données de test. Nous pouvons donc considérer que ce système développé au LIG est à l'état de l'art. Il nous servira donc de système de base pour nos analyses sur les grands corpus du français. .

4.1. Corpus

Le système d'ASR a été développé sur un corpus comprenant différentes émissions extraites de plusieurs corpus : 111h extraites des corpus ESTER1 et ESTER2 , 37h extraites du corpus ETAPE, 54h du corpus REPERE et un sous-ensemble de 41 heures du corpus Quaero[19], constitué dans le cadre d'un « programme collaboratif d'innovation et de recherche industrielle sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues. » Pour plus détails se référer à Elloumi et al. (op).

Nous disposons du corpus d'apprentissage du système brièvement présenté ci-dessus, mais également de deux corpus de test, constitués respectivement de 89h52 et de 8h59 d'enregistrements. Les deux corpus de test que nous nommerons Corpus 1 et Corpus 2 contiennent chacun des types d'émissions différents : le Corpus 1 regroupe des émissions issues de France Inter (émission de nouvelles et Le Téléphone Sonne), de RTM, de France 3, de RFI ainsi que l'émission de LCP Pile et Face. Le Corpus 2 lui, contient des occurrences de TVME, Africa n°1, Ce Soir Ou Jamais, Culture Et Vous, Planète Showbiz, Arte News, La Fabuleuse Histoire, Un Temps de Pauchon et La Place du Village. Les émissions contenues dans le Corpus 2 sont plus compliquées pour les systèmes de reconnaissance automatique de la parole, soit à cause de la présence importante de parole

[19]http://www.quaero.org

spontanée et/ou superposée, soit par des conditions acoustiques difficiles (comme pour Un Temps de Pauchon).

Les deux corpus sont composés comme suit :

Tableau 2. Récapitulatif des données de test du système d'ASR du LIG

	Durée	Nombre de tours de parole	Nombre de fichiers
Corpus 1	89h53min	75031	118
Corpus 2	9h01min	6837	245

Les performances du système sur chacun des corpus nous ont été fournies dans des fichiers CSV contenant chaque tours de parole ainsi que le WER associé.

4.2. Métrique d'évaluation : le WER

Pour évaluer les performances du système, nous utiliserons la métrique du WER. Le WER (*Word Error Rate*) ou taux d'erreur mot est une métrique permettant l'évaluation des systèmes de reconnaissance de la parole. Il est calculé comme la somme des suppressions, ajouts et substitutions divisée par le nombre de mot de la référence.

Partie 3 - Résultats

Chapitre 3. La place des femmes et des hommes dans les corpus médiatiques

Comme expliqué précédemment, les systèmes de traitement automatique de la parole sont principalement basés sur les grands corpus du français que sont ESTER1, ESTER2, ETAPE et REPERE. Or ces quatre corpus sont des corpus médiatiques, et d'après un ensemble d'études, notamment la synthèse fournie par le CSA sur la présence des femmes dans les médias, nous savons que la représentation des hommes et des femmes dans les médias est inégale. Nous pouvons donc supposer que les grands corpus du français ne sont pas équilibrés en terme de genre. C'est ce que nous allons chercher à explorer dans ce chapitre.

1. Comment quantifier la place d'un locuteur ?

Nous avons décidé de nous intéresser à la place des femmes dans les corpus médiatiques. Mais pour ce faire, il est nécessaire de savoir ce que nous entendons par place et comment nous souhaitons la mesurer. En effet, si le nombre de locuteurs et locutrices nous semble rester un indicateur pertinent, il nous semblait aussi important d'étudier la nature des interventions de ces locuteurs. Pour ce faire, nous nous sommes intéressée à deux métriques : le nombre de tours de parole par locuteur, et la longueur des tours de parole.

1.1. La notion de tour de parole

Avant d'étudier la distribution et la longueur des tours de parole, il convient de les définir. On considérera ici les tours de parole comme des unités conversationnelles, pendant lesquels le locuteur peut s'exprimer sans interruption. Comme l'ont écrit Sacks et al. (1974), l'allocation des tours de parole est le résultat « d'une économie de la conversation » existante à l'intérieur d'une organisation sociale. Étudier la distribution des tours de parole nous permet donc d'observer quelle est la place de chaque locuteur dans cette économie. Mais la définition d'un tour de parole est fortement dépendante du type d'interaction. Certaines règles implicites du type « une seule personne parle à la fois »

semblent cependant représenter des constantes (une interruption marquera donc la fin d'un tour de parole). Dans le cadre de la radio par exemple, il n'y a jamais de vrai silence, s'il y a un blanc un peu trop long, le présentateur reprendra immédiatement la main. Il en va de même dans les conversations téléphoniques où la voix constitue le seul canal communicationnel. Ces contextes interactionnels amènent donc à une définition plus stricte des tours de parole. Dans notre travail nous avons utilisé les tours de parole annotés dans les fichiers .trs, comme expliqué dans la partie Méthodologie.

1.2. Dis-moi comment tu parles, je te dirai qui tu es : étude des différentes classes de locuteurs médiatiques.

Si l'on suit l'idée selon laquelle la distribution des tours de parole est une économie, alors on peut supposer qu'il existe différents types d'agents prenant par à cette économie. Intuitivement si l'on cherche à quantifier la place d'un locuteur, on peut se poser les questions suivantes : parle-t-il souvent ? on regardera alors le nombre de tours de parole des locuteurs, et : parle-t-il longtemps ? on s'intéressera ici à la longueur des tours. A partir de ces deux questions de base, on peut potentiellement définir 4 classes de locuteurs : ceux qui parlent peu et peu longtemps (classe 1), ceux qui parlent souvent mais peu longtemps (classe 2), ceux qui parle peu mais longtemps (classe 3) et enfin les locuteurs qui parlent souvent et longtemps (classe 4). La Figure 3 ci-dessous représente l'ensemble des locuteurs de notre corpus en fonction de ces deux paramètres.

On observe un grand nombre de locuteurs qui parlent peu et peu longtemps (classe 1), assez peu de locuteurs qui parlent peu longtemps mais souvent (classe 2) et enfin une distribution assez éparpillée au-delà d'une certaine durée et d'un certain nombre de tours de parole. Nous avons choisi de fixer les limites des classes comme suit : un locuteur parle souvent à partir du moment où il parle plus de 75 fois dans l'ensemble du corpus. Un locuteur parlera donc peu s'il se trouve en dessous de ce seuil. Un locuteur parle longtemps à partir du moment où il cumule 10 min ou plus de parole sur l'ensemble de ses interventions. Ces deux seuils sont représentés par les droites présentes sur la Figure 2.

Ces seuils sont arbitraires, ils ont été choisis visuellement pour séparer les masses, dans le but de mettre en lumière des comportements différents. L'affinage de ces seuils pourrait constituer une piste d'amélioration de ce travail.

On observe donc que les distributions parmi les classes sont loin d'être homogènes (Tableau 3). La très nette majorité des locuteurs (94,50%) intervient de manière ponctuelle (peu et peu longtemps) alors qu'une minorité intervient sur des périodes longues, avec un nombre de tours de parole élevé ou non (classe 3 et 4, respectivement 2,76 % et 2,22 %). Les locuteurs parlant de manière très courte mais avec un nombre de tours de parole important sont quant à eux largement minoritaires (0,52%).

Figure 2. Longueur totale des interventions par le nombre de tours de parole (Un point représentant un locuteur) . Par souci de lisibilité 2 locuteurs extrêmes ne sont pas représentés sur cette figure. Ils se trouvent aux coordonnées (14948;1807) et (14596;6135)

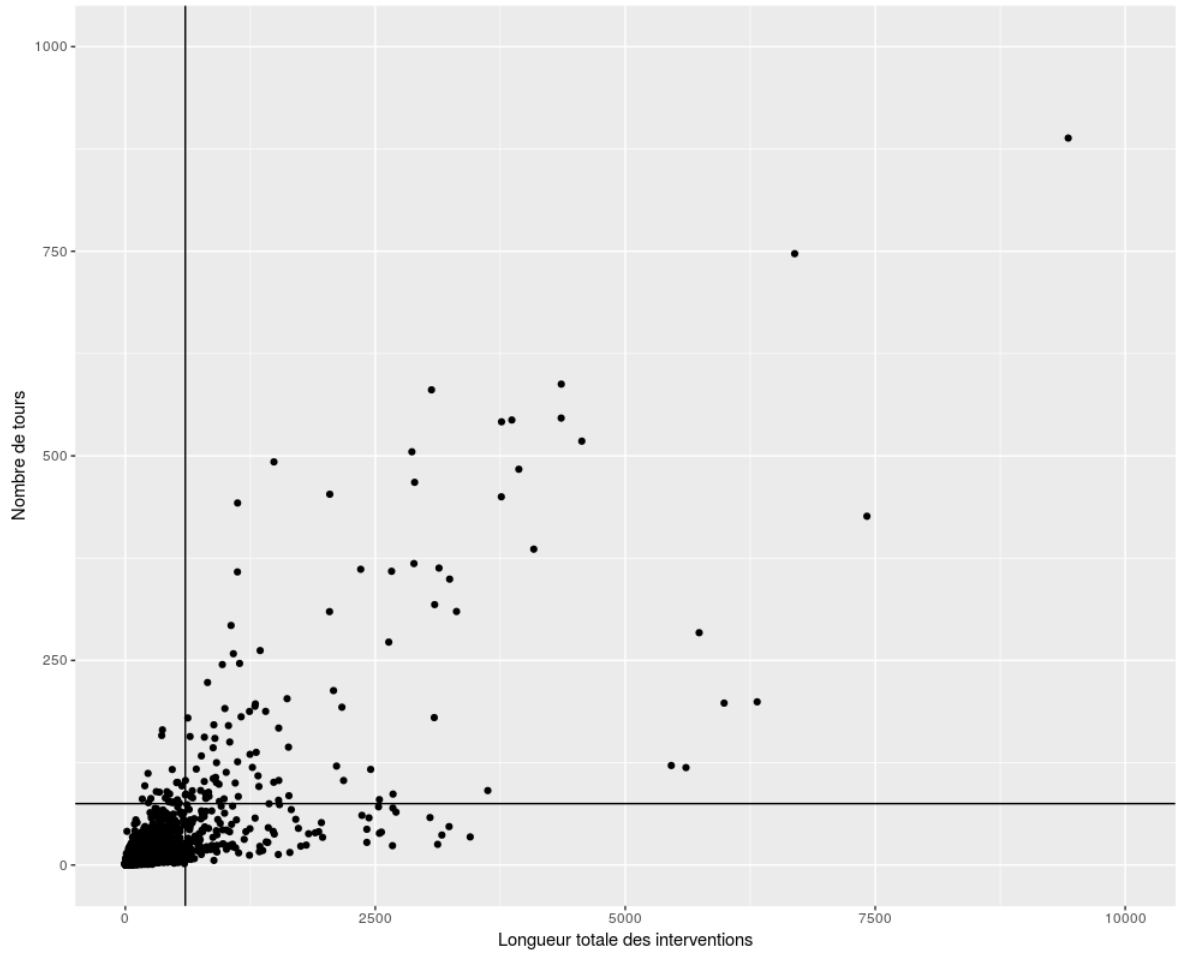


Tableau 3. Répartition des locuteurs par classe

	Classe 1	Classe 2	Classe 3	Classe 4	Total
Nombre de locuteurs	4344	24	127	102	4597
%	94,50 %	0,52 %	2,76 %	2,22 %	100

Ces 4 classes peuvent être interprétées comme une approximation des différents rôles médiatiques possibles. Un locuteur de la classe 1, intervenant peu et sur une durée très courte sera par exemple un auditeur, ou une personne interviewée dans la rue. En revanche une personne appartenant à la classe 3, intervenant peu, mais sur des durées longues, sera plutôt un expert venue commenter un fait d'actualité, ou un auteur venu présenter un livre. Les présentateurs et chroniqueurs d'émissions, les personnes faisant ce qu'on pourrait appeler « la couleur » d'une émission, ainsi que les personnalités publiques ayant une forte présence médiatiques comme les politiques, se retrouveront eux dans la classe 4, correspondant aux locuteurs présents souvent et sur des durées relativement longues. La classe 2 contenant très peu de locuteurs est, elle, assez difficile à interpréter.

2. Une femme pour deux hommes

D'après les chiffres du CSA pour 2017 on se situait à 35 % de femmes pour 65 % d'hommes à l'antenne. Ce sont des proportions que l'on retrouve effectivement dans nos corpus lorsque l'on étudie la répartition des locuteurs (Tableaux 4.a et 4.b, ci-dessous). On remarque néanmoins que l'écart est maximal pour le corpus ESTER2 qui contient des émissions de radio africaines et l'on pourrait supposer que cet écart se justifie par une différence culturelle. En effet, d'après les chiffres du GMMP datant de 2015, les femmes étaient en général moins présentes dans le paysage médiatique en Afrique (22%) qu'elles ne le sont en Europe (25%). On peut également noter que les corpus contenant en partie ou uniquement des émissions de télévision (ETAPE & REPERE) contiennent moins de femmes qu'ESTER1 ne contenant que la radio. De manière semblable aux résultats présentés par le GMMP, on observe que les femmes restent moins présentes que les hommes que ce soit à la radio ou à la télévision.

Tableau 4.a Répartition en fréquence absolue des locuteurs en fonction du genre dans les grands corpus du français (ESTER1, ESTER2, ETAPE et REPERE)

Corpus	Femmes	Hommes	NA	Total
ESTER1	779	1247	26	**2052**
ESTER2	141	329	10	**480**
ETAPE	194	383	19	**596**
REPERE	519	981	3	**1503**
Total	**1599**	**2847**	**53**	**4499**

Tableau 4.b Répartition en pourcentage des locuteurs en fonction du genre dans les grands corpus du français (ESTER1, ESTER2, ETAPE et REPERE)

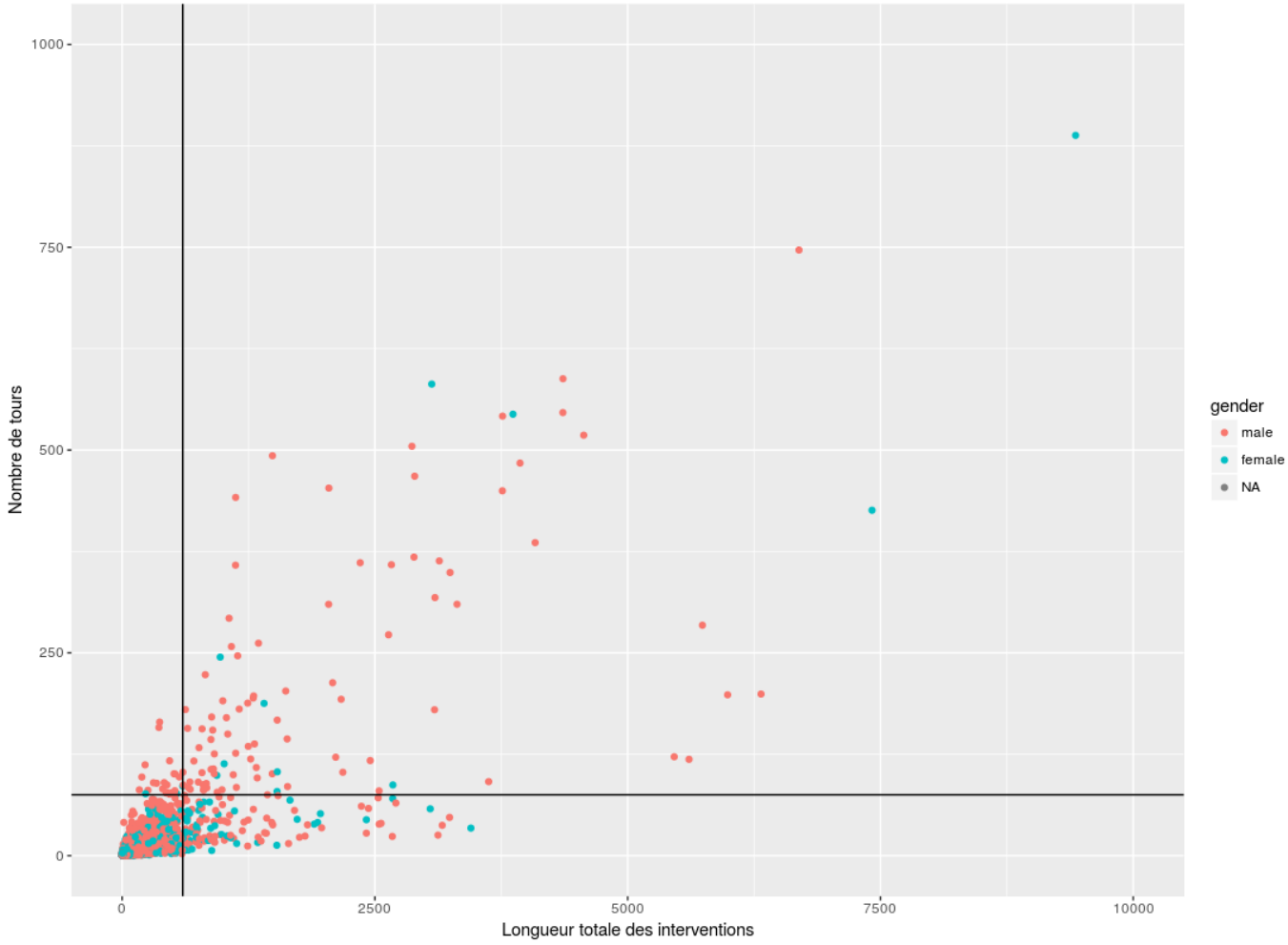
Corpus	Femmes	Hommes	NA
ESTER1	39,96 %	60,77 %	1,27 %
ESTER2	29,38 %	68,54 %	2,08 %
ETAPE	32,55 %	64,26 %	3,19 %
REPERE	34,53 %	65,30 %	0,19 %
Total	**35,54 %**	**63,28 %**	**1,18 %**

Gardons en tête cependant, que nos données étant un peu plus vieilles que celles présentées par le CSA ou le GMMP, nous sommes susceptible d'observer des disparités plus grandes, car le thème de la parité s'est beaucoup démocratisé durant ces dernières années. Mais si le phénomène sociétal est en évolution, ce sont toujours ces données qui sont utilisées dans le développement des systèmes, d'où la nécessité de les étudier.

D'une manière générale donc, on observe une répartition homme/femme de l'ordre de deux tiers/un tiers. Cependant lorsque l'on obtient ces statistiques, elles ne sont pas nécessairement faites en fonction du rôle ou du prototype médiatique dans lequel s'inscrivent les locuteurs. Comme expliqué précédemment, nous avons choisi de considérer 4 classes, comme quatre « comportements » ou rôles prototypiques médiatiques différents. Nous avons donc voulu savoir si la sous-représentation des femmes étaient un phénomène uniforme, auquel cas nous devrions observer cette même proportion dans l'ensemble de nos classes ou si les disparités variaient en fonction du rôle médiatique du locuteur.

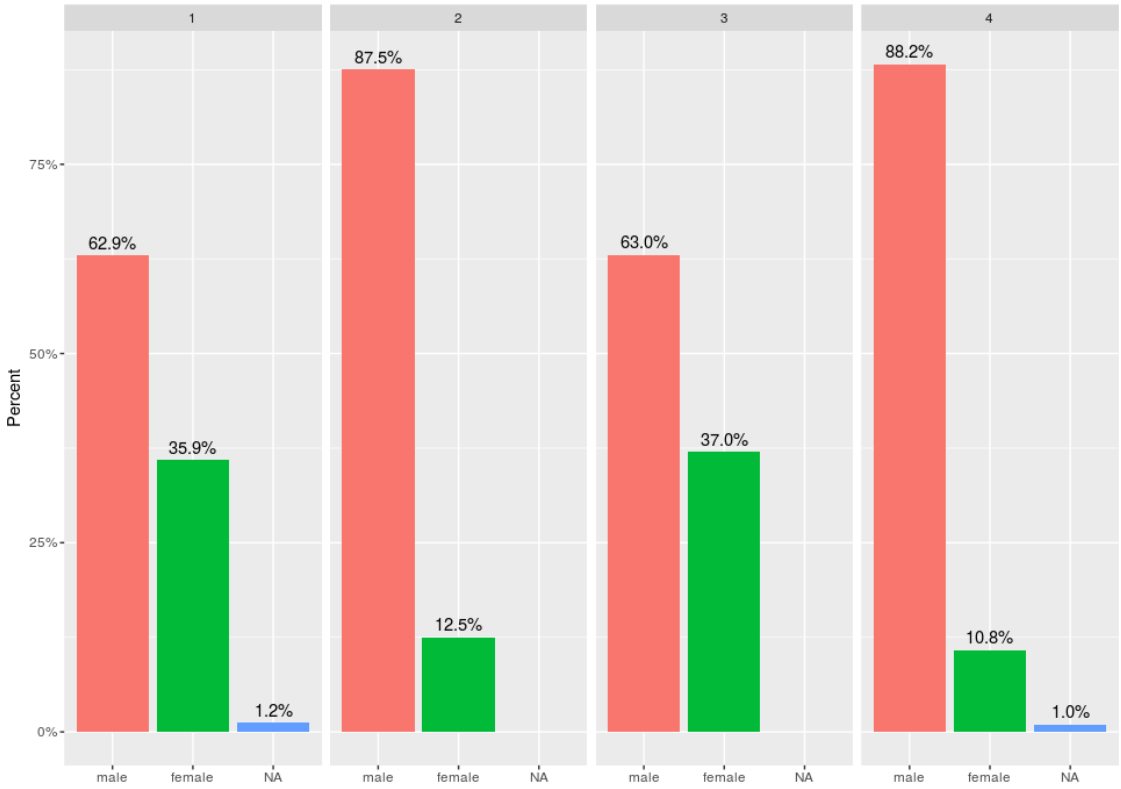
La Figure 3 reprend la Figure 2 précédemment décrite en introduisant cette fois-ci le genre des locuteurs. Or on observe que les femmes semblent être surtout présentes dans la classe 1, ce qui suggère qu'elles sont surtout des intervenantes ponctuelles, et un peu dans la classe 3, celle des locuteurs parlant beaucoup mais rarement.

Figure 3. Longueur totale des interventions en fonction du nombre de tours par genre.Par souci de lisibilité 2 locuteurs extrêmes ne sont pas représentés sur cette figure. Ils se trouvent aux coordonnées (14948;1807 ;male) et (14596;6135 ;NA). Cette dernière valeur représentant le locuteur « multiloc ».



La Figure 4, quant à elle, représente les proportions hommes/femmes dans chacune des quatre classes précédemment décrites. On se rend compte qu'on a bien une femme pour deux hommes dans le cadre des interventions ponctuelles et dans celui des invitées, mais dans la classe contenant les locuteurs intervenant souvent et longuement, la disparité homme femme est beaucoup plus importante se rapprochant plutôt du 1 femme pour 10 hommes. A l'instar du CSA qui observait, dans sa synthèse de 2017, qu'à la télévision les femmes sont encore moins présentes sur les heures à forte audience que sur le temps d'antenne global, on observe ici que les femmes sont également moins présentes dans les rôles médiatiques que l'on pourrait qualifier de principaux. Il semblerait également que les classes 2 et 4 aient des répartitions similaires, de mêmes que les classes 1 et 3.

Figure 4. Proportion d'hommes, de femmes par classe médiatique. Effectifs : classe 1 : 4344, classe 2 : 24, classe 3 : 127, classe 4 ; 102. (Les locuteurs NA représentent les locuteurs pour lesquels le genre n'avait pas été renseigné. Le 1 % présent dans la classe 4 est du au nombre important de tours de parole superposée, regroupé sous un même locuteur non-genrée)



A titre de rappel, les classes 2 et 4, représentent les locuteurs parlant respectivement peu mais souvent et beaucoup et souvent. Ces deux classes représentent donc les locuteurs ayant un grand nombre de tours de parole. Les femmes étant largement moins présentes dans ces classes, alors qu'on retrouve la répartition 65% d'hommes 35 % de femmes dans les classes 1 et 3, on peut donc conclure que d'une manière générale, les femmes semblent comptabiliser moins de tours de parole que les hommes.

3. Des femmes qui prennent la parole deux fois moins que les hommes

3.1. Analyse globale

Nous avons donc vu que les femmes étaient sous-représentées dans nos corpus, mais également que cette sous-représentation n'était pas homogène. Un autre axe d'approche pour évaluer les différences hommes-femmes est celle du nombre de tours de parole par locuteur et la faible proportion de femmes dans les classes représentant les locuteurs parlant souvent semble montrer qu'une différence existe effectivement à ce niveau.

Sur les 60 470 tours de parole de nos corpus, 69,91 % sont produits par des locuteurs masculins, 19,80 % par des femmes et 10,29 % par des locuteurs au genre non-renseigné (dont presque 99 % s'expliquent par des tours de parole superposée). Si on a, en moyenne, 1 femme pour 2 hommes en terme de locuteurs, il semblerait que lorsque l'on s'intéresse à la distribution des tours de parole, nous sommes plus proches du ratio 1 femme pour 5 hommes. (La répartition des tours de parole par genre et par corpus est présentée dans l'Annexe 3, de même que les histogrammes des distributions).

Lorsque l'on superpose les histogrammes de distribution du nombre de tours de parole par locuteur, nous pouvons observer que si les fréquences pour les femmes sont moins élevées, les deux distributions semblent néanmoins suivre des lois similaires. Le nombre de tours de parole par locuteur semble se distribuer selon une loi de Poisson, avec un nombre important d'individus intervenant 1 ou quelques fois et quelques locuteurs ayant un nombre de tours de parole important. Cette distribution vient confirmer les résultats que nous avions trouvés dans la partie précédente, avec un nombre important de locuteurs dans la classe 1 (interventions ponctuelles) et des proportions bien plus faibles de locuteurs plus présents (classes 2 et 4).

Nous avons souhaité quantifier ces différences de distribution et vérifier qu'elles n'étaient pas juste la conséquence de la sous-représentation des femmes dans nos corpus. Pour ce faire, nous avons utilisé des modèles linéaires généralisés (ces modèles sont expliqués dans la Partie 1 de ce travail). Nous avons donc testé notre hypothèse à l'aide du logiciel d'analyse statistique R et les résultats obtenus sont reportés dans l'Annexe 4. Le coefficient obtenu est égal à -0.68, ce qui signifie qu'une locutrice aura en moyenne $e^{-0.68} \approx$ 0,5 fois le nombre de tours d'un locuteur. Une femme aurait donc en moyenne, deux fois moins de tours de parole qu'un homme.

Cette analyse a été effectuée à l'échelle globale du corpus, mais il aurait été intéressant de la réitérer dans chaque classe médiatique, pour regarder si ces différences de nombre de tours moyen par locuteur étaient également observables dans un rôle donné.

3.2. Le cas du débat télévisé

Comme évoqué précédemment, les différences entre hommes et femmes pourraient s'expliquer par la répartition inégales des rôles médiatiques, les femmes seraient

globalement moins présentes dans les rôles qui parlent souvent et cela influencerait le nombre de tour moyen par locuteur. Si tel est le cas les différences que nous observons entre hommes et femmes ne seraient que le résultat d'une sous-représentation des femmes dans ces rôles. Le Tableau 5 présente les 10 locuteurs comptabilisant le plus de tours de parole et on observe effectivement que ceux-ci sont majoritairement des présentateurs, mais également majoritairement des hommes (8 hommes pour 2 femmes). Nous avons fait le choix d'étudier de manière plus approfondie le cas des débats télévisés, en posant l'hypothèse que, dans un débat, l'ensemble des participants incarnait plus ou moins le même rôle, à savoir « membre du débat ». En posant cette hypothèse sur l'uniformité des rôles, nous avons donc souhaité observer si des disparités hommes/femmes étaient toujours présentes.

Tableau 5. Présentation des 10 locuteurs cumulant le plus de tours de tours de parole

	Nom	Genre	Nombre de tours	Rôle
1.	Olivier Truchot	H	2170	Présentateur BFM Story
2.	Christophe Ruaults	H	1072	Présentateur Entre Les Lignes
3.	Arnaud Ardoin	H	1060	Présentateur Ca Vous Regarde
4.	Brigitte Boucher	F	888	Présentatrice LCP Politique Matin
5.	Jean-Pierre Gratien	H	866	Présentateur Pile Et Face
6.	Laurent Neumann	H	836	Invité régulier Entre Les Lignes
7.	Philippe Dufreigne	H	747	Présentateur Culture Et Vous
8.	Claude Weill	H	640	Invité régulier Entre Les Lignes
9.	Patricia Martin	F	581	Présentatrice Matinale France Inter
10.	Jérôme Garcin	H	546	Présentateur Le Masque Et La Plume

Pour étudier les débats, nous avons isolé les épisodes des émissions suivantes : Ça Vous Regarde, Entre les Lignes et Pile et Face. Les effectifs des locuteurs étant très inégaux, 57 femmes pour 252 hommes, nous avons effectué un tirage aléatoire de 57 locuteurs pour pouvoir comparer les nombres de tours de parole. Ce tirage aléatoire a été réitéré 10 fois pour s'assurer de la cohérence de nos résultats.

Les distributions et les résultats des tests statistiques sont tous présentés dans l'Annexe 5 et les coefficients obtenus pour chacun des test ont été rassemblés dans le Tableau 6 ci-dessous (chaque test étant significatif, les p-valeurs n'ont pas été reportées) :

Tableau 6. Récapitulatif des résultats de la régression de Poisson dans l'étude de la distribution des tours en fonction du genre dans les débats télévisés

Tirage	Coefficient	$e^{coefficient}$
1	-1.61175	0.20
2	-1.8468	0.16
3	-1.51398	0.22
4	-1.8288	0.16
5	-1.38195	0.25
6	-150054	0.22
7	-1.91509	0.15
8	-0.73835	0.48
9	-1.33134	0.26
10	-1.29441	0.27

A titre de rappel, les résultats s'interprètent de la manière suivante : pour le tirage un, les femmes ont en moyenne 0.20 fois le nombre de tours de parole des hommes, une femme parle donc 5 fois moins. On observe donc que d'une manière générale une femme parle 4 à 5 fois moins qu'un homme, alors même que les rôles sont censés être communs. Cette valeur est encore plus faible que celle obtenue sur l'ensemble du corpus où l'on avait pu observer qu'une femme parlait en moyenne deux fois moins qu'un homme

4. Des femmes qui parlent moins longtemps que les hommes ?

Après avoir étudié les disparités en terme de rôle et de nombre de tours par locuteur, nous avons choisi dans un troisième temps de nous intéresser à la longueur des tours de parole. Les longueurs moyennes ainsi que les écarts-types sont présentés dans le Tableau 7. Dans ESTER1, la durée des tours de parole semble être sensiblement la même entre les hommes et les femmes. Le fait que les tours de parole des locuteurs non genrés soient si brefs s'explique par le fait que ceux-ci sont majoritairement des tours de parole superposée et donc plus courts. En revanche, une différence notable existe entre la longueur des tours de parole des corpus ESTER1 et ESTER2 et ceux des corpus ETAPE et REPERE. La longueur moyenne est pratiquement trois fois plus faible pour REPERE que pour ESTER1 et ce indépendamment du genre des locuteurs. Cette différence peut

s'expliquer par le fait que dans ces derniers, nous avons majoritairement, voire uniquement pour REPERE, des enregistrements extraits d'émissions télévisées.

Tableau 7. Longueur des tours de parole en fonction du genre dans les grands corpus du français (ESTER1, ESTER2, ETAPE et REPERE)

Corpus	Femmes		Hommes		NA		Tous	
	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$
ESTER1	26,90	39,81	27,63	34,69	3,24	2,96	26,00	35,75
ESTER2	15,31	18,68	17,65	21,74	5,09	8,85	15,33	20,05
ETAPE	7,88	13,51	7,53	11,94	2,46	2,67	7,45	12,11
REPERE	9,84	11,64	8,35	10,65	2,04	2,17	7,53	10,25
Tous	15,30	26,48	12,39	20,54	2,41	3,25	11,94	21,13

Les écarts de durée s'expliqueraient donc par la variable radio/télé. L'ajout du canal visuel peut jouer dans la longueur des tours de parole. On a plus de modalités communicatives, l'information est donc transmise plus rapidement. Ou à l'inverse, il est plus compliqué pour un auditeur de suivre une émission de radio, si les échanges fusent. Cette durée de tours de parole plus courte à la télé, résulte aussi dans un nombre de tours de parole beaucoup plus important dans ces enregistrements (à titre de rappel 27639 tours pour une soixantaine d'heures de parole alors qu'ESTER1 contenait 13269 tours pour une centaine d'heures de parole). On observe également que les écarts-types sont beaucoup plus grands dans ESTER1 et ESTER2 que pour ETAPE et REPERE. Ce qui vient nous conforter dans notre idée que la longueur des tours de parole est directement dépendante du média (radio ou télévision).

Il semblerait également que si les tours de parole des hommes sont plus longs que ceux des femmes à la radio avec une longueur moyenne de 26,90s pour les femmes contre 27,63s pour les hommes dans ESTER1, ce ne soit pas le cas à la télévision où la longueur moyenne est de 9,84s pour les femmes et de 8,35s pour les hommes (chiffres pour REPERE). Le type de média et le genre semblent donc avoir des répercussions sur la longueur des tours de parole.

Contrairement au nombre de tours de parole par locuteur, cas dans lequel chaque locuteur pouvait être considéré comme une instance indépendante des autres, l'étude de la longueur des tours de parole ne peut pas utiliser des tests considérant les variables comme

étant indépendantes, car la longueur des tours est fortement liée à tous les tours de parole ayant été produits précédemment. Pour étudier ces durées, nous avons donc eu recours à des modèles de Cox, présentés dans la partie méthodologie du mémoire.

Dans notre cas, l'événement t que nous cherchons à prédire est la fin du tour de parole. Le genre est considéré comme un effet fixe et le locuteur comme un effet aléatoire. Augmenter le risque revient donc à augmenter la probabilité que la fin du tour de parole survienne plus tôt. Le modèle que nous avons utilisé définit comme facteur fixe le genre du locuteur, le corpus (qui peut se voir comme une opposition de la radio, pour ESTER1 et ESTER2, à la télévision, pour ETAPE et REPERE) et nous avons inséré un effet aléatoire dû au locuteur. Il aurait également pu être intéressant de considérer l'émission comme un effet fixe mais nous n'avons pas eu le temps nécessaire pour effectuer ces analyses. De même, les épisodes auraient pu être considérés comme des effets aléatoires mais R ne permet pas de les multiplier et la variable locuteur nous semblait plus pertinente. Les résultats obtenus pour notre modèle sont les suivants :

```
                                  coef    se(coef)   se2     Chisq  DF   p
turns$gendermale                -0.04854 0.01936  0.01935    6.29   1   1.2e-02
turns$id_corpus2                 0.38054 0.03790  0.03789  100.82   1   1.0e-23
turns$id_corpus3                 0.92374 0.02489  0.02488 1377.41   1   1.7e-301
turns$id_corpus4                 0.69760 0.02266  0.02263  947.74   1   4.1e-208
frailty.gaussian                                            17.57   1   2.8e-05
(turns$id_speaker, df=1)

turns$gendermale:turns$id_corpus2 -0.05384 0.04614  0.04613    1.36     1   2.4e-01
turns$gendermale:turns$id_corpus3  0.08303 0.02816  0.02814    8.70     1   3.2e-03
turns$gendermale:turns$id_corpus4  0.18105 0.02574  0.02570   49.49     1   2.0e-12
```

Si l'on observe les p-valeurs, il semblerait que l'effet du genre soit significatif (p=0.012) ainsi que les effets de chaque corpus. Lorsque l'on s'intéresse à chaque effet fixe indépendamment, on observe que les tours de parole des hommes sont en moyenne plus longs que ceux des femmes, et que les tours sont globalement plus longs dans ESTER1 que dans le reste des corpus. Les interactions entre le genre et le corpus sont également significatives sauf dans le cas d'ESTER2, avec la condition genre égale à homme (p=0.32). Le risque relatif est égal à l'exponentielle des coefficients. Si le coefficient est positif, on aura donc un risque relatif supérieur à 1, et comme dit précédemment, augmenter le risque revient à augmenter la probabilité que le tour de parole soit plus court. A travers l'ensemble des corpus, il semble donc que les tours de parole des hommes soient sensiblement plus longs que ceux des femmes (le risque est égal à $e^{-0.049} \approx 0.95$). On vérifie également ici le fait que les tours de parole sont significativement plus courts à la télé qu'à la radio, car les

effets des corpus ETAPE (id_corpus3) et REPERE (id_corpus4) sont très significatifs (p-valeurs égales respectivement à $1{,}7.10^{-301}$ et $4{,}1.10^{-208}$), avec des coefficients élevés, respectivement 0.92 et 0.70.

Lorsque l'on observe uniquement la variable du genre, on obtient donc un risque plus faible, signifiant que les tours de parole des hommes sont plus longs, mais dans les effets d'interaction significatifs, le risque est toujours plus grand que pour les conditions de base (un locuteur féminin dans le corpus ESTER1). Cela laisserait donc supposer que l'effet du corpus est plus important que celui du genre.

Du fait du nombre important de données que nous étudions l'ensemble de nos tests sont significatifs, en revanche les variations de durée sont faibles entre les hommes et les femmes. Plus qu'une différence de genre, il semblerait donc que ce qui ressorte soit une différence de longueur des tours en fonction du média.

Dans cette partie, nous avons exploré les grands corpus du français via le prisme du genre selon 3 modalités : le rôle médiatique, le nombre de tours de parole par locuteur et la longueur des tours de parole. Les hommes sont globalement plus présents dans le corpus que les femmes avec une proportion générale de 35,54 % de femmes pour 63,28 % d'hommes. On retrouve donc les chiffres du CSA sur la présence des femmes dans les médias. Mais ces proportions sont variables d'un rôle médiatique à l'autre, et les femmes sont encore très absentes dans la classe des locuteurs intervenants souvent et longtemps, où l'on observe un ratio qui est plus de l'ordre d'une femme pour neuf hommes. Nous avons également pu constater que la distribution des tours de parole était inégale entre les genres, les hommes parlant en moyenne deux fois plus que les femmes. Les femmes semblent également parler moins souvent à rôle égal, comme nous avons pu le voir avec le cas du débat télévisé.

Les femmes représentant environ 20 % de l'ensemble des tours de parole de l'ensemble de nos corpus, elles prennent donc moins de place en terme de présence, mais également en terme de nombre d'interventions. Concernant la longueur de ces interventions, il n'y a pas de tendance globale observable entre les genres, car le facteur le plus important semble être le type de média, à savoir radio ou télévision, mais il serait intéressant d'explorer plus avant les différences de genre dans chaque média.

Chapitre 4. Biais et performances : le cas d'un système d'ASR

Dans le chapitre précédent, nous avons analysé les grands corpus du français et montré que ceux-ci présentaient des biais en terme de genre. Or des études comme celles de Buolamwini (op.) ou de Caliskan (op.) ont démontré que les biais contenus dans les données pouvaient se répercuter sur les performances du systèmes. Dans ce chapitre, nous allons donc étudier les performances d'un système de reconnaissance automatique de la parole (ASR) développé sur les grands corpus du français, pour voir dans quelle mesure les biais observés dans le chapitre 3 influencent les performances du système.

1. Analyse du corpus d'apprentissage

Comme expliqué dans la partie méthodologie, le corpus d'apprentissage du système d'ASR étudié est constitué majoritairement d'enregistrements extraits des grands corpus du français. Nous concevons l'analyse de ce système comme une étude prototypique de l'influence des biais dans les corpus pour les performances des systèmes de reconnaissance automatique de la parole. Afin de valider notre choix de prototype nous allons dans un premier temps observer la constitution du corpus d'apprentissage du système pour vérifier qu'elle contient des biais similaires à ceux observer dans l'ensemble des grands corpus du français. Le Tableau 8 présente les proportions en terme de genre de locuteur et de tours de parole

Tableau 8. Répartition des locuteurs et des tours de parole dans le corpus d'apprentissage en fonction du genre.

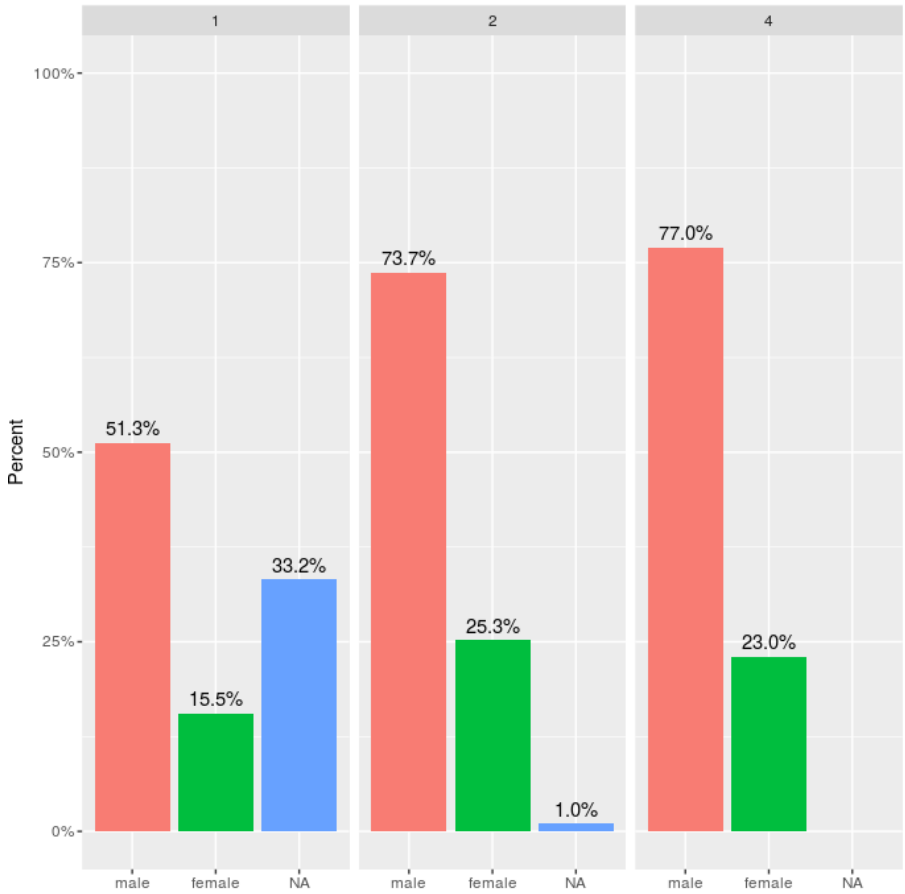
	Locuteurs			Tours de parole		
	F	H	NA	F	H	NA
Freq	327	1068	596	15807	58620	3692
%	16,42	53,64	29,94	20,23	75,04	4,73

Concernant les locuteurs, la proportion d'individus au genre non-identifié est plus importante dans ce corpus que dans les grands corpus du français, mais ces locuteurs représentent peu de tours de parole. Si l'on s'intéresse uniquement aux locuteurs dont le

genre est connu on obtient 75 % d'hommes pour 25 % de femmes. Les proportions en terme de tours de parole, quant à elles, sont comparables à celles observées dans les grands corpus du français (se reporter à l'Annexe 3) avec cependant 5 points de plus pour les hommes, et un pourcentage à 20 % pour les femmes. Le but étant de montrer l'impact du biais, si les disparités sont un peu plus grandes, nous supposons que cela ne fera que faciliter la visualisation de leur impact sur les performances. Nous avons donc pris la décision de considérer les distributions comme globalement similaires.

Une deuxième vérification de la représentativité du corpus d'apprentissage a été faite en observant la distribution des locuteurs parmi les différentes classes. Cette distribution est représentée par la Figure 5 ci-dessous. Avec un effectif de 1791 locuteurs pour la classe 1, 99 locuteurs pour la classe 2 et 100 pour la classe 4, on retrouve bien l'idée selon laquelle la majorité des locuteurs sont des intervenants ponctuels, parlant peu et peu souvent. La classe 3 ne contenant qu'une seule locutrice, nous ne l'avons pas incluse dans notre analyse. On observe également qu'on retrouve bien des proportions similaires entre les classes 2 et 4 (environ 75 % d'hommes pour 25 % de femmes), signifiant que les femmes sont largement sous-représentées dans les classes parlant souvent.

Figure 5. Proportion d'hommes, de femmes par classe médiatique dans le corpus d'apprentissage. Effectifs : classe 1 : 1791, classe 2 : 99, classe 4 : 100). La classe 3 n'a pas été représentée car elle ne contenait qu'un individu.



La forte proportion de locuteurs au genre inconnu dans la classe 1 ne permet pas une comparaison directe aux grands corpus du français, mais les hommes restent majoritaires, comme trouvé précédemment.

Bien que différent de l'ensemble des corpus présentés dans le Chapitre 3, le corpus d'apprentissage du système d'ASR étudié présente bien des biais en terme de genre, avec des femmes moins représentées et qui prennent la parole moins souvent. Nous allons donc maintenant analyser les performances du systèmes au regard de ces données.

2. Analyse des performances

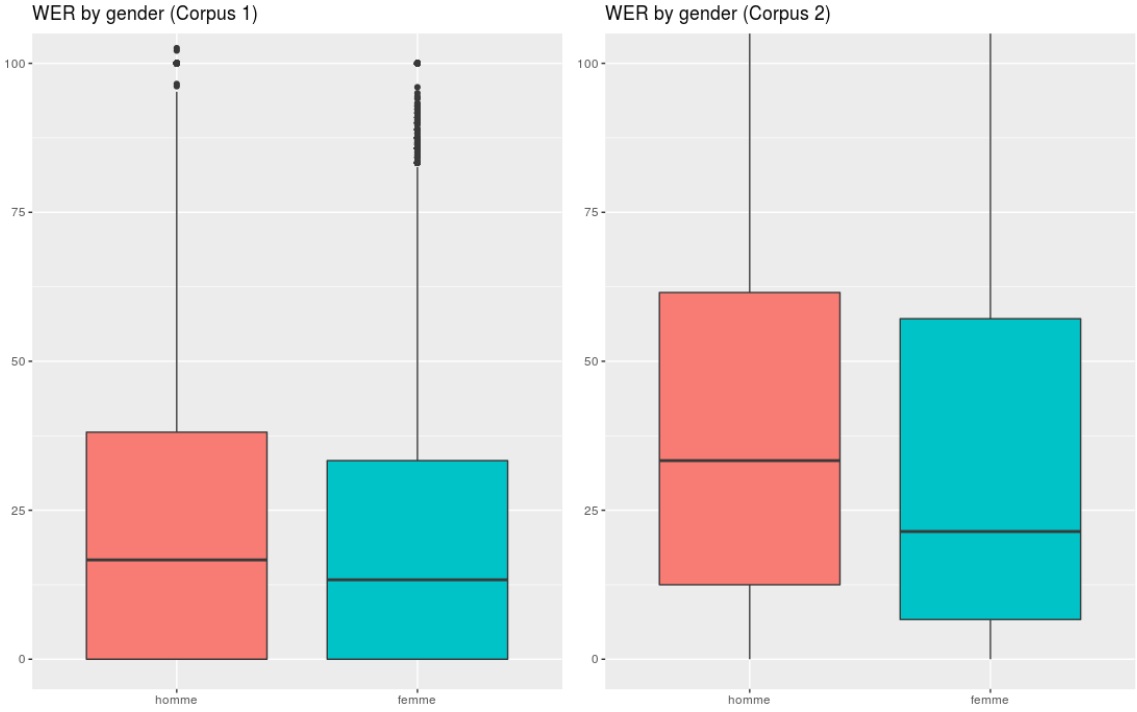
2.1. Analyse générale

Pour étudier les performances nous nous sommes intéressée au taux erreur mot ou WER (Word Error Rate). Sur le Corpus 1 comme le Corpus 2, les performances sont meilleures pour les femmes, avec un WER moyen de 23,52 sur le Corpus 1 et 35,63 sur le Corpus 2 pour les femmes contre un WER moyen à 26,68 pour le Corpus 1 et 41,16 pour le Corpus 2. Les variances étant pratiquement égales, on suppose que cette différence de moyenne est significative. La Figure 6 représente la distribution des WER par genre dans chaque corpus. Ces résultats sont cohérents avec ceux obtenus par Adda-Decker et Lamel (REF). Nous pouvons également observer une grande variation des performances entre les deux corpus. En effet le Corpus 1 contient des données relativement « simples », beaucoup d'émissions contenant de la parole préparée, alors que le Corpus 2 contient des émissions plus compliquées, soit à cause de la présence importante de parole spontanée et/ou superposée, soit du fait de conditions acoustiques difficiles (comme pour Un Temps de Pauchon). On observe également que les différences entre les genres sont plus importantes dans le Corpus 2 que le Corpus 1, laissant supposer que le type d'émission peut être responsable d'un écart de performance plus grand entre les genres. Nous avons décidé d'étudier les performances par type d'émission pour tenter d'expliquer cette variation.

Tableau 9. WER moyen et écart-types, par genre et par corpus.

	WER			
	Hommes		Femmes	
	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$
Corpus 1	26,68	33,11	23,52	32,26
Corpus 2	41,16	37,18	35,63	37,54

Figure 6. Distribution des WER par genre et par corpus. Effectifs pour le corpus 1 : 925 hommes et 440 femmes. Effectifs pour le corpus 2 : 264 hommes et 135 femmes. La distribution dépasse la valeur 100, mais un WER à 100 % signifiant déjà qu'aucun des mots n'a été reconnu correctement nous ne nous intéressons pas aux valeurs supérieures.



2.2. Analyse par émission

Les résultats de l'analyse par émissions sont présentés ci-dessous. Le nombre de locuteurs et les WER moyens sont présentés dans le Tableau 10 pour le Corpus 1 et le Tableau 11 pour le Corpus 2 (une version des mêmes tableaux avec les écarts-types se trouve dans l'Annexe 6). Le Tableau 12 récapitule le nom des émissions ainsi que leur type

médiatique (émissions télévisuelles ou radiophoniques) dans le but de faciliter la lecture des tableaux. Les distributions des WER par émissions et par genre sont présentées dans l'Annexe 7.

Tableau 10. Répartition en locuteurs et WER moyens pour le Corpus 1. (Le type correspond au média, T pour télévision et R pour radio)

Corpus 1		**Nombre de loc.**			**WER moyen**		
Emissions	**Type**	**Hommes**	**Femmes**	**Tous**	**Hommes**	**Femmes**	**Tous**
Pile et Face	T	44	8	52	23,81	19,39	23,55
RFI DGA	R	249	116	365	27,53	22,50	26,08
RFI ELDA	R	139	56	195	23,60	20,68	22,68
FINTER ESTER2	R	146	110	256	39,48	36,72	38,56
FR3	T	21	11	32	31,86	31,84	31,85
TELSONNE	R	252	110	362	30,06	34,44	30,94
QR TELSONNE	R	7	0	7	22,95	NA	22,95
QR FINTER	R	11	2	13	23,73	25,81	24,43
RTM	R	56	27	83	23,22	21,35	22,25

Tableau 11. Répartition en locuteurs et WER moyens pour le Corpus 2. (Le type correspond au média, T pour télévision et R pour radio)

Corpus 2		**Nombre de loc.**			**WER moyen**		
Emissions	**Type**	**Hommes**	**Femmes**	**Tous**	**Hommes**	**Femmes**	**Tous**
tvme	R	26	19	45	24,52	17,51	21,64
africa1	R	31	7	38	31,48	20,94	29,20
csoj.16k	T	13	2	15	36,19	35,36	36,19
CultureEtVous	T	68	46	114	41,93	59,87	49,18
PlaneteShowbiz	T	74	26	100	40,09	71,25	49,68
Arte News	T	14	13	27	16,88	11,88	14,1
Fab Histoire	R	6	14	20	27,42	25,54	26,58
Temps Pauchon	R	14	2	16	42,59	13,51	35,23
TV8 Pl du Village	T	18	6	24	54,80	30,71	53,51

Plusieurs tendances sont observables à partir de ces tableaux : on constate que les performances sont moins bonnes pour les émissions télévisuelles que pour les émissions radiophoniques. Pour la télévision, le WER minimal est celui d'Arte News avec une valeur

de 14,1 % . Le WER maximal, 53,51 %, est obtenu pour la Place du Village, émission de divertissement contenant de la parole accentuée. Concernant la radio, les meilleurs performances sont réalisées sur les nouvelles de TVME, avec un WER égal à 21,64 % et les moins bonnes sont obtenues sur les débats de France Inter avec 38,56 %. Hormis pour le journal d'Arte, les perfomances semblent globalement meilleures pour la radio. Cet écart de performances entre média pourrait s'expliquer par l'impact de la longueur des tours de parole sur le WER. En effet, nous avons démontré précédemment que la durée des tours de parole était plus courte à la télévision Cette durée plus courte des interventions peut s'expliquer par une interaction plus rapide en terme de rythme qui pourrait donc contenir plus de structures syntaxiques incomplètes et donc non-reconnues par les modèles de langue.

D'une manière générale, que ce soit pour la télévision comme pour la radio, on observe également que les performances sont meilleures sur des émissions de type nouvelles (Arte, TVME, Africa1) que sur des émissions de type débat ou de divertissement. Le caractère préparé ou non de la parole est donc aussi un facteur à prendre en compte lors de l'évaluation des performances.

Concernant le genre, la tendance générale semble être que les WER des femmes sont plus bas que ceux des hommes, ce qui nous laisse supposer que la parole médiatique des femmes est peut-être plus normative, comme le laissaient entendre Adda-Decker & Lamel (op) lorsqu'elles parlaient du taux de délétion plus important dans le discours des hommes. **Il ne semble donc pas que la sous-représentation des femmes dans les corpus ait un impact sur les performances.** Mais si globalement les WER sont plus bas pour les femmes ce n'est pas le cas pour 3 émissions : Le Téléphone Sonne, Planète Showbiz et Culture & Vous (en gris dans les Tableaux 10 et 11, nous n'analyserons pas les débats de France Inter contenus dans le Corpus 1 à cause du faible nombre de locuteurs). L'écart est maximal pour Planète Showbiz où le WER moyen des femmes atteints 71,25 %. Ces performances sont vraiment mauvaises et il faudrait explorer plus finement les données pour pouvoir expliquer ces écarts. En parcourant un peu le contenu des données pour cette émissions, nous avons pu remarquer que de nombreux tours de parole aux performances particulièrement mauvaises (entre 100 % et 150 % de taux d'erreur) étaient produits par Pascale De La Tour Du Pin, co-présentatrice de l'émission avec Philippe Dufreigne. Ces tours de parole ne contenaient pour la majorité que le nom de son co-présentateur, qui ne faisait visiblement pas partie du lexique du système et était donc mal reconnu. On pourrait

poser l'hypothèse que dans ce type d'émissions, un rôle souvent tenu par les femmes est celui de « donneuse de parole », la plupart de ses interventions servant à récapituler ce qui a été dit et rendre la parole au présentateur principal. Ce type de rôle, allant de pair avec une utilisation importante d'entités nommées, pourrait expliquer des performances plus mauvaises.

D'une manière générale, nous avons pu observer que le type de média et le type d'émission semblaient être des facteurs plus importants pour les systèmes de reconnaissance automatique de la parole. Sur les émissions de radio, les performances semblent meilleures pour les femmes, démontrant que la sous-représentation des femmes dans les corpus ne semble pas avoir d'impact direct sur les systèmes. Des explications d'ordre sociolinguistiques, comme une parole féminine plus normative seraient donc à creuser pour expliquer ces écarts. Nous avons également vu qu'à la télévision, les performances étaient plus mauvaises, et nous avons pu observer des WER particulièrement hauts pour les femmes. Nous posons l'hypothèse que ces résultats pourraient s'expliquer du fait du rôle de certains locuteurs et qu'il serait intéressant de creuser les performances en fonction des différentes classes médiatiques proposées dans notre chapitre 3 pour étayer notre proposition.

Tableau 12. Récapitulatif du nom et type de média des émissions

Corpus	Nom	Emission	Média
1	rfi-fm-dga	News	Radio
	RFI-ELDA	News	Radio
	RTM-ELDA	News	Radio
	EST2BC-FRE-FINTER-DEBATE	Débats	Radio
	LCP-PileEtFace	Pile et Face	Télévision
	QRBC-FRE-FR-FRANCE3-DEBATE	Débats	Télévision
	QRBC-FRE-FR-FINTER-DEBATE	Débats	Radio
	QRBC-FRE-FR-TELSONNE-POD	Le Téléphone Sonne	Radio
	TELSONNE	Le Téléphone Sonne	Radio
2	tvme	News	Radio
	africa1	News	Radio
	csoj.16k	Ce Soir ou Jamais	Télévision
	BFMTV-CultureEtVous	Culture & Vous	Télévision
	BFMTV-PlaneteShowbiz	Planète Showbiz	Télévision
	ARTE-NEWS	News	Télévision
	FINTER-FABHISTOIRE-POD	La Fabrique de l'Histoire	Radio
	FCULT-TEMPS-POD	Un Temps de Pauchon	Radio
	TV8-LaPlaceDuVillage	La Place du Village	Télévision

Conclusion

Dans ce travail de mémoire nous avons souhaité répondre à 3 grandes questions : est-ce que les femmes étaient sous-représentées dans les grands corpus du français, quel était leur rôle et comment ces disparités de représentation affectaient les performances d'un système développé sur ces données.

Concernant la place des femmes dans les médias, nos conclusions sont multiples : de manière non surprenante, et conformément aux chiffres donnés par le CSA, les femmes sont moins présentes que les hommes. Que ce soit en terme de nombre de locuteurs, mais également en terme nombre d'interventions. On constate un ratio de l'ordre d'une femme pour deux hommes en terme de locuteurs, mais les femmes parlant deux fois moins, on se retrouve avec un ratio d'une femme pour 4 hommes en terme de nombre d'interventions. Ces résultats généraux sont cependant à nuancer. Nous avons proposé 4 classes de locuteurs, censées représenter des comportements différents en terme de nombre de tours de parole et de longueur des interventions. Selon les classes, les répartitions ne sont pas toujours homogènes et on observe que les femmes sont encore plus absentes de la classe des locuteurs influents (qui parlent souvent et longtemps). Les disparités de répartition des interventions sont donc peut-être plus grandes encore dans certaines classes. L'étude de ce phénomène et la recherche d'une définition plus précise de ces différentes classes pourrait constituer une perspective de ce mémoire. Le corpus REPERE, qui contient une annotation en rôle du locuteur, pourrait également permettre de comparer nos seuils avec les rôles annotés.

Nous nous sommes également intéressée à la longueur des tours de parole, mais le genre ne semble pas être un facteur significatif pour cette variable. En revanche, nous avons pu observer que les interventions étaient significativement plus courtes à la télévision qu'à la radio, confirmant l'idée que le type de média fait partie du contexte interactionnel et contribue à définir un type de discours particulier.

Du point de vu des performances, les résultats sont moins tranchés. Globalement les WER sont meilleurs pour les femmes, ce qui pourrait laisser supposer une différence de parole médiatique entre les genres. La parole des femmes serait plus normative, comme le laissaient entendre Adda-Decker et Lamel (op), en constatant le nombre de délétions bien plus important chez les hommes. Une analyse acoustique détaillée ainsi qu'une analyse des transcriptions pourraient permettre de vérifier s'il existe effectivement une explication

sociolinguistique à ces différences. Il serait également intéressant de décomposer le WER en nombre d'insertions, délétions et substitutions pour tenter de reproduire leurs résultats.

Enfin, nous avons également pu constater, que d'une manière générale, les performances du système étaient meilleures sur des enregistrements radiophoniques que sur des émissions télévisuelles. Nous avons émis l'idée que ce résultat était peut-être la conséquence de la différence de durée des tours de parole entre ces deux types médiatiques, mais une fois encore, cette étude constitue une possibilité de perspective pour ce mémoire.

Pour conclure, il semblerait que les biais de genre présents dans les données n'impliquent pas des biais dans les performances. Contrairement à l'article cité dans l'introduction, il semblerait même que les performances soient meilleures pour les femmes. Nous avons esquissé une explication linguistique avec l'idée d'une parole plus normative des femmes, mais une explication technique est aussi envisageable : beaucoup de progrès ont été faits par les systèmes de reconnaissance automatique pour s'adapter au locuteur. Cette adaptation à la personne permettrait donc de réduire les différences de genre, ce qui pallierait les biais de représentation.

Bibliographie

Adda-Decker, M., & Lamel, L. (2005). Do speech recognizers prefer female speakers? In *Ninth European Conference on Speech Communication and Technology*.

Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016). FABIOLE, a Speech Database For Forensic Speaker Comparison. In *Proceedings of LREC*.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (1998). Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech (p. 4). In *Proceedings of LREC*.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349-4357).

Brun, A., Cerisara, C., Fohr, D., Illina, I., Langlois, D., Mella, O., & Smaïli, K. (2004). Ants: le système de transcription automatique du Loria. In *Journées d'Etude sur la Parole-JEP'04* (pp. 4-p).

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (pp. 77-91).

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

Elloumi, Z., Besacier, L., Galibert, O., Kahn, J., & Lecouteux, B. (2018). ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks. *arXiv preprint arXiv:1804.08477*.

Esteve, Y., Bazillon, T., Antoine, J. Y., Béchet, F., & Farinas, J. (2010). The EPAC Corpus: Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of LREC*.

Firth, J. R. (1957). *A synopsis of linguistic theory. Studies in linguistic analysis*. Oxford: Blackwell.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., & Choukri, K. (2006). Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. In *Proceedings of LREC* (Vol. 6, p. 315–320).

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

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., & Quintard, L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *LREC Proceedings*.

Goldwater, S., Jurafsky, D., & 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.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., & Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *Proceedings of LREC*.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K., & Choukri, K. (2004). The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of LREC*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315-3323).

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* (pp. 656-666).

Letué, F., Martinez, M. J., Samson, A., Vilain, A., & Vilain, C. (2018). Statistical Methodology for the Analysis of Repeated Duration Data in Behavioral Studies. *Journal of Speech, Language, and Hearing Research*, *61*(3), 561-582.

Meignier, S., & Merlin, T. (2010). LIUM SpkDiarization: an open source toolkit for diarization. In *CMU SPUD Workshop*. Dallas, United States.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rodger, J. A., & Pendharkar, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. In *International Journal of Human-Computer Studies*, *60*(5-6), 529-544.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7-55).

SSawalha, M., & Abu Shariah, M. (2013). The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds.

Scheffer, N., Istrate, D., Fredouille, C., & Bonastre, J. F. (2005). Les systèmes du LIA pour les tâches de segmentation et de suivi: SES, SRL, SVL. *Atelier ESTER, Avignon*.

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 53-59).

Tatman, R., & Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proceedings of Interspeech 2017*, (pp. 934-938).

Table des annexes

Annexe 1 : Constitution des grands corpus du français

ESTER 1

Sources	Heures transcrites	Heures non-transcrites
France Inter	37	337
France Info	12	643
RFI	27	445
RTM	22	-
France Culture	1	252
Radio Classique	1	-
Total	**100**	**1677**

ESTER 2

Source	Corpus d'apprentissage	Corpus de développement
France Inter	26h	2h
RFI	69h	40 min
Africa n°1	10h	2h20
TVME (ex RTM)	-	1h
Corpus EPAC	13h	-
Total	**118h**	**6h**

REPERE

General repartition	**Development data : 9h Test data : 16 h Learning data : 35h**		
3h sub-corpora TV show repartition			
TV Show	**Channel**	**Dev set (min)**	**Test set (min)**
BFM Story	BFM	60	60
Planete Showbiz	BFM	15	15
Ca vous regarde	LCP	15	15
Entre les lignes	LCP	15	15
Pile et Face	LCP	15	15
LCP Info	LCP	30	30
Top Questions	LCP	30	30
Total	**-**	**180**	**180**

ETAPE

Genre	Train	Dev	Test	Sources
TV News	7h30	1h35	1h35	BFM Story, Top Question (LCP)
TV debates	10h30	2h40	2h40	Pile et Face, Ca vous regarde, Entre les lignes (LCP)
TV amusement	-	1h05	1h05	La Place du Village (TV8)
Radio shows	7h50	3h	3h	Un temps de Pauchon, Service Public, Le Masque et la Plume, Comme on nous parle, Le fou du roi
Total	**25h30**	**8h20**	**8h20**	**42h10**

Annexe 2 : Requêtes SQL pour l'extraction des données

Requête d'extraction des tours de parole

```
SELECT id_turn,
        start_time,
        end_time,
         id_speaker,
        (SELECT gender FROM speaker WHERE id_speaker=turn.id_speaker),
        id_episode,
        (SELECT id_show FROM episode WHERE id_episode=turn.id_episode),
        (SELECT id_corpus FROM episode WHERE id_episode=turn.id_episode)
 FROM turn
```

Requête d'extraction du nombre de tours de parole par locuteur par corpus

```
SELECT id_speaker,gender,
        (SELECT COUNT (turn.id_turn) FROM turn
                WHERE (speaker.id_speaker = turn.id_speaker AND turn.id_episode IN
                        (SELECT id_episode FROM episode WHERE id_corpus =1)))
 FROM speaker
```

La requête ci-dessus permet d'extraire le nombre de tours de parole par locuteur dans le corpus 1, à savoir ESTER 1. Il suffit de changer la variable de id_corpus pour récupérer le nombre de tours dans les autres corpus. Pour le nombre de tours par locuteur sur l'ensemble des corpus la requête est la suivante :

```
SELECT id_speaker,gender,
        (SELECT COUNT (turn.id_turn) FROM turn
                WHERE (speaker.id_speaker = turn.id_speaker))
 FROM speaker
```

Les résultats de nombre de tours par locuteurs par corpus ont ensuite été concaténés manuellement dans un fichier csv de la forme :

id_loc;nb_tours_ESTER1;nb_tours_ESTER2;nb_tours_ETAPE;nb_tours_REPERE;nb_tours_TOTAL

Annexe 3 : Analyse du nombre de tours par locuteurs

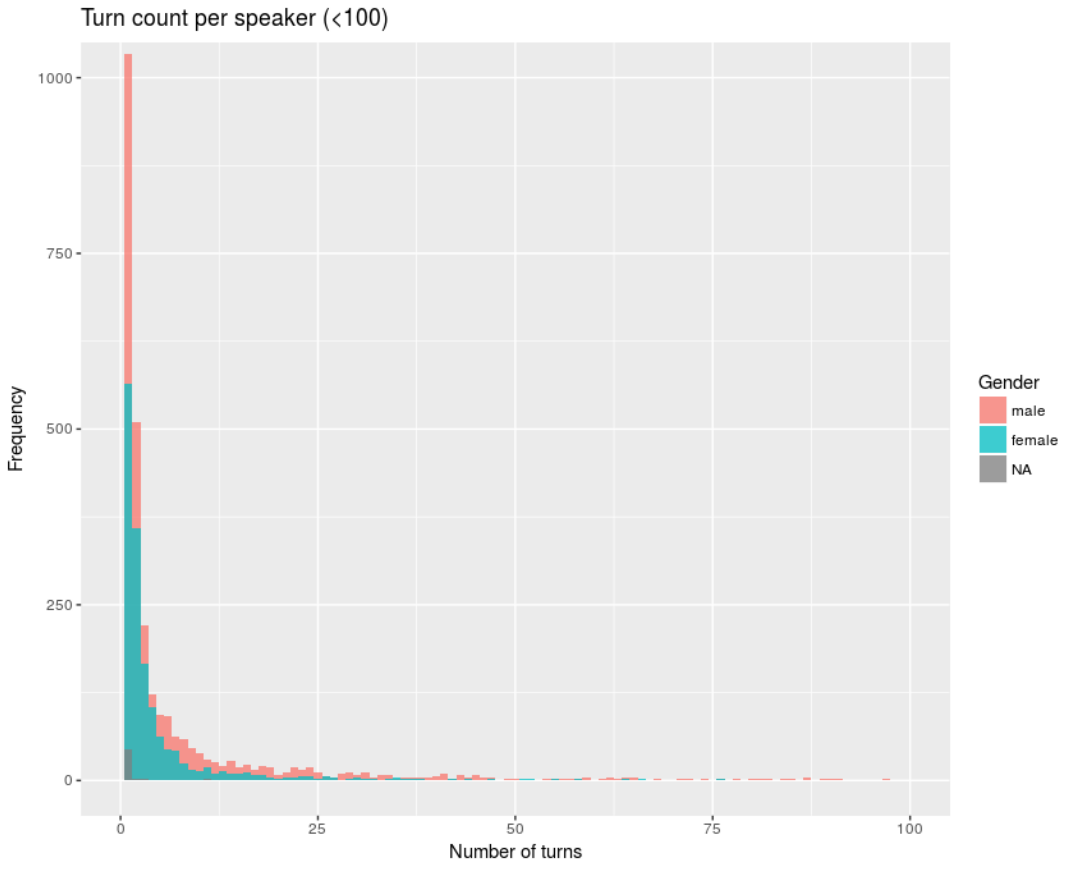
TABLE 2.a) Répartition des tours de parole par corpus en fonction du genre (fréquences)

corpus	female	male	unknown		total
			all	multiloc	
ESTER1 (~100H)	3893	8606	770	735	**13269**
ESTER2 (~124H)	858	1704	384	371	**2946**
ETAPE (~6h)	2923	13237	456	419	**16616**
REPERE (~42h)	4302	18725	4612	4610	**27639**
all	**11976**	**42272**	**6222**	6135	**60470**

TABLE 2.b) Répartition des tours de parole par corpus en fonction du genre (%)

corpus	female	male	unknown	
			all	multiloc
ESTER1	29,34%	64,86%	5,80%	5,54%
ESTER2	29,12%	57,84%	13,03%	12,59%
ETAPE	17,59%	79,66%	2,74%	2,52%
REPERE	15,56%	67,75%	16,69%	16,68%
all	**19,80%**	**69,91%**	**10,29%**	**10,15%**

Figure 1.a Distribution du nombre de tours de parole par locuteur, par genre. *(NB : Nous avons fait le choix de ne représenter que les locuteurs ayant moins de 100 tours de parole par souci de lisibilité, le code R joint permet cependant de visualiser l'ensemble des données)*



Annexe 4 : Test statistique sur le nombre de tours par locuteurs

```
Call:
glm(formula = spk_all$all ~ spk_all$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
 -4.722   -4.204   -2.992   -1.401  131.617

Coefficients:
                            Estimate      Std. Error  z value     Pr(>|z|)
(Intercept)           2.697859      0.004864       554.68    <2e-16 ***
spk_all$genderfemale   -0.684333    0.010352       -66.11            <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 191243  on 4445  degrees of freedom
Residual deviance: 186337  on 4444  degrees of freedom
  (53 observations deleted due to missingness)
AIC: 200025

Number of Fisher Scoring iterations: 7
```

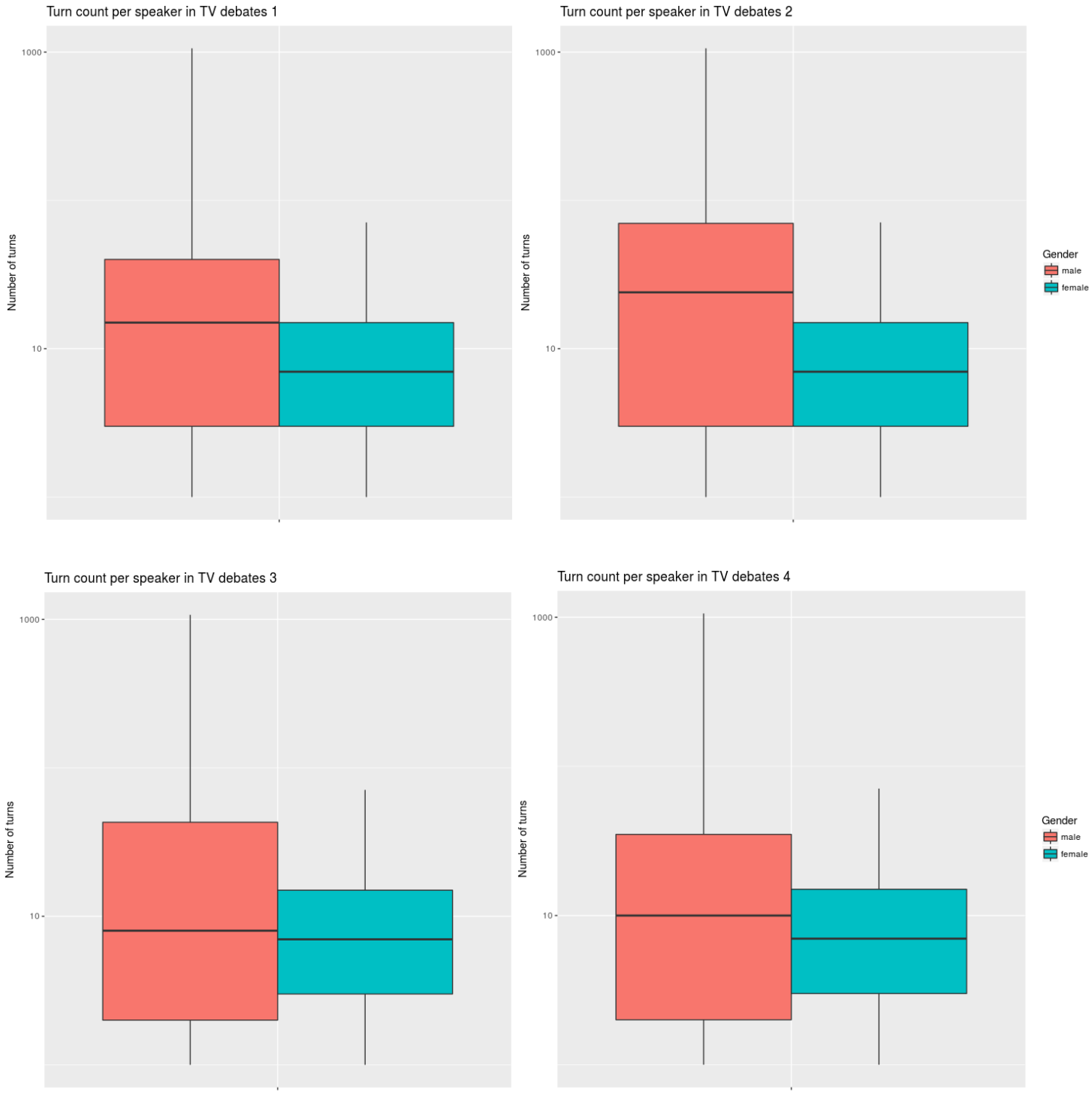
D’après les histogrammes, nous avons fait l’hypothèse que nos variables étaient distribuées selon un loi de Poisson, d’où le choix de la famille poisson pour le modèle.

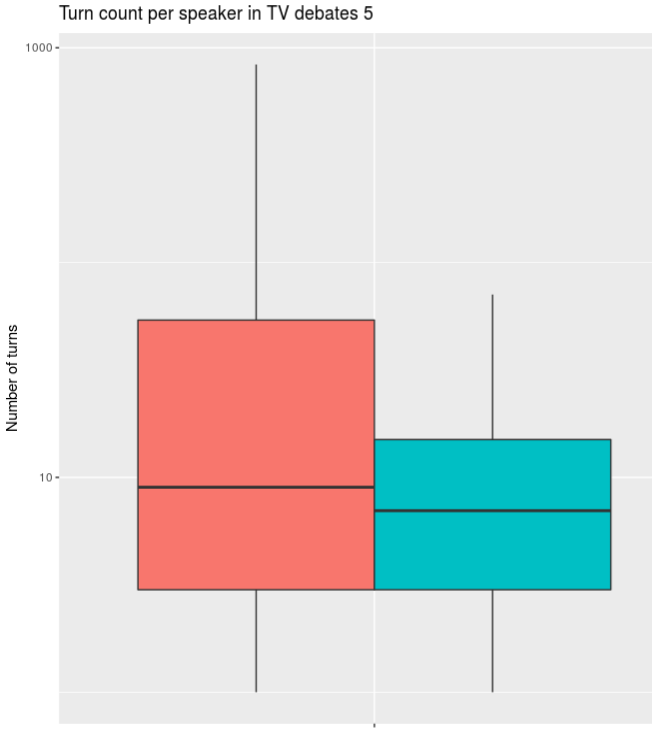
spk_all$all représente le nombre de tours de parole par locuteur, on a donc cherché ici à quantifier la différence entre les distributions selon le genre.

On obtient un coefficient égal à -0.68 avec une p-valeur inférieure à 2.1016. On peut donc en conclure que les distributions sont significativement différentes. Ce coefficient peut s’interpréter de la manière suivante : une locutrice aura en moyenne $e^{-0.68} \approx 0.5$ fois le nombre de tours d’un locuteur masculin.

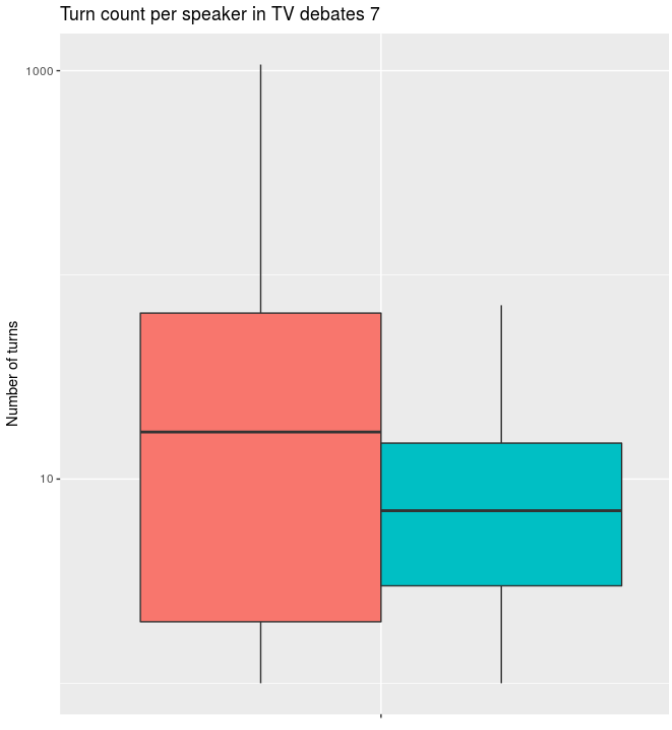
Annexe 5 : Le cas du débat télévisé : distribution et test statistique sur le nombre de tours de parole

Distribution des tours de parole par genre pour les 10 tirages aléatoires :

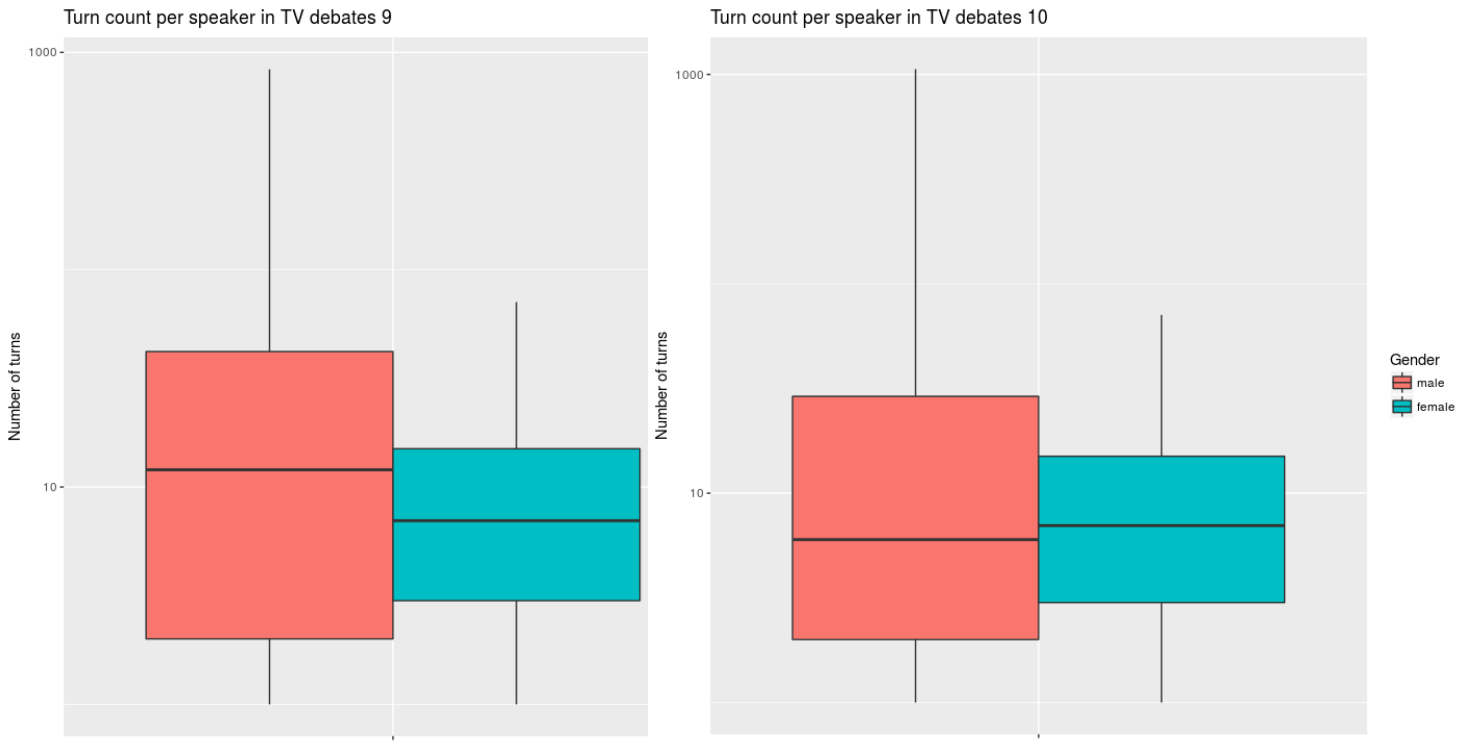



Turn count per speaker in TV debates 5
Number of turns
1000
10


Turn count per speaker in TV debates 6
Number of turns
1000
10
Gender
male
female


Turn count per speaker in TV debates 7
Number of turns
1000
10


Turn count per speaker in TV debates 8
Number of turns
100
10
1
Gender
male
female



Résultats des test statistiques pour chacun des tirages aléatoires

Tirage 1

```
Call:
glm(formula = subset1$nb_turn ~ subset1$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-10.558   -6.999   -2.926   -0.237   63.719

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.10828    0.01698  241.94   <2e-16 ***
subset1$genderfemale -1.61175    0.04163  -38.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 13375  on 113  degrees of freedom
Residual deviance: 11353  on 112  degrees of freedom
AIC: 11821

Number of Fisher Scoring iterations: 7
```

Tirage 2

```
Call:
glm(formula = subset2$nb_turn ~ subset2$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-11.968   -6.871   -2.720    0.452   59.950

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.3434     0.0151  287.68   <2e-16 ***
subset2$genderfemale  -1.8468     0.0409  -45.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 13504  on 113  degrees of freedom
Residual deviance: 10507  on 112  degrees of freedom
AIC: 11000

Number of Fisher Scoring iterations: 6
```

Tirage 3

```
Call:
glm(formula = subset3$nb_turn ~ subset3$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-10.016   -7.906   -2.933    0.452   65.781

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.01052    0.01783  224.93   <2e-16 ***
subset3$genderfemale -1.51398    0.04199  -36.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11558.4  on 113  degrees of freedom
Residual deviance:  9860.7  on 112  degrees of freedom
AIC: 10319

Number of Fisher Scoring iterations: 6
```

Tirage 4

```
Call:
glm(formula = subset4$nb_turn ~ subset4$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-11.855   -9.473   -3.615   -0.486   60.244

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.32541    0.01523  283.93   <2e-16 ***
subset4$genderfemale -1.82888    0.04095  -44.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 17862  on 113  degrees of freedom
Residual deviance: 14950  on 112  degrees of freedom
AIC: 15411

Number of Fisher Scoring iterations: 7
```

Tirage 5

```
Call:
glm(formula = subset5$nb_turn ~ subset5$gender, family = poisson())

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-9.324  -6.786  -2.720   0.796  56.481

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.87848    0.01905   203.6   <2e-16 ***
subset5$genderfemale -1.38195    0.04252   -32.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8967.3  on 113  degrees of freedom
Residual deviance: 7644.8  on 112  degrees of freedom
AIC: 8111.4

Number of Fisher Scoring iterations: 6
```

Tirage 6

```
Call:
glm(formula = subset6$nb_turn ~ subset6$gender, family = poisson())

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-9.944  -7.250  -2.720  -0.333  65.445

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.99707    0.01795  222.66   <2e-16 ***
subset6$genderfemale -1.50054    0.04204  -35.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12606  on 113  degrees of freedom
Residual deviance: 10950  on 112  degrees of freedom
AIC: 11410

Number of Fisher Scoring iterations: 7
```

Tirage 7

```
Call:
glm(formula = subset7$nb_turn ~ subset7$gender, family = poisson())

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-12.409   -8.695   -2.842   -0.113   59.343

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.41163    0.01459  302.37   <2e-16 ***
subset7$genderfemale -1.91509    0.04072  -47.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 17815  on 113  degrees of freedom
Residual deviance: 14476  on 112  degrees of freedom
AIC: 14957

Number of Fisher Scoring iterations: 6
```

Tirage 8

```
Call:
glm(formula = subset8$nb_turn ~ subset8$gender, family = poisson())

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-6.507  -3.723  -1.955   1.254  13.198

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.23489    0.02628  123.10   <2e-16 ***
subset8$genderfemale -0.73835    0.04621  -15.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2755.2  on 113  degrees of freedom
Residual deviance: 2482.3  on 112  degrees of freedom
AIC: 2941.4

Number of Fisher Scoring iterations: 5
```

Tirage 9

```
Call:
glm(formula = subset9$nb_turn ~ subset9$gender, family = poisson())

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-9.071  -5.919  -2.720   0.452  57.184

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.82788    0.01954  195.94   <2e-16 ***
subset9$genderfemale -1.33134    0.04274  -31.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8752.1  on 113  degrees of freedom
Residual deviance: 7555.8  on 112  degrees of freedom
AIC: 8017.3

Number of Fisher Scoring iterations: 6
```

Tirage 10

```
Call:
glm(formula = subset10$nb_turn ~ subset10$gender, family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.889   -7.121   -2.933    0.244   68.555

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              3.79095    0.01990  190.49   <2e-16 ***
subset10$genderfemale -1.29441    0.04291  -30.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9706.2  on 113  degrees of freedom
Residual deviance: 8596.4  on 112  degrees of freedom
AIC: 9045.2

Number of Fisher Scoring iterations: 7
```

Annexe 6 : WER par émissions et par genre

Tableau 12. Répartition en locuteurs,WER moyens et écarts-types pour le Corpus 1. (Le type correspond au média, T pour télévision et R pour radio)

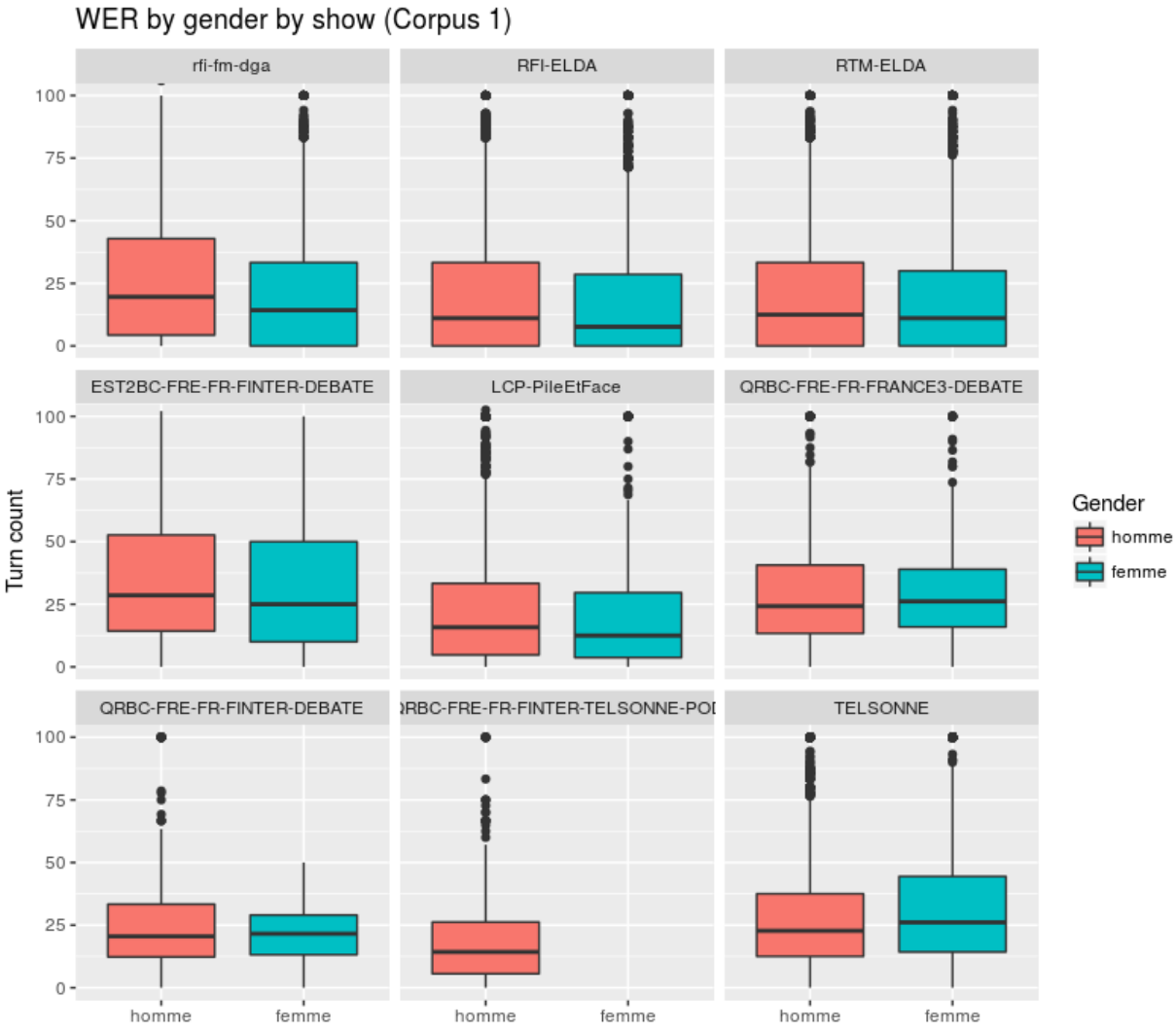
Corpus 1		**Nombre de loc.**			**WER**					
Emissions	**Type**	**Hommes**	**Femmes**	**Tous**	**Hommes**		**Femmes**		**Tous**	
					$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$
Pile et Face	T	44	8	52	23,81	29,14	19,39	28,45	23,55	29,11
RFI DGA	R	249	116	365	27,53	29,45	22,50	27,35	26,08	28,95
RFI ELDA	R	139	56	195	23,60	35,54	20,68	32,42	22,68	34,61
FINTER ESTER2	R	146	110	256	39,48	38,45	36,72	36,64	38,56	37,88
FR3	T	21	11	32	31,86	25,42	31,84	26,66	31,85	25,76
TELSONNE	R	252	110	362	30,06	29,30	34,44	31,62	30,94	29,85
QR TELSONNE	R	7	0	7	22,95	26,56	NA	NA	22,95	26,58
QR FINTER	R	11	2	13	23,73	26,78	25,81	20,65	24,43	24,89
RTM	R	56	27	83	23,22	32,51	21,35	31,98	22,25	32,25

Tableau 13. Répartition en locuteurs , WER moyens et écarts-types pour le Corpus 2. (Le type correspond au média, T pour télévision et R pour radio)

Corpus 2		**Nombre de loc.**			**WER**					
Emissions	**Type**	**Hommes**	**Femmes**	**Tous**	**Hommes**		**Femmes**		**Tous**	
					$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$	$\bar{x}$	$\sigma(x)$
tvme	R	26	19	45	24,52	29,72	17,51	25,58	21,64	28,29
africa1	R	31	7	38	31,48	33,77	20,94	31,04	29,20	33,47
csoj.16k	T	13	2	15	36,19	26,43	35,36	31,04	36,19	26,33
CultureEtVous	T	68	46	114	41,93	31,36	59,87	37,27	49,18	34,97
PlaneteShowbiz	T	74	26	100	40,09	33,09	71,25	38,41	49,68	37,65
Arte News	T	14	13	27	16,88	18,80	11,88	14,70	14,1	16,75
Fab Histoire	R	6	14	20	27,42	24,85	25,54	27,32	26,58	25,98
Temps Pauchon	R	14	2	16	42,59	38,42	13,51	7,44	35,23	35,71
TV8 Pl du Village	T	18	6	24	54,80	41,43	30,71	29,44	53,51	41,23

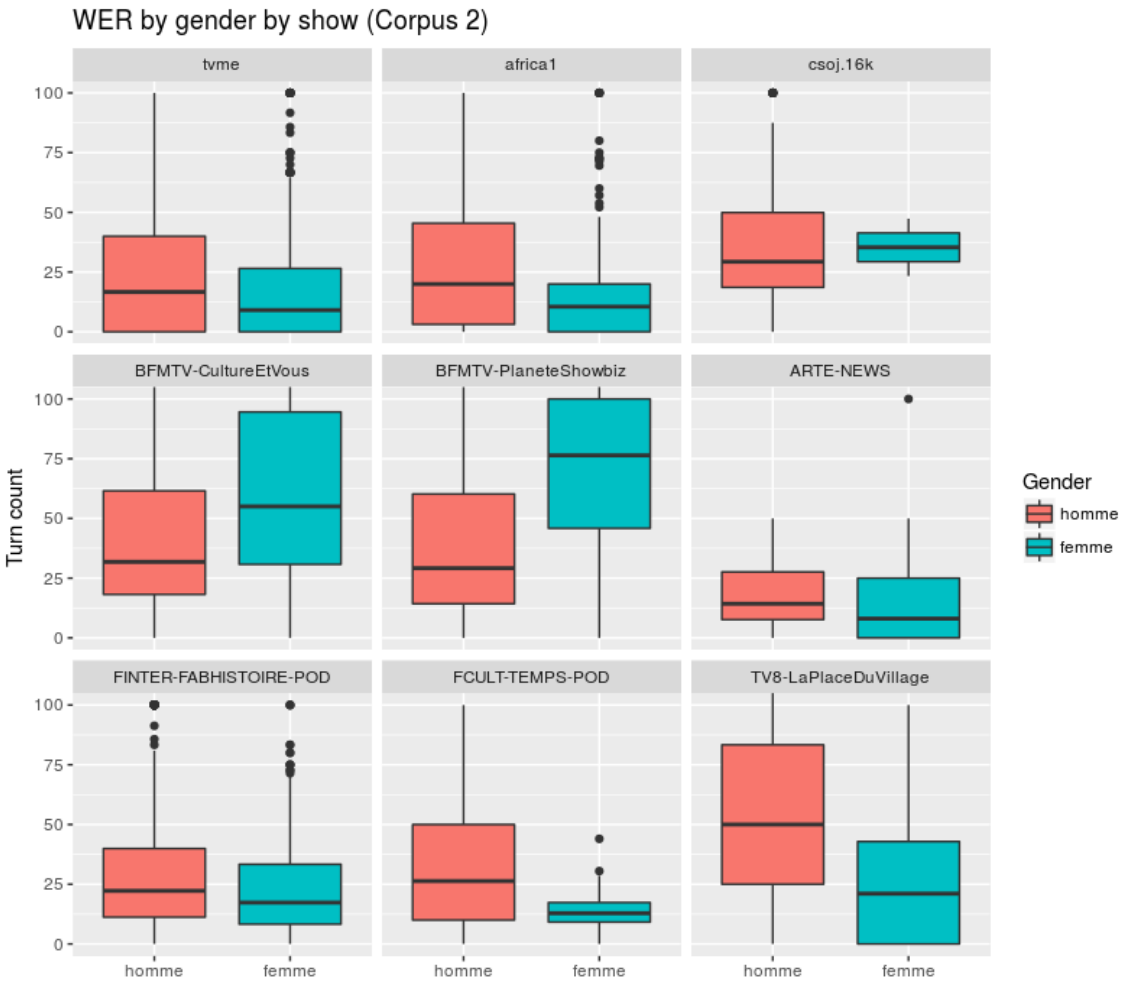
Annexe 7 : Distribution des WER par émissions et par genre

Distributions pour le Corpus 1



Corpus	Nom	Émission	Média
1	rfi-fm-dga	News	Radio
	RFI-ELDA	News	Radio
	RTM-ELDA	News	Radio
	EST2BC-FRE-FINTER-DEBATE	Débats	Radio
	LCP-PileEtFace	Pile et Face	Télévision
	QRBC-FRE-FR-FRANCE3-DEBATE	Débats	Télévision
	QRBC-FRE-FR-FINTER-DEBATE	Débats	Radio
	QRBC-FRE-FR-TELSONNE-POD	Le Téléphone Sonne	Radio
	TELSONNE	Le Téléphone Sonne	Radio

Distribution pour le Corpus 2



Corpus	Nom	Émission	Média
2	tvme	News	Radio
	africa1	News	Radio
	csoj.16k	Ce Soir ou Jamais	Télévision
	BFMTV-CultureEtVous	Culture & Vous	Télévision
	BFMTV-PlaneteShowbiz	Planète Showbiz	Télévision
	ARTE-NEWS	News	Télévision
	FINTER-FABHISTOIRE-POD	La Fabrique de l'Histoire	Radio
	FCULT-TEMPS-POD	Un Temps de Pauchon	Radio
	TV8-LaPlaceDuVillage	La Place du Village	Télévision

MOTS-CLÉS : apprentissage automatique, traitement automatique de la parole, corpus, genre

RÉSUMÉ

Les systèmes d'IA sont développés sur des grands corpus de données et les technologies du traitement automatique de la parole n'échappent pas à cette règle. Mais ces grands corpus de données peuvent contenir des répartitions de genre non-équilibrées qui peuvent conduire au développement d'algorithmes discriminants. Les systèmes d'IA infiltrant de plus en plus notre quotidien, et la voix s'imposant comme la nouvelle interface homme/machine, il devient nécessaire de pouvoir étudier et quantifier l'impact de la répartition homme/femme dans les données d'apprentissage sur les performances des systèmes.

Ce mémoire propose donc dans un premier temps d'étudier la répartition des genres dans les grands corpus du français oral, et dans un second temps, d'évaluer l'impact de cette représentation sur les performances d'un système de reconnaissance automatique de la parole.

KEYWORDS : machine learning, automatic speech processing, corpus, gender

ABSTRACT

AI systems are trained on a huge amount of data, and speech processing technologies are no exception to the rule. However corpora may be statistically imbalanced regarding genders and this can lead to discriminative algorithms. With AI becoming ever more present in our everyday life, it seems more than necessary to be aware of the impact of gender representation in training data on the system's performances.

This masters' thesis proposes to study gender representation in large spoken french corpora and to estimate the impact of this distribution on the performances of an automatic speech recognition system.