lol thats how reddit talks;) : le site américain Reddit comme espace de variation de l'anglais : étude de corpus intersectionnelle et quantitative d'usages non standard, au prisme du genre, de l'âge et de l'ethnicité

Marie Flesch

ATILF-CNRS | École doctorale SLTC (Sociétés, Langages, Temps, Connaissances)

Thèse présentée et soutenue publiquement en vue de l'obtention du titre de docteure de l'université de Lorraine, Mention « Sciences du langage », par Marie Flesch, le 16 décembre 2020

lol thats how reddit talks ;) : le site américain Reddit comme espace de variation de l'anglais.

Étude de corpus intersectionnelle et quantitative d'usages non standard, au prisme du genre, de l'âge et de l'ethnicité

Jury

Président du jury :
Luca Greco, professeur des universités, université de Lorraine

Rapporteuses :
Maria Candea, maitresse de conférences HDR, université Sorbonne Nouvelle
Natalie Kübler, professeur des universités, université de Paris

Examinatrice :
Fabienne Baider, professeure des universités, université de Chypre

Directrice de thèse :
Sophie Bailly, professeure des universités, université de Lorraine

Résumé

Cette thèse étudie les relations entre les pratiques d'écriture en ligne et le genre sur le site communautaire américain Reddit. Elle s'appuie sur un corpus de près de 20 millions de tokens comprenant les commentaires en anglais de 1044 internautes, qui inclut les contributions de 300 personnes transgenres et non binaires. Dans une perspective intersectionnelle, des variables peu souvent prises en compte dans les études sociolinguistiques quantitatives du genre, comme l'âge et l'ethnicité, ont été intégrées aux analyses. Les variables linguistiques étudiées comprennent 11 variations par rapport à la langue écrite standard : 6 procédés d'ajout (émoticônes, émojis, étirements de lettres, étirements de ponctuation, mots en majuscules et interjections) et 5 procédés de réduction (abréviations, graphies phonétiques, *g-droppings*, omissions d'apostrophe et omissions de la majuscule du pronom personnel *I*). En complément de ces analyses linguistiques, la thèse propose une exploration quantitative de l'identité en ligne des Redditors. Elle s'intéresse ainsi aux marqueurs les plus visibles de l'activité des Redditors dans la communauté, dont leurs pseudonymes, leurs centres d'intérêt, leur « karma », la longévité de leurs comptes et la modération de forums. Les analyses, qui s'appuient principalement sur la méthode de la régression multiple, montrent notamment que femmes et hommes transgenres s'alignent rarement sur les femmes et hommes cisgenres. L'intégration de l'ethnicité aux analyses permet par ailleurs de dresser un tableau nuancé des pratiques d'écriture des femmes et des hommes, et montre la pertinence de l'étude de l'interaction du genre avec d'autres variables sociodémographiques. Nos résultats suggèrent ainsi que les femmes afro-américaines et hispaniques jouent un rôle de premier plan dans la diffusion des formes innovantes de la CMC.

Abstract

Title : *lol thats how reddit talks ;)* : variation in English on the American community website Reddit. A quantitative and intersectional study of eleven non-standard variables through the lens of gender, age, and ethnicity

This thesis studies the relationships between non-standard online writing practices and gender on the American community website Reddit. It is based on a corpus of nearly 20 million tokens which contains the comments written in English by 1,044 internet users, including 300 transgender and non-binary people. Using an intersectional sociolinguistic approach, it examines the interaction of gender with age and ethnicity. Eleven non-standard variables were investigated : six additive processes (emoticons, emojis, letter lengthenings, punctuation lengthenings, all caps and interjections) and five reduction processes (abbreviations, phonetic spellings, g-droppings, apostrophe omissions and lower case spellings of the pronoun “I”). In addition to these linguistic analyses, the thesis explores how Internet users construct their virtual identities and occupy the Reddit space, by focusing on the most visible markers of Redditors’ activity in the community : pseudonyms, interests, “karma”, longevity of the accounts and forum moderation. The analyses, which are mainly based on the multiple regression method, provide a nuanced account of the way Redditors use non-standard language to index their gender identity. They show, in particular, that transgender women and men rarely align with cisgender women and men. They also suggest that Hispanic and African-American women play a major role in the spread of non-standard spelling and typography.

Remerciements

Mes remerciements vont tout d'abord à ma directrice de thèse Sophie Bailly, pour son accompagnement bienveillant, pertinent et enthousiaste pendant ces quatre années. Ils vont aussi aux membres du jury, qui ont accepté de lire mon travail. Je remercie ensuite tou·tes les collègues de l'ATILF qui m'ont écoutée, encouragée, et dont les remarques ont permis de faire progresser ma thèse. Je remercie tout particulièrement Bertrand Gaiffe, l'architecte du corpus, dont les conseils m'ont été très précieux ; Alex Boulton, mon directeur de mémoire de Master 2, dont cette thèse est une continuation ; Guillaume Nassau, pour son travail de codage des pseudonymes ; et Véronique Lemoine-Bresson, pour son enthousiasme et son invitation à présenter mon travail à ses étudiant·es. Je remercie également les collègues de l'Espace Quanti et plus particulièrement Simon Paye, pour m'avoir motivée à apprendre R, et Jean-Luc Kop, pour ses explications éclairantes sur la régression. Je remercie mes amies Samantha Ruvoletto pour ses relectures, et Sarah Kremer pour ses conseils sur la mise en page et la typographie. Je remercie aussi tou·tes mes ami·es et mes proches, et en particulier mes parents, mes frères, ma grand-mère, et Philippe, Pénélope et Constantin, pour leur patience et leur soutien tout au long de ce long voyage.

À Hélène et Marie-Paule

Table des matières

Introduction

« How to tweet like a girl » : tel est le titre d'un article paru le 19 février 2013 sur *The Cut* (Stoeffel, 2013), un des sites internet du magazine *New York*. Son auteure nous explique comment « twitter comme une fille ». La recette est très simple. Elle tient en 285 mots, et en huit instructions : il faut, entre autres, exprimer ses émotions, utiliser abondamment les émoticônes, allonger ses mots, multiplier les interjections — et, surtout, éviter les jurons. L'article de *The Cut* a été inspiré par la parution d'une étude (Bamman et al., 2013, republiée en 2014) que la journaliste n'a pas jugé nécessaire de consulter, s'appuyant à la place sur un autre article de la presse populaire (Heaney, 2013). À quoi bon, de toute façon ? Le texte de Bamman et al., souligne ironiquement Stoeffel, est une « groundbreaking study from the University of Duh », une « étude révolutionnaire réalisée par l'Université des Enfonceurs de Portes Ouvertes ». Les femmes et les hommes twittent différemment ? C'est une évidence !

La preuve ? Les spécialistes du traitement automatique des langues peuvent aujourd'hui prédire le genre d'un internaute uniquement à partir de ses tweets. C'est ce que l'on appelle la stylométrie, ou l'*authorship attribution*. Et cela semble fonctionner assez bien, avec des taux de réussite qui tournent souvent autour des 80 % (Burger et al., 2011). Notons toutefois que, quand on mélange les genres, c'est-à-dire quand on analyse des corpus hétérogènes, la tâche est beaucoup plus compliquée (Overdorf & Greenstadt, 2016). Les méthodes utilisées dans ces recherches sont toutes aussi variées que les applications possibles, qui vont des enquêtes policières (*forensics linguistics*) au marketing (Rangel et al., 2015), et peuvent poser évidemment, des problèmes éthiques importants.

Dans une perspective plus descriptive, le mariage de la sociolinguistique à la linguistique de corpus et à la linguistique computationnelle s'est avéré extrêmement fertile. Aujourd'hui, les chercheur·es ne sont plus obligé·es, pour des raisons pratiques, de se contenter de corpus composés des interventions de quelques personnes. Grâce à l'essor des réseaux sociaux, les corpus de CMC (*computer-mediated communication*, ou communication médiée par ordinateur) comptent des millions, voire des milliards de mots, permettant d'analyser les pratiques d'écriture des femmes et des hommes à une échelle sans précédent.

Tout comme les spécialistes du traitement automatique des langues, les sociolinguistes de la CMC semblent donner raison à Stoeffel. Reprenons un à un les ingrédients du « style féminin » cités par son article. Les mots asso-

ciés aux émotions, comme *love* ou *babe*, et les interjections comme *haha* ou *aww* sont plus fréquemment utilisés par les femmes sur Twitter (Burger et al., 2011 ; Coats, 2017b). Les mots étirés (*soooooo*, *riiiight*) semblent l'apanage des femmes (Coats, 2017b ; Rao et al., 2010), et les hommes jurent plus que les femmes (Gauthier, 2017 ; Thelwall, 2008). Quant aux émoticônes, qui sont sans conteste les stars des études du genre et de la CMC, elles sont majoritairement produites par les femmes (Del-Teso-Craviotto, 2008 ; Oleszkiewicz et al., 2017 ; Tossell et al., 2012).

Notons, évidemment, que les résultats des études quantitatives du genre et de la CMC sont souvent nuancés, et parfois très contrastés. Ce qui l'est moins, c'est la façon dont ces travaux conçoivent le genre. Ils adoptent, pour leur immense majorité, la perspective essentialiste de la sociolinguistique variationniste « classique », qui catégorise le genre comme un phénomène fixe et binaire, et qui, de plus, décide des variables linguistiques à observer en amont, trouvant inévitablement ce qu'elle cherche. Ce parti pris a été nécessaire : l'internet grand public n'existe que depuis les années 1990, et les grands réseaux sociaux ont une quinzaine d'années. Pour explorer ce terrain émergent, qui efface le corps et impose de nouvelles contraintes, il a fallu établir une fondation. Les études quantitatives essentialistes permettent de connaitre les façons dont la variation et le changement linguistique s'y affichent. Ce qu'écrivent Meyerhoff et Stanford (2015) sur les langues peu étudiées pourrait donc s'appliquer à la CMC : pour comprendre les significations sociales de la variation, il faut savoir quels aspects de la variation ont une signification sociale. Le travail variationniste essentialiste est ainsi « a necessary first step [...] for the more socially situated analysis of a third wave approach » (p. 11). En établissant des corrélations entre les choix linguistiques des internautes et les catégories macrosociales, les sociolinguistes quantitatif·ves de la CMC ouvrent la voie à des études qualitatives plus fines de la performance du genre.

Toutefois, le travail quantitatif n'a pas à être contraint par une perspective essentialiste. C'est ce que montre l'étude de Bamman et al. (2014), qui, contrairement à ce que laisse penser l'article de *The Cut*, n'a pas trouvé qu'il existe une façon de twitter « comme une fille » ou « comme un garçon ». Les chercheurs adoptent une approche innovante. Employant la technique du *clustering*, ils regroupent 14 000 utilisateur·trices de Twitter en fonction de leur style d'écriture et de leurs centres d'intérêt. Cela leur évite de comparer un groupe monolithique de femmes avec un groupe monolithique d'hommes, et de se focaliser tout particulièrement sur celles et ceux qui transgressent les normes linguistiques du genre et de la CMC. Ils montrent ainsi que, sur Twitter, le genre émerge dans la manière dont les internautes se positionnent par rapport aux personnes avec qui elles et ils interagissent et aux sujets dont elles et ils parlent, et que, parfois, elles et ils s'écartent délibérément d'une vision normative du genre, par leurs réseaux et leurs pratiques linguistiques.

Notre projet est, dans son objectif, similaire à celui de Bamman et al. : nous souhaitons explorer la façon dont les internautes construisent leur identité de genre par leurs pratiques d'écriture en anglais, tout en évitant

d'opposer systématiquement femmes et hommes. Nous adoptons une démarche résolument quantitative, principalement basée sur la méthode inférentielle de la régression, utilisée en sociolinguistique variationniste depuis les années 1970. Nous nous intéressons à des phénomènes d'écriture non standard (par rapport à l'anglais écrit enseigné à l'école) souvent décrits comme étant caractéristiques de la CMC, comme les émoticônes, les émojis, les acronymes, les omissions d'apostrophe ou encore les étirements graphiques.

L'originalité de notre thèse repose sur deux grandes orientations, qui ont rarement été mises en application dans les études sociolinguistiques quantitatives. Tout d'abord, nous avons décidé de briser la binarité avec laquelle le genre est généralement étudié dans la CMC en nous intéressant à cinq groupes d'internautes : les femmes et les hommes cisgenres, les femmes et les hommes transgenres, et les personnes non binaires. L'objectif était double : d'une part, nous souhaitions intégrer à nos analyses des catégories de la population fréquemment écartées, ou rendues invisibles, dans les études sociolinguistiques quantitatives du genre. De l'autre, comme l'a montré le travail de Zimman (2017) sur la voix des hommes transgenres, étudier les personnes transgenres permet de complexifier et d'enrichir notre compréhension de la façon dont les éléments du langage deviennent genrés. La manière dont ces « transgresseur·es » construisent leur identité, en ligne, peut nourrir la recherche sur le genre et le langage en général.

La seconde orientation que nous avons prise est celle de l'intersectionnalité. En effet, si certaines variables linguistiques sont genrées, elles sont utilisées par des personnes qui ne sont pas uniquement des femmes ou des hommes. L'approche intersectionnelle établit des liens avec d'autres facettes de l'identité des internautes, en étudiant leurs intersections. Ici, nous nous intéressons principalement aux intersections entre genre, âge, et ethnicité (ce que le contexte nord-américain auquel nous nous intéressons nous permet). La statistique nous fournit un outil précieux pour étudier ces intersections : la notion d'interaction entre deux variables (ou plus), qui, seule, est à même d'explorer la diversité des identités de façon quantitative. Intégrée aux modèles de régression, elle ne considère pas les femmes et les hommes comme des groupes homogènes ; elle révèle les dynamiques complexes de la langue et de l'identité, et rend « visible l'invisible » (López et al., 2018). En sociolinguistique, l'intersectionnalité est une approche relativement nouvelle, qui a principalement été appliquée à des études qualitatives (Levon, 2015). Nous souhaitons montrer, par cette thèse, qu'elle a également une pertinence dans des méthodologies sociolinguistiques quantitatives. En alliant cette approche à l'étude de la non-conformité de genre, nous espérons contribuer, à notre niveau, au projet de recherche intersectionnel quantitatif tel qu'il a été décrit par Bowleg et Bauer (2016, p. 340) :

> « Conducting rigorous quantitative intersectionality research transcends technique and methodology, alone. It involves, we believe, a commitment to methodologies that undo the empirical and evidentiary erasure of multiple marginalized groups. »

Pour mettre notre projet en œuvre, il nous fallait un corpus. Pour le

créer, nous avons choisi Reddit, un espace qui, contrairement à Twitter, a encore assez peu attiré l'attention des linguistes. Comme le site de microblogage, Reddit a l'avantage d'être public, et de permettre aux chercheur·es d'utiliser ses données librement. Mais Reddit n'est pas un réseau social ; c'est un « site web communautaire » composé de plus de deux million de forums. On ne va pas sur Reddit pour « suivre » des célébrités ou se faire des amis. Reddit est un site où on échange, où on questionne, où on raconte, et où on argumente avec une communauté d'internautes anonymes (ou, pour être plus exacte, « pseudonymes »), le tout avec une grande liberté de ton.

Reddit est né en 2005 de l'idée de deux étudiants de l'université de Virginie, deux « geeks » qui souhaitaient permettre à leurs pairs de partager facilement des liens vers des contenus intéressants. La suite de l'histoire ressemble, de loin, à bien des *success story* de la Silicon Valley : les deux jeunes entrepreneurs sont devenus multimillionnaires, l'un d'eux a épousé une star du tennis, et leur site est consulté quotidiennement par des millions d'internautes du monde entier. De près, les choses sont moins reluisantes. Imprégné par une culture geek libertaire, Reddit a longtemps refusé de modérer les discussions qui s'y déroulent, devenant un terreau fertile pour les haines. C'est un des berceaux du mouvement *incel* (« célibataires involontaires »), profondément misogyne, et de la pensée d'extrême droite *redpill*. Le racisme, l'homophobie, la transphobie, et la violence contre les femmes et les minorités s'épanouissent dans de nombreux forums, et les théories du complot prolifèrent. Le *doxxing* (la révélation d'informations personnelles) menacent celles et ceux qui élèvent la voix face à la haine, et les *trolls* font volontiers irruption dans les forums dédiés aux femmes pour semer le chaos. En même temps, par son fonctionnement décentralisé, Reddit a permis l'éclosion de nombreux espaces qui promeuvent la diversité : les féministes, les internautes issu·es de minorités et les personnes transgenres et non binaires ont créé des forums dans lesquels elles et ils échangent des informations, s'entraident, se conseillent, et se racontent.

C'est cette liberté d'écriture, cette multiplicité d'identités et de points de vue, affichés sans complexe, qui nous a permis de constituer un corpus diversifié, capable de répondre à nos objectifs de recherche. En construisant ce corpus, que nous avons baptisé « RedditGender », nous avons recueilli des données sociodémographiques sur les internautes, mais aussi des informations sur leur identité de « Redditors ». Cela nous a ouvert de nouvelles perspectives : ne pas uniquement étudier comment les Redditors écrivent, mais aussi quels sont leurs itinéraires sur le site et leurs centres d'intérêt, quel est leur statut dans la communauté, et comment elles et ils s'inventent une identité dans le cyberespace.

Cette thèse s'articule en quatre parties. La première partie est notre cadre théorique. Son premier chapitre commence par faire le point sur la notion de genre, puis dresse un état des lieux de la recherche linguistique sur le genre et sur la sociolinguistique variationniste, avant de s'intéresser à l'approche intersectionnelle. Le second chapitre est consacré à la CMC. Il présente cet objet d'étude ainsi que les méthodes utilisées par les chercheur·es pour recueillir des informations sociodémographiques sur des in-

ternautes souvent anonymes et pour étudier leurs centres d'intérêt ; il décrit ensuite les principales caractéristiques de la langue utilisée dans la communication électronique, en se concentrant plus en détail sur les types d'écarts par rapport à l'anglais standard que nous avons choisi d'étudier dans cette thèse. Le cadre théorique se conclut par une présentation de Reddit, qui passe en revue son historique, son fonctionnement, et sa dimension « geek ».

La partie « Méthodologie » est elle aussi structurée en trois parties. Tout d'abord, elle décrit le processus d'échantillonnage, la construction du corpus et les outils utilisés pour son exploitation. Ensuite, elle liste les trois types de variables que nous avons intégrées à nos analyses : les variables sociodémographiques qui servent de variables explicatives, les variables de l'identité en ligne (que nous avons appelée « Redditentité »), et les variables linguistiques. Enfin, cette partie se termine par un exposé des techniques statistiques descriptives et inférentielles utilisées dans la thèse.

La troisième partie est intitulée « Identités et itinéraires ». Elle explore, à partir des données non linguistiques recueillies lors de la construction du corpus, la façon dont les Redditors investissent le site. Dans un premier temps, nous nous intéressons à l'identité virtuelle des internautes en étudiant leurs pseudonymes, la longévité de leurs comptes Reddit, leur choix d'être ou non modérateurs ou modératrices bénévoles du site, et leur « karma », symbole de statut sur le site. Dans un second temps, nous essayons de savoir comment ils occupent l'espace sur le site par l'étude de leurs centres d'intérêt, de leur mobilité, et de la longueur de leurs commentaires.

La quatrième partie est consacrée aux analyses linguistiques. Elle commence par proposer une analyse de l'ensemble des phénomènes non standard étudiés. Elle se concentre ensuite sur six procédés d'« ajout » : émoticônes, émojis, étirements de lettres et étirements de ponctuation, mots en majuscules, et interjections. Puis, elle explore cinq procédés de « réduction » : abréviations, graphies phonétiques, g-droppings, omissions d'apostrophe et omissions de la majuscule du pronom personnel *I*. Enfin, dans un dernier chapitre, nous proposons une synthèse et une interprétation des résultats de nos analyses linguistiques. La thèse se clôt par une conclusion générale.

Chaque chapitre se termine par une courte synthèse intitulée, en clin d'œil à notre objet d'étude,« tl ;dr », l'acronyme de *too long, didn't read*, utilisé sur Reddit et dans les forums de discussion pour résumer de longs commentaires.

Note sur l'écriture inclusive. Nous avons adopté dans cette thèse plusieurs principes de l'écriture inclusive. Nous utilisons le point médian pour accorder noms et adjectifs en genre et en nombre, de la façon suivante : « Américain·es », « modérateur·trices », « auteur·es ». Nous dédoublons également généralement les pronoms (« elles et ils », « eux et elles »). Nous employons le terme anglais « Redditor » comme nom épicène désignant les utilisateur·trices de Reddit, quelle que soit leur identité de genre.

Première partie

Cadre théorique

Chapitre 1

L'approche intersectionnelle du genre et du langage

Dans ce chapitre, nous présentons l'approche intersectionnelle du genre et du langage que nous adoptons dans nos analyses et dans leur interprétation. Il commence par une plongée dans la biologie complexe du sexe pour ensuite explorer la construction sociale du genre et la non-conformité de genre. Il propose ensuite un bref historique des recherches sur le genre et le langage, de la sociolinguistique variationniste, et de la contribution de la linguistique de corpus à l'étude du genre. La troisième section est consacrée à l'intersectionnalité ; elle se focalise sur son application en linguistique et sur les problématiques de la recherche intersectionnelle quantitative. La quatrième et dernière section porte sur les deux variables sociales dont nous avons choisi d'étudier les interactions avec le genre : l'âge et l'ethnicité, ici présentée dans le contexte nord-américain.

1.1 La construction du genre

Parce qu'il est profondément imbriqué dans les institutions, les actions et les croyances, le genre apparait comme un phénomène naturel et évident. En réalité, il est le résultat d'un processus de construction : ce n'est pas une cause, mais un effet (Eckert & McConnell-Ginet, 2003, p. 9), qui commence avec l'exagération des différences biologiques entre femmes et hommes.

1.1.1 La notion de sexe

Si la notion de sexe est ancrée dans la biologie, ce n'est pas pour autant que le sexe est simple, binaire, et qu'il divise l'humanité en deux groupes homogènes.

Du biologique au social

La complexité du sexe biologique a été mise en évidence pour la première fois par le psychologue John Money et ses collègues Joan Hampson et John Hampson (1972), qui étaient pionnier·es dans l'étude des personnes intersexe. Fausto-Sterling (2012), se basant sur leurs travaux, explique qu'à la naissance, le sexe du bébé est composé de cinq « couches » : le sexe chromosomique, le sexe fœtal, le sexe gonadique, le sexe hormonal et le sexe reproductif interne (p. 5). À la puberté se rajoutent deux composantes supplémentaires : le sexe hormonal pubertaire et le sexe morphologique pubertaire. Parfois, une de ces « couches » du sexe se développe indépendamment des autres, et un bébé intersexe nait. On considère ainsi que 1.7 % des bébés naissent avec un corps différent du corps féminin ou masculin standard, à cause de particularités chromosomiques (deux chromosomes XX pour un garçon, par exemple) ou hormonales (l'insensibilité aux androgènes), ou parce qu'ils présentent des combinaisons variées d'organes génitaux ou reproductifs. La fréquence de l'intersexualité varie en fonction des populations ; l'hyperplasie congénitale des surrénales (qui provoque une masculinisation des organes génitaux d'enfants aux chromosomes XX) est par exemple relativement fréquente en Alaska, et l'hermaphrodisme est fréquent en Afrique du Sud (Blackless et al., 2000).

Quand nait un bébé qui ne présente pas un sexe féminin ou masculin standard, les adultes tentent de le faire conformer le plus possible à une des deux catégories, par la chirurgie ou l'endocrinologie. Les standards sont plus stricts pour les bébés nés avec un appareil génital masculin, qui est souvent féminisé sans prendre en compte les autres caractéristiques sexuelles (Eckert & McConnell-Ginet, 2003). Tout cela montre que les catégories biologiques de « femme » et d'« homme » ne sont pas déterminées par un seul et unique critère ; elles sont basées sur une combinaison de critères biologiques, mais aussi sur des croyances culturelles. Au bout du compte, la décision d'assigner un sexe masculin ou féminin à une personne est un acte social :

> « Labeling someone a man or a woman is a social decision. We may use scientific knowledge to help us make the decision, but only our beliefs about gender — not science — can define our sex » (Fausto-Sterling, 2000, p. 3).

Dimorphisme versus différence

Certain·es chercheur·es ont tendance à considérer les différences comportementales ou anatomiques entre femmes et hommes comme des dimorphismes, c'est-à-dire des divergences importantes entre les deux groupes, qui auraient très peu de points communs (McCarthy & Konkle, 2005, cité par Fausto-Sterling, 2012). Dans sa méta-analyse d'études portant sur les différences entre filles et garçons et femmes et hommes, la psychologue Janet Hyde (2005) a montré que c'était loin d'être vrai. Dans l'immense majorité des cas, les différences sont réduites, à l'exception de l'agressivité physique, de la distance à laquelle on peut lancer des objets, et de l'atti-

tude par rapport aux relations sexuelles occasionnelles. On ne peut donc pas parler de dimorphisme sexuel, mais plutôt de différences faibles entre les sexes (Fausto-Sterling, 2012). Le problème, pour Eckert et McConnell-Ginet (2003), c'est que les différences d'origine biologique entre femmes et hommes occupent une place prédominante dans la recherche scientifique et sociale, et que ces différences sont ensuite amplifiées par les médias, créant une dichotomie entre les femmes et les hommes, et effaçant les similitudes entre les deux catégories :

> « Sex difference is being placed at the center of activity, as both question and answer, as often flimsy evidence of biological difference is paired up with unanalyzed behavioral stereotypes. And the results are broadcast through the most august media as if their scientific status were comparable to the mapping of the human genome » (p. 13).

1.1.2 Le genre

De la notion de sexe, nait le genre. Le lien entre les deux n'a toutefois rien de naturel : les chromosomes, l'anatomie et les hormones ne déterminent pas le type de vêtements que l'on porte, les couleurs que l'on préfère, le métier que l'on pratique ou la façon dont on parle. Le genre est créé de façon sociale par l'exagération des différences biologiques entre femmes et hommes.

Le développement de l'expression de genre

En plus de mettre en lumière la complexité du sexe biologique, le travail de Money a permis de faire émerger la notion de genre. Le psychologue a remarqué qu'à la naissance, quand les adultes identifient le sexe du bébé en se basant sur son anatomie externe, ils amorcent le processus de la socialisation de genre (*gender socialisation*). Aujourd'hui, évidemment, ce processus commence souvent avant la naissance, avec l'échographie qui permet de déterminer le « sexe » du bébé en se basant sur son sexe génital, et avec l'acte linguistique du choix du prénom (Eckert & McConnell-Ginet, 2003). Fausto-Sterling (2012) note que la réaction des adultes face au sexe génital des bébés est intense, et se manifeste notamment par les vêtements et les jouets que l'on achète ; c'est ce qu'elle appelle la « gender fortification » (Fausto-Sterling, 2012, p. 10). Au début de la vie d'un enfant, ce sont donc les adultes qui font le « travail du genre » pour elle ou lui, en la ou le traitant comme une fille ou comme un garçon, et en interprétant ses actions au travers du prisme du genre. Ensuite, l'enfant prend le relais en effectuant sa propre socialisation de genre, et en participant à construire celle des autres (Eckert & McConnell-Ginet, 2003).

Ainsi, dès l'âge de 18 mois, la plupart des petites filles commencent à s'intéresser aux poupées, et les petits garçons aux petites voitures, dans certains pays en tout cas. Ces préférences ne sont pas universelles : une étude a ainsi montré que les petit·es Hollandais·es avaient des choix de jouets moins stéréotypiquement masculins ou féminins que les petit·es Italien·nes (Zammuner, 1987, cité par Nelson, 2005). À partir de l'âge de trois

ans, les enfants peuvent répondre correctement à la question « est-ce que tu es une fille ou un garçon ? », et se tournent vers des activités « appropriées » à leur genre, comme jouer à la dinette (Fausto-Sterling, 2012). Tout cela semble souvent normal aux adultes ; par exemple, le fait que les petites filles préfèrent le rose, et les garçons le bleu, s'impose comme une évidence.

Et pourtant, les couleurs sont l'exemple parfait du caractère arbitraire de la socialisation de genre. Jusqu'au début du 20ème siècle, c'était l'inverse : le bleu était considéré comme une couleur délicate, et on habillait les petits garçons en rose, couleur « forte » (Fausto-Sterling, 2012). Les couleurs sont également un bon exemple de l'asymétrie avec laquelle le genre est renforcé. Une étude (Rust et al., 2000, cités par Fausto-Sterling, 2012) a montré que les garçons de 4 à 11 ans préféraient, entre autres, le noir, le bleu et le brun, et les filles le rose et le violet. Toutefois, les filles choisissaient également parfois le bleu, tandis que les garçons choisissaient très rarement le rose : c'est le résultat de la dévalorisation des comportements liés aux filles (Fausto-Sterling, 2012). À cause de cette asymétrie liée à la valorisation des comportements masculins et à la dévalorisation des comportements féminins, les filles ont souvent plus de latitude que les garçons dans leurs choix de vêtements ou de jeux. Elles peuvent adopter des activités et des attitudes jugées comme étant « masculines » sans remontrances de la part des adultes et de leurs pairs ; elles sont alors vues comme des « garçons manqués ». Les garçons, en revanche, ne peuvent pas aussi librement adopter des comportements considérés comme « féminins » ; ceux-ci sont « marqués », et réservés uniquement à une partie de la population, là où les comportements « masculins » sont vus comme étant « non marqués » et « normaux » (Eckert & McConnell-Ginet, 2003).

1.1.3 La non-conformité de genre

Identité de genre et non-conformité de genre

Dans la sous-section précédente, nous avons parlé de la façon dont le genre est construit au travers, notamment, de l'expression de genre, c'est-à-dire de la manière dont on se comporte pour exprimer le genre (Stryker, 2017). Nous abordons ici le concept d'identité de genre. Celle-ci apparait après l'expression de genre, vers l'âge de 3 ans, et se solidifie dans les années qui suivent (Fausto-Sterling, 2012). Dans certains cas, l'identité de genre d'une personne ne correspond pas au sexe qui lui a été assigné à la naissance ; cela se produit chez des personnes qui présentent une forme d'intersexualité comme chez des personnes n'en présentent pas. Ce sentiment de décalage peut émerger très tôt (dès trois ans), et concernerait, aux États-Unis 0.9 % des garçons et 1.7 % des filles (Zucker & Cohen-Kettenis, 2008, p. 381). Appelé « dysphorie de genre », il provoque de l'inconfort et une anxiété due au fait de ne pas vivre dans son vrai genre (Braun, 2019). Quand il persiste, les personnes qui en souffrent sont dites « transgenres ».

La formation de l'identité de genre chez les personnes transgenres reste un processus mystérieux. Certain·es scientifiques pensent qu'elle a une origine biologique, et qu'elle résulte notamment des effets des hormones

prénatales sur le cerveau ; cela n'a toutefois pas été prouvé par les études réalisées post-mortem sur les cerveaux de femmes transgenres (Fausto-Sterling, 2012). D'autres chercheur·es émettent l'hypothèse que le décalage entre le sexe assigné à la naissance d'une personne et son identité de genre pourrait être causé par des interactions sociales pendant la petite enfance et l'enfance. Il n'y a donc pas de consensus, et il est tout à fait possible que le sentiment d'inadéquation entre l'identité de genre et le sexe assigné ait des origines différentes chez différentes personnes. Fausto-Sterling émet l'hypothèse qu'il est dû à des variations dans le développement neurosensoriel, en combinaison avec des psychodynamiques familiales (Fausto-Sterling, 2012, p. 67).

Les mots du genre

C'est dans les années 1990 que le terme « transgenre » a commencé à être utilisé pour désigner les personnes ressentant une dissonance entre leur sexe assigné à la naissance et leur identité de genre. Auparavant, il était employé pour parler des travestis (qui changent leur expression de genre par la façon dont ils s'habillent) et les transsexuels (qui transforment leur corps par la prise d'hormones ou par la chirurgie) (Stryker, 2017). Le mouvement transgenre a notamment gagné en visibilité avec la parution d'autobiographies, comme celle de Kate Bornstein, *Gender Outlaw* (1994). Il est complexe, et regroupe plusieurs identités de genre : les femmes et les hommes transgenres, mais aussi les personnes agenres, qui ont le sentiment de ne pas avoir d'identité de genre, et les personnes non binaires et genderqueer, qui ne se reconnaissent pas dans une vision binaire du genre, selon laquelle il n'existe que deux genres. Notons que toutes les personnes non binaires ne s'identifient pas comme des personnes transgenres (Stryker, 2017). Les personnes non binaires peuvent être AFAN (assignées filles/femmes à la naissance, traduction de l'anglais *AFAB*, *assigned female at birth*) ou AGAN/AHAN (assignées garçons/hommes à la naissance, traduction d'*AMAB*, *assigned male at birth*) (Braun, 2019). Depuis le début du 21ème siècle, on utilise le terme « cisgenre » comme synonyme de « non transgenre ». Ce terme est aujourd'hui souvent employé par celles et ceux qui se considèrent comme des « alliés » des personnes transgenres et non binaires (Stryker, 2017).

Le rôle essentiel d'internet pour les personnes transgenres

Dans les années 1990, internet a joué un rôle fondamental dans l'émergence de l'activisme transgenre aux États-Unis : il a permis aux personnes transgenres de communiquer facilement, de construire des réseaux (Whilchins, 2004), et de sortir ainsi de leur statut de simples « objects of medicalization » (Shapiro, 2004). Shapiro (2004) recense dès 2002 plus de 800 000 sites internet, listes de diffusion et chat rooms dédiés aux personnes transgenres, et souligne que le web est une source d'information sans précédent. Il permet d'affronter les défis administratifs et médicaux, et de diffuser à grande échelle les cas de violence et de discrimination dans le monde du tra-

vail (Whittle, 1998). Internet a aidé les activistes transgenres à s'organiser, en dépassant les contraintes financières et géographiques. L'anonymat et la sécurité relative offerts par le cyberespace ont permis une désinhibition du mouvement activiste transgenre : les internautes transgenres ont pu s'engager sans peur de l'« outing » ; en fait, ils n'avaient même pas besoin d'être « out » pour militer.

Enfin, internet est un espace précieux sur le plan identitaire. Il permet aux personnes transgenres qui cachent leur identité de genre dans le monde réel ou qui souffrent de harcèlement de vivre pleinement en tant que femmes et hommes (Shapiro, 2004). Dans son analyse de newsgroups transgenres israéliens, Marciano (2014) note qu'internet est pour les personnes transgenres une « sphère alternative », « a parallel world that provides its inhabitants with different and sometimes contradictory experiences from those available in the offline world » (p. 830). Parmi ces expériences, il cite le fait de vivre son identité de genre et de construire des relations romantiques. Il observe que certaines femmes transgenres affirment cacher leur statut transgenre en ligne, ce qui leur permet de se sentir comme de « vraies femmes biologiques », ce que la chirurgie de réattribution sexuelle ne leur offre bien souvent pas entièrement. Whittle (1998) souligne le caractère paradoxal de la situation :

> « Ironically, the cyberworld in which others have to learn how to manage their virtuality, is a world in which the transgender person's actual identity can thrive » (p. 392).

Il va jusqu'à dire qu'internet permet aux personnes transgenres de pleinement comprendre leur identité de genre, écrivant que « the mechanics of the new identity formation that has taken place in the community could not have existed outside of cyberspace » (p. 405). Shapiro (2004) note par ailleurs que la facilité avec laquelle les personnes transgenres peuvent se rassembler et échanger en ligne peut faire émerger une fausse impression de sécurité et d'acceptation.

Internet, en combinaison avec la visibilité grandissante des personnes transgenres dans l'actualité et la culture populaire, a provoqué un changement générationnel. Dans leur étude réalisée à grande échelle aux États-Unis, Rankin et Beemyn (2012) ont analysé plus de 3500 questionnaires et de 400 entretiens réalisés avec des personnes transgenres et non binaires. Les chercheur·es remarquent que les identités de genre sont aujourd'hui de plus en plus diversifiées, grâce au pouvoir informationnel du web. Ils notent que les participant·es qui ont grandi avant les années 1980 n'avaient pas les mots pour décrire leur identité de genre ; un homme transgenre qu'ils ont interrogé pensait par exemple qu'il était une lesbienne « butch ». Plus des deux tiers des participant·es les plus âgé·es ne connaissaient pas d'autres personnes transgenres avant de s'identifier en tant que transgenre, tandis que 69 % des participant·es les plus jeunes connaissaient une ou plusieurs personnes transgenres avant de faire leur *coming out*. Pour Spack et al. (2012), la croissance de la communauté transgenre en Amérique du Nord est également due au fait que de plus en plus d'enfants transgenres sont élevés dans des environnements compréhensifs.

La non-binarité

La non-binarité est installée depuis longtemps dans certaines cultures. C'est par exemple le cas chez plusieurs peuples amérindiens, dont la tribu sioux Lakota, les Navajos et les Zunis, qui reconnaissent l'existence de « two-spirit people », un terme générique relativement récent qui désigne plusieurs identités et rôles de genre. Dans la République Dominicaine, il existe une troisième catégorie, appelée « guevedoche » ou « machihembra », qui désigne des enfants nés avec des caractéristiques féminines mais qui développent des traits masculins à la puberté. En Inde, les hijras, qu'elless soient intersexes ou nés AGAB, constituent également un troisième groupe de genre, associé à des croyances religieuses. Citons aussi les kathoey de Thaïlande, des personnes AGAN qui s'identifient soit comme femmes soit comme une troisième catégorie de genre, les warias de la Sulawesi du sud-ouest, en Indonésie, qui sont des personnes AGAN qui adoptent diverses identités de genre féminines, et les machis du peuple Mapuche, au Chili et en Argentine, des guérisseurs qui épousent des identités de genre fluides (Richards et al., 2017).

La non-binarité n'a que récemment gagné en visibilité dans les pays occidentaux (Bosson et al., 2019). Aujourd'hui, plusieurs pays autorisent leurs citoyen·nes, même lorsqu'ils ou elles ne sont pas intersexes, à indiquer un genre « X » (ni féminin ni masculin) dans leurs documents officiels. C'est par exemple le cas du Canada depuis 2017 (Canada.ca, 2017) et de l'Islande depuis janvier 2020 (Fisher, 2019). Aux États-Unis, plusieurs états permettent à leurs résident·es de choisir le genre « X » sur leurs permis de conduire et/ou actes de naissance, dont l'Oregon, la Californie, le Nevada, l'Utah, l'Illinois et le Maine (Norwood, 2019). Des célébrités s'identifient ouvertement en tant que non binaires, comme l'actrice et mannequin Cara Delevingne (Foster, 2018) et Jonathan Van Ness, une des stars de l'émission *Queer Eye* de Netflix (Tirado, 2019).

Internet a également aidé l'émergence d'identités de genre multiples. De 2014 à 2015, Facebook a ainsi permis à ses membres américain·es et britanniques de choisir parmi plus de 50 identités de genre différentes, dont *bigender*, *gender fluid*, *gender nonconforming*, *trans person*, *two-spirit*, *transgender man* ou *woman*, *cisgender* man ou *woman* (Zimman, 2015) (aujourd'hui, les internautes de différents pays, dont la France, peuvent personnaliser leur identité de genre dans un champ « libre »). La plateforme de microblogage Tumblr, qui a la réputation d'être ouverte à la culture queer, encourage la déconstruction de la binarité femmes/hommes d'une façon encore plus forte. Le site propose un système de création des identités en ligne fluide et hautement personnalisable, par la combinaison de biographies, de tags et de pages « About me » (Oakley, 2016).

1.2 La recherche sur le genre et la langue

L'étude linguistique du genre est un champ de recherche fécond, qui concerne de nombreuses disciplines, de la phonétique à l'analyse conver-

sationnelle et la sémantique en passant par l'anthropologie linguistique et la sociolinguistique, et utilise des méthodes qualitatives comme quantitatives. Nous proposons ici un bref historique des recherches, qui présente les différentes approches du genre adoptées par les linguistes, les vagues successives de la sociolinguistique variationniste, l'état de la recherche en France, et les apports de la linguistique de corpus.

1.2.1 Historique

Naissance d'un objet de recherche

Les premiers travaux sur la langue et le genre remontent au début du 20ème siècle, avec *Language : Its nature, development, and origin* du linguiste danois Jespersen (1922). S'appuyant sur des exemples tirés de dictons, proverbes et citations de diverses langues, comme le japonais, les langues caraïbes et l'anglais, il décrit la langue féminine comme étant vide et superficielle (Bailly, 2008). Il soutient notamment que les femmes ont un vocabulaire plus limité que les hommes et utilisent des phrases à la construction plus simple, des déficiences qui sont pour lui dues aux différences biologiques entre femmes et hommes.

Pendant près d'un demi-siècle, les théories de Jespersen n'ont pas été remises en question. Il faudra attendre 1973, avec la publication de « Language and woman's place », l'article de Robin Lakoff (1973) qui paraitra ensuite sous forme de livre, pour que naisse véritablement la recherche linguistique sur le genre (Eckert & McConnell-Ginet, 2003). Influencée par les féministes de la deuxième vague féministe, qui « analysent le patriarcat comme un système général d'oppression masculine » (Dagorn, 2011), Lakoff examine les manifestations linguistiques de la domination masculine. Elle considère le langage des femmes comme étant « faible » par comparaison avec celui des hommes. Toutefois, pour elle, la source de cette différence ne se trouve pas dans la biologie, mais dans l'inégalité entre femmes et hommes. La langue des femmes, pour Lakoff, est caractérisée par des stratégies conversationnelles reflétant leur position de dominées dans la société, comme les marqueurs d'atténuation (*hedges*), les *tag questions*, l'intonation montante (*uptalk*) ou encore les exagérations (Lakoff, 1973).

Même s'il reste encore très influent, comme le montre la réédition augmentée de *Language and Woman's Place* en 2004, le travail de Lakoff a été abondamment critiqué. Plusieurs reproches lui ont été faits, dont sa théorie de l'« infériorité » du langage féminin. Celle-ci se base sur l'idée que certains procédés langagiers connotent la faiblesse, alors que les linguistes ont montré depuis que les connotations sociales des phénomènes phonologiques et grammaticaux ne sont pas fixes et ont plusieurs facettes. On a également critiqué son ethnocentrisme, ainsi que le manque de scientificité de sa méthode, qui reposait sur l'analyse de textes littéraires et sur une approche introspective héritée de sa formation en linguistique formelle (Wolfram & Schilling, 2016). En revanche, l'ouvrage de Lakoff a eu le mérite de montrer que les femmes sont souvent dans une situation contradictoire : si elles adoptent des traits linguistiques emblématiques de leur domination, l'ac-

cès au pouvoir leur est nié ; mais, si elles ne le font pas, elles sont critiquées pour ne pas se conformer aux normes de la féminité (Meyerhoff & Ehrlich, 2019).

Le paradigme de la domination

Le travail de Lakoff a eu un effet catalyseur, et a déclenché une vague d'études empiriques du genre et de la langue qui explorent les façons dont se manifeste la domination masculine. Selon cette approche, les différences linguistiques entre femmes et hommes sont dues aux dynamiques de pouvoir dans la société, et au fait que les femmes n'y sont pas les égales des hommes. Les chercheur·es définissent un style communicatif masculin marqué par la confrontation et l'absence de coopération, et un style féminin caractérisé par l'écoute et la solidarité.

Parmi les travaux emblématiques du paradigme de la domination parus dans la sphère anglophone, citons ceux de Fishman (1978), qui s'est intéressée au « shitwork » (p. 405) conversationnel des femmes. Pour elle, une conversation entre un homme et une femme est asymétrique. C'est aux femmes qu'il revient de faire le travail conversationnel nécessaire pour interagir avec les hommes, en posant par exemple plus de questions qu'eux ; il pèse donc sur elles une charge supplémentaire. West et Zimmerman (1983) ont également analysé des conversations entre femmes et hommes, identifiant les interruptions comme une manifestation de la domination linguistique masculine. O'Barr et Atkins (1998) ont étudié un contexte différent : celui des tribunaux, examinant la façon dont les témoins parlent pendant les procès en Caroline du Nord, aux États-Unis. Se focalisant sur les stratégies linguistiques associées aux femmes par Lakoff, ils mettent en évidence un « powerless language ». Ils soulignent toutefois qu'il n'est pas spécifique aux femmes, mais plus présent chez elles parce qu'elles sont plus fréquemment en position de dominées.

Le paradigme de la domination remet également en question le fait que les problèmes de communication sont liés à des différences fondamentales entre femmes et hommes. Il souligne que, souvent, les hommes tirent parti, consciemment ou inconsciemment, de ce qu'ils perçoivent comme étant une faiblesse pour continuer à asseoir leur domination. Mendoza-Denton (2012) a montré comment, dans une affaire de harcèlement sexuel, des sénateurs adaptaient leurs stratégies linguistiques à leur interlocuteur ; ils ont par exemple posé des questions fermées à l'homme accusé (le juge de la Cour suprême des États-Unis Clarence Thomas), qui lui permettait de donner une impression d'honnêteté, réservant les questions nécessitant des réponses complexes à la femme qui l'accusait (Anita Hill). De cette façon, les usages linguistiques des hommes perpétuent les inégalités de pouvoir entre femmes et hommes.

Le paradigme de la différence culturelle

Le paradigme de la différence culturelle est l'autre approche qui a dominé l'étude du genre et de la langue pendant des années. Elle tire son

origine dans le travail de Maltz et Borker (1982) ; influencé·es, notamment, par les théories de Gumperz (1982) sur la « communication interethnique », Maltz et Borker soutiennent que les problèmes de communication des femmes et des hommes sont liés à une différence culturelle entre les deux genres, due à des processus de socialisation différents. Cette théorie a été popularisée par le best-seller de Tannen (1990), *You just don't understand : Women and men in conversation*, qui soutient que les différences entre les styles conversationnels viennent du fait que, dans de nombreuses sociétés, filles et garçons grandissent dans des sous-cultures différentes. Les groupes de filles valorisent la coopération, l'égalité et les amitiés, tandis que dans les groupes de garçons, qui sont fortement hiérarchisés, il est important de montrer sa domination. Par conséquent, les filles développent un style conversationnel qui repose sur la coopération, et les garçons un style caractérisé par la compétition. Ces différences persistent à l'âge adulte, ce qui explique, selon Tannen, les problèmes de communication entre femmes et hommes. L'ouvrage de Tannen a eu beaucoup de succès, chez les sociolinguistes comme auprès du grand public ; il a notamment inspiré le best-seller de Gray, *Men are from Mars, Women from Venus* (1992), « un guide pratique pour améliorer la communication et la relation de couple ».

L'approche de la différence culturelle a été abondamment critiquée par les linguistes défendant la théorie de la domination. Pour Uchida (1992), elle est simpliste et obscurcit le contexte patriarcal dans lequel ont lieu les interactions sociales. Henley et Kramarae (1991) montrent que les différences mises en lumière par Maltz et Borker (1982) et les problèmes de communication soulignés par Tannen peuvent tous être reliés à des écarts de pouvoir entre femmes et hommes. Ils expliquent par ailleurs que la théorie de la différence culturelle peut avoir des conséquences néfastes ; elle peut par exemple être utilisée pour justifier un viol comme résultant d'un malentendu, l'homme interprétant de façon erronée le refus d'une femme.

Malgré les débats parfois intenses qui les ont opposés, le paradigme de la domination et celui de la différence avaient beaucoup en commun. Cameron (2005) souligne qu'ils considéraient tous les deux les femmes et les hommes comme des groupes homogènes inscrits dans une dichotomie bien définie, et qu'ils expliquaient les différences comme le résultat de la socialisation pendant l'enfance. Ils se concentraient également tous les deux sur un certain type de femmes et d'hommes, des personnes blanches, hétérosexuelles et monolingues issus de la classe moyenne d'Amérique du Nord, généralisant souvent leurs résultats à d'autres catégories de la population.

1.2.2 La linguistique variationniste et le genre

La première vague variationniste

La vision essentialiste qui caractérisait les études sur la langue et le genre jusque dans les années 1990 a également longtemps imprégné un autre champ de recherche, qui a pris son essor peu avant la parution de l'ouvrage de Lakoff : la sociolinguistique variationniste, dont le point de départ est la publication, en 1966, de *The social stratification of English in*

New York City. Labov y pose la fondation du variationnisme : l'étude du changement linguistique par l'analyse de variables généralement phonétiques ou phonologiques qui sont mises en lien avec des variables macrosociales, par l'utilisation d'enregistrements et de méthodes statistiques inférentielles. Le travail de Labov a initié un tournant quantitatif et empirique, selon lequel les individus sont vus comme des « human tokens – bundles of demographic characteristics » (Eckert, 2012). Les sociolinguistes variationnistes cherchent à établir des corrélations entre la variation linguistique et l'âge, la classe sociale, l'ethnicité et le genre. Ils et elles ont souvent montré que la langue des femmes était plus standard que les hommes ; c'est notamment le cas chez Wolfram (1969), dans son étude de l'anglais afro-américain à Détroit, chez Trudgill (1974) à Norwich, en Angleterre, et chez Macaulay (1977) en Écosse.

La « conformité linguistique » des femmes a conduit Labov à les considérer comme les principales actrices du changement linguistique : elles impulsent un changement depuis le vernaculaire vers le standard. Pour lui, l'utilisation d'une variante standard est un phénomène conscient, qu'il appelle le « change from above », et qui est motivé par le statut socioéconomique d'un individu. Le conservatisme linguistique des femmes serait donc expliqué par leur désir d'atteindre une position plus élevée dans la hiérarchie sociale. Cet argument est en contradiction avec le second constat fait par Labov, qui est que les femmes sont également les plus innovatrices en matière de « changes from below », c'est-à-dire de changements qui s'effectuent de manière inconsciente et qui s'écartent du standard : c'est ce qu'il appelle le « gender paradox » (Labov, 2001). La position de Labov a par la suite été critiquée, entre autres à cause de sa caractérisation simpliste de la notion de « standard » : le paradoxe du genre n'en est un que si on considère que les formes non standard et les prononciations innovantes ont le même sens social (Eckert, 2012). Bauvois (2002) a également souligné que, pour les femmes de la classe ouvrière, utiliser des formes prestigieuses à la place du vernaculaire employé par les hommes n'est pas une marque de conservatisme, mais une innovation. De plus, de nombreuses exceptions aux principes de Labov ont été mises en lumière, y compris par lui-même dans ses travaux les plus récents, qui montrent que le genre n'a pas un effet uniforme sur la variation. Il a ainsi trouvé une interaction entre genre et classe sociale : les femmes de classe moyenne supérieure utilisent davantage de formes standard que les hommes, mais les femmes de la classe ouvrière en emploient moins qu'eux (Labov, 2001).

La seconde vague variationniste

Au début des années 1980, le variationnisme prend un tournant ethnographique, donnant naissance à une « deuxième vague » de recherches. Celle-ci s'intéresse également au changement linguistique en utilisant des méthodologies quantitatives, mais elle délaisse les catégories macrosociologiques des travaux précédents (les femmes, les hommes, la classe ouvrière, la classe moyenne, etc.). À la place, elle se focalise sur des groupes identifiés par le travail de terrain des chercheurs (Drummond & Schleef, 2016),

et considère l'utilisation du vernaculaire comme l'expression d'une identité locale ou de classe. Milroy (1980) a initié cette seconde vague, avec son travail sur des communautés ouvrières de Belfast, qui a mis en évidence une corrélation entre la densité et la complexité des réseaux sociaux et les variables phonologiques vernaculaires.

Cette approche a notamment été adoptée par Cheshire (1982) dans son étude ethnographique de groupes d'adolescents de la classe ouvrière à Reading, en Angleterre, qui a montré un lien entre l'utilisation de formes morphosyntaxiques non standard et une culture antiautoritaire. Eckert (1989a) s'est également intéressée au rôle de la classe sociale dans les usages linguistiques des adolescent·es de plusieurs lycées de Détroit ; au lieu d'utiliser les catégories de « classe moyenne » et « classe ouvrière », elle identifie deux groupes d'adolescent·es, les « jocks » et les « burnouts ». Cela lui a permis de montrer que la variation adolescente n'est pas forcément expliquée par la classe sociale des parents, mais plutôt par l'affiliation des adolescent·es à des groupes sociaux. La seconde vague variationniste a eu le mérite de faire entrer l'identité dans le champ de recherche ; toutefois, comme la première vague, et comme les recherches sur le genre et la langue dans les paradigmes de la domination et de la différence, elle adopte une vision essentialiste : elle se focalise sur des catégories statiques, et considère l'identité comme quelque chose de fixe et stable (Eckert, 2012).

La troisième vague variationniste

Là où les chercheur·es de la deuxième vague variationniste voient le langage comme un reflet des catégories macrosociales, ceux de la troisième vague considèrent que les identités sont construites par le langage (Drummond & Schleef, 2016). Les formes linguistiques n'indexent plus des catégories sociales ou des identités, mais des significations sociales. Par exemple, la variable (ING), dont la réalisation apicale non standard a été notamment liée aux hommes et à la classe ouvrière par les variationnistes de la première vague, a acquis plusieurs significations sociales : aux États-Unis, la variante vélaire (*talking*) est associée avec l'intelligence, l'éducation, à la prononciation soignée, et aux locuteurs du nord du pays. La variante apicale (*talkin*) est quant à elle liée à l'informalité et aux accents du Sud (Campbell-Kibler, 2007).

Une fois qu'une variable est associée à des significations sociales, les individus peuvent faire leurs choix entre les variantes, et combiner plusieurs variables pour indiquer l'appartenance à un groupe social ou pour indexer des caractéristiques et des stances associées à une population. Par exemple, les adolescents blancs et asiatiques utilisent parfois des formes issues de l'anglais afro-américain pour indexer une identification à une masculinité urbaine et « cool » (Bucholtz, 1999b ; Cutler, 1999). Ainsi, pour étudier les significations sociales des formes linguistiques, les sociolinguistes ne se focalisent plus sur des variables isolées, mais s'intéressent au « style » des locuteurs en examinant leurs pratiques langagières en combinaison avec d'autres pratiques sociales. En mettant l'accent sur la pratique stylistique dans laquelle sont constamment engagés les individus, la troisième vague

variationniste ne les voit pas comme des « porteur·ses passifs » de dialectes, mais comme des « agent·es stylistiques » qui se servent de la langue pour construire leur identité :

> « It has become clear that patterns of variation do not simply unfold from the speaker's structural position in a system of production, but are part of the active—stylistic— production of social differentiation » (Eckert, 2012, p. 98).

Le concept de « communautés de pratique » créé par l'anthropologue sociale Jean Lave et le théoricien de l'éducation Étienne Wenger (1991) et introduit en sociolinguistique par Eckert et McConnell-Ginet (2012) a contribué à révolutionner la recherche sur la langue et le genre (Bucholtz, 1999a). Appliqué à la linguistique, ce concept montre que les pratiques langagières sont imbriquées dans les communautés et les contextes locaux. Il a permis de dépasser les généralisations sur les femmes et les hommes par l'étude non pas des styles communicatifs genrés, mais de la façon dont le genre et la langue interagissent dans les pratiques sociales de communautés locales et bien définies (les communautés de pratique) (Meyerhoff & Ehrlich, 2019). Eckert et McConnell-Ginet (2012) exhortent les chercheurs à « penser de façon pratique et à s'intéresser au local » :

> « To think practically and look locally is to abandon several assumptions common in gender and language studies : that gender can be isolated from other aspects of social identity and relations, that gender has the same meaning across communities, and that the linguistic manifestations of that meaning are also the same across communities » (p. 462).

Eckert (2000) a appliqué cette méthodologie à son travail sur les « jocks » et les « burnouts » de Détroit. Elle a montré, en conciliant analyses quantitatives et qualitatives, que les jeunes filles de ces deux communautés de pratique avaient des pratiques langagières différentes à cause de leur relation différente à la culture lycéenne.

1.2.3 Le paradigme de la performance

Au moment où les variationnistes introduisent une vision dynamique de l'identité, la recherche sur la langue et le genre abandonne elle aussi les catégories homogènes et dichotomiques des paradigmes de la domination et de la différence. Sous l'influence de la philosophe Judith Butler qui publie son ouvrage *Gender Trouble* en 1990, le genre n'est plus considéré comme quelque chose que l'on « a », mais comme quelque chose que l'on « fait » :

> « Gender is the repeated stylization of the body, a set of repeated acts within a highly rigid regulatory frame that congeal over time to produce the appearance of substance, of a natural sort of being » (Butler, 2006, p. 45).

Dans cette vision performative, les pratiques langagières ne découlent pas tout naturellement d'un genre préexistant, mais elles créent le genre. Pour les chercheur·es de la performance, il ne s'agit donc plus d'examiner

les différences langagières entre femmes et hommes, mais d'étudier comment les individus utilisent (ou pas) les ressources associées à la masculinité et à la féminité pour construire leur identité de genre, et de mettre en avant la diversité de ces identités (Meyerhoff & Ehrlich, 2019). « La différence des genres laisse la place à leur diversité, voire leur prolifération se déclinant avec la classe, l'âge ou la race » (Greco, 2014, p. 12). Comme la troisième vague variationniste, le paradigme de la performance considère l'identité comme étant fluide, et pouvant donc changer en fonction des contextes et des communautés de pratique dans lesquels une personne évolue.

Les études de l'identité sexuelle et de la transgression de genre

Le travail de Butler (1991) a pointé le fait que la performance du genre fait partie de la performance de l'identité sexuelle, et vice versa (Cameron, 2014). Avec la montée du mouvement de la « gay liberation », les chercheur·es ont commencé à s'intéresser à la langue des minorités sexuelles, tout d'abord par l'identification des marqueurs linguistiques de l'identité gay (W. Leap, 1996), puis par l'étude des styles discursifs gay, avec, par exemple, l'étude de l'intonation « gay » (Gaudio, 1994). Les travaux sociolinguistiques sur l'identité lesbienne sont toutefois moins nombreux, peut-être parce que celle-ci est attachée à peu de stéréotypes langagiers, par contraste avec l'identité gay (Cameron, 2011). L'ouvrage de Cameron et Kulick (2003) a fait entrer l'hétérosexualité dans le champ de recherche ; Kitzinger (2005) a par exemple utilisé l'analyse conversationnelle pour montrer comment l'hétérosexualité s'affiche dans les interactions du quotidien. Eckert (2014) souligne que les études sociolinguistiques de la sexualité l'ont souvent considérée comme un phénomène binaire, effaçant la diversité des pratiques ; elle remarque également que l'identité sexuelle a peu été intégrée dans les études de corpus, parce qu'elle n'est pas facile à identifier.

La pratique du drag a également été évoquée par Butler comme étant un exemple frappant de la dimension construite et performative du genre : « drag is an example that is meant to establish that "reality" is not as fixed as we generally assume it to be » (Butler, 2006, p. XXV). Des sociolinguistes influencé·es par la queer theory se sont saisi·es de cet exemple ; Barrett (1999) s'est par exemple intéressé à la façon dont les drag queens afro-américaines utilisent la langue des femmes blanches de la classe moyenne comme une forme de résistance contre le racisme et l'homophobie de la société américaine. Greco (2018) a étudié les ateliers de Drag Kings de Bruxelles pour montrer comment ils construisent des identités plurielles et politiquement subversives par le langage.

Des chercheur·es ont examiné comment les individus qui sont à la marge des catégories du sexe et du genre manipulent le système grammatical pour construire leur(s) identité(s). Borba et Ostermann (2007) ont montré que les travestis brésiliens n'emploient pas uniquement des formes féminines, mais utilisent des formes grammaticales masculines pour, notamment, parler de leurs relations familiales et se distinguer d'autres travestis. Saisuwan (2016) s'est penché sur la façon dont les kathoey, des per-

sonnes qui sont considérées en Thaïlande comme des femmes transgenres (même si elles peuvent aussi être vues comme une « troisième catégorie », ni hommes, ni femmes) utilisent le système complexe de pronoms personnels sujets du thaï pour s'aligner avec la féminité, mais de façon sélective. McGlashan et Fitzpatrick (2018) ont mis en évidence la difficulté posée par le système des pronoms anglais pour des jeunes queer et transgenres néozélandais, qui l'utilisent de façon variable pour refléter la fluidité de leur identité de genre. Bershtling (2014) a montré comment des personnes genderqueer israéliennes font preuve de créativité dans leur usage des catégories grammaticales « genrées » de l'hébreu, utilisant des pronoms associés avec l'« autre » genre, en combinant des formes masculines et féminines, et en évitant les formes genrées. Zimman (2014) a étudié l'utilisation du vocabulaire du corps par des hommes transgenres dans une communauté en ligne, montrant comment ceux-ci brisent non seulement la binarité du genre, mais aussi la naturalisation du sexe, en utilisant les mêmes mots (*dick*, *cock*) pour désigner les pénis et clitoris.

Le travail de Zimman (2016, 2017, 2018) sur la voix des hommes transgenres apporte un contrepoint nécessaire aux travaux des orthophonistes, qui ont abondamment abordé le sujet (ce qui est lié au fait qu'ils sont sollicités pour aider à « féminiser » ou à « masculiniser » des voix). Ceux-ci adoptent souvent une perspective déterministe, partant du principe que les différences entre femmes et hommes sont principalement ou exclusivement d'origine biologique. Or, Zimman, qui a étudié l'évolution de la fréquence fondamentale de la voix des hommes transgenres avec la prise de testostérone, montre que le genre d'une voix est le résultat d'une construction complexe et fluide qu'il appelle « stylistic bricolage » (Zimman, 2017, p. 342).

Certain·es sociolinguistes se sont intéressé·es à la façon dont des personnes transgenres utilisent des variables souvent étudiées par le variationnisme, et dont les réalisations possibles sont associées à la féminité ou à la masculinité. Hazenberg (2015) a étudié la production phonétique de [s] ainsi qu'une variable lexicale (les adverbes d'intensité *so* et *pretty*) dans l'Ottawa Trans Corpus, comparant des personnes transgenres à des personnes cisgenres gay et hétérosexuelles. Il a remarqué que les femmes et les hommes transgenres ne s'alignaient pas sur les femmes et les hommes cisgenres, et qu'ils évitaient les extrêmes, « choosing a path that is neither markedly feminine nor markedly masculine, but nevertheless falls within the acceptable ranges of both » (p. 289).

Gratton (2016) a réalisé une étude sociophonétique qualitative de la variable (ING) (→ p. 82) dans une communauté non binaire de Toronto. Elle remarque que, en fonction des situations, les personnes non binaires utilisent la variante [ɪn] (souvent associée aux hommes) ou la variante [ɪŋ] (souvent associée aux femmes) pour se distancier de leur genre assigné à la naissance. La tendance à la distanciation était plus forte dans les lieux publics que dans les lieux privés, considérés comme des « safe spaces », « where marginalized groups can feel secure expressing themselves without being subject to mainstream norms and stereotypes » (p. 55). Elle note ainsi

que, malgré leur désir de créer un système non binaire, ces individus réinstallent une binarité dans leurs pratiques linguistiques.

1.2.4 Le variationnisme et les études du genre en France

En France, la sociolinguistique variationniste n'a pas connu le même succès que dans les pays anglo-saxons (Boutet, 2017). Le travail de Labov a été traduit dès 1976 (William et al., 1976) et défendu par Encrevé (Encrevé, 1976), mais il n'a provoqué qu'un intérêt superficiel (Gadet, 2003). Le contexte politique et théorique français n'a jamais été réellement propice à l'étude de la diversité et de la variation dans la langue ; le désir d'uniformité politique s'est traduit par la marginalisation des langues régionales et le déni de la variation individuelle. De plus, contrairement au Canada et aux États-Unis, la France n'a jamais souhaité baser ses politiques sociales sur des enquêtes sociolinguistiques (Gadet, 2003). Pour toutes ces raisons, il n'a pas réellement existé de sociolinguistique française s'inscrivant dans la tradition labovienne. Celle-ci a toutefois inspiré quelques travaux, comme Laks (1977), dans une analyse du /r/ chez des adolescents de Villejuif, Houdebine-Gravaud (1978), qui a étudié la variation phonologique dans le Poitou, ou Encrevé (1988), qui s'est intéressé au changement dans la liaison. Il a paru peu de manuels français traitant de sociolinguistique ; il existe un dictionnaire de la sociolinguistique (Moreau, 1997), mais il a été réalisé par une linguiste belge. Les travaux variationnistes portant sur le genre en français ont vu le jour hors des frontières françaises, comme Bauvois (2002), en Belgique, ou G. Sankoff et Thibault (1977) et Tousignant (1987) au Québec.

L'étude du genre et de la langue (non variationniste) a mis du temps à s'installer en France, malgré le fait que de célèbres féministes aient reconnu le rôle du langage dans la construction du genre (Greco, 2014). Cela est dû au fait qu'il n'existe pas d'équivalents français des « Gender and Language Studies », ou recherches linguistiques sur le genre. Le genre a donc mis du temps à s'imposer comme un centre d'intérêt pour les linguistes, en dépit des travaux de Yaguello (1979) et Houdebine (1979), qui font figure de pionnières en France, de Aebischer et Forel (1983), qui posent les jalons d'une nouvelle linguistique féministe, de Baider (2004) en linguistique de corpus, et de Beeching (2002) sur la politique et les particules pragmatiques. Pour Greco (2014), ces études ont contribué « à poser les pièces d'un ensemble de travaux qui ne connaîtra jamais ni une véritable reconnaissance ni une institutionnalisation scientifiques » (p. 14). Depuis les années 2000, le genre attire davantage l'attention des sociolinguistes français·es, avec par exemple, l'ouvrage collectif de Chetcuti (2012), qui examine les débats autour du langage en tant qu'outil de construction du genre et de lutte contre la domination masculine, ou encore une présentation du champ de recherche par Bailly (2008).

1.2.5 Linguistique de corpus et genre

Avec la linguistique de corpus, qui a pris son essor grâce à la démocratisation des ordinateurs, l'étude du genre et du langage a changé d'échelle. Il est désormais possible d'examiner les productions orales, mais aussi écrites, de milliers de femmes et d'hommes sans sortir de chez soi. La linguistique de corpus a été décrite comme étant un ensemble de procédures et de méthodes permettant d'étudier la langue dans une perspective quantitative à l'aide d'outils informatiques (McEnery & Hardie, 2012). Ceux-ci permettent d'explorer de grands corpus, qu'il serait impossible (ou extrêmement chronophage) de lire intégralement ou d'analyser manuellement. Les linguistes de corpus utilisent plusieurs techniques, dont la concordance, qui est une liste de toutes les occurrences d'un terme dans un corpus, présentée avec son contexte (Baker et al., 2006), et qui est « the single most important tool available to the corpus linguist » (McEnery & Hardie, 2012, p. 35). Les *keywords*, des mots qui apparaissent significativement plus fréquemment dans un corpus que dans un autre, sont également centraux en linguistique de corpus ; ils ont été abondamment utilisés pour étudier les différences entre femmes et hommes.

Le premier corpus informatisé date des années 1960 ; il s'agit du Brown Corpus, qui contient 1 million de mots en anglais américain écrit (Francis & Kucera, 1964). Une deuxième génération de corpus, plus grands, a vu le jour dans les années 1990, avec le British National Corpus (BNC), qui contient 100 millions de mots d'anglais britannique écrit et oral (Leech, p. d.). Ce corpus a servi de support à plusieurs études du genre et du langage. Rayson et al. (1997) ont étudié uniquement sa composante conversationnelle (4.5 millions de mots) ; ils ont identifié les 25 mots les plus caractéristiques des conversations des femmes (*she*, *her*, *lovely*, *nice*, *oh*) et des hommes (*fucking*, *yeah*, *aye*, *right*, *mate*), et ont mis en évidence l'effet de l'âge et de la classe sociale. H. J. Schmid (2003) est parti du travail de Tannen (1990) pour savoir si le corpus oral du BNC (plus de 8 millions de mots) reflétait les supposés différents styles communicatifs des femmes et des hommes, en se focalisant notamment sur les « women's words », c'est-à-dire les marqueurs d'atténuation, les hésitations et les questions. À première vue, ses résultats semblent corroborer les thèses du paradigme de la différence, mais Schmid a également mis en lumière des nuances. Il s'est par exemple rendu compte que, si les hommes jurent plus que les femmes, celles-ci utilisent plus fréquemment certains jurons (*bloody*, *shit*).

La linguistique de corpus a aussi permis aux chercheur·es s'intéressant au genre d'examiner l'écrit beaucoup plus facilement qu'auparavant. L'écriture épistolaire, composée d'interactions écrites, a fait l'objet de plusieurs études des styles communicatifs des femmes et hommes. Palander-Collin (1999) a analysé le Corpus of Early English Correspondence, qui contient des lettres écrites au 16ème et au 17ème siècle ; elle a trouvé que les femmes utilisent davantage les pronoms de la 1ère et de la 2ème personne que les hommes, ce qui est pour elle le signe que les femmes ont un style plus « interactif » que les hommes. Ce corpus a également fait l'objet d'une rare étude (Nevalainen, 1996) à s'être intéressée à l'orthographe, dont les di-

mensions sociales et culturelles ont longtemps été négligées par les linguistes (Sebba, 2007). Elle a montré que les femmes innovaient davantage que les hommes dans le choix de certaines graphies, adoptant par exemple plus volontiers la graphie *has* à la place de *hath*.

Les écrits formels ont été étudiés par Argamon et al. (2003), qui se sont penchés sur l'utilisation de procédés lexicaux et syntaxiques par les femmes et les hommes dans des textes de fiction et de non-fiction, toujours dans le British National Corpus. Encore une fois, cette étude quantitative trouve de nombreuses différences entre femmes et hommes. S'appuyant sur le travail de Biber (1995) sur les registres, elle décrit le style des femmes comme étant caractéristique de la dimension « involved » de l'écrit (par l'utilisation « extraordinairement » plus fréquente de pronoms personnels, p. 326), et celui des hommes comme étant typique de la dimension « informative » (avec l'utilisation plus fréquente de noms). Argamon et al. en concluent que, même dans des textes dénués de dimension interactionnelle, les différences entre femmes et hommes persistent, procédant sans doute à une généralisation hâtive.

Dans son ouvrage consacré à l'étude du genre par les techniques de la linguistique de corpus, Baker (2014) relativise ces résultats. Il souligne que, dès que l'on compare deux groupes, qu'il s'agisse d'un groupe de femmes et d'un groupe d'hommes ou de deux groupes de femmes, on trouve forcément des différences entre eux. Il met également en évidence l'importance des contextes dans lesquels les productions orales ont été recueillies, ainsi que le fait que bon nombre des différences relevées (dans l'emploi du mot *fucking*, par exemple) sont dues à la surutilisation d'un terme par quelques locuteur·trices. Pour lui, dans les études de corpus qui se focalisent sur les différences, « the atypical has become stereotypical » (p. 41). Il propose aux linguistes de corpus plusieurs pistes pour éviter que leurs recherches ne perpétuent les stéréotypes sur le genre, comme mesurer la similarité, et non la différence, entre le lexique de deux corpus en utilisant la distance de Manhattan plutôt que des tests de corrélation. Il suggère également d'étudier la dispersion des mots dans un corpus pour prendre en compte la variation individuelle, ou encore de comparer un corpus de productions de femmes et un corpus de productions d'hommes à un troisième corpus.

Avec l'avènement de la CMC et les progrès des outils informatiques, la linguistique de corpus est entrée au 20ème siècle dans une nouvelle ère. Il n'a jamais été aussi facile de créer de grands corpus, mais aussi de les analyser avec des techniques statistiques poussées, qui permettent en grande partie de dépasser les limites soulignées par Baker ; nous verrons plus en détail ce changement de paradigme dans le chapitre suivant (→ p. 52).

1.3 L'intersectionnalité : une nouvelle approche du genre

1.3.1 Définition

Ces vingt dernières années, la recherche sociolinguistique s'est nourrie de la théorie de l'intersectionnalité. Immensément populaire aux États-Unis, dans le monde universitaire comme dans les productions culturelles et artistiques, les médias et l'arène politique (Coaston, 2019), l'intersectionnalité a été décrite comme étant « the most important theoretical contribution that women's studies, in conjunction with related fields, has made so far » (McCall, 2005, p. 1771). L'intersectionnalité n'est pas une théorie homogène (Levon, 2015). Elle a été définie comme une théorie, une perspective, un concept, un type d'analyse, et une approche méthodologique.

Cette « constellation dynamique d'idées et de pratiques » (Collins & Chepp, 2013, p. 60) est née d'une critique de la recherche sur le genre et l'ethnicité. Elle postule que prendre en compte une seule catégorie sociale ne suffit pas pour étudier la diversité des comportements et des expériences (Levon, 2015). En effet, le genre, l'ethnicité et d'autres variables sociales comme l'âge, l'orientation sexuelle ou la classe sociale, se « croisent » pour créer des réalités distinctes. Elles ne peuvent donc pas être analysées isolément (Collins & Chepp, 2013). Il est ainsi impossible de comprendre l'expérience d'une femme afro-américaine en s'appuyant sur les études du genre d'un côté, et sur les études de l'ethnicité de l'autre, parce que les premières se focalisent généralement sur les femmes blanches, et les secondes sur les hommes afro-américains (McCall, 2005) : « The intersectional experience is greater than the sum of racism and sexism », écrit Crenshaw dans le texte fondateur de l'intersectionnalité (Crenshaw, 1989, p. 140).

Émergence de l'intersectionnalité

L'approche intersectionnelle entretient un lien étroit avec l'activisme politique noir américain : elle tire son origine chez les féministes afro-américaines des années 1950, 1960 et 1970, dans le contexte du mouvement des droits civiques. Certain·es la relient plus précisément au Combahee River Collective Statement de 1977 (« The Combahee River Collective Statement », p. d.), un document clé du Black Feminism rédigé par une organisation de femmes lesbiennes afro-américaines de Boston (Ferguson, 2012). L'intersectionnalité s'est donc nourrie des expériences des féministes noires américaines, qui subissaient à la fois le sexisme, le racisme et l'exploitation sociale de la société américaine, et qui ne se retrouvaient pas dans le discours féministe blanc. Celui-ci se focalisait par exemple sur le droit à l'avortement, sans faire de place aux luttes spécifiques aux femmes de couleur, victimes de campagnes de stérilisation forcée (Collins & Chepp, 2013).

La sexualité était également un thème important dans la lutte des féministes afro-américaines lesbiennes et bisexuelles, qui considéraient l'hétéronormativité comme un outil de dominance (Collins & Chepp, 2013). Dans

ce contexte, Angela Davis (1983), Bonnie Thornton Dill (1983) et d'autres féministes noires américaines ont exprimé la nécessité d'adopter de nouvelles approches de l'analyse de l'oppression. En même temps, des femmes afro-américaines, amérindiennes, et d'origine mexicaine et chinoise parviennent aux mêmes conclusions. Ensemble, elles posent la fondation de ce qui deviendra ensuite l'approche intersectionnelle (Collins & Chepp, 2013).

C'est le travail de la juriste Kimberlé W. Crenshaw qui donnera un nom à l'intersectionnalité. Dans son article « Demarginalizing the intersection of race and sex : A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics » (1989), elle examine plusieurs procès opposant des femmes noires à leurs employeurs. Elle souligne que dans les cas de discrimination fondés sur l'ethnicité, l'intérêt est focalisé sur les hommes et sur les personnes de classes sociales privilégiées, tandis que les procès d'anti-discrimination sexiste s'intéressent essentiellement aux femmes blanches. Ce fonctionnement exclut les femmes noires et d'autres catégories de personnes qui subissent des formes de discrimination combinées.

L'approche intersectionnelle a fait son chemin dans le monde universitaire dans les années 1980 et 1990 pour devenir extrêmement populaire au début du 21ème siècle (Collins & Chepp, 2013). De par ses origines, elle entretient des liens étroits avec la recherche sur le genre et les sciences politiques. Aux États-Unis, c'est aujourd'hui l'approche dominante dans les études de genre et les « women's studies » (Collins & Chepp, 2013). Elle est employée dans de nombreuses disciplines : la sociologie, la psychologie, l'économie, la littérature, l'histoire, les sciences politiques, mais aussi les domaines du droit et de la santé publique (Ferguson, 2012).

L'intersectionnalité en France

En France, l'intersectionnalité a été importée à partir de la deuxième moitié des années 2000 (Jaunait & Chauvin, 2012). Cette apparition tardive s'expliquerait par le fait que la France n'a pas la même tradition du droit antidiscriminatoire que les États-Unis. La politique anti-discrimination mise en place par l'Union européenne dans les années 2000 aurait donc contribué à l'apparition de l'intersectionnalité en France (Jaunait & Chauvin, 2012). On peut également attribuer la lente adoption de l'approche intersectionnelle dans l'Hexagone aux différences fondamentales entre le féminisme américain, né au sein du mouvement des droits civiques et la lutte contre les discriminations raciales, et le féminisme français, qui a émergé dans un contexte de luttes ouvrières et de pensée marxiste (Jaunait & Chauvin, 2012). Parmi les travaux intersectionnels réalisés en France, on peut citer les études de Kergoat sur les ouvrières (1978, 2000), le travail de Baudelot et Establet sur les filles à l'école (2006), et l'ouvrage collectif dirigé par Dorlin (2009).

1.3.2 Intersectionnalité et linguistique

De la nécessité d'une étude intersectionnelle du genre en linguistique

La nécessité d'intégrer plusieurs variables dans les études sociolinguistiques a été mise en lumière par Epstein (1986). Elle soutient qu'étudier le genre isolément a également pour effet de minimiser l'impact d'autres variables comme le contexte, la classe sociale, la variation géographique, l'âge, la race et l'ethnicité (p. 37). Par conséquent, selon elle, de nombreuses études du genre et du langage ont eu tendance à généraliser leurs résultats trop hâtivement, concluant qu'il existe un langage des hommes et un langage des femmes. Eckert (2014) souligne quant à elle qu'intégrer plusieurs catégories sociales dans les études sociolinguistiques permet d'aller au-delà de la « binarité du genre et de la sexualité ».

Levon (2015) appelle ainsi à incorporer davantage la théorie intersectionnelle dans les études linguistiques, estimant qu'elle y est encore peu présente. Selon lui, les travaux sociolinguistiques ont tendance à être trop compartimentés : certains s'intéressent uniquement au genre et à la sexualité, d'autres à l'ethnicité, à la variation géographique, ou à la classe sociale. Cette séparation a pour effet de faire disparaitre l'influence que ces catégories ont les unes sur les autres, et de masquer la complexité des processus sociaux. Il exhorte donc les sociolinguistes qui étudient le genre et la sexualité à prendre en compte l'ethnicité, la classe sociale et d'autres variables sociales (p. 304), parce que :

> « There is no 'gender effect' to be discovered and analyzed ; there is only the effect of gender in relation to class, race, etc. » (Levon, 2015, p. 298).

L'intersectionnalité a toutefois mis du temps à être adoptée par les sociolinguistes ; après quelques travaux fondateurs réalisés dès la fin des années 1970, elle a mis une vingtaine d'années à se faire une place dans le domaine.

Les pionnier·es de l'approche intersectionnelle en linguistique

Dès la fin des années 1970, des linguistes ont commencé à défendre l'idée que les variables sociales n'ont pas un effet uniforme sur le langage (Levon & Mendes, 2015). Nichols (1978) a par exemple montré que l'utilisation de certaines structures morphosyntaxiques par des femmes afro-américaines de Caroline du Sud était influencée à la fois par le genre et par le statut socioéconomique. Comme nous l'avons vu dans la sous-section 1.2.2, Milroy (1980) s'est intéressée à l'intersection du genre, des réseaux professionnels et sociaux et de l'ethnicité à Belfast, et à son impact sur les changements vocaliques. Eckert (1989b) est également une pionnière de l'approche intersectionnelle en linguistique, notamment avec son étude du « Northern Cities Chain Shift », des changements dans la prononciation de certaines voyelles dans la région des Grands Lacs aux États-Unis. Elle y souligne que le sexe (elle n'utilise pas le terme « genre » dans ce texte) n'a pas un effet uniforme dans la variation phonologique, parce qu'il est imbriqué dans des

pratiques sociales complexes. Son analyse de données recueillies pendant deux ans dans un lycée d'une banlieue de Détroit montre ainsi que genre et catégorie sociale ne sont pas des variables indépendantes l'une de l'autre et qu'elles interagissent, les filles de la classe moyenne se comportant différemment des filles de la classe populaire.

Travaux ouvertement intersectionnels

Malgré l'importance de ces travaux fondateurs, ce n'est qu'à la fin des années 1990 que des linguistes ont commencé à étudier de façon plus explicite les interactions entre le genre et d'autres variables sociales (Levon, 2015). Par exemple, Bucholtz a exploré l'intersection du genre et de la race avec son travail sur la « whiteness » (1999, 2001, 2010) et l'argot des jeunes immigrants mexicains (2009). Mendoza-Denton (2014) adopte aussi une perspective intersectionnelle dans son étude longitudinale du parler de Latinas appartenant à des gangs en Californie, réalisée dans les années 1990.

Citons également le travail de Morgan sur les femmes afro-américaines (1996, 2004, et celui de Pilcher (2009), qui s'est intéressée à l'intersection du genre avec la classe sociale et l'ethnicité dans son étude qualitative de conversations de trois groupes d'adolescentes blanches et d'origine bengalie de Londres. Certain·es chercheur·es ont pris en compte, avec le genre, la sexualité ou l'orientation sexuelle ; c'est le cas de Levon, qui l'a fait dans le contexte israélien, travaillant notamment sur l'argot (2012) et la prosodie (2014). Enfin, Podesva et Van Hofwegen (2014) ont mené un travail sociophonétique dans le nord de la Californie, intégrant comme variables à la fois le genre, la sexualité et les notions de ruralité et urbanité. Notons que tous ces travaux, à l'exception de ceux de Eckert (1989b) et de Mendoza-Denton (2014), sont des études qualitatives. L'approche intersectionnelle reste donc relativement rare en sociolinguistique quantitative.

1.3.3 Intersectionnalité et méthodes quantitatives

Les origines racistes de la statistique

La démarche intersectionnelle a majoritairement été appliquée à des travaux qualitatifs, et n'a que très progressivement adopté des méthodologies quantitatives (Rouhani, 2015). Cette réticence s'explique en partie par le lourd passé de la statistique, qui a entretenu pendant des décennies un lien étroit avec les théories et les politiques eugénistes. Les concepts de corrélation et de régression ont ainsi été développés par Francis Galton, cousin de Charles Darwin et inventeur de l'eugénisme. Galton utilisait ces techniques pour démontrer la « supériorité » intellectuelle de la « race » européenne et des classes sociales bourgeoises et aristocratiques britanniques (Zuberi, 2001). Les Britanniques Ronald Fisher, créateur du test F, de l'analyse de variance et du concept de l'hypothèse nulle, et Karl Pearson, le père de la valeur p et du test du χ^2, défendaient eux aussi des théories eugénistes.

Dès sa naissance, la statistique a ainsi été liée à la recherche sur les

différences humaines, dans le but de mettre en lumière une prétendue supériorité de la race blanche (Zuberi, 2001). Ces théories et ces techniques ont nourri l'idéologie nazie, mais ont également motivé, aux États-Unis, des politiques eugénistes comme les campagnes de stérilisation forcée de minorités et de populations pauvres (Rivard, 2014). Les tests standardisés utilisés aujourd'hui dans le système éducatif américain, comme le SAT Reasoning Test, qui détermine en partie l'admission dans les universités, perpétuent l'héritage de la statistique eugéniste. Ces tests censés mesurer l'intelligence et la performance ont été conçus pour exclure les personnes de couleur du système éducatif ; ils continuent à jouer un rôle important dans le racisme structurel américain (Levy, 2019).

Comment rendre la statistique intersectionnelle ?

Même si certain·es pensent que les méthodes qualitatives sont plus adaptées à l'intersectionnalité, ces quinze dernières années, des techniques quantitatives ont commencé à être intégrées aux travaux intersectionnels. Cela a notamment été fait dans les domaines de la sociologie de la santé (Veenstra, 2011, Warner et Brown, 2011), de l'épidémiologie (Marcellin et al., 2013) de la psychologie (Stirratt et al., 2008) et de l'éducation (Covarrubias, 2011).

À cause du passé de la statistique, cette adoption du quantitatif s'accompagne d'une réflexion indispensable sur l'éthique. D'un côté, les chercheur·es s'interrogent sur les techniques les plus adaptées à la démarche intersectionnelle, soulignant que les structures mathématiques des méthodes quantitatives ne sont pas neutres sur le plan théorique (N. A. Scott & Siltanen, 2017) et qu'il convient de les choisir avec soin et de les utiliser avec une grande rigueur. De l'autre, ils mettent en évidence la nécessité d'utiliser la statistique pour défaire l'effacement des groupes marginalisés (Bowleg & Bauer, 2016) et pour « rendre visible l'invisible » (López et al., 2018). Pour Bowleg et Bauer (2016), cela ne peut se faire que dans le cadre d'une « mixed-methods approach », qu'elles définissent comme :

> « A distinct methodology focused on rigorously combining statistical approaches with in-depth culturally grounded narratives and meanings from qualitative approaches to gain a deeper understanding than either method could provide alone » (p. 338).

Les interactions, un outil statistique fondamental pour la recherche intersectionnelle quantitative

Prenant en exemple son travail de recherche sur les lesbiennes afro-américaines, la psychologue Lisa Bowleg (2008) souligne les défis méthodologiques de la recherche intersectionnelle quantitative. Pour elle, les méthodes quantitatives (et qualitatives) sont souvent basées sur une approche « additive » qui considère l'effet de chaque variable comme indépendant, ce qui est contraire aux principes de la recherche intersectionnelle. Pour mesurer les discriminations auxquelles sont confrontées les lesbiennes afro-américaines, il ne suffit pas d'additionner l'expérience des femmes, des les-

biennes et des Afro-Américaines : l'expérience de ce groupe n'est pas la simple somme de ces trois variables.

La régression multiple, une méthode inférentielle aujourd'hui largement utilisée dans les sciences sociales, apporte une réponse à ce problème (N. A. Scott & Siltanen, 2017). Elle permet d'étudier l'effet de plusieurs variables sur un phénomène donné. Toutefois, dans une approche intersectionnelle, il convient de ne pas étudier uniquement les effets principaux (ou isolés) des variables. Il faut, à la place, intégrer aux modèles de régression ce que l'on appelle des « interactions » entre deux variables (ou plus), de façon à mettre en lumière les effets parfois différents des combinaisons entre, par exemple, genre et ethnicité, ou genre et catégorie socioprofessionnelle. Cela permet de changer de perspective : dans les modèles de régression sans interaction, pour étudier par exemple la façon dont les femmes afro-américaines utilisent tel ou tel procédé linguistique, il faut additionner l'effet de la variable principale « femme » (qui est le même pour toutes les femmes qu'elles soient blanches, afro-américaines ou autres) et de la variable principale « afro-américain·e » qui est le même pour les femmes et les hommes). Sans les interactions, l'analyse statistique peut masquer complètement une partie de la réalité.

Un article, paru en 1999 aux États-Unis dans le *New England Journal of Medicine* (Schulman et al., 1999) et souvent cité en exemple (Bowleg & Bauer, 2016), illustre l'importance des interactions. Il étudie l'impact des préjugés des médecins sur leur propension à envoyer des patients souffrant de douleurs thoraciques vers un spécialiste, en prenant en compte trois variables : l'âge, le genre, et l'ethnicité (blanc·hes et Afro-Américain·es). Les chercheurs n'ont inclus que les effets principaux de ces variables dans leur analyse, et non leur interaction. Ils en ont conclu que la probabilité qu'un·e patient·e soit envoyé·e vers un spécialiste était 40 % moins importante pour les femmes blanches et les patient·es afro-américain·es que pour les hommes blancs, et 60 % moins importante pour les femmes afro-américaines. Les auteur·es de cette étude ont créé un second modèle de régression avec interactions et en ont présenté les résultats, sans toutefois les interpréter, se focalisant sur le modèle avec effets principaux. L'analyse de ce second modèle révèle une réalité différente. En fait, les femmes afro-américaines sont le seul groupe à souffrir de discrimination dans ce cas ; elles sont 60 % moins susceptibles d'être envoyées vers un spécialiste que les autres patient·es, y compris les hommes afro-américains et les femmes blanches.

Cet exemple montre bien à quel point étudier seulement les effets principaux peut être trompeur. L'analyse des interactions est donc centrale dans la recherche intersectionnelle quantitative, quel que soit son objet d'intérêt, car « the experiences of those at a particular intersection cannot be understood as a sum of their parts » (Bowleg & Bauer, 2016, p. 339). En aucun cas les modèles contenant uniquement les effets principaux ne peuvent être utilisés dans une perspective intersectionnelle : il ne suffit pas d'étudier plusieurs variables, encore faut-il savoir si et comment elles interagissent.

Contraintes

Si l'analyse des interactions est un outil très puissant, elle a également des contraintes de poids. Les difficultés commencent avec la constitution des échantillons. Ceux-ci doivent être plus grands que les échantillons nécessaires pour réaliser des analyses incluant uniquement les effets principaux. En effet, lorsque l'on intègre une interaction à un modèle de régression (pour reprendre l'exemple cité ci-dessus), on ne peut plus comparer toutes les femmes à tous les hommes (quelle que soit leur ethnicité), ou tou·tes les Afro-Américain·es à tou·tes les blan·ches. On compare des sous-groupes : les femmes afro-américaines aux femmes blanc·hes, aux hommes afro-américains et aux hommes blancs. Plus il y a de niveaux dans une variable (c'est-à-dire de catégories), plus l'échantillon doit être important.

Les difficultés augmentent lorsque l'on souhaite étudier des interactions dites de *high order*, c'est-à-dire qui prennent en compte plus de deux variables. En théorie, il est tout à fait possible d'analyser l'effet de trois ou quatre variables (genre + âge + ethnicité + orientation sexuelle, par exemple). En pratique, cela pose des problèmes considérables, car, à chaque fois que l'on ajoute une variable dans l'interaction, on divise la taille des sous-groupes. On a donc besoin d'échantillons de taille importante. Un nombre trop réduit d'observations dans chaque sous-groupe entraine des problèmes de puissance statistique : les tests statistiques n'atteignent souvent pas le niveau de significativité (Bauer, 2014). Il a été proposé, comme solution, d'augmenter le seul de significativité. Celui-ci est conventionnellement fixé à 0.05, signifiant qu'il y 5 % de chances d'obtenir les résultats constatés si l'hypothèse nulle (celle de l'absence de relation entre les variables, par exemple le genre et l'âge, sur le phénomène d'intérêt). Le faire passer à 0.10 pourrait, selon certains chercheurs, pallier les problèmes de puissance statistique (N. A. Scott & Siltanen, 2017). Notons également que les interactions de plus de deux variables sont extrêmement difficiles à interpréter, comme nous le montrons dans le chapitre consacré aux méthodes statistiques (→ p. 165).

Le cas de la sociolinguistique

Si ces problèmes concernent l'ensemble de la recherche intersectionnelle, ils touchent de façon plus forte certains domaines, comme la linguistique. Les sociologues ont accès à des bases de données issues de sondages ou de recensements, qui leur permettent d'étudier de gros échantillons combinant plusieurs variables. Ces données ne sont évidemment pas adaptées à des recherches sociolinguistiques variationnistes ou de corpus. Lorsque l'on veut réaliser une étude de corpus, il faut non seulement recueillir les informations sociodémographiques des personnes (en multipliant les variables, et en ne se contentant pas de l'identité de genre ou de l'ethnicité), mais aussi leur production orale ou écrite. Cela n'est pas forcément difficile si on peut se contenter d'un petit échantillon de quelques personnes, dans une optique qualitative, mais peut être extrêmement long et coûteux quand on souhaite étudier plusieurs centaines ou milliers de sujets. Cela explique

sans doute la faible présence de l'approche intersectionnelle en linguistique quantitative.

La CMC : un nouveau terrain pour la sociolinguistique intersectionnelle

La majorité des études sociolinguistiques (intersectionnelles ou non) se sont intéressées à la langue orale. Face à la difficulté de constituer de grands échantillons accompagnés des annotations sociodémographiques nécessaires, la CMC s'impose comme un terrain fertile pour les sociolinguistes (→ chapitre 2). D'énormes quantités de données sont librement accessibles, et, comme nous le verrons dans le chapitre où nous présentons la construction de notre corpus (→ p.109), il est possible de recueillir des informations démographiques de manière relativement fiable.

1.4 Les intersections du genre, de l'âge et de l'ethnicité

Cette section est consacrée à l'âge et à l'ethnicité, les deux variables sociodémographiques que nous étudions en interaction avec le genre dans nos analyses.

1.4.1 L'âge

La construction de l'âge

L'âge est une composante centrale de l'expérience humaine. Il a une dimension biologique, avec le vieillissement du corps, et une dimension sociale (Fournier, 2010). C'est donc à la fois une expérience individuelle et collective, la construction de l'histoire personnelle d'un individu et de sa place dans la société (Eckert, 1998). Même si l'âge est un phénomène universel, les systèmes d'âge varient selon les cultures. Dans les sociétés industrialisées, l'âge chronologique, calqué sur le calendrier, est la mesure officielle. Certains anniversaires sont toutefois plus marquants que d'autres, comme le « sweet sixteen » aux États-Unis, ou la majorité. L'entrée dans une nouvelle décennie peut aussi avoir une signification sociale particulière.

La vie est également ponctuée par des évènements qui ne sont pas forcément liés à l'âge chronologique. La sociologie utilise le terme de « cycle de vie » pour désigner ces évènements de la vie, qui s'enchainent de façon similaire pour tous les individus. Ce cycle de vie est composé de « l'entrée dans l'institution scolaire, dans la sexualité, dans le travail, dans l'indépendance économique, dans la conjugalité, dans la maternité-paternité, dans l'inactivité… » (Fournier, 2010, p. 103). Ces évènements sont connectés à de grandes étapes de la vie : l'enfance, l'adolescence, l'âge adulte, la vieillesse, des concepts que l'on invoque souvent pour expliquer les différences de comportements entre les individus (Eckert, 1998).

Les étapes de la vie s'accompagnent parfois de pressions normatives sur les individus, comme la pression pour les femmes de se marier avant d'atteindre un certain âge. Sur le plan linguistique, ces pressions poussent par exemple les adolescent·es à utiliser une langue fortement vernaculaire, et les adultes à adopter des pratiques langagières plus conservatrices (Eckert, 1998). Toutefois, les étapes du cycle de vie sont avant tout des constructions idéologiques : « Just as gender does not unfold naturally from biology, neither do life-stages such as childhood, adolescence, adulthood, or old age. » (Eckert, 2003, p. 381). L'adolescence, par exemple, a été inventée récemment. Elle n'existait pas avant le 20ème siècle, et est un produit des changements du monde du travail : avec l'industrialisation, il était devenu nécessaire que les jeunes ne puissent pas entrer immédiatement sur le marché du travail (Eckert, 2003). Récemment, on a vu l'apparition d'une nouvelle catégorie, créée par le marketing : les préadolescents, ou « tweens », à cheval entre enfance et adolescence (Mitchell & Reid-Walsh, 2005).

L'âge social se différencie enfin de l'âge biologique lorsque l'on prend en compte les évènements historiques qui marquent la vie d'un individu et le différencient des personnes d'autres générations (Fournier, 2010). Les « digital natives », c'est-à-dire les personnes nées à partir des années 1980, ont grandi dans le monde d'internet, de la diffusion de l'anglais à l'échelle mondiale et de l'essor de l'urbanisation, des faits structurants qui influencent leurs usages langagiers (Tagliamonte, 2016b).

Les études variationnistes regroupent généralement les individus en fonction de leur âge chronologique, qui est la mesure officielle de l'âge dans les sociétés occidentales. L'âge n'est toutefois pas considéré comme une variable numérique (19 ans, 20 ans, 21 ans, etc.). Pour mettre en valeur des différences significatives sur le plan statistique, il est nécessaire de grouper les individus (Eckert, 1998). Deux approches sont possibles. L'approche étique consiste à grouper les personnes arbitrairement dans des classes d'âge, comme les décennies, comme l'a par exemple fait Labov (2006). L'approche émique les regroupe en fonction de phases du cycle de vie, comme l'enfance, l'adolescence, ou l'âge mûr, ou d'une expérience historique partagée (les baby boomers, les millenials).

Âge et changement linguistique

La question principale qui a retenu l'attention des sociolinguistes est le lien entre les différences intergénérationnelles et le changement linguistique (Thibault, 1997). Les études synchroniques du changement linguistique cherchent ainsi à identifier des corrélations entre l'âge des locuteurs et la fréquence de certaines variables. La présence d'une corrélation peut être expliquée par deux phénomènes : elle peut refléter un changement linguistique en cours dans la communauté, ou être due au fait que les individus ne parlent pas de la même façon à des époques différentes de leur vie (gradation d'âge). Les adolescent·es, par exemple, utilisent davantage de variantes vernaculaires que les adultes. Toute la question est de savoir quel phénomène explique les différences constatées entre les classes d'âge.

C'est un projet difficile, parce qu'on ne peut pas se contenter de compa-

rer des tranches d'âge à un moment donné. Deux méthodologies permettent d'isoler la variation liée au cycle de vie d'une personne et la variation liée au changement linguistique. La première (*trend study*) consiste à reproduire une même étude dans une même communauté à des époques différentes, en s'intéressant à chaque fois à plusieurs classes d'âges différentes. Cette méthodologie a été adoptée, par exemple, pour reproduire l'étude réalisée par Labov en 1966 (2006 pour la seconde édition) sur la prononciation du /r/ à New York (Guy, 2018 ; Mather, 2012). La *panel study* est la seule méthodologie qui peut réellement montrer comment les usages linguistiques varient au cours de la vie. Elle est plus complexe à mettre en œuvre, car il faut suivre les mêmes individus tout au long de leur vie (G. Sankoff, 2006). Pour cette raison, la plupart des études du changement linguistique ont été des *trend studies* (Eckert, 1998).

Changement au cours du cycle de vie

Pour Labov (2001), le langage change, de la naissance à l'âge adulte, selon un modèle composé de trois étapes. Tout d'abord, l'enfant acquiert le langage en le modelant sur la façon de parler de la personne qui s'occupe de lui le plus souvent (sa mère, généralement). C'est la phase d'acquisition. Quand il entre à l'école, il commence à utiliser de nouvelles variantes au contact de la communauté. C'est ce que Labov appelle l'incrémentation. Cette phase se poursuit jusqu'à la fin de l'adolescence. Ensuite vient la phase de stabilisation : les individus cessent d'intégrer de nouvelles formes à leur façon de parler. L'hypothèse de Labov est que leur façon de parler ne change plus au cours du reste de leur vie ; le conservatisme de l'âge adulte serait dû à la pression d'utiliser la langue standard dans le monde du travail. Notons toutefois qu'aujourd'hui, le jeunisme exerce une autre pression, et des adultes s'approprient parfois le langage des jeunes (Branca-Rosoff, 2018).

Cette perspective, couplée au fait que l'étude de l'âge a longtemps été majoritairement motivée par la volonté d'identifier les changements linguistiques, généralement phonologiques, a conduit de nombreux sociolinguistes à considérer l'âge comme étant une force immuable, face à laquelle les individus n'ont aucun choix (contrairement à la classe sociale, par exemple, où la mobilité est possible). Pour ces linguistes, l'âge n'a pas vraiment de dimension sociale, dans le sens où il n'est pas construit par les individus. Coupland (2001) regrette le fait que beaucoup ne considèrent pas l'âge adulte et l'âge mûr comme étant « sociolinguistically crucial » (p. 188). Pour lui, l'âge est le parent pauvre de la sociolinguistique, par rapport à d'autres variables abondamment étudiées comme le genre, la classe sociale et l'ethnicité. Eckert fait le même constat : « There has been no concerted study of variation from a life-course perspective » (Eckert, 1998, p. 152).

L'adolescence

L'adolescence est la phase de la vie qui a été la plus abondamment étudiée par les sociolinguistes. Pendant cette période, spécifique à l'époque

moderne et aux sociétés industrialisées, les adolescent·es s'écartent de la sphère familiale pour devenir « elles-mêmes » ou « eux-mêmes ». En même temps, on leur refuse l'entrée dans le monde des adultes en les isolant dans les collèges et les lycées. L'école devient une « hothouse for the construction of identities » (Eckert, 1998, p. 163), où les changements intellectuels, sociaux et physiques s'accélèrent. Par conséquent, c'est à l'adolescence que les usages non standard sont les plus importants. Consciemment ou inconsciemment, les adolescent·es utilisent des mots que leurs parents n'emploient pas, pour s'affranchir de leur autorité. Le langage est également un moyen pour les adolescent·es de marquer la solidarité avec leurs pairs (Tagliamonte, 2016b). Il leur permet d'indiquer à quel groupe ils appartiennent (par exemple, chez Eckert, les « burnout » et les « jocks »). Comme elle est une période particulièrement non conformiste, l'adolescence est une phase clé pour l'étude du changement linguistique. Certains des usages adolescents influencent durablement l'ensemble de la langue d'une communauté : « adolescents act as major agents of linguistic change » (Eckert, 2003, p. 391).

Âge et genre

Pour Eckert (2003), l'étude sociolinguistique de l'âge est indissociable de celle du genre, mais aussi d'autres variables comme la classe sociale et l'ethnicité. Elle appelle les sociolinguistes à étudier le cycle de vie et la façon dont les différentes étapes de la vie sont « genrées ». L'interaction du genre a notamment été mise en évidence en Finlande par Paunonen (1994) (cité par Eckert, 1998), qui a montré, en combinant une *trend study* et une *panel study* que les usages des femmes devenaient moins normatifs avec le temps, contrairement à ceux des hommes.

1.4.2 L'ethnicité

Note : dans la section qui suit, nous utilisons le terme « race » pour désigner le mot anglais « race ». Cependant, dans les autres sections et chapitres de la thèse, nous utilisons le terme « ethnicité » pour désigner l'anglais « race », à cause du caractère controversé du terme français.

1.4.3 Race et biologie

Historiquement, la race a été définie sur une base biologique, par la classification des êtres humains en fonction de leurs traits physiques et génétiques. La division traditionnelle en trois races (blanche, noire, asiatique) tire son origine dans la pensée européenne médiévale, à l'époque où les Européens connaissaient uniquement l'existence de l'Europe, de l'Afrique et du Proche-Orient (Haney-Lopez, 1994). Cette conception de la race ne s'appuie cependant sur aucune base scientifique solide, puisque les différences génétiques à l'intérieur des groupes raciaux sont plus importantes que les différences entre les groupes raciaux. De plus, la variation génétique est graduelle (il n'y a pas de différence nette entre les groupes), et corrélée avec

la géographie. Il y a ainsi une plus grande distance génétique entre les Espagnol·es et les Suédois·es qu'entre les Espagnol·es et les Nord-Africain·es (Haney-Lopez, 1994).

Par ailleurs, la science a montré que les caractéristiques physiques, comme la couleur de la peau ou des cheveux, ne sont pas fortement corrélées avec la variation génétique. Cela explique que des populations peuvent être très proches sur le plan génétique et avoir des caractéristiques différentes, comme c'est par exemple le cas pour les Européen·nes et les Indien·nes du nord de l'Inde. L'inverse est également vrai : des populations qui se ressemblent physiquement peuvent avoir un profil génétique très différent. Pour toutes ces raisons, le concept de race biologique est aujourd'hui rejeté par la communauté scientifique. En même temps, parce qu'elle semble découler du bon sens et a une apparence de scientificité, cette conception de la race reste très répandue (Haney-Lopez, 1994).

1.4.4 Une construction sociale

Comme le genre, la race est une construction sociale, basée sur la culture, et non sur la biologie. Elle est créée par des processus sociopolitiques et sociopsychologiques (Kolko et al., 2000). Haney-Lopez (1994) raconte comment une race a été inventée ; au début du 19ème siècle, les Américains ont assigné aux peuples d'Amérique latine des nationalités et des races. Un Mexicain pouvait donc être blanc, noir, ou asiatique. Toutefois, dans les années 1840, la guerre américano-mexicaine a modifié la perception que les Américains avaient des Mexicains. Le mépris et l'hostilité qu'ils leur manifestaient se sont traduits par la transformation de la nationalité mexicaine en « race » mexicaine. Les races sont donc des constructions fluides : les concepts de « whiteness » et de « blackness » varient selon les époques et les pays. Au début du 20ème siècle, par exemple, Italiens et Irlandais n'étaient pas considérés comme des blancs. Et, aujourd'hui, « noir » n'a pas la même signification aux États-Unis et au Royaume-Uni, où le terme peut désigner des personnes originaires du Pakistan (Kolko et al., 2000) (alors que les personnes d'origine pakistanaise sont considérées comme étant asiatiques aux États-Unis).

Pour Haney-Lopez, la construction raciale a quatre caractéristiques principales. Les races sont créées par les êtres humains ; elles font partie d'un tissu social qui intègre également les relations de genre et de classe ; les systèmes de pensée qui les entourent changent rapidement ; et les races sont construites par opposition les unes aux autres plutôt qu'isolément. Il souligne que les notions de masculinité et de féminité sont indissociables de la construction raciale : « the archeology of race soon becomes the excavation of gender and sexual identity » (Haney-Lopez, 1994, p. 32). En France, Dorlin a retracé les articulations entre genre, sexualité et race. Elle met en évidence leur « rapport génétique, c'est-à-dire leur engendrement réciproque », soulignant que sexisme et racisme sont « inextricablement liés d'un point de vue historique » (Dorlin, 2014, p. 12).

1.4.5 Race et ethnicité

L'ethnicité est un concept distinct de celui de *race* ; elle a longtemps été considérée comme son pendant social, par opposition à une prétendue base biologique de la *race*. L'ethnicité est aujourd'hui définie de diverses façons par les anthropologues et les sociologues (Fought, 2006). Pour Marger (2014, p. 490), un groupe ethnique est : « A group within a larger society that displays a unique set of cultural traits and a sense of community among members ». Même s'il n'y a pas de consensus sur la définition d'ethnicité, les chercheur·es soulignent généralement trois caractéristiques fondamentales. Tout d'abord, comme la race, l'ethnicité est une construction sociale, qui n'est pas définie par des critères mesurables. D'autre part, elle ne peut pas être comprise ou étudiée isolément ; elle est intimement liée à d'autres variables sociales comme le genre et la classe sociale (Fought, 2006). Comme Bucholtz l'écrit, « any performance of ethnicity is always simultaneously a performance of gender » (Bucholtz, 1995, p. 364). Enfin, l'ethnicité est construite à la fois par l'individu et par la communauté (Fought, 2006).

1.4.6 Ethnicité et langage

Note : À partir de cette section, nous utilisons le terme « ethnicité » pour désigner le concept de « race », comme expliqué p.41

Comme le genre, l'ethnicité est construite en partie par les choix linguistiques des individus. Pour ce faire, ils ont plusieurs ressources à leur disposition, dont leur langue maternelle, le *code-switching*, des procédés discursifs, des procédés prosodiques, des variables phonétiques, syntaxiques et lexicales, ou encore l'emprunt de procédés linguistiques utilisés par des individus d'autres ethnicités (Fought, 2006). Les relations entre langue et ethnicité ont commencé à être étudiées par les sociolinguistes américains dans les années 1960, avec les descriptions de plusieurs variétés d'anglais comme, par exemple, l'anglais afro-américain (Baugh, 1983 ; Labov, 1972), l'anglais chicano (Fought, 2003), et l'anglais parlé par les Amérindiens (W. L. Leap, 1993). Plutôt que de décrire les diverses variétés d'anglais, d'autres chercheur·es ont exploré la manière dont les identités ethniques sont construites par le langage. C'est le cas de Zentella (1997), qui a étudié les enfants portoricains bilingues à New York, ou de Cutler (1999) et Bucholtz (1999b), qui se sont intéressées à la façon dont des adolescents blancs utilisent des éléments typiques de l'anglais afro-américain et de la culture hip-hop.

Une grande partie de ce que nous avons écrit sur l'étude du genre et de la langue vaut également pour l'étude de l'ethnicité et de la langue. Wolfram et Schilling (2016) soulignent qu'il est difficile d'établir des corrélations simples entre les usages linguistiques et la race, car l'ethnicité est un concept fluide et multidimensionnel, qui est souvent façonné en grande partie par la langue. Lorsque l'on étudie l'ethnicité, il ne faut pas oublier que la construction identitaire est complexe, et que l'ethnicité n'en est qu'une des facettes, qui peut être plus ou moins mise en valeur selon le contexte

(Fought, 2006).

Les Afro-Américain·es

Les Afro-Américain·es occupent une position unique dans la société américaine. Ils sont présents aux États-Unis depuis plus longtemps qu'aucun autre groupe d'immigrants ; victimes d'une immigration involontaire, ils ont subi (et continuent à subir) des discriminations tout au long de leur histoire. Ils sont, sans aucune ambiguïté, « américain·es », mais n'ont pas atteint le même statut que d'autres immigrant·es, et se sentent plus à l'écart de la société blanche que d'autres minorités ethniques (Phinney & Onwughalu, 1996). Depuis les années 1990, de « nouveaux » Afro-Américain·es sont apparus, avec d'importantes vagues migratoires venues d'Afrique subsaharienne et des Caraïbes. Contrairement aux Afro-Américain·es issus de l'esclavage, qui sont présents dans tout le pays, ces « nouveaux » Afro-Américain·es sont concentrés sur la côte Est. De plus, même s'ils souffrent également de la ségrégation, ils jouissent d'un statut socioéconomique supérieur à celui des descendants des esclaves (Logan, 2007).

L'anglais afro-américain a été abondamment étudié ; il a fait l'objet de cinq fois plus de publications qu'aucun autre dialecte américain (Wolfram & Schilling, 2016). Depuis les années 1960, il a reçu différentes appellations, dont *Negro Dialect*, *Nonstandard Negro English*, *Black Communications*, *Black Folk speech*, *Black English*, *Vernacular Black English*, *Afro-American English*, *African American Vernacular English*, et *African American Language* (Green, 2002). Le nom qu'on lui a donné a souvent reflété le contexte sociopolitique américain. Certaines appellations lui confèrent un statut de langue à part entière (*African American Language*), mais il est généralement considéré comme un dialecte (Green, 2002). La distinction entre langue et dialecte est motivée par des critères extralinguistiques, et, dans ce cas précis, est compliquée par l'histoire et le statut des Afro-Américain·es. Les premières descriptions de l'anglais afro-américain, comme celle de Labov (1972), se sont intéressées à la langue des jeunes afro-américain·es des quartiers défavorisés des centres-villes, qui parlent une variété caractérisée par l'emploi important d'argot (appelée *African American Vernacular English* par Labov). Elles se sont focalisées sur les structures grammaticales et phonologiques qui diffèrent le plus de l'anglais américain standard, comme le *be* invariable (*they always be playing*), l'absence de copule (*she nice*), la réduction des clusters consonantiques (*Wes' Africa* au lieu de *West Africa*) ou la prononciation [f] de *th* (*baf* pour *bath*). Ces descriptions ont mis en avant une version « idéalisée » et uniforme de la langue des Afro-Américain·es, au détriment de la variation sociale et géographique (Wolfram & Schilling, 2016).

En réalité, les structures de l'anglais afro-américain sont présentes dans d'autres dialectes d'anglais. Ce qui rend l'anglais afro-américain unique, c'est la façon dont il combine ces structures. De plus, l'anglais afro-américain n'est pas utilisé par tou·tes les Afro-Américain·es (Green, 2002). Et quand il l'est, il ne l'est pas forcément pleinement : de nombreux·ses Afro-Américain·es indexent leur ethnicité par la langue sans employer toutes les res-

sources phonologiques et grammaticales de l'anglais afro-américain. La variation régionale est importante ; sous l'influence du dialecte new-yorkais, les Afro-Américain·es de New York prononcent ainsi moins fréquemment le *r* en position finale que les communautés afro-américaines des Appalaches (Wolfram & Schilling, 2016).

Il n'y a pas de consensus sur les origines de l'anglais afro-américain. Plusieurs hypothèses ont été avancées : pour certain·es, l'anglais afro-américain refléterait l'anglais britannique parlé dans les colonies américaines, pour d'autres, il serait le résultat de l'évolution d'un créole né des premiers contacts entre Africain·es et Européen·nes. L'émergence de nouvelles sources, dont les mémoires, lettres et écrits d'anciens esclaves, montrent toutefois qu'au 19ème siècle, l'anglais afro-américain n'était pas si différent des dialectes parlés par les Américain·es d'origine européenne. L'hypothèse la plus probable est que l'anglais afro-américain a été influencé par un substrat africain et par le créole parlé sur le littoral d'Afrique de l'Ouest et dans les îles des Caraïbes où séjournaient les esclaves avant d'arriver sur le continent. Le développement de la langue afro-américaine a été influencé par l'esclavage, les lois Jim Crow, la ségrégation « de facto » qui persiste aujourd'hui aux États-Unis, et par la grande migration afro-américaine du milieu du 20ème siècle, depuis le sud rural vers le nord urbain. Il est également dû au fait que les Afro-Américain·es ont, depuis longtemps, une identité culturelle et ethnique forte ; dans la deuxième moitié du 20ème siècle, on a ainsi vu apparaitre une norme suprarégionale, composée de traits linguistiques présents partout où l'anglais afro-américain est utilisé (Wolfram & Schilling, 2016).

De nombreux Afro-Américain·es indexent leur identité ethnique par la langue, en se distanciant d'autres groupes, et notamment des blanc·hes. Pour certain·es, adopter l'anglais américain standard est l'équivalent d'« acting white », c'est-à-dire de se comporter comme un·e blanc·he (Wolfram & Schilling, 2016). Évidemment, la façon dont chaque personne puise dans les ressources linguistiques de l'anglais afro-américain dépend de nombreux facteurs. Il peut y avoir une interaction avec l'âge : Van Hofwegen et Wolfram (2010) a montré, dans une étude longitudinale, qu'il y a des pics et des creux dans la fréquence des caractéristiques de l'anglais afro-américain pendant l'enfance et l'adolescence. Wolfram (1969) a mis en évidence une interaction avec le genre : dans son étude menée à Détroit, les femmes afro-américaines utilisaient une langue plus standard que les hommes. La fréquence des structures de l'anglais afro-américain dépend également du contexte des discussions, et peut varier, dans une même conversation, en fonction du sujet évoqué (Schilling-Estes, 2004).

Même si l'anglais afro-américain continue à être rejeté par les institutions et le monde du travail, il est aujourd'hui dynamisé par la place importante qu'il occupe dans la culture populaire, aux États-Unis et dans le monde entier. Certain·es locuteur·trices non afro-américain·es utilisent des éléments de la langue afro-américaine pour indexer une affiliation à la culture urbaine et « cool » du hip-hop. De nombreux termes d'argot américain ont ainsi une origine dans l'anglais afro-américain, même si la façon

dont les Américain·es blanc·hes se les approprient a rarement été explorée (Bucholtz, 2012).

Les Hispaniques

Les Hispaniques sont, après les Amérindien·nes, les plus anciens habitants de l'Amérique ; ils sont également le plus important groupe d'immigrants récents aux États-Unis. Entre 1980 et les années 2010, la population d'Hispaniques a quadruplé, pour atteindre près de 17 % de la population (Wolfram & Schilling, 2016). La migration hispanique est très diverse. Des personnes d'origine mexicaine vivent depuis des siècles au Texas, au Nouveau-Mexique, en Californie et dans l'Arizona ; des populations portoricaines vivent dans le Nord-Est depuis des générations, et le sud de la Floride a vu arriver de nombreux réfugiés cubains. Dans d'autres régions, l'immigration hispanique est plus récente ; ces dernières années, des communautés rurales et urbaines du Midwest, du Sud-Est et du Nord-Ouest ont par exemple accueilli des vagues d'immigrant·es d'Amérique centrale (Wolfram & Schilling, 2016).

Plusieurs termes sont utilisés pour désigner les populations hispaniques. Fought (2003) a remarqué que les personnes qu'elle a étudiées à Los Angeles préfèrent les termes « Mexican American », « Chicano » et « Latino » à « Hispanic », qui est pour eux « a white person's word » (p. 17). Le terme « hispanique », qui a été employé pour la première fois dans un recensement en 1980 (Salinas Jr., 2019), semble effectivement être plutôt utilisé par les institutions (Wolfram & Schilling, 2016). Notons qu'il n'a pas le même sens que le terme latino : « hispanique » désigne les personnes issues de pays qui ont l'espagnol comme langue principale, tandis que « latino » se réfère aux personnes venues du Mexique, des pays d'Amérique centrale et du sud, et des Caraïbes, même si l'espagnol n'y est pas parlé. Ces deux termes ont pour point commun de désigner une ethnicité, et non une « race ». Une personne hispanique peut ainsi être blanche, noire, ou amérindienne. Aujourd'hui, le terme non genré « Latinx », apparu en 2004 dans les communautés queer, est de plus en plus utilisé, notamment sur les réseaux sociaux (Salinas Jr., 2019). Le terme « Chicano », enfin, désigne les populations d'origine mexicaine vivant dans le Sud-Ouest des États-Unis (Wolfram & Schilling, 2016).

À cause de la grande diversité des populations hispaniques, on ne peut pas parler de variété unique de « Latino English ». En effet, certaines communautés latino ne parlent que l'anglais, d'autres parlent principalement l'espagnol, et d'autres sont bilingues, à différents degrés (Wolfram & Schilling, 2016). L'interaction entre l'ethnicité et d'autres facteurs socioculturels, sociohistoriques et sociopsychologiques explique la diversité de l'anglais latino. Il existe toutefois des caractéristiques phonologiques et grammaticales communes à de nombreux locuteur·trices de l'anglais latino, comme la prononciation « pleine » du schwa (qui, dans le mot *because* devient par exemple un [i]), et le fait que les syllabes tendent à avoir toute la même durée, ce qui n'est pas le cas en anglais américain standard (Wolfram & Schilling, 2016).

L'anglais latino semble par ailleurs être plus perméable à l'anglais afro-américain que l'anglais parlé par les populations blanches et non hispaniques. Wolfram (1974) a montré que des Portoricains de New York ont adopté deux caractéristiques grammaticales de l'anglais afro-américain : l'absence de copule (*he tired*) et le *be* habituel, qui indique une action récurrente (*he be tired*). Ces caractéristiques ont également été mises en évidence dans l'anglais chicano (Fought, 2003 ; Santa Ana & Bayley, 2004), et chez des adolescents de Caroline du Nord (Carter, 2013). Wolfram remarque que l'anglais latino peut être influencé par l'anglais afro-américain même en cas d'interactions limitées entre les deux communautés. Fought (2003) précise qu'il est difficile de déterminer l'étendue de l'influence de l'anglais afro-américain dans le développement de l'anglais latino.

Deux études ont mis en évidence une interaction entre le genre et l'ethnicité dans l'anglais latino. Poplack (1978) a montré que les adolescentes portoricaines utilisaient davantage le système vocalique de l'anglais de Philadelphie (associé avec les Européen·nes-Américain·es) que les garçons portoricains. Eux, par contraste, adoptaient plus volontiers des éléments de l'anglais afro-américain. Carter (2013) partage ce dernier constat, d'après son étude d'adolescent·es latinos de Caroline du Nord. Pour lui, l'utilisation de caractéristiques de l'anglais afro-américain par les garçons n'était pas nécessairement une façon d'indexer la masculinité latino. Il émet l'hypothèse que cela peut être lié au fait que les garçons avaient davantage de contacts avec des adolescents non latinos, alors que les filles préféraient rester entre elles et parler espagnol.

Les Asiatiques

En Amérique du Nord, la catégorie « asiatique » est extrêmement hétérogène. Elle comprend des individus originaires de trois grandes régions : l'Asie de l'Est (Japon, Chine et Corée), l'Asie du Sud-Est (Vietnam, Laos et Cambodge, entre autres) et l'Asie du Sud-Ouest (Inde, Pakistan et Sri Lanka). Les Asiatiques sont le groupe ethnique qui a le plus augmenté aux États-Unis entre 2000 et 2010 ; il compte aujourd'hui près de 20 millions d'individus (Bureau, p. d.). En même temps, la catégorie « asiatique » est fortement racialisée aux États-Unis. Les Asiatiques affrontent un type de racisme différent de celui vécu par les Afro-Américains. D'un côté, ils sont considérés comme la minorité modèle. On les représente comme des individus travailleurs, intelligents et particulièrement doués pour les sciences et les mathématiques. Leur réussite universitaire et économique est exagérée ; elle est perçue comme étant le produit de leur culture de l'effort, et non de l'aide financière de l'État (Pyke & Dang, 2003).

Même s'il est positif, ce stéréotype est dangereux. Tout d'abord, faire des Asiatiques un exemple à suivre est une façon de dévaloriser les autres groupes raciaux, qui ne feraient pas assez d'efforts pour réussir. Ensuite, le stéréotype n'est pas forcément si positif que cela. Les Asiatiques sont souvent représentés comme des *nerds* (« intellos »), mal à l'aise en société, et *uncool*. De plus, il ne tient pas compte de la diversité des expériences du groupe asiatique, étant basé sur les immigrants de pays d'Asie de l'Est et

du Sud-Ouest. Les Asiatiques souffrent également d'un autre stéréotype, celui des « forever foreigners » : ils sont d'éternels étrangers qui ne sont pas perçus comme des Américains (Tuan, 1998). Cela s'accompagne d'une glottophobie : on pense souvent qu'ils ne parlent pas anglais et la culture populaire se moque souvent de leur façon de parler (Wolfram & Schilling, 2016). Un troisième stéréotype a émergé, celui de la minorité à problème ; il est accolé aux immigrant·es et aux descendant·es d'immigrant·es d'Asie du Sud-Est, souvent défavorisés sur le plan économique, qui ne sont pas vu·es comme des « nerds » mais comme des « gangsters dangereux » (Bucholtz, 2004). Pour toutes ces raisons, et malgré le fait qu'ils et elles sont parfois considérés comme des « honorary whites », les Asiatiques occupent une position intermédiaire entre les blanc·hes et les Afro-Américain·es depuis le milieu du 19ème siècle (Pyke & Dang, 2003).

Contrairement à l'anglais afro-américain et aux variétés d'anglais latino, l'anglais des Asiatiques n'est pas une variété « marquée » (Bucholtz, 2004), même si certaines études ont mis en évidence des différences avec l'anglais américain standard (Hall-Lew, 2009 ; Hanna, 1997). En effet, l'immense diversité culturelle, économique et linguistique des Asiatiques a empêché la formation d'un « Asian American English » (Bucholtz, 2004). Les Asiatiques de la classe moyenne parlent donc en général l'anglais américain standard de la région où elles et ils vivent. Sans doute pour cette raison, les Asiatiques ont fait l'objet de relativement peu d'études sociolinguistiques, comparés aux autres groupes ethniques (Wolfram & Schilling, 2016). Depuis les années 2000, les Asiatiques commencent toutefois à attirer l'attention des sociolinguistes. Les études réalisées se focalisent généralement sur un contexte urbain ou rural particulier, et non pas sur les « Asian Americans » de manière générale, comme la communauté asiatique de Sunset District à San Francisco (Hall-Lew, 2009), des lycéennes d'origine laotienne (Bucholtz, 2004), ou encore des New-Yorkais d'origine chinoise (Wong, 2007).

Toutefois, ce n'est pas parce qu'il n'existe pas de variété marquée d'« anglais asiatique » que les Asiatiques ne peuvent pas construire linguistiquement leur identité ethnique, et notamment par l'utilisation, ou la non-utilisation, de l'argot et d'éléments issus de l'anglais afro-américain (Bucholtz, 2004). Chun (2001) a ainsi observé comment un jeune Coréen-Américain du Texas utilise le lexique d'une version « imaginée » de l'anglais afro-américain pour construire une identité masculine et coréenne, en relation à la fois avec la « whiteness » et la « blackness ». Cette même appropriation de l'anglais afro-américain a été étudiée par Reyes (2005), qui a mis en valeur le « capital linguistique » représenté par cette variété pour des jeunes issus de l'immigration du sud-est de l'Asie, et par Bucholtz (2004) qui a examiné comment des jeunes filles d'origine laotienne adoptent ou pas des caractéristiques de l'anglais afro-américain pour se positionner par rapport aux stéréotypes attachés aux Asiatiques (« nerd » ou « gangster »).

tl ;dr

Notre approche intersectionnelle et quantitative du genre et du langage est ancrée dans une réflexion sur le sexe et le genre, nourrie par les apports de la biologie et des sciences sociales. Elle reconnait le fait que le sexe est un phénomène complexe, et que le genre est une construction basée sur l'exagération des différences biologiques entre femmes et hommes.
Elle délaisse la perspective essentialiste utilisée dans les premiers temps de la recherche sur le genre et le langage au profit d'une vision du genre comme étant créé, en partie, par le langage. Elle remet en question certains résultats des chercheur·es des paradigmes de la domination et de la différence, et s'inspire des travaux de la troisième vague variationniste. Elle prête une attention toute particulière à la non-conformité de genre et à ses expressions linguistiques. Elle montre qu'il est possible d'explorer la langue de façon quantitative dans de grands corpus, à condition d'adopter des méthodes statistiques qui, comme la régression avec interactions, permettent d'examiner les intersections du genre avec d'autres variables.
Notre projet a pu être concrétisé grâce à la richesse textuelle et la diversité sociodémographique du site internet Reddit ; il met en lumière les interactions du genre avec l'âge et l'ethnicité, qui sont également des constructions sociales, étudiées ici dans le contexte nord-américain.

Chapitre 2

La CMC : un terrain fertile pour la sociolinguistique

Ce chapitre commence par introduire brièvement la CMC, ou *Computer Mediated Communication*. Il explore ensuite trois des aspects qui intéressent la sociolinguistique, et qui font l'objet des analyses présentées dans cette thèse. « Qui sont les internautes ? » s'intéresse à deux facettes de l'identité en ligne. Cette section se penche tout d'abord sur la question du recueil de données sociodémographiques, compliqué par le fonctionnement des plateformes de CMC mais crucial pour la sociolinguistique, puis montre comment les internautes construisent leur identité par le choix de leurs pseudonymes. « De quoi parlent les internautes ? » fait le point sur les recherches sur les centres d'intérêt des femmes et des hommes en ligne. Enfin, « Comment les internautes parlent-ils ? » décrit la langue d'internet, et plus particulièrement les 11 phénomènes linguistiques que nous analysons dans la partie IV de cette thèse.

2.1 Un nouvel objet d'étude

2.1.1 Qu'est-ce que la CMC ?

Le terme *Computer Mediated Communication* (CMC) a été utilisé pour la première fois par Kiesler et al. (1984), dans un article qui examine les effets sociopsychologiques de la communication par les outils numériques. À l'époque, le World Wide Web n'existait pas, et la CMC était limitée à un usage professionnel au sein du réseau ARPANET, un système développé par le département de la défense des États-Unis. Près de quarante ans plus tard, la CMC s'est démocratisée, et internet est utilisé au quotidien par des milliards de personnes. La recherche sociolinguistique sur la CMC a évolué avec les changements technologiques. Dans les années 1990, avant l'essor de l'internet grand public, elle s'est concentrée sur les listes de diffusion, les newsgroups (ancêtres des forums), et l'Internet Relay Chat (discussions en groupes sur des canaux spécifiques). Aujourd'hui, les sociolinguistes s'intéressent principalement au web public participatif, étudiant notamment les

blogs, les forums, les réseaux sociaux et les wikis (Androutsopoulos, 2014).

2.1.2 Un médium hétérogène

La CMC n'est pas monolithique : elle est composée d'une vaste gamme de registres utilisés sur le web, les réseaux sociaux, les SMS, la messagerie instantanée, le chat, etc. Elle est donc aussi variée que les technologies qui la composent. Les différents types de CMC ont toutefois un point commun : ils permettent tous la communication écrite par les nouvelles technologies (Tagliamonte, 2016b). Dans les études sociolinguistiques, la langue de la CMC est souvent appelée tout simplement « CMC ». Certains chercheurs font toutefois la distinction entre le *Netspeak*, la langue d'internet, qui s'affiche publiquement sur le web et les réseaux sociaux, et le *Textese* (Crystal, 2008), la langue des SMS et de la messagerie instantanée, utilisée dans des interactions privées entre utilisateur·trices. Dans cette thèse, nous utilisons le terme « Netspeak » pour désigner l'ensemble des procédés non standard (émoticônes, acronymes, émojis, etc.) que nous étudions.

2.1.3 De nouvelles perspectives

La CMC a ouvert de nouvelles perspectives à la sociolinguistique. Jusque-là principalement intéressée par la langue orale, la discipline s'est tournée vers l'écrit, que l'essor du web et des réseaux sociaux ont rendu facilement accessible. De plus, l'étude du langage a changé d'échelle. Jamais les productions langagières n'avaient été si accessibles aux linguistes. Il est désormais possible de créer des corpus contenant des milliards de mots et les contributions de plusieurs milliers, voire même de millions d'individus. Par exemple, le corpus de Twitter de Eisenstein (2015) comprend 114 millions de messages mis en ligne par 2.77 millions d'internautes ; le corpus de Facebook d'Oleszkiewicz et al. (2017) est composé des interventions de plus de 86 000 utilisateur·rices ; le Wikicorpus (Reese et al., 2010) contient 750 millions de mots et le corpus du forum de discussion Reddit de Baumgartner (p. d.) 800 millions de commentaires. Ces très grands corpus présentent plusieurs avantages : d'une part, ils permettent d'étudier des phénomènes relativement rares et, de l'autre, leur taille limite la sensibilité des analyses statistiques aux valeurs aberrantes qui peuvent fausser les résultats (Eisenstein, 2015).

Les sociolinguistes qui étudient la CMC s'intéressent à plusieurs aspects, dont la variation, le changement linguistique, les contraintes imposées par les supports numériques sur les interactions, l'identité et les relations interpersonnelles en ligne, le multilinguisme et le *code-switching*, ou encore les relations entre langage et mondialisation (Androutsopoulos, 2014). Ces recherches qualitatives et quantitatives puisent dans les diverses traditions sociolinguistiques, comme la sociolinguistique variationniste, la sociolinguistique interactionnelle et l'analyse du discours. Toutefois, le changement de paradigme introduit par la CMC a majoritairement profité au variationnisme, qui a une longue tradition de recherche

quantitative : les chercheur·es utilisent des outils informatiques et des méthodes statistiques poussées comme la régression logistique depuis les années 1970 (Cedergren & Sankoff, 1974).

La CMC a également entrainé de nombreuses innovations dans la linguistique computationnelle, un domaine qui, comme la linguistique variationniste, fait appel à des outils statistiques. Traditionnellement, cette discipline étudiait des corpus principalement composés d'articles scientifiques et de journaux, et ne s'intéressait que très peu aux dimensions sociales de la langue et à la langue informelle (Nguyen et al., 2016). Les corpus de CMC ont changé la donne, avec l'apparition de nouvelles méthodes et de nouveaux objectifs ; par exemple, les spécialistes du traitement automatique des langues essaient aujourd'hui de mettre au point des modèles permettant de déterminer automatiquement le genre ou l'âge des internautes. Aujourd'hui, on assiste ainsi à l'émergence d'un nouveau champ de recherche : baptisé « computational sociolinguistics » par Nguyen et al. (2016), il étudie les phénomènes sociaux avec des méthodes de traitement automatique des langues. Les travaux réalisés dans cette perspective ont montré l'intérêt de la démarche. Par exemple, en reconnaissant la dimension performative du langage, Bamman et al. (2014) et Nguyen et al. (2014) dépassent la vision simpliste du genre adoptée par la linguistique computationnelle et soulignent les limites des méthodes d'identification automatique du genre.

2.1.4 Les difficultés

Si la CMC est un terrain extrêmement fécond pour la sociolinguistique, son étude n'est pas toujours aisée. Tout d'abord, tous les types de communication informatisée ne se prêtent pas facilement à la création de corpus ; ainsi, même si des milliards de SMS sont produits chaque année (Arcep, 2020), il reste difficile d'en recueillir, pour des raisons éthiques et techniques (Nguyen et al., 2016). Les contenus des réseaux sociaux et des forums en ligne sont, par contraste, très accessibles ; toutefois, là encore, des questions éthiques se posent, car les internautes évoquent souvent des sujets privés et intimes. Par ailleurs, la langue de la CMC a très souvent un caractère informel, que les outils de linguistique computationnelle et de linguistique de corpus (comme les outils d'étiquetage morphosyntaxique) ne savent pas encore vraiment prendre en charge (Nguyen et al., 2016). Enfin, le recueil d'informations sociodémographiques, essentiel pour l'étude sociolinguistique, est complexe. Sur les réseaux les plus ouverts, comme Twitter (par contraste avec Facebook, où les profils sont moins souvent publics), les internautes indiquent rarement leur âge, leur identité de genre ou leur ethnicité. Sur les forums de discussion, comme Reddit, où l'anonymat ou le pseudonymat est la règle, savoir qui écrit quoi est encore plus difficile. C'est ce point que nous allons maintenant développer.

2.2 Qui sont les internautes ?

2.2.1 Le recueil d'informations sociodémographiques

Les sociolinguistes et les spécialistes du traitement automatique des langues utilisent plusieurs types de solutions, souvent ingénieuses, pour associer des caractéristiques sociodémographiques aux contenus du web.

Les approches manuelles

Par « approches manuelles », nous désignons ici les modes de recueil de données sociodémographiques qui ne mettent pas en œuvre des techniques automatiques. La première méthode consiste à contacter les internautes pour obtenir les informations d'intérêt. À notre connaissance, elle est rarement utilisée dans le cadre d'études du web public. Finlay (2014) a trouvé une alternative : il a utilisé les réponses des internautes à un sondage lancé sur le subreddit r/atheism par un Redditor qui s'interrogeait sur la composition démographique du forum, ce qui lui a permis de déterminer l'identité de genre de 734 Redditors.

Les linguistes peuvent également se baser sur les informations fournies par les internautes, comme les avatars et les pseudonymes. Dans son étude du forum geek Slashdot, Bucholtz (2002) a ainsi examiné les pseudonymes des participant·es pour tenter de déterminer leur identité de genre. Elle note cependant que cette méthode n'est pas fiable, parce que certain·es internautes « jouent » avec le genre lorsqu'ils ou elles se choisissent un pseudonyme. Il est également possible d'explorer un à un les profils et les productions des internautes pour en tirer les informations désirées ; c'est un processus chronophage et qui, dans de nombreux cas, ne permet pas d'obtenir les données nécessaires. C'est la méthode pour laquelle nous avons opté, car nous avons découvert que Reddit s'y prête mieux que d'autres plateformes, notamment grâce à son organisation en communautés. Nous décrivons en détail notre processus de recueil des données dans le chapitre suivant (→ p. 111). Enfin, Androutsopoulos (2014), dans la lignée des recherches de la deuxième et de la troisième vague variationnistes, propose de ne pas recueillir d'informations sociodémographiques. Il suggère, à la place, d'examiner des catégories propres aux plateformes étudiées (par exemple, les modérateur·trices d'un site et les utilisateur·trices lambda, les expert·es et les novices, etc.), ou de se focaliser sur les pratiques discursives par lesquelles les internautes créent leur identité sociale en ligne et en attribuent une aux autres.

Les approches automatiques

L'inférence automatique utilise les pseudonymes et les noms d'utilisateur·trices pour déterminer l'identité de genre des internautes. Cette méthode consiste généralement à comparer, par des méthodes automatiques, les noms des internautes à une liste de prénoms. Elle est décrite par Mislove et al. (2011), qui ont tenté de déterminer la composition démogra-

phique d'un échantillon de plus de 3 millions d'utilisateur·trices de Twitter américain·es. Utilisant des listes des 1000 prénoms féminins et masculins les plus populaires chaque année aux États-Unis entre 1900 et 2009, ils ont pu identifier le genre de 64.2 % des auteur·es de leur corpus. Cette procédure a été utilisée par d'autres chercheur·es, comme Coats (2017b), qui a réussi à identifier le genre des auteur·es de 62.5 % des tweets finlandais et de 37.5 % des tweets américains de son corpus, ou Cunha et al. (2014), qui ont inféré le genre des auteur·es des tweets de leur corpus brésilien avec une liste de prénoms populaires au Brésil, assignant un genre aux auteur·es de près de 4 millions de tweets, soit 34.7 % de leur corpus. L'inférence du genre à partir des noms d'utilisateur·trices a également été utilisée sur Reddit, avec moins de succès, parce que les noms des Redditors sont plus créatifs que sur Twitter. Thelwall et Stuart (2018), dans leur étude basée sur une partie du corpus de Baumgartner (Baumgartner, p. d.), se sont servi d'une liste des 10 000 prénoms les plus courants aux États-Unis pour attribuer un genre aux internautes, y parvenant dans seulement 4.9 % des cas.

L'ethnicité des internautes a été inférée de deux façons principales. Mislove et al. (2011) ont utilisé les noms de famille des utilisateur·trices de Twitter pour identifier leur ethnicité. Ils les ont comparés à des données recueillies lors du recensement américain de 2000, qui indiquent la distribution ethnique de chacun des noms de famille portés par plus de 100 personnes aux États-Unis. Eisenstein (2015) s'est servi des données de GPS contenues dans certains tweets pour inférer la composition ethnique de son corpus. Il s'est pour cela appuyé sur les statistiques géographiques du recensement de 2010, qui indiquent la composition raciale de chaque comté américain.

La fiabilité de la *gender inference* automatisée a été testée par De Choudhury et al. (2017). Ils ont comparé l'inférence automatique au codage manuel réalisé par deux personnes, obtenant un taux d'agrément de 79 % sur 100 noms d'utilisateur·trices. Ce résultat indique que, dans leur cas, plus d'un utilisateur sur cinq a été mégenré. Thelwall et Stuart (2018) signalent quant à eux que la méthode automatique génère une part non négligeable d'erreurs, dues à l'utilisation de prénoms inhabituels et de diminutifs non repérés par le script qu'ils ont utilisé. Ils estiment que leurs résultats surreprésentent la présence des hommes de 3.5 %, à cause de ces erreurs. L'identification de l'ethnicité et de l'origine ethnique comporte également une part importante d'ambiguïté (Mislove et al., 2011).

On peut aussi noter que les internautes peuvent tout à fait utiliser un prénom dont le genre ne correspond pas au genre avec lequel ils s'identifient. Mislove et al. (2011) soulignent par ailleurs qu'il est possible qu'il y ait une corrélation entre le fait d'être un homme ou une femme, et le choix de révéler son genre en ligne. Burger et al. (2011) remarquent ainsi que, dans leur corpus de Twitter, les femmes semblent plus susceptibles de fournir des informations explicites sur leur genre dans leur description de profil.

2.2.2 Bref aperçu des études sociolinguistiques du genre, de l'âge et de l'ethnicité dans la CMC

Comme le genre est la variable sociodémographique la plus « facile » à recueillir, c'est aussi celle qui a le plus attiré l'attention des chercheur·es, qu'ils ou elles viennent du monde de la linguistique computationnelle ou de la sociolinguistique. Les premières études du genre et de la CMC remontent aux années 1990, avec, notamment, le travail de la linguiste américaine Susan C. Herring, qui s'est intéressée aux styles communicatifs des femmes et des hommes en ligne. Elle a identifié un style masculin caractérisé par le dénigrement, des assertions fortes et provocatrices, le sarcasme et la vantardise, et un style féminin caractérisé par la solidarité et l'atténuation (Herring, 1994, 1996). Elle note que l'internet reproduit « le statu quo sociétal » : le genre n'y est pas invisible, et les différences communicatives remarquées dans des études de la langue orale y sont présentes (Herring, 2003, p. 218). Pour elle, femmes et hommes constituent ainsi, dans le cyberspace, « deux cultures possédant des normes et des pratiques communicatives différentes » (Herring, 2015, p. 151). Certain·es chercheur·es font le même constat (Colley & Todd, 2002 ; Cunha et al., 2012 ; R. Thompson & Murachver, 2001), tandis que d'autres ont des conclusions plus nuancées (Park et al., 2016). Les relations entre genre et lexique, graphies non standard et émoticônes ont également été abondamment étudiées. Il a par exemple été noté que les femmes utilisent davantage de pronoms personnels (Bamman et al., 2014 ; Schwartz et al., 2013), de graphies typiques de la CMC, comme les étirements graphiques et les acronymes (Bamman et al., 2014 ; Coats, 2017b) et d'émoticônes (Oleszkiewicz et al., 2017), et que les hommes emploient plus de mots grossiers (Schwartz et al., 2013). Nous fournissons davantage de précisions sur les résultats qui nous intéressent plus bas, dans la section 2.4. Notons par ailleurs que la plupart des travaux quantitatifs sur le genre et la CMC considèrent femmes et hommes comme des entités homogènes, ce qui contribue à perpétuer les stéréotypes de genre.

L'âge a aussi été étudié, et surtout les usages adolescents (Hilte, 2019 ; Tagliamonte & Denis, 2008). De nombreuses études ont mis en lumière une corrélation négative entre l'âge et la fréquence des éléments non standard : les internautes les plus jeunes utilisent davantage d'étirements graphiques (Rao et al., 2010), d'acronymes (Rosenthal & McKeown, 2011 ; Tagliamonte, 2016a), ou de mots en majuscules (Rosenthal & McKeown, 2011). Ils écrivent par ailleurs des messages plus courts, par exemple sur Reddit (Finlay, 2014) et sur Twitter (Nguyen et al., 2013).

L'ethnicité a été beaucoup moins étudiée avec une perspective quantitative que le genre et l'âge, à cause de la complexité du recueil des données démographiques. De manière générale, aux États-Unis, ce sont principalement les usages des utilisateur·trices afro-américain·es qui ont attiré l'attention des chercheur·es. En effet, la CMC a permis la diffusion de formes écrites de l'anglais afro-américain, auxquelles les Afro-Américain·es, tout comme les autres Américain·es, n'étaient pas exposé·es auparavant (Jones,

2015). Eisenstein et al. (2011) ont montré que, sur Twitter, les personnes qui habitent dans des comtés fortement peuplés par des Afro-Américain·es utilisent davantage d'acronymes et de graphies non standard que les personnes vivant dans des lieux où la population blanche est surreprésentée. Ils suggèrent ainsi que l'anglais afro-américain joue un rôle important dans la langue des réseaux sociaux. Jones (2015) souligne que de nombreux termes spécifiques à l'anglais afro-américain utilisés par des internautes afro-américain·es sur les réseaux sociaux sont ensuite empruntés par les internautes blanc·hes. Ilbury (2020) montre que cette appropriation dépasse le cadre américain, avec une étude qualitative de la façon dont des utilisateurs de Twitter gays et britanniques utilisent des éléments de l'anglais afro-américain pour se construire une image de « Sassy Queens ».

Notons, enfin, qu'à cause de la difficulté du recueil d'informations sociodémographiques, les études quantitatives de la CMC se focalisent dans l'immense majorité des cas sur une seule variable démographique (Nguyen et al., 2016) et que, quand elles prennent en compte plusieurs variables, elles n'examinent que rarement leurs interactions. L'approche intersectionnelle est donc très peu présente dans la sociolinguistique variationniste quantitative de la CMC.

2.2.3 Les pseudonymes

Maintenant que nous avons passé en revue les solutions utilisées par les chercheur·es pour savoir qui sont les internautes, nous allons nous intéresser à la façon dont ceux-ci créent leur identité en ligne, par l'étude du premier mot qu'ils écrivent sur un forum ou un réseau social : le pseudonyme, ou nom d'utilisateur·trice (appelé *username* sur les forums et réseaux sociaux, *handle* sur Twitter ou encore *nick* dans les chat rooms). Les pseudonymes feront l'objet d'une analyse dans le chapitre 7 de cette thèse (→ p. 176).

Le pseudonyme comme identité

Le choix d'un pseudonyme pour interagir en ligne a été décrit comme un « acte rituel » (Crystal, 2006), puisque, dans la plupart des cas, à l'exception de sites véritablement anonymes comme 4chan (où tous les internautes s'appellent « Anonymous »), il est impossible de participer à des conversations en ligne sans pseudonyme. Crystal souligne que se créer un nom d'écran est un acte extrêmement complexe, guidé par la culture de la communauté dont les internautes désirent faire partie (p. 165). Stommel (2007) écrit ainsi que les pseudonymes peuvent être considérés comme des « emblems of self construction on the internet » (p. 142).

Les pseudonymes ont sans doute beaucoup plus d'importance sur les forums de discussion et dans les chat rooms que sur les réseaux sociaux où il est possible de lier son identité en ligne avec son identité réelle (sur Twitter ou Instagram, par exemple). Ce n'est pas le cas dans les chat rooms ou sur des forums comme Reddit, où les profils restent très sommaires, et où aucun lien n'est directement établi entre l'identité de la personne

« hors ligne » et son identité en ligne. Dans ces conditions, le pseudonyme est « almost the total embodiment of a user » (Rintel et al., 2001). C'est l'identité électronique d'une personne (Crystal, 2006), sa seule représentation visuelle sur les forums, et donc le seul indice qu'ont à leur disposition les internautes pour se reconnaitre entre eux (Campbell, 2014). Le choix d'un pseudonyme n'est donc certainement pas arbitraire : il reflète la façon dont les internautes souhaitent être perçus par les autres. Cornetto et Nowak (2006) soulignent que le choix du pseudonyme est tout aussi important, dans la construction de l'identité en ligne, que les choix linguistiques effectués dans les messages, comme l'utilisation d'émoticônes.

Les autres fonctions des pseudonymes

Dans les forums et les chat rooms, les pseudonymes jouent un rôle important dans la gestion des interactions. Ils permettent de maintenir une cohérence dans les échanges dans des fils de discussion à la structure parfois complexe. Utiliser le nom d'un autre internaute permet ainsi de lier les messages aux autres (en précisant à quel·le auteur·e un·e internaute s'adresse) ; ce processus peut être comparé au rôle du regard et du langage corporel dans les conversations en face à face (Crystal, 2006). Les pseudonymes agissent également comme une invitation à l'échange (Crystal, 2006). Les internautes qui se connaissent ou se reconnaissent se saluent souvent en utilisant les pseudonymes des autres. À l'inverse, des formules de politesse plus traditionnelles comme « hello » agissent comme une mise à distance et sont utilisées quand les personnes ne se connaissent pas (Rintel et al., 2001).

Création des pseudonymes

Le choix d'un pseudonyme est une tâche plus complexe qu'il n'y parait, notamment à cause de contraintes techniques. Sur les chat rooms et dans des forums comme Reddit, les pseudonymes doivent être constitués d'une seule chaine de caractères (Crystal, 2006). Le nombre de noms « réels », comme Lucy or MrSmith, est limité, car deux utilisateurs ne peuvent pas avoir le même pseudonyme. Les internautes jouent donc avec la typographie et la morphologie pour créer leurs noms (Crystal, 2006). Nous présentons plus en détail les contraintes associées au choix des pseudonymes sur Reddit dans le chapitre suivant (→ p. 90).

Informations contenues dans les pseudonymes

Les pseudonymes peuvent contenir plusieurs types d'informations sur les internautes. Bechar-Israeli (1995) a remarqué que les informations liées à des caractéristiques physiques et à la personnalité sont les plus fréquentes. Dans son étude de 260 pseudonymes utilisés sur l'Internet Relay Chat, il a trouvé que 45 % des pseudonymes sont en lien avec un trait physique ou de caractère, une profession, une passion, un âge ou un lieu. Parmi les types d'informations que les pseudonymes peuvent véhiculer, le genre est la plus

commune (Stommel, 2007). Le fait que les pseudonymes soient genrés ou non dépend généralement du contexte dans lesquels ils sont utilisés. Campbell (2014) a trouvé que, dans son échantillon de pseudonymes recueillis sur des chat rooms gay, la moitié des pseudonymes dénotent explicitement ou implicitement une identité masculine et/ou gay. Dans les groupes de discussion fréquentés uniquement par des femmes, l'utilisation de pseudonymes féminins est implicitement conseillée par la « female-gendered netiquette » (Hall, 1996). En revanche, sur le forum Hungrig-Online, consacré aux troubles alimentaires, indiquer son genre est important, mais n'est pas obligatoire (Stommel, 2007).

Dans les chat rooms dédiées aux rencontres amoureuses, de nombreux internautes optent pour des pseudonymes « gender-transparent » (Subrahmanyam et al., 2004). Ces noms ont souvent des connotations sexuelles : ils servent à créer des corps sexualisés, dotés de caractéristiques stéréotypées (Del-Teso-Craviotto, 2008, p. 258). Dans une communauté de gaming, les femmes choisissent des noms d'entités féminines monstrueuses, réelles ou fictives, qui évoquent la force et le courage, qualités traditionnellement associées aux hommes (Kennedy, 2006). Dans une chat room, Campbell (2014) note que les pseudonymes peuvent également véhiculer des informations sur l'ethnicité des internautes (« BlackMuscle » ou « asian_cub »), et remarque que les pseudonymes indiquent une ethnicité essentiellement quand leurs détenteurs ne sont pas blancs.

Anonymat et pseudonymes

En ligne, la notion d'anonymat recouvre des réalités différentes selon les plateformes. Dans certains forums, comme 4chan, qui a été créé en 2003 et qui est calqué sur le modèle des *imageboards* japonais, l'anonymat est total. Sur Facebook, l'identité en ligne de l'utilisateur doit refléter son identité « réelle ». Et, sur les forums et les chats rooms, un internaute qui écrit un message est en théorie anonyme : il est uniquement identifié par son pseudonyme, et non pas par une photographie de profil ou un prénom. Toutefois, quand les internautes socialisent dans ces communautés en ligne, ils ne sont pas toujours véritablement anonymes. Campbell (2014) explique que dans les chat rooms, certains internautes peuvent devenir « connus » des autres. Il en va de même sur Reddit ; certain·es Redditors jouissent d'une certaine notoriété dans la communauté, et d'autres sont même « Reddit famous » (Hoffa, 2017). L'identité en ligne peut ainsi devenir une véritable identité, incarnée en premier lieu par le pseudonyme (Cornetto & Nowak, 2006, p. 379).

2.3 De quoi parlent les internautes ?

2.3.1 Hors ligne

La question des différences d'intérêts entre femmes et hommes a été étudiée par différentes disciplines dont la psychologie, la psychologie sociale

et la sociologie. Certain·es chercheur·es ont exploré les différences entre femmes et hommes en analysant leurs conversations. La plupart d'entre eux se sont toutefois davantage intéressés à la façon dont les conversations sont menées, mettant en lumière les styles communicatifs des hommes et des femmes, plutôt qu'aux sujets évoqués (Dunbar et al., 1997).

L'étude de Moore (1922) est considérée comme un travail pionnier sur les centres d'intérêt par l'observation des conversations. Chaque soir, pendant plusieurs semaines, Moore s'est promené lentement sur l'avenue de Broadway à New York. Lors de ses déambulations, il a recueilli des fragments de conversations qu'il a ensuite analysés. Il a remarqué que quand les femmes parlaient avec des femmes, leurs conversations tournaient autour des hommes ou des vêtements, de leurs appartements ou maisons et de la décoration intérieure. Les hommes, quant à eux, parlaient davantage d'argent, de travail et de loisirs.

En 1990, Bischoping (1993) a reproduit l'expérience de Moore à l'université du Michigan, chargeant ses étudiant·es d'écouter des conversations dans une salle de classe, dans la salle du syndicat étudiant et dans plusieurs cafétérias. Elle souhaitait comparer ses résultats avec ceux de son prédécesseur et savoir si les sujets de conversation avaient évolué. Les 262 conversations écoutées par ses étudiant·es révèlent qu'en 1990, les femmes parlaient davantage de travail, d'argent et de loisirs que 88 ans auparavant. Elle note toutefois que les différences entre les sexes persistent, même si elles se sont atténuées. Dans cette même étude, Bischoping analyse les résultats de huit autres travaux sur les sujets de conversation des hommes et des femmes (Carlson et al., 1936 ; Kipers, 1987 ; C. Landis, 1927 ; M. Landis & Burtt, 1924 ; Meil, 1984 ; Sleeper, 1930 ; Stoke & West, 1930 ; Watson et al., 1948, tous cités par Bischoping). Elle en conclut que, dans toutes les études, les femmes parlent plus des gens, des relations et de l'apparence physique que les hommes, et que les hommes évoquent davantage des sujets liés au travail, à l'argent et aux questions politiques et sociales. Elle note que les hommes parlent davantage des loisirs, mais que, dans plusieurs cas, cette différence peut être attribuée au seul fait que les hommes semblent s'intéresser davantage au sport que les femmes. Par leur objectif (la recherche de différences) et leur conception du genre (essentialiste), ces études amplifient bien souvent les différences entre femmes et hommes. Bischoping (1993) relativise par ailleurs les résultats des études réalisées, soulignant que la façon dont les catégories (travail, famille, etc.) sont créées et la façon dont une conversation est placée dans une catégorie peuvent être subjectives et donc influencer les résultats.

Dunbar et al. (1997) ont utilisé une méthodologie similaire, observant à leur insu les conversations d'étudiant·es dans les universités de Londres et Liverpool. À la différence de leurs prédécesseur·es, ils ont tenté de déterminer avec davantage de précision la proportion de temps consacrée à chaque thème. Leurs résultats diffèrent considérablement de ceux des études précédentes. Ils révèlent une seule différence significative : les hommes parlaient plus de travail et d'études que les femmes. Cette différence était amplifiée quand les hommes étaient en présence de femmes. Les chercheur·es

notent également une différence dans les conversations sur des sujets techniques, mais indiquent qu'elle est obscurcie par un grand nombre d'observations zéro (c'est-à-dire de personnes qui n'en parlent pas). Aucune différence n'a été relevée dans les thèmes liés aux expériences personnelles et aux questions sociales. L'étude montre, enfin, que les personnes plus âgées parlent moins de leurs expériences personnelles que les plus jeunes, et que cette tendance est particulièrement marquée chez les hommes.

Les centres d'intérêt des femmes et des hommes ont également été étudiés par le biais de questionnaires, les chercheurs demandant aux sujets de quels thèmes ils parlent le plus souvent, et à quelle fréquence. Haas et Sherman (1982) se sont intéressé·es aux conversations entre personnes de même genre ; ils notent que les sujets de conversation privilégiés dépendent des personnes à qui on s'adresse, et que ceux-ci sont les mêmes pour les femmes et les hommes (le travail avec un·e collègue, la famille avec un parent ou un enfant, les personnes de « l'autre sexe » avec un·e ami·e). Les différences concernent la musique et le sport, privilégiés par les hommes, et les relations, la cuisine et la mode, privilégiés par les femmes. En utilisant une méthode similaire, Aries et Johnson (1983) ont trouvé que les hommes parlent moins que les femmes, et discutent des sujets liés aux domaines « extérieurs » davantage qu'elles. Sehulster (2006) a également utilisé un questionnaire pour étudier les sujets de discussion de 515 étudiant·es américain·es. Il remarque que les hommes disent parler moins que les femmes. Ils parlent davantage de loisirs qu'elles, tandis qu'elles préfèrent les sujets liés à la famille, aux relations amoureuses et aux ami·es.

2.3.2 En ligne

Avec internet et l'essor considérable des forums et réseaux sociaux, les méthodes utilisées pour étudier les centres d'intérêt ont évolué. Ces plateformes offrent une opportunité inédite d'étudier les centres d'intérêt des hommes et des femmes, en donnant aux chercheurs l'accès à une immense masse de données. Plusieurs techniques quantitatives et qualitatives ont été utilisées pour explorer les sujets de discussion. De manière générale, les tendances remarquées par les chercheurs hors ligne semblent également exister en ligne. Les femmes y parleraient davantage de la maison et des relations sociales que les hommes, qui évoqueraient davantage des thèmes liés au travail, au sport, à la politique et à la religion (Schwartz et al., 2013 ; Thelwall & Stuart, 2018 ; Wang et al., 2013).

Twitter, qui donne librement accès à ses données publiques à qui souhaite les exploiter (« About Twitter's APIs », p. d.) est un des espaces privilégiés par les chercheur·es qui explorent les centres d'intérêt genrés. Holmberg et Hellsten (2015) se sont intéressées aux échanges sur le changement climatique sur Twitter en comparant le contenu des tweets des femmes et des hommes ainsi que les hashtags utilisés ; elles ont trouvé que les hommes se focalisent plus sur la politique, l'économie et la science, et les femmes sur les effets sociaux de changement climatique. H. Evans (2016) a utilisé l'analyse de contenu (en classant les tweets par thèmes) pour détecter les

différences entre les thèmes évoqués par des femmes et hommes politiques sur Twitter ; elle a trouvé que si les femmes évoquent davantage des sujets « féminins » que les hommes, leur sujet de prédilection est le « business ».

Wang et al. (2013) ont analysé un million de statuts Facebook en examinant des clusters de mots qui apparaissent souvent ensemble et qui ont des sens similaires. Ils ont découvert que les femmes parlent beaucoup plus de relations et de leur vie personnelle que les hommes, qui ont tendance à évoquer davantage la politique et la religion, et à parler plus de sport. Ils remarquent par ailleurs que les différences entre femmes et hommes sont moins marquées chez les adolescent·es que chez les adultes. Schwartz et al. (2013) se sont également intéressés à Facebook, utilisant l'analyse de mots clés (les mots qui sont significativement utilisés plus fréquemment que les hommes, et vice versa) pour examiner les messages de 75 000 volontaires. Les résultats montrent que les femmes utilisent davantage de termes liés aux processus sociaux et à la maison, tandis que les hommes parlent plus fréquemment de travail et de leurs réussites. Thelwall et Stuart (2018) ont analysé à la fois les taux de participation des femmes et des hommes dans les 100 forums les plus populaires de Reddit et les mots les plus fréquemment utilisés par chaque groupe. Ils ont montré que les hommes participent davantage aux forums consacrés à l'humour, aux jeux vidéo, à l'actualité et à la politique.

2.4 Comment écrivent les internautes ?

Cette section présente les principales caractéristiques du Netspeak et du Textese (→ p. 52), puis décrit plus en détail les procédés que nous avons choisi d'étudier dans cette thèse.

2.4.1 Une nouvelle langue ?

Très vite, les chercheur·es ont eu besoin de définir le langage de la CMC. Au début des années 2000, Crystal a forgé le terme Netspeak pour désigner la langue d'internet. Il écrit qu'internet a entrainé l'apparition d'un nouveau mode de communication linguistique, « a genuine third medium » (Crystal, 2006, p. 52), ce qu'il qualifie d'évènement extrêmement rare dans l'histoire de l'humanité. Même si cette nouvelle langue peut être vue comme de l'écrit qui tire vers l'oral, elle n'est ni vraiment de « l'écrit parlé », ni de « l'oral écrit » (Crystal, 2006). C'est une langue hybride, qui mêle les caractéristiques de la langue écrite et de la langue orale ainsi que des éléments novateurs qui lui sont propres (Thurlow, 2014).

Cette langue se caractérise par sa créativité (Crystal, 2006 ; Thurlow, 2014), qui vient en partie des contraintes de la CMC écrite. Le corps s'efface, et, avec lui, les indices visuels et prosodiques qui participent à une communication efficace. Pour communiquer à l'écrit avec un smartphone ou un ordinateur, les utilisateur·trices doivent trouver des moyens créatifs de retranscrire le rire, l'hésitation, l'ironie ou l'émotion (Herring, 2012). La CMC aurait par ailleurs un effet « désinhibiteur » (Cougnon & François,

2010). Elle favorise, entre autres, l'utilisation de régionalismes, de néologismes, d'argot et d'abréviations, et « s'émancipe de l'écriture conventionnelle » (Boutin, 2012, p. 1). Comme les internautes aiment expérimenter, utiliser de nouveaux mots et jouer avec la typographie, l'internet est un espace où le changement linguistique va très vite (Crystal, 2006).

Toutefois, les linguistes se sont vite rendu compte que les caractéristiques du Netspeak ne sont pas propres à l'internet, et ne sont pas si novatrices que cela. Les représentations du rire, comme *haha* ou *hehe*, remontent à l'Antiquité (« ha-ha | Origin and meaning of ha-ha by Online Etymology Dictionary », p. d.), tandis que l'initialisme *OMG* est attesté depuis 1917 (« Admiral Lord Fisher to Churchill », 2012) (→ p. 71). Les graphies non standard et les abréviations existent elles aussi depuis des siècles (Tagliamonte, 2016a).

Le Netspeak est donc en partie un « recyclage », écrit Shortis : « such respelling is not new, but recycles popular but relatively undocumented practices » (2009, p. 225). Il note que les graphies non standard abondaient dans quatre domaines bien avant l'invention de l'internet : les noms de marques commerciales, la culture populaire (graffiti, chansons, films et jeux), la littérature pour enfants et la sténographie (Shortis, 2009). Tagg et al. (2012) ajoutent à cette liste les textes publicitaires, les dialogues reproduisant le parler de personnages dans les romans, et l'écriture épistolaire. Shortis (2009) souligne ainsi que, même si elles n'étaient pas entrées dans les dictionnaires, les abréviations et autres graphies non standard existaient dans la conscience collective bien avant la CMC. Toutefois, si la CMC n'a pas conduit à l'invention de graphies et de principes orthographiques radicalement différents, elle a permis leur diffusion et leur amplification. Ce que Shortis (2009) nomme « unregimented writing » (p. 227) n'a pas révolutionné l'orthographe, mais a changé la façon dont on conçoit l'orthographe, par la légitimation de formes et de pratiques sous-représentées dans les textes universitaires et les médias, qui sont dominés par l'anglais standard.

2.4.2 La variation orthographique de la CMC

Le caractère intentionnel de la variation orthographique

La variation orthographique, dans la CMC, est telle que l'on peut se demander où est la frontière entre fautes de frappe et graphies délibérément non standard. Les linguistes qui ont étudié la langue des SMS et d'internet soulignent son caractère intentionnel :

> « The deviant spellings we see in text messaging give the impression of people consciously manipulating the writing system, rather than making inadvertent errors. » (Crystal, 2008, p. 48).

Dans leur étude d'un corpus de 11 000 SMS produits au Royaume-Uni de 2004 à 2007, Tagg et al. (2012) ont trouvé peu de formes orthographiques qu'ils ont considéré être des « erreurs » ; il s'agissait, par exemple, d'erreurs de frappe comme *adn* pour *and* ou de manquements aux normes orthographiques comme *definately* pour *definitely*. Ces erreurs représentaient seulement respectivement 1.75 % et 0.92 % de l'ensemble des tokens non

standard relevés dans leur corpus. Varnhagen et al. (2010) notent également que les fautes de frappe sont rares dans leur corpus, arrivant loin derrière les abréviations, acronymes et autres graphies non standard.

Les trois propriétés de la variation orthographique de la CMC

Tagg et al. (2012) décrivent la variation orthographique de la CMC comme suivant trois principes : elle est « functional, principled and meaningful » (p. 369). Les formes non standard sont fonctionnelles dans la mesure où elles sont des réponses à des « exigences fonctionnelles immédiates » (Tagg et al., 2012, p. 369). Elles émergent dans des contextes spécifiques, dans lesquels sont développées des pratiques communes qui sont comprises par tou·tes les participant·es. La tendance à l'abréviation, par exemple, nait du besoin de répondre rapidement à un message, ou de respecter une limite de nombres de caractères. L'utilisation minimaliste de capitalisation et de ponctuation émerge elle aussi souvent (mais pas toujours, comme nous le verrons p. 79) d'un besoin de vitesse (Thurlow, 2003), répondant ainsi à un principe d'économie linguistique.

Par ailleurs, la variation orthographique de la CMC est « principled » (raisonnée) parce qu'elle reflète des tendances présentes dans l'orthographe de textes historiques et contemporains, qui n'est pas gouvernée par le hasard mais offre des alternatives logiques aux formes standard (Shortis, 2009 ; Tagg et al., 2012). On peut ainsi établir des parallèles entre les variantes observées dans les corpus de CMC et dans d'autres textes. Par exemple, les textes publicitaires et les noms de marques ont vu depuis le début du 20ème siècle émerger de nombreuses graphies non standard, comme l'utilisation de la lettre homophone *u* (pour *you*) dans le nom du produit « Uneedabiscuit » (cité par Shortis, 2009), ou celle du *k* phonétique de Kwik Mart (exemples cités par Tagg et al., 2012). Des variantes orthographiques non standard sont également présentes dans les livres pour enfants et les dialogues des romans ; *eye dialect* (→ p. 82) et graphies phonétiques sont utilisés pour refléter les parlers régionaux ou le statut « non éduqué » de personnages (Picone, 2016). Dans les films, la musique et les jeux vidéo, les graphies non standard sont courantes, en partie à cause de l'influence de l'anglais afro-américain (Shortis, 2009). Les échanges épistolaires informels contiennent également des variantes orthographiques. Ainsi, dans leur étude de lettres d'amour envoyées par des jeunes filles londoniennes au 19ème siècle, Kesseler et Bergs (2003) ont trouvé de nombreuses occurrences de graphies non standard, similaires à celles de la CMC : abréviations, absences de capitalisation, omissions d'apostrophes ou encore utilisations de symboles (le fameux *xxx* qui signifie « kisses »).

Enfin, pour Tagg et al. (2012), les variantes orthographiques de la CMC sont « meaningful », ou significatives, parce que l'orthographe est une « pratique sociale » (Sebba, 2007). Les pratiques orthographiques et typographiques contribuent à la construction de l'identité sociale ; c'est un des outils que les internautes utilisent pour se présenter aux autres et pour se positionner dans la communauté (Tagg et al., 2012). Les restitutions paralinguistiques (*I'm parked next to A MINI ! !*) et les approximations phono-

logiques (*I kinda remember wot that is hmmmm*) (exemples tirés de Tagg et al., 2012) indexent ainsi un langage informel proche de la langue parlée. Elles peuvent indiquer une décontraction et une intimité entre les interlocuteur·trices, l'émotion de la personne qui écrit et celle qu'elle souhaite susciter chez celle qui lit, ou être utilisées pour indexer « a sense of play and fun » (Tagg, 2012, p. 183). Les variantes orthographiques peuvent également indiquer l'appartenance à un groupe ou à une communauté.

2.4.3 Catégoriser les procédés du Netspeak

De la difficulté de la catégorisation

Du fait de la variation orthographique et du changement linguistique rapide qui caractérisent la CMC (Crystal, 2006, p. 98), catégoriser ses phénomènes linguistiques est complexe, surtout quand la quantité de données est importante (Tagg et al., 2012). Bien souvent, il n'y a pas une seule variante non standard d'une graphie standard : dans son étude de SMS écrits par des étudiant·es australien·nes, Kemp (2010) relève ainsi 11 graphies différentes de *because* : *bc*, *bcoz*, *coz*, *cos*, *bcos*, *cause*, *bcuz*, *cuz*, *cus*, *bcause* et *bcaz* (p. 63). Par ailleurs, certains éléments participent à la fois de l'orthographe et de la typographie, ou de l'orthographe et de la morphologie (Herring, 2012). Une forme peut parfois être placée dans plusieurs catégories. *2nite*, par exemple, peut à la fois être considéré comme un nombre homophone et une graphie phonétique de *tonight* (Tagg et al., 2012). D'autres graphies résistent en revanche à être mises dans une catégorie (Tagg et al., 2012) parce qu'elles peuvent être dérivées de plusieurs formes différentes : *ur* peut ainsi être une forme non standard de *you're* et de *your* (Tagg et al., 2012).

Pour toutes ces raisons, le processus de catégorisation comporte une part inévitable de subjectivité : dans une certaine mesure, les linguistes imposent leur point de vue sur les données (Tagg et al., 2012). Les chercheur·es se basent souvent sur une catégorisation établie par d'autres, la complétant ou l'adaptant en fonction de ce qu'ils ou elles remarquent dans leur propre corpus. Les catégorisations des formes de la CMC varient donc d'une étude à l'autre, en fonction des corpus, de la variété d'anglais étudiée, et des objectifs des auteur·es.

Exemples de classifications

Les classifications des phénomènes orthographiques et typographiques de la CMC sont plus ou moins détaillées. Danet (2001) liste 9 « caractéristiques de l'écriture digitale » (p. 17) : ponctuation multiple, graphies « excentriques », astérisques, utilisation de capitales, transcription du rire, smileys, descriptions d'actions, abréviations et absence de majuscules. Crystal (2006) fait la distinction entre néologismes, abréviations, procédés typographiques, graphies non standard, ponctuation (émoticônes), et variation grammaticale. Thurlow (2003) classe les phénomènes orthographiques en 10 catégories : réductions (*def* pour *definitely*), contractions (*txt* pour *text*),

g-clippings (*goin* pour *going*), acronymes (*DI* pour *detective inspector*), initialismes (*ASAP* pour *as soon as possible*), chiffres et lettres homophones (*Y* pour *why*), fautes de frappe (*esay* pour *easy*), graphies non standard (*fone* pour *phone*) et graphies phonétiques (*dat* pour *that*). La catégorisation de Tagg et al. (2012), basée sur un corpus de 11 000 SMS produits au Royaume-Uni entre 2004 et 2007, est plus détaillée. Elle comprend 17 catégories : les lettres homophones, nombres homophones, réductions, omissions d'apostrophe, *eye dialect*, *colloquial contractions* (*whaddya*), absences d'espaces (*u2*), omissions de voyelles (*pls*), fautes de frappe (*adn* pour *and*), fautes d'orthographe (*your* pour *you're*), omissions de voyelles ou consonnes doubles (*stil*), morphèmes visuels (*Lunch@12*), erreurs due à la frappe prédictive (*of* pour *if*), graphies régionales (*dis* pour *this*), autres types d'abréviations (*checkd*), graphies ambiguës, et éléments impossibles à placer dans les catégories précédentes. D'autres chercheur·es préfèrent classer les caractéristiques du Textese et du Netspeak en plusieurs grandes familles ; c'est le cas de Shortis (2009), qui distingue graphies « économiques », comme l'omission de voyelle, les acronymes et les réductions, graphies qui imitent la langue parlée (*eye dialect*, graphies phonétiques, étirements graphiques), et jeux graphiques et typographiques.

2.4.4 Les caractéristiques non standard de la CMC étudiées dans notre thèse

Cette section présente les phénomènes non standard que nous avons choisi d'étudier dans cette thèse. Elle n'a donc pas pour prétention de faire une description exhaustive des procédés du Netspeak. Par « non standard », nous désignons les graphies et procédés qui s'écartent des normes de l'anglais écrit, tel qu'il est enseigné à l'école ou qu'il s'affiche dans les médias.

Émoticônes et émojis

Les émoticônes sont composées de plusieurs caractères (chiffres, lettres, symboles, signes de ponctuation, etc.), et représentent souvent des expressions faciales : un sourire (**:-)**), un clin d'œil, (**;)**), un visage qui rit (**XD**). On ne sait pas exactement quand les émoticônes sont apparues ; elles ont été mentionnées pour la première fois en 1982 par Scott Fahlman, informaticien à l'Université Carnegie-Mellon, sur un *bulletin board system*, sorte d'ancêtre du web (« Original Bboard Thread in which :-) was proposed », p. d.). Le mot émoticône, ou *emoticon* en anglais, serait apparu vers 1987 (« Definition of emoticon », p. d.). C'est un mot-valise, composé des termes anglais *emotion* et *icon*. Émoji, quant à lui, vient du japonais 絵 (e, ou image) et 文字 (moji, ou caractère). C'est donc un hasard si les deux mots se ressemblent (« FAQ - Emoji & pictographs », p. d.).

Les émojis sont nés au Japon il y a un peu plus de vingt ans. En Asie de l'Est, des émoticônes plus élaborées que les smileys traditionnels se sont développées dans les années 1980 et 1990, comme **(")(-_-)(")**, qui représente un personnage au visage contrarié qui lève les mains (« FAQ - Emoji & pictographs », p. d.). NTT DoCoMo, le plus gros opérateur de téléphonie mobile

japonais, a commencé à remplacer ces émoticônes complexes par des images qu'on pouvait insérer directement dans un message et a lancé en 1999 une série de 172 pictogrammes. Ces caractères sont rapidement devenus extrêmement populaires auprès des utilisateur·trices de téléphones mobiles japonais (« FAQ - Emoji & pictographs », p. d.). Au début des années 2000, seuls les téléphones japonais pouvaient utiliser des émojis. Il y avait une solution à ce problème : Unicode, le standard informatique qui permet aux caractères de s'afficher sur tous les systèmes d'exploitation partout dans le monde, fondé en 1991 en Californie. En 2007, le consortium Unicode, une organisation composée, notamment, de représentants de Google, Microsoft, Apple et Yahoo ! a créé, sous l'impulsion de Google, l'Unicode Emoji Subcommittee, qui est devenu l'organisme de référence en matière d'encodage des émojis. Les émojis ont été intégrés à Unicode avec Unicode 6.0, qui a été lancé en octobre 2010. Le jeu de base « emoji contient alors 722 émojis, dont 608 nouveaux caractères et 114 caractères déjà présents dans Unicode 5.2, redéfinis, de façon rétroactive, comme des émojis. Les caractères ont été intégrés en masse dans Unicode avec la version 7.0, publiée en 2014 (M. Davis & Edberg, 2020). Depuis, chaque année, de nouveaux émojis sont ajoutés à Unicode.

Même si émoticônes et émojis sont parfois placés dans la même catégorie, ils diffèrent par plusieurs aspects. Les émoticônes sont composées par les utilisateur·trices, à la différence des émojis qui sont contrôlés par les membres du consortium Unicode, c'est-à-dire par les géants du web. Ainsi, même s'il augmente chaque année, le nombre d'émojis est limité. Dans la version 10.0 d'Unicode, lancée en juin 2017, il y avait 1144 émojis, soit 56 de plus que dans Unicode 9.0. Dans les cas des émoticônes, en revanche, le nombre de combinaisons est en théorie infini. De plus, là où les émoticônes sont composées de plusieurs caractères, chaque émoji est un caractère à part entière. Il faut donc appuyer sur plusieurs touches d'un clavier pour faire une émoticône, mais il suffit en général d'un seul clic (sur un ordinateur) ou d'un seul « tap » (sur un smartphone) pour produire un émoji comme 👌 ou 😉.

Des phénomènes abondamment étudiés Émoticônes et émojis sont sans conteste les « stars » des études de la CMC, sans doute à cause de leur caractère novateur. De nombreuses études, qualitatives et quantitatives, se sont intéressées aux émoticônes dès les années 90 ; au début, elles se sont souvent basées sur de petits corpus de messages recueillis dans des forums et des listes de diffusion. Par exemple, Rezabeck et Cochenour (1995) ont recensé les émoticônes produites dans plus de 1500 messages envoyés dans quatre listes de diffusion, et Witmer et Katzman (1997), Wolf (2000) et Provine et al. (2007) ont exploré des corpus de messages publiés dans des forums. Les études des émoticônes utilisées dans les SMS et la messagerie instantanée ont souvent pris pour participant·es des étudiant·es. C'est le cas du travail de Baron (2004) et de Garrison et al. (2011) sur la messagerie instantanée, ou de D. Thompson et Filik (2016) et Tossell et al. (2012) sur les SMS. Les émoticônes ont également été étudiées dans le contexte

de l'email (Skovholt et al., 2014), des chat rooms (Del-Teso-Craviotto, 2008), et des blogs (Huffaker & Calvert, 2005 ; Kavanagh, 2016). Sur Twitter, les émoticônes ont fait l'objet d'études par T. Schnoebelen (2012), Pavalanathan et Eisenstein (2016) et Coats (2017b). Tsou (2016) s'est quant à lui penché sur l'utilisation des émoticônes sur Reddit, et Oleszkiewicz et al. (2017) sur Facebook. D'autres chercheur·es ont utilisé des questionnaires pour interroger les internautes sur leur usage des émoticônes (Derks et al., 2008 ; Kaye et al., 2016).

Les émoticônes et émojis les plus fréquents Les études qui se sont penchées sur les émoticônes montrent que, s'il en existe plusieurs centaines, seul un petit nombre est fréquemment utilisé. Tossell et al. (2012) ont remarqué que les trois émoticônes les plus fréquentes (**:)**, **:(** et **:D**) représentent 70 % de l'ensemble des émoticônes recensées dans leur corpus de SMS. Chez Pavalanathan et Eisenstein (2015), 90 % de toutes les occurrences d'émoticônes sont dues aux 20 émoticônes les plus fréquentes. Dans le corpus de statuts Facebook d'Oleszkiewicz et al. (2017), les quinze émoticônes les plus populaires représentent 99.6 % de l'ensemble des émoticônes répertoriées, et les cinq émoticônes les plus fréquentes 88 % de toutes les émoticônes du corpus.

L'émoticône la plus populaire semble être, dans tous les corpus et sur toutes les plateformes, le smiley ou « happy face ». Il arrive en tête de la liste de fréquence chez Wolf (2000), avec 93 % de toutes les émoticônes recensées, et chez Huffaker et Calvert (2005), où il représente 53 % des émoticônes étudiées. Rezabeck et Cochenour (1995), Skovholt et al. (2014), Oleszkiewicz et al. (2017), et Provine et al. (2007) font le même constat. T. Schnoebelen (2012) remarque également que **:)** est la seule émoticône utilisée par plus de 21 000 des 102 304 internautes de son corpus de tweets. De manière générale, les émoticônes positives sont plus fréquentes que les émoticônes négatives (Huffaker & Calvert, 2005 ; Oleszkiewicz et al., 2017 ; Skovholt et al., 2014). L'émoticône « sad », ou **:(** , est toutefois aussi une des plus fréquentes : elle arrive en troisième position chez Oleszkiewicz et al. (2017).

Concurrence entre émoticônes et emojis Dans leur étude longitudinale de deux corpus de Twitter créés en février 2014 et août 2015, Pavalanathan et Eisenstein (2016) se sont interrogés sur l'impact de l'introduction des émojis sur les émoticônes. Ils ont montré que les utilisateur·trices qui ont adopté les émojis ont tendance à utiliser moins d'émoticônes par rapport à ceux qui n'emploient pas d'émojis. Les auteurs notent que la fréquence des émoticônes souriantes, comme **:)**, et malicieuses, comme **:p** déclinait à un rythme plus rapide que celle des émoticônes tristes comme **:(** . Ils émettent ainsi l'hypothèse que l'introduction des émojis a provoqué, sur Twitter, une diminution de la diversité des émoticônes produites par les internautes.

Fonction des émoticônes et émojis Émoticônes et émojis sont nés d'un besoin de compenser l'absence de paralangage de la CMC écrite (Provine et

al., 2007). Ils servent ainsi à remplacer les gestes, les expressions du visage, l'intonation ou les variations de hauteur et de volume de la voix typiques des conversations en face à face. Ils constituent donc non pas un langage, mais un paralangage : ils ne sont généralement pas destinés à être « lus », et fournissent des indications sur l'intention du locuteur.

Les émoticônes sont souvent considérées comme des marqueurs d'émotion et d'affect. Dans une des premières études consacrées aux émoticônes, Rezabeck et Cochenour (1995) écrivent que le smiley **:-)** exprime les sentiments de la personne qui envoie le message, et que les émoticônes sont un « moyen permettant de mieux préciser les émotions et l'intention derrière une phrase ou un énoncé » (p. 372). Provine et al. (2007) définissent quant à eux l'émoticône comme « un symbole de valence émotionnelle » (p. 306). Cette vision des émoticônes est toutefois considérée comme réductrice par certains, et tou·tes les chercheur·es ne s'accordent pas à dire qu'il s'agit de la fonction principale de ces pictogrammes. Coats (2017b) note que concevoir les émoticônes uniquement comme des marqueurs d'émotion est problématique, parce que de nombreuses études ont montré que les émoticônes peuvent ne pas avoir de portée affective, et remplir plusieurs fonctions discursives différentes.

Les émoticônes peuvent par exemple également indiquer le sarcasme à l'écrit là où, à l'oral, on utiliserait une intonation particulière (Kunneman et al., 2015), et notamment les émoticônes **;)** et **:p** (D. Thompson & Filik, 2016). Provine et al. (2007) comparent les émoticônes aux rires enregistrés des séries télévisées, qui ont pour rôle de signifier au public qu'une blague a été dite. Pour Dresner et Herring (2010), les émoticônes ont avant tout un rôle pragmatique, et doivent donc être comprises en termes linguistiques plutôt que paralinguistiques (p. 250). Skovholt et al. (2014) voient ainsi les émoticônes comme des marqueurs discursifs qui servent à contextualiser ou à modifier un énoncé. Elles peuvent renforcer un message, quand elles sont placées après un remerciement ou une salutation, ou l'adoucir, quand elles viennent après des requêtes ou des corrections (Skovholt et al., 2014, p. 792), jouant ainsi le même rôle que les *hedges* (marqueurs d'atténuation) (Kavanagh, 2016).

Effet des variables sociales Le genre est la variable sociale qui a été le plus abondamment étudiée en relation avec les émoticônes. La majorité des études indique que les émoticônes semblent être un marqueur de féminité et un « acte genré » (Del-Teso-Craviotto, 2008). Dès les années 1990, les chercheur·es ont fait état de différences dans la fréquence d'utilisation entre hommes et femmes. Dans leur étude de 3000 messages postés dans des newsgroups, Witmer et Katzman (1997) ont remarqué que les femmes utilisaient davantage d'émoticônes que les hommes. Il faut toutefois noter que leur échantillon contenait une minorité de femmes (16.4%). La tendance se retrouve dans de nombreuses études, basées sur diverses plateformes et types de contenu. Les femmes utilisent plus fréquemment des émoticônes dans le corpus de messagerie instantanée de Baron (2004), dans les corpus de chat de Del-Teso-Craviotto (2008) et de Fullwood et al. (2013),

et dans celui de SMS de Holtgraves (2011). Dans le corpus de 158 098 SMS de Tossell et al. (2012), les femmes ont utilisé deux fois plus d'émoticônes que les hommes. Le constat est le même sur Facebook (Oleszkiewicz et al., 2017) et sur Twitter (Coats, 2017b).

Les études par questionnaires révèlent aussi des différences entre femmes et hommes. Ogletree et al. (2014), qui ont interrogé 183 étudiant·es américain·es, notent que les émoticônes semblent être pour eux un marqueur de féminité. Dans l'étude de Prada et al. (2018), les femmes déclarent utiliser les émojis plus fréquemment que les hommes, mais pas les émoticônes. Toutefois, certaines études dressent un tableau plus nuancé. Dans un corpus d'articles de blogs écrits par des adolescents, Huffaker et Calvert (2005) ont trouvé que les émoticônes étaient plus nombreuses dans les productions des garçons, ce qui indique une possible interaction du genre avec l'âge. Wolf (2000) n'a pas noté de différence dans le nombre d'émoticônes utilisées par femmes et hommes.

Il y a également des différences dans le type d'émoticônes utilisées par femmes et hommes. Chez Wolf (2000), femmes et hommes emploient la même quantité d'émoticônes, mais les femmes préfèrent les émoticônes qui indiquent l'humour, tandis que les hommes ont une prédilection pour les émoticônes qui expriment le sarcasme et la moquerie. Oleszkiewicz et al. (2017) constatent aussi que les femmes utilisent davantage d'émoticônes positives que les hommes. Huffaker et Calvert (2005) notent que les garçons emploient plus d'émoticônes « flirty » (aguicheuses) et tristes que les filles.

Plusieurs études montrent qu'il y a une corrélation négative entre l'âge et la fréquence d'utilisation d'émoticônes. Les internautes les plus jeunes ont tendance à utiliser davantage d'émoticônes que les moins jeunes. C'est par exemple le cas dans les corpus de Facebook de Schwartz et al. (2013), de Settanni et Marengo (2015), et Oleszkiewicz et al. (2017). Dans leur étude par questionnaire d'internautes portugais, Prada et al. (2018) ont trouvé que les participant·es les plus jeunes indiquent utiliser davantage d'émoticônes et d'émojis que les plus âgé·es. Les plus jeunes ont également une attitude plus positive par rapport aux émoticônes et émojis. Chez Fullwood et al. (2013), en revanche, l'âge n'a pas d'impact sur la fréquence d'utilisation d'émoticônes. Enfin, Eisenstein et al. (2011) ont établi un lien entre ethnicité et usage des émoticônes ; il semble que les internautes blanc·hes utilisent davantage ces pictogrammes que les internautes afro-américain·es et hispaniques.

Formes longues et abrégées des émoticônes Les études de la CMC réalisées dans les années 1990 indiquent qu'à l'époque, le smiley avec un nez, symbolisé par un trait d'union (**:-)**) était plus fréquent que la variante sans nez (**:)**) (Rezabeck & Cochenour, 1995). Avec le temps, il semble que la variante sans nez, ou abrégée, ait pris le pas sur la version avec nez. D. Thompson et Filik (2016), qui ont étudié un corpus de SMS produits par des étudiant·es écossais, ont remarqué, que, dans les 1184 cas où une émoticône pouvait être écrite avec ou sans nez, la variante sans nez l'emportait dans

76 % des cas. Dans le corpus de Facebook créé par Oleszkiewicz et al. (2017), le smiley sans nez est utilisé à une fréquence 18 fois plus importante que le smiley sans nez.

T. Schnoebelen (2012) a étudié les 28 émoticônes les plus utilisées en anglais américain dans un grand corpus de Twitter. Il s'est rendu compte que les internautes qui utilisent le plus d'émoticônes sont aussi ceux qui ont le plus tendance à omettre les nez, et que l'omission du nez ne semble pas motivée par un besoin de concision. Les internautes qui utilisent des nez ont en effet tendance à écrire des messages plus longs que ceux qui n'en utilisent pas. Schnoebelen a remarqué que les internautes qui utilisent les émoticônes abrégées sont plus jeunes que ceux qui préfèrent les formes longues. Ils et elles emploient également davantage d'étirements graphiques, omettent plus souvent les apostrophes, et utilisent davantage de graphies phonétiques. Les émoticônes avec nez, qui sont apparues avant les émoticônes sans nez, seraient donc associées à des formes standard, tandis que les émoticônes sans nez seraient liées à un anglais moins standard, et préférées par les internautes les plus jeunes.

Abréviations

Les abréviations ont été décrites comme étant un des aspects les plus marquants de la CMC (Crystal, 2006, p. 90). Cette pratique n'est pourtant pas nouvelle : l'utilisation d'abréviations remonte à l'Antiquité, avec par exemple les acronymes latins *SPQR* (*Senatus populusque romanus*) et *INRI* (*Iesus Nazarenus Rex Iudaeorum*). Au 15ème siècle, plusieurs dictionnaires d'abréviations ont été rédigés, le plus fameux étant le *Modus Legendi Abreviaturas* (1475 ?), où on trouve des abréviations comme *gle* (*generale*) et *nobc* (*nobiscum*) (Cannon, 1989). L'acronyme *OMG* (*Oh my god*) aurait quant à lui plus d'un siècle : sa première utilisation attestée date de 1917, dans une lettre adressée par l'amiral britannique Lord Fisher à Winston Churchill (« Admiral Lord Fisher to Churchill », 2012).

À partir de la fin du 19ème siècle, le nombre d'abréviations a considérablement augmenté dans la langue anglaise, à cause du besoin croissant de vocabulaire précis dans les domaines de la médecine, du droit, de la politique et du commerce. Cette accélération s'est poursuivie au 20ème siècle, avec la Seconde Guerre Mondiale et l'apparition de l'informatique et de ses nouveaux termes et concepts, puis avec l'émergence des SMS et de l'internet grand public (Mattiello, 2013). De nombreuses abréviations nées dans des domaines techniques sont entrées dans le langage courant, comme *chemo* (*chemotherapy*) ou *mono* (*monucleosis*). D'autres abréviations viennent du langage courant, comme *sis* (*sister*), *varsity* (*university*) ou *hubby* (*husband*).

Même si elles sont de plus en plus nombreuses en anglais, les abréviations restent assez peu étudiées par les linguistes, par rapport aux autres processus morphologiques. Elles sont en effet considérées comme des processus « extra-grammaticaux ». Contrairement à l'affixation et à la composition, l'abréviation ne modifie généralement pas le sens des mots source, et, en anglais tout du moins, le processus d'abréviation est souvent irrégulier et imprévisible (Mattiello, 2013, p. 66). De plus, les abréviations sont

parfois ambiguës : une même abréviation peut par exemple désigner deux termes source distincts (*sub* pour *submarine* ou *substitute*) ou appartenant à deux catégories grammaticales (*dif* pour *difference* ou *different*). Un mot source peut également donner naissance à deux abréviations différentes (*absent without leave* → *A.W.O.L.* ou *awol*), caractéristique qui est elle aussi symptomatique de l'extra-grammaticalité et l'irrégularité des abréviations (Mattiello, 2013). De fait, il existe d'importantes divergences terminologiques dans la littérature. Comme Mattiello (2013), nous avons choisi d'utiliser le mot « abréviation » comme un terme générique désignant deux phénomènes liés : d'un côté, les réductions (*clippings*), et de l'autre, les acronymes et initialismes.

Réductions (*clippings*) Les réductions sont le résultat du raccourcissement d'un mot, qui ne conserve plus qu'une, deux, ou plus rarement trois syllabes, avec une perte phonétique importante (Mattiello, 2013). Les mots ainsi abrégés sont généralement des substantifs (*auto* ← *automobile*), mais peuvent également être des adjectifs (*fave* ← *favorite*), des verbes (*prep* ← *prepare*), et des adverbes (*def* ← *definitely*). Les conjonctions (*'cos* ← *because*), et interjections (*lor* ← *lord*) sont plus rarement abrégées (Mattiello, 2013). Le type de réduction la plus courante, en anglais et dans de nombreuses langues, est l'apocope, qui consiste à supprimer la fin d'un mot (*bro* ← *brother*, *amp* ← *amplifier*). Ce raccourcissement peut s'accompagner de modifications orthographiques destinées à mieux représenter la prononciation du mot source (*biz* ← *business*, *cuke* ← *cucumber*, *Jeez* ← *Jesus*) ou de l'ajout de suffixes (*comfy* ← *comfortable*, *preggers* ← *pregnant*).

Parfois, c'est le début d'un mot qui disparait, un phénomène appelé en français aphérèse (*choke* ← *artichoke*, *tude* ← *attitude*) (Fradin, 2003). Il est éventuellement accompagné d'ajustements orthographiques (*leet* ← *elite*, *nuff* ← *enough*). Dans d'autres cas, seule la partie médiane d'un mot est supprimée (*ana* ← *anorexia*, *cortisone* ← *corticosterone*) ou conservée (*flu* ← *influenza*, *lax* ← *relaxed*), donnant naissance à des termes moins transparents (Mattiello, 2013). Certaines réductions ne retiennent que certaines lettres, généralement des consonnes, apparemment au hasard (*dlr* ← *dollar*, *Jpn* ←, *pls* ← *please*). Elles continuent toutefois, contrairement aux cas précédents, à être prononcées sous leur forme entière, et sont donc utilisées majoritairement à l'écrit, et non à l'oral.

Des frontières floues Mattiello (2013) souligne que la frontière entre les réductions et d'autres processus morphologiques est parfois très floue. Les réductions peuvent être confondues avec des dérivations régressives (*troncation* ou *back-formation* en anglais) du type *edit* (*editor*) ou *burgle* (*burglar*). Seulement, à la différence des dérivations progressives, les réductions ne modifient pas la catégorie grammaticale d'un mot. Par ailleurs, les réductions qui présentent un fort degré de raccourcissement, comme *c.* (*century*) ressemblent à des acronymes ou initialismes. La différence entre réductions et alphabétismes est encore plus ténue dans le cas où deux ou trois lettres non voisines sont conservées, comme *ID* (*identification*) et *TV*

(*television*). Mattiello (2013, p. 72) note que *ID* et *TV* sont des réductions et non des acronymes parce que les lettres conservées proviennent d'un seul mot, et non de plusieurs. Il existe par ailleurs des abréviations purement graphiques (*etc* ← *et cetera*, *Dr.* ← *doctor*), que Mattiello ne considère pas comme des réductions, car le mot entier continue à être prononcé. Certaines abréviations sont hybrides et peuvent à la fois être vues comme des abréviations graphiques ou des réductions. C'est le cas des mots *Dem* (*democrat*) ou de *Jan* (*January*), qui peuvent être prononcés dans leur forme entière ou tronquée.

Enfin, Mattiello fait état d'une dernière catégorie, que l'on peut considérer comme relevant à la fois de la réduction et de l'acronymie, et qui est spécifique aux SMS et à internet. Les syllabes d'un mot sont remplacées par une lettre, un groupe de lettres ou un chiffre homophone (*B4* ← *before*, *THX* ← *thanks*, *U* ← *you*). L'auteure les considère comme des abréviations « purement graphiques, qui ne présentent pas de réduction phonétique » (Mattiello, 2013, p. 87). Cette catégorie est tout particulièrement importante dans l'étude des abréviations de la CMC.

Il existe malgré tout des régularités dans la formation d'abréviations, comme la préférence pour une structure monosyllabique se terminant par une consonne (Kreidler, 2000 cité par Mattiello, 2013). Par ailleurs, même si on considère que les abréviations n'entrainent pas de changements sémantiques, le processus crée parfois un glissement ou un rétrécissement de sens, voire une coupure totale avec le terme source. Les mots *pants* (*pantaloons*) et *mob* (*mobile vulgus*) sont aujourd'hui complètement dissociés de leurs origines, à tel point qu'ils ne sont plus considérés comme des abréviations. Les abréviations ajoutent par ailleurs souvent une connotation de familiarité avec les objets auxquels elles se réfèrent ou avec les personnes auxquelles le locuteur ou la locutrice s'adresse, et sont généralement utilisées dans des contextes moins formels que leurs versions longues (Adams, 1973, p. 135). Les abréviations ont aussi une fonction pragmatique ; elles expriment l'attitude du locuteur, et placent le niveau stylistique du discours à un niveau moins formel (Mattiello, 2013).

Acronymes et initialismes La littérature établit généralement une distinction entre acronymes et initialismes sur la base de leur prononciation (Mattiello, 2013 ; Plag, 2003). Les premiers sont orthoépiques, c'est-à-dire qu'ils se prononcent comme des mots à part entière (*AIDS* → *Acquired Immune Deficiency Syndrome*). Les seconds s'épellent lettre par lettre (*BBC* → *British Broadcasting Corporation*). Dans les deux cas, seules les initiales d'une séquence de mots, qui peut être un titre, une liste, une expression ou un mot composé, sont retenues (Mattiello, 2013). Il existe des cas hybrides, qui peuvent être prononcés de deux façons (*AKA* ou *aka*, *ASAP* ou *asap*), ou qui combinent les deux prononciations (*MS-DOS*, *JPEG*). Plag (2003) note également que seuls les initialismes peuvent comporter des points (*U.F.O.* ← *unidentified flying object*), même si la tendance est à les omettre. Plusieurs graphies sont souvent possibles (*UFO*, *Ufo*, *ufo*) ; cela peut indiquer des degrés divers de lexicalisation, c'est-à-dire « le fait qu'une expression

linguistique accède au statut d'entité codée » (Fradin, 2003, p. 101). Quand les acronymes et initialismes sont pleinement lexicalisés, ils ne sont plus considérés comme des abréviations : la construction devient un seul mot, effaçant la distinction entre les morphèmes (*laser* ← *light amplification by stimulated emission of radiation*).

Les abréviations en CMC Mattiello (2013) cite les SMS comme étant un des médiums responsables de la prolifération des abréviations en anglais et dans d'autres langues, à cause du besoin de concision qu'il impose. Dans le cas des SMS, la taille réduite de l'écran et du clavier, ainsi que la limite du nombre de caractères autorisés (Crystal, 2006), seraient en grande partie responsables de l'apparition d'un langage bien plus abrégé que celui des groupes de chat et des mondes virtuels. Utiliser des abréviations permet également d'écrire plus vite (Herring, 2012). Baron (2004) et Tagliamonte (2016a) ont trouvé que les abréviations, et plus spécifiquement l'acronyme *Lol*, sont la catégorie lexicale caractéristique du Netspeak la plus utilisée dans leurs corpus de messagerie instantanée. Cougnon et François (2010) ont quantifié la réduction due aux abréviations dans un corpus de SMS de français de Belgique ; ils ont noté une réduction moyenne de 9.4 % de la longueur des messages par rapport à leur version non abrégée. Dans l'étude d'un corpus de SMS, Thurlow (2003) a trouvé que 18.75 % du contenu des messages est abrégé.

Les études des abréviations de la CMC semblent indiquer que la proportion d'abréviations est plus importante en anglais que dans d'autres langues. Dans son étude contrastive de SMS anglais et allemands, Bieswanger (2007) a recensé 5.57 abréviations par SMS en anglais, contre 0.86 seulement pour l'allemand. En italien, Herring et Zelenkauskaite (2008) ont compté 1.79 abréviation par SMS envoyés par des femmes, et 1.18 pour ceux écrits par des hommes.

Dans les études de la CMC, la définition d'abréviations varie selon les auteur·es. Herring et Zelenkauskaite (2008), qui les appellent « deletions », incluent les acronymes et réductions, mais aussi l'omission de ponctuation et les graphies phonétiques. Kemp (2010) utilise le terme « textism » pour désigner les abréviations graphiques de type *r u* (*are you*) et les acronymes. Ce que Mattiello (2013) appelle « abréviations graphiques », comme *c u* (*see you*) est parfois appelé « rebus writing » (Danet & Herring, 2007).

Études sociolinguistiques des abréviations de la CMC Plusieurs études ont trouvé que l'identité de genre est corrélée avec le nombre d'abréviations. Baron (2004) a remarqué que les femmes étaient plus susceptibles que les hommes d'utiliser des mots non abrégés. Cougnon et François (2010) notent que, dans leur corpus de SMS de français de Belgique, les messages des femmes sont plus longs que ceux des hommes, mais aussi qu'ils contiennent plus d'abréviations. Herring et Zelenkauskaite (2009), dans un corpus de SMS italiens, font le même constat. En revanche, dans un corpus de chat néerlandais, Peersman et al. (2016) n'ont trouvé aucun effet du genre sur la probabilité de produire des abréviations et acronymes. L'âge

semble également être corrélé avec le taux d'abréviation des messages, les utilisateur·trices les plus jeunes employant plus d'abréviations que les plus âgés (Cougnon & François, 2010). Tagliamonte et Denis (2008) notent que la fréquence de *lol*, l'acronyme le plus fréquent dans leur corpus de messagerie instantanée, diminue avec l'âge : il est moins utilisé par les jeunes adultes de 19 à 20 ans que par les adolescents de 15 à 18 ans. Leur étude date d'il y a plus de 10 ans, et il est possible que *lol* soit aujourd'hui moins utilisé par les adolescents.

Eisenstein et al. (2011) ont trouvé, en s'appuyant sur des données de géolocalisation, que les Afro-Américain·es utilisent des acronymes innovants plus fréquemment que les autres internautes. Ils listent notamment *lml* (*laughing mad loud*), *smh* (*shaking my head*), *smfh* (*shaking my fucking head*), *lls* (*laughing like shit*) et *lmao* (*laughing my ass off*), ainsi que des réductions comme *sis* (*sister)*. Les personnes vivant dans des comtés américains à la forte population hispanique sont également associées, dans une moindre mesure, à ces graphies. Bamman et al. (2014), qui a groupé les utilisateur·trices de son corpus par clusters en fonction des mots qu'ils et elles utilisent le plus fréquemment, a identifié un groupe de personnes qui se distingue du reste de son corpus par une utilisation fréquente de structures de l'anglais afro-américain (*finna* pour *fixing to*, par exemple), et d'acronymes (*lls*, *lmao*).

Étirements graphiques

L'étirement de lettres consiste à répéter la même lettre plusieurs fois pour allonger un mot, comme dans *noooooo*, *gawwwwd* ou *niceeeeee*. C'est un procédé d'insertion qui nécessite l'utilisation de davantage de caractères que dans la forme standard d'un mot, par contraste avec les procédés d'abréviation fréquents dans la CMC (Herring & Zelenkauskaite, 2009). Ce procédé a été appelé « expressive lengthening » (J. Schnoebelen, 2012), « alphabetic character repetition » (Rao et al., 2010), « expansion » (Holtgraves, 2011), « vocal spelling » (Lin, 2016 ; Riordan & Kreuz, 2010), ou encore « étirement graphique de lettres » (Overbeck, 2015). Il peut concerner des mots « de dictionnaire » (*cooooolll*), des interjections (*ohhhhhh*), des acronymes (*omgggggggg*), et des onomatopées (*booooooommm*). Les étirements graphiques sont également possibles avec la ponctuation (*??????!!!!!*) ou avec les émoticônes (*:)))))))*)

Fonctions Pour les chercheur·es qui se sont intéressé·es à ce phénomène, la fonction principale de l'étirement graphique est d'imiter la prosodie de la langue parlée (Riordan & Kreuz, 2010 ; J. Schnoebelen, 2012). Dans leur étude de l'Enron Corpus (composé d'emails), Kalman et Gergle (2014) notent que plus de 94 % des répétitions de lettres sont « prononçables » (*freeeeeezing*). Ils soulignent leur caractère « dynamique et changeant » (p. 192), et listent leurs fonctions : indiquer un changement de hauteur de voix (*sweeeet*), signaler une pause (*hmmmmmmmm*), exprimer des sons (*vvvrrrroooooommmmm*), reproduire une intonation musicale (*happy birthday to youuuuuu*), indiquer un cri (*WOOOOOOO*) ou exprimer des rires et sons gutturaux.

Les étirements graphiques n'imitent toutefois pas toujours des sons, comme le montre la présence de termes que l'on ne peut pas prononcer, principalement des acronymes et termes caractéristiques de la langue électronique (*afffff* pour *as fuck*) (Coats, 2017b). Pour Riordan et Kreuz (2010), les étirements graphiques sont donc aussi une façon d'exprimer de l'émotion et de mettre de l'emphase sur certains termes.

Types d'étirements graphiques Les types de mots les plus fréquemment étirés semblent varier selon les corpus. Dans l'Enron Corpus, ce sont les interjections (Kalman & Gergle, 2014) qui sont le plus souvent étirées ; dans le corpus de tweets américains de Coats (2017b), ce sont les mots « de dictionnaire » qui arrivent en première position en termes de fréquence ; et, dans le corpus de messages de forums produits par des adolescent·es de Lin (2016), ce sont les adverbes d'intensité. Kalman et Gergle (2014), Riordan et Kreuz (2010) et Coats (2017b) montrent tous les trois qu'en anglais, les voyelles sont beaucoup plus susceptibles que les consonnes d'être répétées, et que la voyelle la plus sujette à l'étirement graphique est le *o*.

Fréquence des étirements graphiques Il semble que les étirements graphiques de lettres soient de plus en plus fréquents dans la CMC. Kalman et Gergle (2014) notent qu'ils sont peu présents dans l'Enron Corpus, qui date de 2002. Les étirements graphiques sont plus nombreux dans le corpus de blogs, d'email et de chat examiné par Riordan et Kreuz (2010), ainsi que dans le corpus de Twitter de Coats (2017b). Kalman et Gergle (2014) ont réalisé une étude longitudinale exploratoire, recherchant les occurrences de plusieurs étirements graphiques dans des blogs. Comparant des blogs écrits entre 1998 et 2002 d'un côté, et entre 2002 et 2012 de l'autre, ils notent une augmentation considérable de la fréquence relative d'un certain nombre d'étirements de lettres ; l'étirement graphique de *please* (*pleeease*, *pleeeease*, *pleeeeease* et *pleeeeeeeease*) est ainsi 5.29 fois plus fréquent dans les blogs de la période la plus récente. Commentant ces résultats, Coats (2017b) souligne que les techniques de recherche utilisées ont pu jouer un rôle. Kalman et Gergle (2014) n'ont en effet pris en compte que les étirements graphiques issus de mots « de dictionnaire » et non de séquences dérivées de termes non standard, ce qui donne une image partielle de la réalité. Coats ajoute que le type de plateforme peut également expliquer les importantes différences de fréquence constatées entre les études ; plus formel que les SMS ou Twitter, l'email se prêterait moins aux étirements graphiques de lettres.

Études sociolinguistiques des étirements graphiques Les études des étirements graphiques sur diverses plateformes indiquent qu'ils sont plus utilisés par les femmes. C'est notamment le cas sur Twitter (Bamman et al., 2014 ; Coats, 2017b ; Rao et al., 2010), et dans les corpus de SMS (Herring & Zelenkauskaite, 2009 ; Holtgraves, 2011). Pour Bamman et al. (2014), l'étirement graphique est donc un « marqueur féminin » (p. 142).

Coats (2017b) note par ailleurs une différence dans le type de formes allongées ; les femmes utilisent plus de marqueurs d'affect et d'interaction (interjections, injures et marqueurs de politesse) que les hommes, qui étirent davantage des noms de lieux, des noms et des adjectifs. Ces résultats doivent toutefois être interprétés avec précaution, car son échantillon est de petite taille. L'étude de Coats montre aussi une différence culturelle dans l'utilisation d'étirements graphiques. Comparant un corpus de tweets américains à un corpus de tweets finlandais écrits en anglais, il note que les deux utilisent des étirements de mots et d'interjections prononçables. En revanche, les Finlandais ont tendance à utiliser davantage de termes que l'on ne peut pas prononcer, notamment les acronymes typiques de l'anglais d'internet (*lmaoooo*) et les émoticônes.

Interjections

Les interjections sont, depuis l'Antiquité, un sujet controversé (Buridant, 2006). Classées par les grammairiens grecs dans la catégorie des adverbes, elles étaient considérées comme une partie du discours par les Romains, pour qui elles exprimaient des émotions ou des sentiments (Ameka, 1992). Elles sont aujourd'hui toujours vues comme une partie du discours, qui a pour particularité d'être indépendante sur le plan grammatical (Biber et al., 1999).

Interjections primaires et secondaires On fait généralement la distinction entre deux catégories d'interjections : les interjections primaires et les interjections secondaires. Les premières sont des mots courts ou des *non-words* qui sont uniquement utilisés en tant qu'interjections (*ouch*, *wow*, *ugh*), tandis que les secondes sont des mots appartenant à d'autres parties du discours, qui peuvent être aussi être utilisés en tant qu'interjections (*goodness*, (*thank you*, (*please*) (Ameka, 1992). Les interjections primaires appartiennent à une classe de mots ouverte, ne sont pas sujettes à l'inflexion ou à la dérivation, et sont souvent composées de sons ou de graphies qui dévient du système phonétique ou orthographique d'une langue. En anglais, il s'agit par exemple du son [x] de l'interjection *ugh*, de mots sans voyelles (*mhm*, *psst*), ou de graphies qui n'ont pas de relation directe avec les formes phonétiques correspondantes, comme *whew*, qui représente un son d'expiration (Norrick, 2009). Pour ces raisons, les interjections sont souvent considérées comme des éléments paralinguistiques, et ont été assez peu étudiées (Ameka, 1992).

Fonction des interjections Ameka (1992) attribue aux interjections trois fonctions principales, qui peuvent être combinées (p. 106). Les interjections expressives sont des « vocal gestures » qui reflètent l'état d'esprit de la personne qui les utilise (*yuk* pour le dégoût, *wow* pour le plaisir ou la surprise, *ouch* pour la douleur). Les interjections conatives sont utilisées en interaction, pour obtenir une réaction du destinataire (*eh ?* quand le ou la locuteur·trice veut une précision, ou *sh* quand il ou elle demande le

silence). Enfin, les interjections phatiques servent à maintenir le contact communicatif (*mhm*, *uh huh*).

Les interjections dans la CMC L'étude des interjections dans les corpus oraux et écrits est compliquée par le fait que bon nombre d'entre elles n'ont pas de forme orthographique ou phonétique fixe (Aijmer, 2009). Pour cette raison, et sans doute également à cause de leur statut périphérique dans la langue, les interjections de la CMC ont été relativement peu étudiées. Une étude (Coats, 2017a) a mis en évidence une association entre fréquence des interjections et genre : dans un corpus de tweets produits en anglais par des internautes danois, suédois, islandais, norvégiens et finlandais, les interjections sont plus fréquemment utilisées par les femmes que par les hommes. En revanche, dans leur corpus de messages produits par des étudiant·es, Guiller et Durndell (2007) n'ont pas trouvé de différence significative entre femmes et hommes. Verheijen (2017) a comparé l'utilisation des interjections sur diverses plateformes et a trouvé qu'elles sont plus fréquentes dans la messagerie instantanée que dans les SMS et sur Twitter, et qu'elles sont également plus utilisées par les adolescent·es (12 à 17 ans) que par les jeunes adultes (18 à 23 ans).

Typographie non standard

La CMC est caractérisée par un usage riche et souvent non standard de la typographie. Certains procédés, comme l'omission de majuscules et les mots entièrement en majuscules, sont largement utilisés par les internautes. D'autres procédés, souvent moqueurs et ironiques, sont plus confidentiels et viennent des sous-cultures d'internet. C'est le cas de l'alternance aléatoire de majuscules et de minuscules (*WoRdS LikE tHiS*), qui indique l'ironie et tire son origine d'un mème mettant en scène le personnage de dessin animé Bob l'éponge, (Kircher, 2017), ou des mots contenant un espace entre chaque lettre (*l i k e t h i s*), qui tirent leur origine dans le mouvement musical et artistique vaporwave (Ferson, 2016). Nous nous intéressons ici à deux procédés : les mots en majuscules et l'omission de majuscules.

Mots en majuscules L'écriture de mots entièrement en majuscules, appelée *all caps* (« All caps », p. d.) en anglais, est généralement considérée comme un marqueur d'émotion. Elle est souvent associée à la colère et au fait de « crier » par écrit (*shouting*) (Turnage, 2007), mais peut aussi exprimer la joie, l'excitation, la tristesse ou encore la peur (Parkins, 2012). Les majuscules peuvent également signaler l'emphase, en mettant un mot en relief (Hilte, 2019). Ce procédé a rarement été étudié par les travaux sociolinguistiques sur la CMC. Une étude d'un corpus de posts publiés sur Facebook et Twitter par des utilisateur·trices australien·nes a trouvé que les mots en majuscules étaient plus fréquemment utilisés par les femmes (Parkins, 2012). Une autre (Rosen et al., 2010) en vient aux mêmes conclusions, non pas à partir de l'analyse d'un corpus, mais des réponses d'internautes de 18 à 25 ans à une enquête sur leurs pratiques d'écriture en ligne.

Omission de majuscules L'omission de majuscules fait partie des procédés « économiques » de la CMC, qui vise en premier lieu à gagner du temps. Comparée à d'autres procédés du Netspeak, elle a été relativement peu étudiée. Repérer l'absence de majuscules là où on les attendrait normalement selon les conventions orthographiques (en anglais, notamment, en début de phrase, dans le pronom personnel *I* ou dans les noms propres) nécessite en effet l'utilisation de techniques de traitement automatique des langues. Comme Tagliamonte et Denis (2008), nous avons décidé d'étudier un procédé facile à repérer dans un corpus : l'omission de la majuscule du pronom personnel sujet. Pour Tagliamonte et Denis (2008), cette variable est typique de la langue de la messagerie instantanée, puisque 74 % des occurrences du pronom personnel sont sous la forme minuscule. Ils notent par ailleurs que la majorité des 72 adolescent·es de leur corpus a exclusivement utilisé la forme *i*, et qu'une minorité a utilisé principalement *I*. Les chercheur·es en concluent que la variable *I*/*i* se comporte comme une variable sociolinguistique « typique », et que la sélection d'une des deux variantes est dictée par les choix stylistiques des individus.

L'omission de majuscules a souvent été stigmatisée par les médias, notamment parce qu'elle est considérée comme un signe de paresse. Toutefois, l'apparition des claviers intuitifs n'a pas éliminé l'omission de majuscules dans les messages écrits sur smartphones. Il semble que certain·es internautes tiennent à écrire en minuscules, et désactivent la saisie intuitive. Dans ce cas, l'utilisation de minuscules à la place de majuscules est parfois vue comme un comportement passif-agressif ou rebelle. Pour McCulloch (2019), l'omission de majuscules participe à la création d'une typographie ironique, qui introduit une dissonance dans les messages, et sont à l'opposé d'une typographie « polie » qui consiste à faire un effort supplémentaire, par l'utilisation de majuscules et d'étirements de ponctuation.

Omission d'apostrophe

Ce procédé consiste à omettre l'apostrophe dans une contraction (*dont* pour *don't*, *Im* pour *I'm*) ou dans une construction possessive (*my dads car* pour *my dad's car*). Les premières études de l'omission d'apostrophe en anglais sont bien antérieures à la CMC. Au début du 20ème siècle, Johnson (1917) (cité par Connors et Lunsford, 1988) note que l'omission d'apostrophe est la cinquième « erreur » la plus fréquente dans un corpus de devoirs écrits par des étudiant·es. Certaines études se sont focalisées sur la question des apostrophes possessives. En 1922, Lester (cité par Hokanson & Kemp, 2013), a trouvé que 8.2 % des fautes d'orthographe produites par des étudiant·es de 17 et 18 ans étaient liées à l'omission, à l'utilisation abusive ou au mauvais placement d'apostrophes possessives. Dans un corpus de 3000 essais écrits par des étudiant·es américain·es, les erreurs d'utilisation de l'apostrophe possessive représentent 5.1 % du total des graphies non standard, et la confusion entre *it's* et *its* 1 % des graphies non standard (Connors & Lunsford, 1988).

L'utilisation de l'apostrophe possessive est particulièrement problématique parce que celle-ci est apparue tardivement en anglais, ne devenant

la norme qu'au 19ème siècle (Sklar, 1976). Aujourd'hui, elles posent un problème particulier aux enfants comme aux adultes (Hokanson & Kemp, 2013). Ces difficultés sont dues au fait qu'on voit ces apostrophes, mais qu'on ne les entend pas (Cop & Hatfield, 2017). Bien les utiliser suppose donc de comprendre le principe morphologique qui les gouverne sans s'appuyer sur des indices phonologiques (Hokanson & Kemp, 2013). Le fait qu'elles semblent relativement rares n'arrange pas les choses ; dans le Brown Corpus (1 million de mots), les possessifs apparaissent 1857 fois au singulier et 334 fois au pluriel (Francis et al., 1982, cité par Cop et Hatfield, 2017).

La paire *it's*/*its* est également problématique pour de nombreuses personnes, parce qu'elle crée une confusion entre le concept de possession et l'utilisation de l'apostrophe. *Its* dénote la possession mais ne comporte pas d'apostrophe, et *it's* est la contraction de *it is*. Pour bien utiliser *its* et *it's*, il faut donc comprendre que, contrairement aux noms, les pronoms possessifs ne prennent pas d'apostrophe (Hokanson & Kemp, 2013).

L'omission d'apostrophes dans la CMC Dans la CMC, l'omission d'apostrophes semble relativement fréquente, d'après les rares études qui lui ont été consacrées. Chez Tagg et al. (2012), elle constitue 10.56 % des occurrences de variantes orthographiques, et arrive en quatrième position de fréquence sur les 17 catégories identifiées par les auteur·es. Drouin et Driver (2014) ont noté une proportion similaire dans un corpus de SMS : l'omission d'apostrophes est la troisième catégorie de procédés non standard la plus fréquente après l'absence de capitalisation et les graphies phonétiques, représentant 10 % des variantes non standard. Dans le corpus de SMS produits par des adolescent·es et jeunes adultes australien·nes de De Jonge et Kemp (2012), l'omission de l'apostrophe arrive en deuxième position en termes de fréquence derrière l'omission de capitalisation, et représente 16 % des graphies non standard. L'étude de Squires (2012) révèle que la probabilité d'une omission d'apostrophe est plus forte dans certains mots. Dans son corpus de messagerie instantanée, les contractions de *am*, *are*, *is* et *not* génèrent davantage d'omissions que *will*, *would* et le possessif.

L'omission d'apostrophes, choix ou « erreur » ? On peut se demander par quoi l'omission d'apostrophes est causée dans la CMC ; est-elle un choix, ou le fruit d'une méconnaissance des règles de l'orthographe anglaise ? Tagg et al. (2012) s'interrogent sur ces questions. Partant du principe que la variation orthographique de leur corpus semble majoritairement intentionnelle, ils concluent que l'omission d'apostrophe peut être considérée comme un choix délibéré, qui permet de gagner du temps et de l'espace, de faire moins d'efforts, ou de créer un effet particulier. Il semble en effet peu probable que les utilisateur·trices ne sachent pas qu'il faille utiliser une apostrophe dans les contractions *Im*, *thats* ou *hes*. Les auteur·es ajoutent toutefois que l'apostrophe est peut-être une catégorie à part dans la variation orthographique, à cause des difficultés qu'elle pose. Dans certains cas épineux, comme celui de l'apostrophe possessive et de la paire *its*/*it's*, il est possible que l'omission ne soit pas intentionnelle. Tagg et al. soulignent

donc le fait qu'il est difficile de savoir quand l'omission est une erreur d'orthographe, et quand elle devient intentionnelle.

Omission de l'apostrophe et variables sociales Les études sociolinguistiques de l'omission d'apostrophe indiquent des corrélations possibles entre cette variable linguistique et le genre, la littératie, et la réussite universitaire. Dans leur étude de SMS écrits par des étudiant·es américain·es, Drouin et Driver (2014) n'ont trouvé aucune différence significative entre femmes et hommes dans la fréquence des omissions d'apostrophes. En revanche, Squires (2012) met en lumière une corrélation entre genre et omission d'apostrophes dans un corpus de 16 conversations de messagerie instantanée produites par 26 étudiant·es américain·es. Utilisant des modèles de régression, elle constate que l'omission d'apostrophes est plus fréquente chez les hommes. Ses résultats révèlent également l'existence d'une interaction entre le genre de la personne qui écrit et celui de son destinataire. Les hommes utilisent ainsi moins d'apostrophes quand ils écrivent à des femmes que lorsqu'ils interagissent avec des hommes. Même si cette interaction est significative, Squires invite à l'interpréter avec prudence, son corpus étant de petite taille. Pour la chercheure, l'apostrophe est une variante orthographique qui participe à la construction d'une identité sociale genrée : omettre l'apostrophe est prestigieux pour les hommes, et le contraire est vrai pour les femmes. Elle émet également l'idée que l'omission d'apostrophe serait devenue ou serait en train de devenir une forme « standard » dans la messagerie instantanée. Si c'était le cas, écrit-elle, les hommes s'aligneraient avec l'usage standard, tandis que les femmes auraient tendance à préférer un usage non standard ou « superstandard » (Bucholtz, 2001, p. 84).

Dans leur étude d'un corpus de SMS écrits par des étudiant·es américain·es, Drouin et Driver (2014) ont analysé les corrélations entre leurs compétences en écriture, orthographe et vocabulaire et les différentes catégories de Textese (langue des SMS). Ils font la distinction entre les catégories « négatives », dans lesquelles ils placent l'omission d'apostrophe et l'absence de capitalisation, et les catégories « positives », comme l'*accent stylization* et l'utilisation de symboles. Ils ont trouvé que la seule catégorie qui est corrélée négativement avec la littératie, et plus spécifiquement avec les compétences en lecture, est l'omission d'apostrophe ; il faut toutefois noter que, si les résultats sont significatifs, la corrélation est faible (-0.23). Les catégories « positives » sont quant à elles corrélées positivement avec la littératie.

Cop et Hatfield (2017) ont étudié l'omission de l'apostrophe possessive dans un corpus de 1414 tests linguistiques administrés à des étudiant·es néo-zélandais·es. Leur objectif était de savoir s'il existe une corrélation entre cette variable linguistique et la réussite de ces étudiants à l'entrée dans les programmes universitaires de médecine, pharmacie et physiothérapie. En utilisant des modèles de régression logistique, ils ont trouvé que les étudiant·es qui utilisaient correctement l'apostrophe possessive étaient 38 % plus susceptibles d'être accepté·es dans le programme de médecine

que les autres.

Graphies phonétiques et *eye dialect*

Plusieurs autres types de graphies non standard ont été identifiés par les chercheur·es. Tagg et al. (2012) font ainsi la distinction entre les « letter/number homophones », des lettres ou nombres qui sont combinés ou non pour représenter des mots (*ur* pour *your*, *gr8* pour *great* ou *y* pour *why*), les graphies non standard, des « irregular spellings, usually phonetic » (*xcitd* pour *excited*, *ryt* pour *right*) (p. 379), l' « accent stylization », qui représente de l'argot ou des prononciations familières (*wanna*, *didja*), et les erreurs typographiques et fautes de frappe (*htat's*, *phsychology*).

Parfois, le terme « eye dialect » est utilisé pour désigner des formes qui ne reflètent pas la façon dont les mots seraient prononcés dans la langue parlée informelle (comme *gud*, *wots*) (Tagg et al., 2012). Ce concept a été défini par Krapp (1925) (cité par Picone, 2016), pour qui il est utilisé pour donner l'impression qu'un locuteur utilise un dialecte. Picone (2016) fait la distinction entre l'*eye dialect*, qui reflète pour lui la langue informelle, et les *pronunciation respellings*, des graphies qui visent à représenter une prononciation dialectale (comme *dat* pour *that*).

Dans la littérature, *eye dialect* et *prononciation respellings* sont souvent utilisés pour représenter la langue de communautés stigmatisées, comme les Afro-Américain·es et les « Hillbillies » des Appalaches. Le but est souvent de donner l'impression de locuteur·trices ignorant·es et peu éduqué·es ; cela montre, pour Picone, que l'on peut se moquer de leur façon de parler en toute impunité (Preston, 2000). Dans la CMC, *eye dialect* et *pronunciation respelling* ont en commun le fait qu'ils ne sont pas de simples contractions, mais qu'ils sont généralement utilisés pour indexer un certain type d'identité (« rebellious, careless, unconcerned » (Tagg et al., 2012, p. 381).

À cause de l'ambiguïté entre les différentes catégories de graphies non standard, celles-ci ont rarement fait l'objet d'études sociolinguistiques dans la CMC. Il semble toutefois, d'après Eisenstein et al. (2011), que les graphies phonétiques et dialectales soient davantage utilisées par les internautes afro-américain·es que par les blanc·hes et les Hispaniques. Il s'agit notamment de *ova* (*over*), *wat* (*what*), *ya* (*you*), *dats* (*that's*) ou encore *skool* (*school*), et *yall* (*y'all*). Notons enfin que, dans cette thèse, nous utilisons le terme « graphies phonétiques » pour désigner à la fois *eye dialect* et *pronunciation respellings*.

G-droppings

Les g-droppings dans la langue orale Le terme « g-dropping » désigne deux phénomènes liés. À l'écrit, il se réfère à l'omission du *g* final d'un mot se terminant par *ing* ; à l'oral, il désigne la prononciation alvéolaire [n] à la place de la prononciation vélaire [ŋ] dans les syllabes finales en *-ing*, ce qui est aussi appelé « la variable (ING) ». Le (ING) alvéolaire ne se produit que sur des syllabes non accentuées, et n'affecte jamais les mots d'une seule syllabe, comme *thing* ou *sing*. De plus, les syllabes qui portent un

accent secondaire sont moins susceptibles d'être terminées par [ɪn]. Des mots comme *everything* et *anything* présentent donc moins de g-droppings que *nothing* et *something* (Yuan & Liberman, 2011). Par ailleurs, la prononciation [ɪn] n'affecte pas les noms propres terminés par *-ing*, comme *Flushing*, *Reading* ou *Harding* (Labov, 2001). Au début des années 1980, le travail d'étudiant·es de l'université de Pennsylvanie a révélé que la variable (ING) était également conditionnée par la catégorie grammaticale des mots. Les g-droppings sont les plus fréquents dans les formes progressives et les participes, moins fréquents dans les adjectifs et les gérondifs, et encore moins fréquents dans les noms. Ces contraintes grammaticales ont depuis été montrées dans toutes les études réalisées sur la variable (Labov, 1989).

La variable (ING) a été abondamment étudiée par les sociolinguistes depuis les années 1950 (Labov, 2001). Elle a été la première variable sociolinguistique à faire l'objet d'une étude quantitative, avec le travail de Fischer (1958) sur son utilisation par 24 enfants d'un village de la Nouvelle-Angleterre, aux États-Unis. Cette analyse a révélé une association du g-dropping avec le genre et le statut socioéconomique : le g-dropping était plus fréquent chez les garçons et chez les enfants des classes sociales les plus modestes. La variable (ING) a été décrite par Labov (2001) comme la variable sociolinguistique la plus stable et la plus uniforme de l'anglais. Il souligne qu'elle fonctionne d'une façon similaire depuis le début du 19ème siècle au moins, ne semble pas être sujette au changement diachronique, et est associée aux mêmes variables sociales dans l'ensemble du monde anglophone. La quasi-totalité des études de la variable (ING) ont ainsi trouvé que la prononciation [ŋ] est un marqueur de prestige, qui est plus souvent utilisé par les femmes et dans des contextes formels. La prononciation [ɪn] est plus fréquente chez les hommes, dans les classes sociales modestes, et dans des contextes informels (Labov, 1989). Elle indexerait un tempérament dur et tourné vers la confrontation. Cela a notamment été montré par Trudgill (1974) à Norwich, en Angleterre, par Mock (1979) dans une communauté rurale du Missouri, et par Woods (1979) à Ottawa, au Canada. Labov (2001) a aussi mis en évidence une interaction du genre avec la classe sociale. Dans son étude menée à Philadelphie, il a trouvé que les différences dans l'utilisation de (ING) par les femmes et les hommes sont en général modérées. Elles sont toutefois plus marquées dans la classe moyenne que dans la classe ouvrière supérieure, dans les contextes formels comme dans les contextes informels.

Récemment, certaines études de la variable (ING) ont mis en lumière la façon complexe dont les locuteurs l'utilisent pour construire leur identité, et notamment leur identité de genre. Kiesling (1998), qui a étudié une fraternité étudiante américaine, souligne que ses sens sont variés et interconnectés de façon complexe. Selon lui, la prononciation [ɪn] n'indexe pas forcément la masculinité en soi, mais peut prendre des sens variés en fonction du contexte et du rôle des locuteurs, comme la camaraderie, et les valeurs de la classe ouvrière. Gratton (2016), dans son étude de deux personnes non binaires canadiennes, a montré que la variable (ING) était une

des ressources utilisées par ces locuteur·trices pour indexer des positions de résistance à la cisnormativité féminine et masculine.

Les g-droppings dans la CMC À l'écrit, la prononciation[m] est généralement transcrite par le remplacement du *g* final par une apostrophe, comme dans *walkin'* ou *nothin'* (Yuan & Liberman, 2011). Cette graphie a été décrite comme étant une représentation écrite emblématique de la langue informelle orale (Davies, 2015). Le g-dropping est parfois mentionné par les études des graphies non standard de la CMC (par exemple, chez De Jonge et Kemp, 2012), mais a rarement été étudié en détail. Les chercheur·es qui s'y sont intéressé·es ont remarqué que cette variable est associée à des contextes et des variables sociales similaires à ceux auxquels la variable phonologique (ING) est liée. Dans un corpus de SMS écrits par des étudiant·es, Holtgraves (2011) a ainsi trouvé que les hommes ont produit significativement plus de g-droppings que les femmes.

Dans le corpus de tweets américains d'Eisenstein (2015), les g-droppings sont relativement fréquents, et généralement conditionnés par des contraintes phonétiques et syntaxiques. Le mot *goin* est le 292$^{\text{ème}}$ token le plus fréquent, avant *live* et *might*. Eisenstein note que, comme dans le cas de la prononciation [m], les verbes sont plus susceptibles d'être affectés par des g-droppings que les noms et les adjectifs. Il remarque une exception notable à ce phénomène : *fucking* et son euphémisme *freaking*, plus souvent utilisés en tant qu'adjectifs que verbes, et qui présentent un taux très élevé de g-droppings. Utilisant des données de géolocalisation pour déterminer la composition démographique de son corpus, il établit un lien entre l'utilisation de g-droppings et l'ethnicité : la suppression du *g* est moins fréquente dans les comtés principalement habités par des blanc·hes, et plus fréquente dans les comtés où vit une importante population afro-américaine.

Le g-dropping a également été étudié dans le contexte du Mock Ebonics, une parodie de l'anglais afro-américain utilisée à des fins racistes, (Ronkin & Karn, 2002). Dans leur corpus de pages web créé en 1997, les chercheuses ont trouvé que le g-dropping est fréquemment utilisé par les auteur·es racistes. Elles en concluent que pour eux, le g-dropping est une des graphies qui symbolisent leurs préjugés sur l'anglais afro-américain et ses locuteur·trices. Enfin, dans son étude de tweets écrits par Sarah Palin alors qu'elle était candidate à la vice-présidence des États-Unis, Davies (2015) note que la femme politique ne tweete pas comme elle parle : elle utilise peu de g-droppings à l'écrit, alors qu'elle prononce fréquemment les syllabes finales en *-ing* [m]. Davies note aussi que les g-droppings se produisent dans des contextes précis ; par exemple, dans l'expression *just sayin…*, qui minimise la portée d'une critique ou d'un reproche (« (I'm) just saying », p. d.), dans l'expression *times, they r a'changin*, citation d'une chanson de Bob Dylan, ou encore dans une référence possible au slogan de McDonald's *lovin it*. Palin utilise également parfois des g-droppings quand elle parle des médias et de Barack Obama, qui sont pour elle des cibles de choix. Davies conclut que la femme politique utilise les g-droppings de façon stratégique, que ce soit pour faire référence à des valeurs populaires, ou pour signifier

une position négative vis-à-vis du sujet qu'elle évoque.

tl ;dr

La CMC ouvre de nouvelles perspectives à l'étude du genre et de la langue. L'essor des réseaux sociaux s'est accompagné d'une explosion de données, souvent facilement accessibles, qui permettent d'explorer la langue écrite informelle à une échelle sans précédent.
Le genre est la variable sociale qui a le plus bénéficié de l'intérêt des sociolinguistes et des spécialistes du traitement automatique des langues. Leurs travaux quantitatifs, qui adoptent généralement une perspective essentialiste, révèlent des différences d'usage entre femmes et hommes. L'âge et l'ethnicité ont bénéficié d'un intérêt plus relatif. Comme il est difficile de recueillir des informations démographiques de façon fiable, peu d'études quantitatives de la CMC prennent en compte l'interaction entre plusieurs variables.
Les pseudonymes sont un précieux point de départ pour explorer la manière dont les internautes créent leur identité en ligne. Le type d'informations qu'ils contiennent varie selon les plateformes ; les pseudonymes indexent souvent une identité de genre, par l'utilisation de prénoms ou d'attributs genrés. Les centres d'intérêt des internautes, quant à eux, semblent refléter la division traditionnelle entre centres d'intérêt dits « féminins » et « masculins ».
La dernière section de ce chapitre s'est intéressée à certaines des formes non standard de la CMC, qui, pour beaucoup, ne sont pas réellement nouvelles, mais dont internet a favorisé l'usage. L'importante variation orthographique de la CMC complexifie le traitement des données par les outils informatiques, et les tentatives de classification des procédés comportent forcément une part de subjectivité. Les études réalisées sur les variables qui nous intéressent montrent des corrélations entre genre, âge, et fréquence des formes non standard, même si leurs résultats sont parfois contrastés. De manière générale, les jeunes internautes, les femmes et l'anglais afro-américain semblent jouer un rôle important dans la promotion des innovations langagières en ligne.

Chapitre 3

Reddit

Ce troisième et dernier chapitre de notre cadre théorique est consacré au terrain que nous explorons dans cette thèse : le site internet Reddit, un ensemble de forums dédiés à toutes sortes de thématiques, aujourd'hui extrêmement populaire aux États-Unis. Après une brève description du site et de sa composition démographique, nous présentons les aspects essentiels de son fonctionnement. Nous replaçons ensuite le site dans le contexte de la culture geek ; nous montrons en quoi il est emblématique des « toxic technocultures » d'internet (Massanari, 2017), et comment il a été conçu de façon à exclure certain·es internautes. Nous décrivons, enfin, la nouvelle orientation prise récemment par le site.

3.1 Présentation de Reddit

Reddit est un site internet qui a été créé en 2005 par Alexis Ohanian et Steve Huffman, deux amis alors tout juste sortis de l'université de Virginie (« Alexis Ohanian - Wikipedia », p. d.). Ils souhaitaient en faire « the front page of the internet », (« The Heartbreaking Backstory To The Founding Of Reddit », p. d.), et l'ont baptisé d'un nom issu du jeu de mots « I read it on reddit » (« faq - reddit.com », p. d.). Le site est composé d'un ensemble de forums, appelés subreddits, dont le nom est toujours précédé du préfixe r/. Ces forums abordent une immense variété de thèmes, de la politique à la science en passant par la culture populaire, l'humour, la pornographie, la cuisine ou le bricolage. Certains de ces subreddits ne comptent qu'une poignée d'inscrit·es, ou sont rapidement abandonnés ; d'autres, comme r/funny, r/AskReddit et r/gaming, ont plus de 25 millions de membres. Contrairement à d'autres forums, qui sont organisés par thèmes, Reddit n'a pas une structure hiérarchique, et les subreddits sont tous indépendants les uns des autres (Cole et al., 2017) ; le site adopte ainsi un fonctionnement « fluide et décentralisé » (Massanari, 2017, p.340).

Plutôt qu'un réseau social, Reddit a été décrit comme étant une « communauté de communautés », qui représente une multitude de cultures (Massanari, 2017, p.331). Il ressemble davantage aux premiers forums en ligne, appelés « message boards », et aux premiers sites communautaires comme

le WELL qu'à des réseaux sociaux comme Facebook ou Twitter (Massanari, 2017). Sur Reddit, les profils sont basiques, et les connexions entre membres sont moins importantes que les discussions qui prennent place publiquement sur les forums (Thelwall & Stuart, 2018).

Reddit a beaucoup changé depuis sa création en 2005. Initialement, le site était uniquement un agrégateur de contenu : on ne pouvait pas créer des subreddits ou commenter les liens mis en ligne par les Redditors. La possibilité d'écrire des commentaires a été introduite en 2005, et celle de créer des subreddits en 2008. Tout le monde peut consulter Reddit, mais il faut avoir un compte pour participer à des discussions. D'autres fonctionnalités ont été ajoutées au fur et à mesure, dont la possibilité de mettre en ligne images et vidéos directement sur Reddit, les applications et outils mobiles, le flair (une petite étiquette qui apporte des précisions sur l'identité d'un Redditor et qui apparait à côté de son pseudonyme sur les forums), les fils de discussion en direct, et le crossposting (qui consiste à mettre en ligne un post sur plusieurs subreddits) (Amg137, 2019).

3.1.1 Une des plus grandes communautés en ligne du monde

Longtemps fréquenté par une minorité d'internautes initiés et geek, Reddit est peu à peu devenu un site incontournable du paysage numérique américain. Il a gagné en popularité en août 2012 avec le « AMA » fait par Barack Obama, alors président des États-Unis, qui s'est plié à un des rituels du site : une séance de questions-réponses sur le subreddit IAmA, sur lequel on peut échanger avec des personnalités ou des internautes ayant eu une expérience étonnante (« Barack Obama surprises internet with Ask Me Anything session on reddit », 2012). Selon le site Alexa.com, qui propose des statistiques sur le web, Reddit était, en mai 2017, lorsque nous avons créé le corpus, le 29ème site le plus fréquenté au monde et le 9ème site le plus populaire aux États-Unis. Plus de la moitié des visiteurs venaient des États-Unis. Les autres pays les plus représentés sur Reddit étaient alors le Royaume-Uni, l'Inde, le Canada et l'Australie (Alexa, p. d.).

3.1.2 Composition démographique de Reddit

Selon une étude réalisée en janvier et février 2018 par le centre de recherche indépendant Pew Research Center (« Demographics of Social Media Users and Adoption in the United States », 2019), Reddit est utilisé par 11 % des adultes américain·es. Les hommes y sont majoritaires : 15 % des hommes américains fréquentent le site contre 8 % des femmes, alors que la proportion d'hommes et de femmes dans l'ensemble de la population américaine est très similaire (« U.S. Census Bureau QuickFacts », p. d.). Une autre étude du Pew Research Center (Barthel et al., 2016a) indique qu'en 2016, 71 % des Redditors américain·es étaient des hommes, contre 29 % de femmes. À titre de comparaison, en 2018, 74 % des femmes américaines et 62 % des hommes américains utilisaient Facebook. Les femmes étaient

également majoritaires sur Instagram (utilisé par 39 % des femmes et 30 % des hommes). Sur Twitter, les proportions de femmes et d'hommes étaient à peu près équivalentes (23 % des femmes et 24 % des hommes l'utilisaient) (« Demographics of Social Media Users and Adoption in the United States », 2019). Reddit est donc davantage dominé par les hommes que les autres plateformes numériques populaires.

Les utilisateur·trices américain·es de Reddit sont jeunes : en 2018, 22 % des 18-29 ans fréquentaient le site, contre 14 % des 30-49 ans et 6 % des 50-64 ans. La proportion de Redditors blanc·hes et hispaniques est plus forte que la proportion d'Afro-Américain·es ; toujours selon la même étude, 12 % des Blanc·hes et 15 % des Hispaniques utilisaient le site en 2018, contre 4 % des Afro-américain·es (« Demographics of Social Media Users and Adoption in the United States », 2019). En 2016, 64 % des Redditors avaient entre 18 et 29 ans. Quarante-quatre pour cent d'entre eux se disaient de tendance politique « liberale », et 19 % s'identifiaient en tant que conservateurs. Cette enquête a également révélé que les Redditors utilisent davantage internet que la moyenne des Américain·es : 97 % d'entre eux ont déclaré aller en ligne tous les jours, contre 71 % pour l'ensemble des Américain·es.

3.2 Fonctionnement du site

3.2.1 Les comptes

Création de comptes

Les personnes qui sont abonnées à des subreddits et qui y commentent sont appelées « Redditors », terme qui est la contraction de « Reddit » et « editor » (Kidd, 2018). Les Redditors sont anonymes, et choisissent un pseudonyme pour interagir sur le site. Ils jouissent d'une grande liberté : ils peuvent lancer des fils de discussion, répondre aux commentaires des autres, créer des subreddits, et « upvoter » et « downvoter » les interventions des autres internautes (c'est-à-dire leur donner un vote positif ou négatif).

Créer un compte est obligatoire pour écrire des commentaires. Le processus est très simple et rapide : aucune donnée personnelle ou adresse email n'est requise. Il est par ailleurs possible de créer et de supprimer autant de comptes que l'on veut. Reddit encourage même ses utilisateur·trices à créer plusieurs comptes (« faq - reddit.com », p. d.), à condition qu'ils ne s'en servent pas pour « upvoter » leurs propres commentaires. Cette liberté a donné naissance au phénomène des *throwaway accounts*, ou *throwaways* (« comptes jetables »), une pratique bien connue dans la communauté. Un *throwaway* est un compte créé spécialement pour participer à un fil de discussion ou à un subreddit particulier sans laisser de traces. Il permet aux utilisateur·trices de créer une ou plusieurs autres identités, dissociées de leur compte principal, et de profiter d'une couche supplémentaire d'anonymat. Les comptes jetables sont donc essentiellement utilisés dans les forums où les internautes évoquent des sujets intimes et personnels (Leavitt, 2015). La durée de vie d'un compte jetable est généralement courte, et bon

nombre d'entre eux ne sont utilisés qu'une seule fois. Les femmes utilisent davantage de comptes jetables que les hommes, sans doute à cause de la domination parfois menaçante des Redditors hommes dans de nombreux subreddits (Leavitt, 2015).

Il arrive que des comptes Reddit soient suspendus (c'est-à-dire qu'ils existent encore, mais ne sont plus accessibles ni par leurs propriétaires, ni par les autres Redditors). Selon la section d'aide de Reddit (« Account status », p. d.), les suspensions de compte ne peuvent être décidées que par des employés de Reddit. Elles se produisent quand un internaute ne respecte pas la politique du site (« Reddit Content Policy », p. d.), qui interdit notamment les contenus illégaux, la pornographie involontaire, les spams, les menaces et l'incitation à la violence.

Les pseudonymes

Reddit fonctionne sous le système du pseudonymat. Pour créer un compte sur le site et participer à des discussions, les internautes doivent obligatoirement utiliser un pseudonyme, appelé *username* sur le site. Il est extrêmement rare que les internautes choisissent leurs propre prénom et nom de famille ; ceux qui le font sont en général des personnes publiques comme, par exemple, Bill Gates (u/thisisbillgates) et l'astrophysicien Neil Degrasse Tyson (u/neiltyson). Bien souvent, les Redditors choisissent des pseudonymes ludiques, avec des jeux de mots et des références à la culture populaire ou à la culture de Reddit. Les pseudonymes sont importants, car, au sein d'une discussion, c'est bien souvent la seule information dont dispose la communauté pour se faire une idée des personnes qui interagissent sur les forums (il faut cliquer sur un pseudonyme pour accéder aux autres informations d'un compte). Le pseudonyme est donc à la fois le premier mot que les Redditors écrivent sur le site, et la marque la plus visible de leur identité.

Pour choisir leurs pseudonymes, les Redditors doivent respecter plusieurs règles. Chaque pseudonyme doit être unique (il est impossible d'utiliser un pseudonyme qui existe déjà sur le site) ; il ne peut pas y avoir d'espace dans un nom d'utilisateur·trice ; et seuls sont autorisés les caractères suivants : lettres de A à Z en majuscules ou en minuscules, chiffres de 0 à 9, tiret bas (_) et tiret (-) (kinsi55, 2014). Une fois créé, un pseudonyme ne peut pas être modifié.

Les flairs

Sur certains subreddits, les pseudonymes des Redditors sont accompagnés de courts textes ou de symboles, que l'on appelle *flairs*. Les flairs sont spécifiques aux subreddits qui les utilisent. Un·e Redditor peut ainsi avoir plusieurs flairs, selon les subreddits auxquels il ou elle est abonné ; le flair affiché à côté de son nom dépend du subreddit où il ou elle écrit un message (« How do I get flair (the text/image next to my username) ? », p. d.). Les Redditors ne sont pas obligé·es de choisir un flair dans les subreddits qui les utilisent. S'ils décident de le faire, ils peuvent sélectionner un flair à partir d'une liste prédéterminée, et/ou, dans certains cas, ajouter un texte

personnalisé. Les flairs, dont nous présentons des exemples dans la figure 3.1, ont généralement un lien avec le thème des subreddits qui les utilisent. Dans r/WeddingPlanning, par exemple, les Redditors peuvent indiquer le lieu et la date de leur mariage ; les abonné·es à r/GirlGamers sont invité·es à indiquer le nom de leur plateforme de jeux en ligne ou de leur console préférée. Sur r/CFB, un subreddit consacré au football américain universitaire, les abonné·es utilisent les logos des équipes qu'ils supportent.

Certains subreddits consacrés au genre utilisent également des flairs. C'est le cas de r/AskMen et r/AskWomen. Ces deux subreddits permettent aux Redditors de choisir entre plusieurs symboles : le symbole féminin, le symbole masculin, le symbole transgenre et le symbole agenre. Les abonné·es à r/AskMen peuvent également ajouter un texte à côté du symbole choisi. Les flairs utilisés dans les subreddits r/AskMenOver30 et r/AskWomenOver30 permettent, en plus, de connaitre l'âge des Redditors. Les subreddits dédiés aux transgenres, comme r/asktransgender, r/transpostive, r/ftm, r/mtf ou r/genderqueer proposent quant à eux davantage de catégories sur l'identité de genre et l'orientation sexuelle. Souvent, dans ces forums, les Redditors indiquent également leur prénom, leur âge, leur lieu de résidence ou la date du début de leur transition.

3.2.2 Le système de vote et le karma

Les *Upvotes* et *downvotes*

Reddit organise les forums et les fils de discussion par un système d'*upvotes* (votes positifs) et de *downvotes* (votes négatifs), qui permet à la communauté de choisir les commentaires et contenus les plus pertinents. En cliquant sur une petite flèche qui va vers le haut ou vers le bas, chaque Redditor peut donner son avis sur les contenus. Les fils de discussion et commentaires qui reçoivent le plus de votes positifs s'affichent par défaut en haut de la page d'accueil ou du fil de discussion. Il est possible de changer les paramètres de Reddit pour modifier l'ordre des commentaires et afficher les commentaires controversés, qui sont cachés par défaut. Ce système de vote est central dans le fonctionnement de Reddit : d'une part, parce qu'il a un impact direct sur la visibilité des contenus, en mettant en valeur les contributions les plus appréciées des Redditors, et, de l'autre, parce qu'il permet aux Redditors d'obtenir des points de « karma » et d'augmenter leur statut sur le site.

Le karma

Chaque Redditor reçoit un score de *comment karma* et de *post karma*. Le *post karma* (karma de post) est le score global attribué par la communauté aux *self posts* d'un Redditor (c'est-à-dire aux messages qui commencent un fil de discussion). Le *comment karma* (karma de commentaire) est le score global des commentaires d'un Redditor (au sein d'un fil de discussion). Ces deux scores figurent en caractères gras sur la page de profil des Redditors, juste en dessous de leur pseudonyme (figure 3.2).

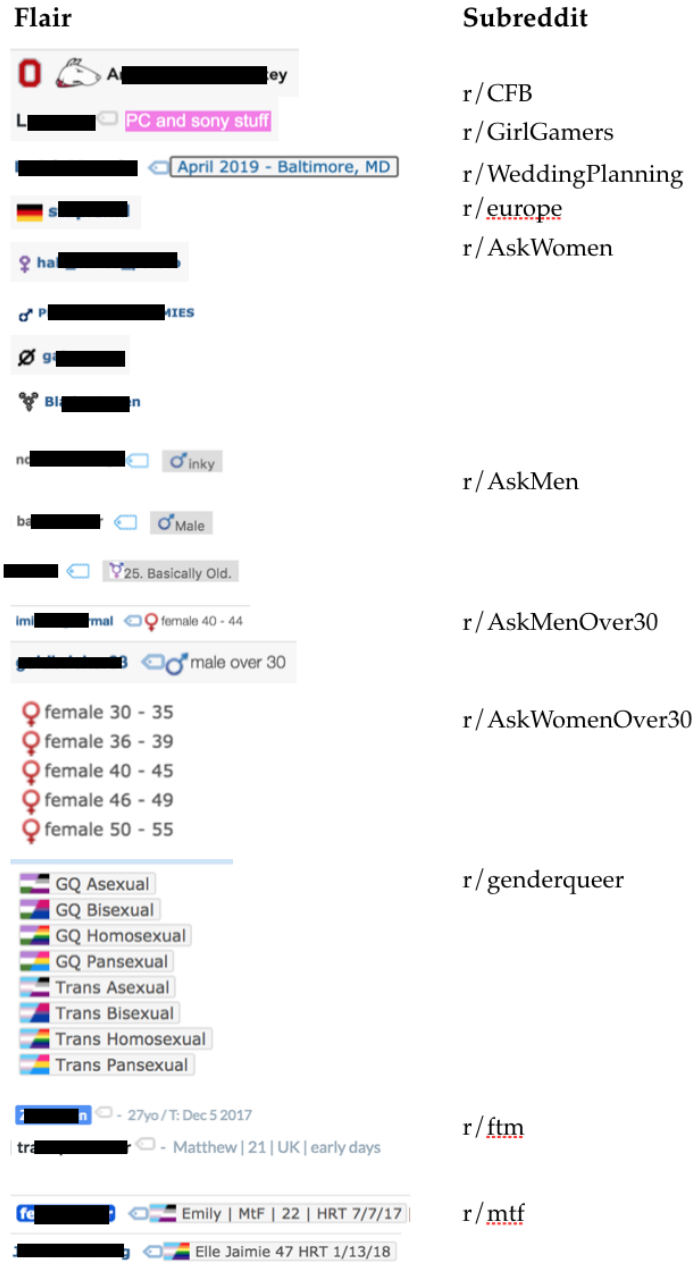

FIGURE 3.1 – Exemples de flairs sur Reddit (nous avons masqué les noms d'utilisateur·trices)

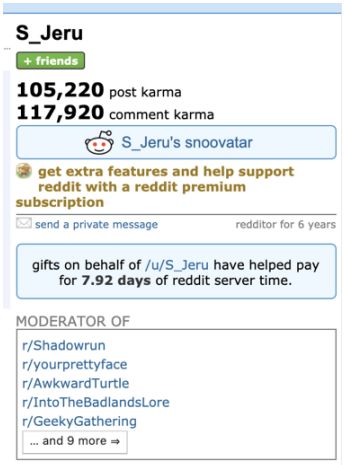

FIGURE 3.2 – Capture d'écran d'une partie du profil d'un·e Redditor montrant ses scores de karma

Le karma est censé refléter la pertinence des contributions d'un·e internaute à la communauté de Reddit (« How do I get karma ? | Reddit Help », p. d.). Il s'inscrit dans la volonté de Reddit de décentraliser la gestion du contenu : ce sont les membres de la communauté, et non les dirigeants du site, qui jugent de la qualité du contenu (Richterich, 2014). Selon le site d'aide de Reddit, « being on-topic, relevant, funny, interesting, or engaging are great ways to earn comment karma. Being insulting, rude, or abusive is not. Snark can go either way. » (« What is karma ? | Reddit Help », p. d.).

Avoir un karma élevé est un symbole de statut sur Reddit, puisque le karma mesure le succès d'une personne dans la communauté (Finlay, 2014, p. 20). La pratique controversée du *karmawhoring*, c'est-à-dire le fait de produire du contenu sur le site uniquement pour accumuler du karma, a toutefois changé la signification du karma. Aujourd'hui, il y a une fracture entre des nouveaux utilisateur·trices avides de karma et des Redditors qui fréquentent le site depuis longtemps et qui sont réfractaires à l'accumulation de karma. Pour elles et eux, cette pratique est emblématique du changement de statut de Reddit, qui, d'un espace dédié aux sous-cultures du web, est devenue une communauté extrêmement populaire (Richterich, 2014).

3.2.3 La modération

Reddit offre la possibilité à ses membres de modérer un ou plusieurs subreddits. Tous les forums du site sont ainsi modérés par des internautes bénévoles. Les modérateur·trices peuvent effectuer plusieurs actions : elles ou ils peuvent modifier le titre, la description et l'apparence d'un subreddit, en restreindre l'accès à certain·es internautes, mettre en place des filtres à spam, ou encore supprimer des commentaires. Le fait qu'un·e Redditor ait choisi d'être modérateur indique qu'il ou elle connait bien les règles du site, y passe du temps, et s'implique dans une ou plusieurs communautés.

3.2.4 Écrire sur Reddit

Cette sous-section présente trois points qui peuvent influencer la façon dont les internautes s'expriment sur Reddit : son interface de rédaction, dont le fonctionnement peut être obscur pour certain·es Redditors, son code langagier, dominé par l'utilisation d'acronymes, et le type de terminaux utilisé par les Redditors pour rédiger des messages (smartphones ou ordinateurs).

L'interface de rédaction des commentaires

Sur l'Old Reddit, l'ancienne interface du site, qui était majoritairement utilisée lorsque nous avons créé le corpus en 2017, il n'est pas possible de mettre en forme un texte comme on le fait sur les logiciels de traitement de texte ou les éditeurs de texte des messageries email, en cliquant simplement sur un bouton. Il faut, à la place, utiliser des balises. L'éditeur de texte du site fait en effet appel à un langage propre à Reddit, qui est basé sur Markdown (« markdown - reddit.com », p. d.). Markdown est un langage de balisage qui utilise une syntaxe simple (« Markdown », 2019) et qui est utilisé dans de nombreuses applications. La figure 3.3 montre des exemples de balises de la variante de Markdown utilisée sur Reddit. Celles-ci permettent notamment de mettre du texte en italiques et en gras, d'utiliser une police de code, de barrer du texte, et d'insérer des hyperliens ou des listes numérotées.

Written	Rendered
`_italic_` or `*italic*`	*italic*
`__bold__` or `**bold**`	**bold**
`___bold-italic___` or `***bold-italic***`	***bold-italic***
`~~strikethrough~~`	~~strikethrough~~
`>!spoilers!<`	
`^superscript` or `^(superscript)`	superscript
`` `code` ``	`code`

FIGURE 3.3 – Capture d'écran du wiki de Reddit montrant des examples de syntaxe du langage de balisage de Reddit (« markdown - reddit.com », p. d.)

Ce fonctionnement confère aux internautes familier·es avec les langages de balisage et de programmation, souvent des geeks, une plus grande liberté et créativité. On peut le voir comme un obstacle de plus qui bloque l'entrée du site aux non-initié·es. Avec le New Reddit, né d'une envie de démocratiser la plateforme (→ p. 103), cette barrière a été levée : la nouvelle interface, présentée dans la figure 3.4 offre, par défaut, un éditeur de texte plus classique (« Submit to Reddit », p. d.). L'éditeur Markdown reste toujours accessible aux Redditors qui le souhaitent.

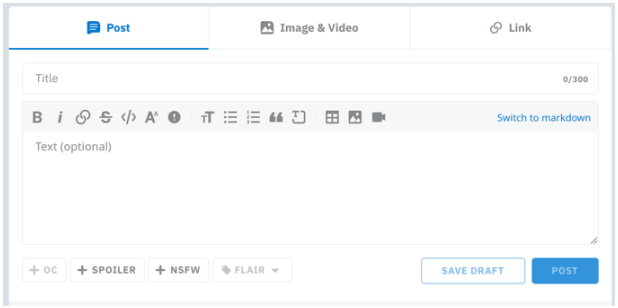

FIGURE 3.4 – Capture d'écran montrant l'éditeur de texte du New Reddit (« Submit to Reddit », p. d.)

La langue de Reddit

Une des caractéristiques de Reddit est son code langagier, que les nouveaux utilisateur·trices ont souvent du mal à comprendre. On peut ainsi y trouver de nombreux fils de discussion où des Redditors s'interrogent sur le sens des acronymes utilisés sur le site, avec des questions comme « Can someone help me understand Reddit's… lingo, jargon, language, whatever you want to call it ? » (simply-hopeless, 2019), « What are some reddit slang everyone should know ? » (omnikan, 2017), ou « I am new to Reddit. Tell me the acronyms ! » (daisygirl_79, 2013).

Les Redditors qui ont répondu à ces questions listent notamment des acronymes comme *ITT* (*in this thread*), *FTFY* (*fixed that for you*, utilisé pour rectifier un commentaire mis en ligne par un·e autre Redditor, souvent de manière humoristique), *AFAIK* (*as far as I know*, pour autant que je sache), *ELI5* (*explain like I'm 5*, nom d'un subreddit et acronyme utilisé lorsque l'on souhaite avoir une réponse simple à une question portant sur un sujet complexe), *SJW* (*social justice warrior*, « guerrier·e de la justice sociale », expression péjorative utilisée pour désigner des internautes défendant des causes féministes ou antiracistes, « Social justice warrior », 2020) ou encore *IANAL* (*I am not a lawyer*, utilisé en guise de clause de non-responsabilité lorsque l'on donne un conseil portant sur un aspect juridique, « IANAL », 2019). D'autres expressions sont mentionnées, comme *circle jerk*, nom d'une pratique sexuelle devenue synonyme de chambre d'écho (« What Does circle jerk Mean ? », p. d.), ou \s, un marqueur de sarcasme inspiré par les balises html (« Urban Dictionary », p. d.).

Ces termes ne sont pas tous propres à Reddit ; *IANAL* tire par exemple son origine d'une publicité pour la marque Vicks (« IANAL », 2019), tandis que *SJW* a été utilisé pour la première fois sur Twitter (« Social justice warrior », 2020). C'est toutefois bien souvent sur Reddit et sur d'autres forums comme 4chan qu'ils ont été popularisés. Les acronymes utilisés sur Reddit étant opaques pour les novices, de nombreux glossaires sont disponibles sur le web. Redditthelp.com, le site officiel d'assistance de Reddit, propose une rubrique intitulée « What do all these acronymes mean ? » (« What do all these acronyms mean ? », p. d.), où l'on trouve des acronymes comme *GTFO* (*get the fuck out*), *IRL* (*in real life*, par opposition à « sur internet »), *OC*

(*original content*, utilisé pour indiquer que l'auteur·e d'un message a créé la vidéo ou la photographie à laquelle elle ou il fait référence), ou encore *TIFU* (*today I fucked up*, nom d'un subreddit et employé pour parler d'une erreur que l'on a faite).

Sur Reddit, on trouve d'autres glossaires créés par les Redditors, comme « A New User's Guide to Reddit » sur r/RedditGuides (meemersbarnhart, 2013), ou le « Glossary » créé par le subreddit r/theoryofReddit, qui préface sa liste d'acronymes en expliquant que « when new users started using Reddit they can find the lingo used to be somewhat alienating » (« glossary - TheoryOfReddit », p. d.). D'autres sites internet proposent également des explications sur le code langagier de Reddit, comme « Learn how to speak Reddit » sur The Daily Dot (« Learn how to speak Reddit », 2012), « What does SMH mean ? All the Reddit slang you pretend to understand actually explained » sur Pink News (Flood, 2018) ou encore « The TL ;DR Guide to Reddit Lingo » de Mashable (Koerber, 2014), qui précise que « To a new user, Reddit may seem like a confusing jumble of text, numbers, pictures and acronyms ». Toutes ces questions, messages et contenus montrent bien que, pour comprendre les messages mis en ligne sur Reddit et écrire comme un Redditor, il faut apprendre les codes de la communauté langagière du site. La courbe d'apprentissage est sans doute moins prononcée pour les internautes qui fréquentent déjà d'autres sites où ce type d'acronymes est utilisé, comme Twitter ou 4chan.

Les terminaux utilisés pour accéder à Reddit

Nous terminons cette partie sur la rédaction de commentaires par la question de l'accès au site. Le type de terminal utilisé pour accéder au site peut en effet avoir un impact sur les commentaires des Redditors. Par exemple, les émojis sont disponibles sur le clavier virtuel d'un smartphone, mais pas depuis le clavier physique d'un ordinateur : pour les produire, il faut ouvrir un clavier virtuel, ce qui nécessite une manipulation supplémentaire et peu pratique (l'interface de Reddit ne « transforme » pas automatiquement les émoticônes en émojis, comme c'est le cas sur Facebook ou WhatsApp). Les correcteurs orthographiques et la frappe prédictive proposés par les systèmes d'exploitation des téléphones peuvent également affecter le type de graphies utilisées.

Dans un message mis en ligne en 2017 sur r/announcements, Steve Huffman, le cofondateur et dirigeant du site, écrit sous son pseudonyme spez que 40 % des visites sur le site s'effectuent depuis l'application mobile de Reddit (spez, p. d.). Il n'existe apparemment pas de statistiques officielles plus récentes. Toujours en 2017, 15 Redditors ont répondu à la question « What percentage of your reddit usage is desktop vs mobile ? » posée par datterHFX sur r/AskReddit. Dix d'entre disent utiliser à 100 % ou en vaste majorité leur smartphone, et 5 préfèrent l'ordinateur (datterHFX, 2017). En août 2018, l'utilisateur·trice thendofthebeginning pose une question similaire sur le même subreddit (thendofthebeginning, 2018). Cinq Redditors disent préférer l'ordinateur, et 7 le mobile. Un Redditor explique qu'il se sert majoritairement son ordinateur quand il consulte Reddit au travail,

et un autre dit utiliser son smartphone quand il est en déplacement.

Il est également possible que la façon dont on accède au site dépende des subreddits que l'on fréquente. Répondant à un message annonçant le lancement de nouvelles fonctionnalités mobiles de Reddit, le Redditor TheEnigmaBlade écrit que moins de 5 % des visiteurs du subreddit r/leagueoflegends y accèdent depuis un smartphone. Il précise que ce chiffre peut être lié au fait que League Of Legends est un jeu vidéo pour PC (Drunken_Economist, p. d.), et donc que les personnes qui fréquentent r/leagueoflegends sont bien souvent déjà sur leur ordinateur lorsqu'ils ou elles souhaitent consulter Reddit.

Il faut noter que les Redditors se plaignent souvent de la mauvaise qualité de l'application et de l'interface mobile du site, avec des fils de discussion intitulés « Is Reddit making its mobile site worse on purpose to force people to use the app ? » (Ungratefulpanda, 2018), « Why is Reddit mobile a piece of shit ? » (xauxiheo, 2019) ou encore « Mobile Reddit is terrible ! I need to rant » (SwillFish, 2017). Il est possible que les problèmes rencontrés poussent certain·es Redditors à préférer l'ordinateur au smartphone pour écrire des messages sur le site.

3.3 Reddit, un espace emblématique de la culture geek

3.3.1 Les origines de la culture geek

La culture geek a émergé dans les années 1960 et 1970, avec l'ouverture des premiers programmes de computer science des universités américaines de Harvard, MIT, Carnegie Mellon et Stanford (Varma, 2007). Dans ces filières, fréquentées quasi exclusivement par des hommes, est née une culture extrêmement compétitive et individualiste, qui valorise les compétences intellectuelles au détriment du corps (Woodfield, 2000). Dans les années 1980, on a commencé à utiliser le terme *geek* pour désigner ces jeunes hommes ; le mot, apparu aux États-Unis au début du 20ème siècle, désignait originellement les *sideshow freaks*, les personnes faisant partie des *freak shows* (des expositions itinérantes d'êtres humains) (« geek », p. d.). En changeant de domaine, le terme conserve une connotation négative, évoquant cette fois-ci une obsession pour les ordinateurs et les nouvelles technologies et un comportement antisocial.

À quelques nuances près, le geek est proche du *nerd* (terme qui aurait été inventé par l'écrivain de science-fiction Philip K. Dick, Konzack, 2014). Le nerd est associé à la science, informatique ou non, et le geek uniquement à l'informatique. Dans les années 1990, avec l'émergence de l'internet, le mot *geek* a perdu une partie de sa connotation négative, tandis que *nerd* est resté un terme péjoratif. Le nerd est encore moins à l'aise en société que le geek ; il est traditionnellement représenté avec une chemise à manches courtes, une cravate et des lunettes à monture épaisse (Kendall, 2011). Les nerds sont considérés comme étant « asexual, intellectual, wimpy, uncool »

(Kendall, 2011, p. 515) : ce sont des « intellectual overachievers and social underachievers » (Bucholtz, 2001, p. 85).

3.3.2 Les geeks et la masculinité

Aujourd'hui incarnée par les complexes rutilants de la Silicon Valley, la culture geek est plus « cool » et « glamour » qu'elle ne l'a jamais été (Margolis & Fisher, 2002). Mais, même si la société a changé de regard sur les geeks, leur idéologie reste la même. Les geeks, maitres des ordinateurs, sont rationnels, compétitifs, déterminés, et ont un esprit entrepreneurial (Varma, 2007). Toutefois, ils continuent à se considérer comme des marginaux, à cause de leur manque d'aptitudes sportives ou parce ce qu'ils ne se voient pas comme étant désirables. Caractérisée par l'intelligence et l'expertise, la masculinité geek contraste avec la masculinité hégémonique (ou dominante), qui est virile et agressive (Bury, 2011). Ces deux conceptions de la masculinité ont cependant des points communs. Elles sont toutes les deux hétéronormatives et se construisent en opposition à la féminité, liée à l'empathie, à la sociabilité et à la domesticité (Bury, 2011). Pour Varma (2007, p. 362), « geek culture rearticulates very old notions of male and female in a new context ».

3.3.3 Un club fermé

La culture geek est donc un univers d'hommes, dont les femmes sont exclues. Elles ont pourtant joué un rôle fondamental dans l'histoire de l'informatique : à la fin du 19ème siècle, elles étaient les premiers « computers » de l'histoire, effectuant des calculs complexes à la main. Une fois les machines inventées, ce sont elles qui ont donné naissance à l'art de la programmation (C. L. Evans, 2018). Après la Seconde Guerre mondiale, elles ont été peu à peu évincées du secteur de l'informatique par les hommes, et elles continuent à l'être aujourd'hui. Ainsi, en 2015, aux États-Unis, 18 % des « bachelor's degrees » (licences) en informatique étaient obtenus par des femmes, contre 35.7 % en 1986 et 25.1 % en 2004 (« Digest of Education Statistics », 2016). Les étudiantes en informatique disent se sentir isolées, ne pas se reconnaitre dans la culture geek (Varma, 2007), et perdent rapidement confiance en elles (Margolis & Fisher, 2002). Parmi les femmes qui persévèrent et trouvent un emploi, le taux d'attrition est élevé, car elles se heurtent fréquemment au plafond de verre (Ashcraft et al., 2016).

Ce *gender gap* tire son origine de la socialisation des enfants (on offre aux garçons des jeux de construction qui leur permettent d'inventer des choses), mais aussi dans la puissance des stéréotypes attachés au genre : les femmes sont perçues comme étant émotives, et les hommes rationnels. Cela explique qu'elles ont pu s'imposer dans d'autres secteurs autrefois hautement masculins, comme le droit et la médecine, domaines faisant appel à des qualités considérées comme féminines (les relations humaines et les soins aux autres), alors qu'elles restent à la marge des nouvelles technologies (Varma, 2007). Notons que, récemment, des femmes ont commencé à

s'approprier le label de geek : c'est « an emergent, hybridized alternative feminine identity » (Bury, 2011, p. 37).

3.3.4 Une identité racialisée

L'univers geek n'est pas seulement fortement genré ; il est aussi racialisé. Les geeks et les nerds sont, avant tout, des hommes blancs (Bucholtz, 2001). Les femmes ne sont donc pas les seules exclues de l'univers geek ; en Amérique du Nord, les hommes afro-américains, amérindiens et hispaniques le sont aussi. Il y a donc, proportionnellement à la distribution raciale de la population américaine, encore moins de personnes de couleur que de femmes dans les professions de l'informatique, et les femmes de couleur souffrent encore plus que les femmes blanches de l'hostilité du monde geek (Varma, 2007).

Ainsi, au terme de l'année scolaire 2016-2017, en Amérique du Nord, seuls 0.9 % des masters en informatique (*computer science*) ont été décrochés par des Afro-Américain·es (Zweben & Bizot, 2018) alors que ceux-ci représentent 13 % de la population américaine (Office, 2011) et 2.2 % de la population canadienne (« Black Canadians », 2020). Les Asiatiques, quant à eux, vivent une situation différente en Amérique du Nord. Aux États-Unis, ils sont considérés comme « the "model minority" – that is, the racialized group that most closely approaches "honorary" whiteness » (Bucholtz, 2001, p. 87). Ils sont donc plus facilement acceptés dans les espaces geeks. Pour ne donner qu'un exemple, en 2015, 18 % des développeur·ses de logiciel américain·es étaient des femmes ; 9 % étaient des femmes blanches, 7 % étaient asiatiques, 1 % étaient hispaniques ou latinas, et 1 % étaient afro-américaines (Ashcraft et al., 2016, p. 8). Le fait que les femmes et les personnes de couleur soient exclues de la culture geek protège le statut économique et technologique supérieur des hommes blancs (Kendall, 2011) : ce sont eux qui continuent à créer les logiciels et les outils que nous utilisons au quotidien, et qui construisent ainsi le monde numérique de demain.

3.3.5 Les espaces geek toxiques

Sur internet, la culture geek s'exprime sous sa forme la plus « toxique » dans plusieurs espaces : sur le forum anonyme 4chan, sur les groupes Usenet, sur le darknet (un réseau qui utilise des protocoles spécifiques et qui est anonyme) sur Twitter, dans les jeux vidéo, et sur Reddit (Massanari, 2017). Certains internautes y partagent des idées réactionnaires sur le genre, l'identité sexuelle et l'ethnicité, se positionnant ouvertement contre la diversité et le multiculturalisme. Ils utilisent comme arguments leurs interprétations, souvent erronées, de la psychologie évolutionniste, brandissant leur masculinité comme étalon de la rationalité (Massanari, 2017). Les femmes, pour eux, sont uniquement des objets sexuels, des intruses, ou les deux à la fois (Varma, 2007).

Certains de ces espaces ne sont pas aisément accessibles à l'ensemble des internautes, parce qu'ils requièrent une certaine expertise technologique (dans le cas du darknet) ou parce qu'il faut connaitre les sous-cultures

d'internet pour comprendre les mèmes et les « inside jokes » qu'on y trouve. C'est le cas de 4chan, d'une partie de Twitter, et, dans une certaine mesure, de Reddit (Massanari, 2017). La langue est également une façon d'empêcher les « outsiders » de participer aux espaces geek. Même si certaines caractéristiques de la langue du web, comme les émoticônes, sont aujourd'hui largement répandues, ce qu'écrivait Bailey en 1996 est encore vrai :

> « The Net nation deploys shared knowledge and language to unite against outsiders : Net jargon extends beyond technical language to acronyms both benign (BTW, 'By the way') and snippy (RTFM, 'Read the fucking manual'). It includes neologisms, text-graphical hybrids called emoticons, and a thoroughgoing anti-'newbie' snobbery. Like any other community, it uses language to erect barriers to membership » (Bailey, 1996, p. 22).

De plus, les communautés geeks toxiques affichent une philosophie « techno/cyberlibertarian » (Massanari, 2017), qui valorise l'individualisme. Pour les geeks qui les fréquentent, le manque de diversité n'est pas lié à la présence de barrières d'entrée : si les femmes et les personnes de couleur n'y participent pas, ce n'est pas parce qu'elles n'y sont pas les bienvenues, mais c'est parce qu'elles ne veulent pas y participer.

3.3.6 La masculinité geek de Reddit : un terreau propice à la haine

Sur Reddit, comme dans les autres espaces geek, l'éthos geek justifie le harcèlement envers les femmes. Certains Redditors hommes pensent qu'ils ont un droit sur le corps des femmes, et considèrent que leurs désirs sexuels sont plus importants que le respect de la vie privée des femmes (Marwick, 2017). Cette vision du monde s'est notamment exprimée sur Reddit lors de deux évènements, The Fappening et GamerGate, et continue à imprégner de nombreux subreddits. Le 31 août 2014, un internaute anonyme a mis en ligne sur le site 4chan près de 500 photographies et selfies de femmes célèbres, dont l'actrice Jennifer Lawrence, qui les montraient souvent dénudées, dans des contextes intimes et personnels. Ces images avaient été obtenues par des hackers par diverses techniques, dont le phishing. Les images se sont rapidement répandues sur le web pour atteindre Reddit. Des utilisateurs du site ont créé un subreddit baptisé r/TheFappening, mot-valise formé de « The Happening » et de « fap », un terme de l'argot du web synonyme de masturbation. Le CEO de Reddit à l'époque, Yishan Wong, a annoncé une semaine plus tard que le subreddit ne serait pas interdit, défendant la liberté d'expression des Redditors (yishan, 2014). Le même jour, le subreddit a toutefois été interdit, pour des raisons légales et non éthiques. Il avait généré suffisamment d'argent (par l'achat par ses membres de « Reddit Gold », une fonctionnalité premium payante) pour payer les frais de serveurs du site pendant un mois (Marwick, 2017). Plusieurs dizaines de subreddits sur le même thème ont rapidement vu le jour, dont r/TheFappenin et r/JenniferLawrenceLeaks (Massanari, 2017). Le même mois a eu lieu un autre évènement misogyne : GamerGate, un mouvement anonyme dont le

but était de révéler la « corruption » régnant dans le milieu du journalisme sur les jeux vidéo, et de montrer que les féministes essaient de détruire l'industrie des jeux vidéo (Chess & Shaw, 2015). Il s'agissait, en fait, d'une véritable campagne de harcèlement envers les femmes journalistes et critiques de jeux vidéo, organisée en partie sur le subreddit r/KotakuInAction (Marwick, 2017).

De nombreux subreddits, comme r/MGTOW, r/badwomensanatomy ou r/Braincels, font partie de ce qui a été décrit comme la « manosphere » (Farrell et al., 2019), c'est-à-dire des espaces qui encouragent la misogynie et la violence envers les femmes. Ils utilisent les arguments du « men's rights activism » et des « incels » (*involuntary celibates*, des hommes qui considèrent leur célibat comme la faute des femmes), qui visent à mettre en lumière les « discriminations » dont souffrent les hommes (Farrell et al., 2019). Cette rhétorique a été liée à des faits divers violents, comme la tuerie d'Isla Vista commise par Elliot Rodger en Californie en mai 2014, et l'attaque à la voiture-bélier de Toronto, qui a tué 10 personnes en avril 2018 (Farrell et al., 2019).

3.3.7 Un site conçu pour exclure

Par la façon dont il a été conçu et dont il est géré, Reddit promeut une « technoculture toxique ». Massanari (2017) cite cinq éléments qui perpétuent cette culture dans le site : le karma, le fait que Reddit est un agrégateur de contenu, la facilité avec laquelle on peut créer des comptes et des subreddits, la structure de gouvernance du site, et sa politique sur les contenus choquants (p. 329).

Le karma

Le système du karma semble démocratique, car il permet à la communauté de décider quels commentaires et fils de discussion sont les plus pertinents et intéressants. Il a toutefois des effets pervers. Il encourage la publication de commentaires et de posts susceptibles de rapporter un maximum de points de karma à ses auteurs, c'est-à-dire, souvent, des contenus qui reflètent le point de vue geek en matière de relations entre femmes et hommes, de technolibertarianisme ou de thèmes geeks (science, jeux vidéo, etc.). Le karma pousse aussi certain·es Redditors à republier des contenus populaires sur de nombreux subreddits, ce qui a notamment contribué à la diffusion rapide des images du scandale « The Fappening » sur le site. Pour Massanari, ce système de vote contribuerait au maintien de la philosophie geek de Reddit.

Le système d'agrégation de contenu

La page d'accueil de Reddit, r/all, présente des fils de discussion issus de divers subreddits, leur offrant une précieuse visibilité. Pour Massanari (2017), c'est « a kind of barometer of the community as a whole » (p. 337). Un algorithme choisit les fils de discussion qui apparaissent sur cette page,

mais ceux-ci proviennent d'une liste déterminée par les administrateurs, disponible sur le wiki du subreddit r/ListOfSubreddits (« defaults - ListOfSubreddits », p. d.). On y trouve, en juillet 2020, 49 subreddits, dont des subreddits sur la science (r/askscience, r/science, r/space, r/dataisbeautiful), sur l'actualité (r/worldnews), sur la culture populaire (r/television, r/movies, r/Music) ou sur le gaming (r/gaming) (en plus, les internautes qui ont créé un compte sur Reddit voient s'afficher les subreddits dont ils sont membres). Les subreddits qui figurent par défaut sur la première page sont généralement de gros forums, et reflètent majoritairement l'orientation geek de Reddit.

En mai 2014, les administrateurs de Reddit, conscients de ce problème, ont introduit de nouveaux subreddits à cette page d'accueil, dont le subreddit principalement fréquenté par des femmes r/TwoXChromosomes. De nombreux Redditors ont exprimé leur colère, disant qu'ils ne voulaient pas voir des discussions sur les agressions sexuelles, sur les règles ou sur la *body image* sur leur page d'accueil. D'autres se sont demandé pourquoi r/mensrights, un forum dédié au mouvement des droits des hommes, n'était pas non plus ajouté sur r/all (Massanari, 2017). Par ailleurs, en 2017, Reddit a introduit une alternative à r/all : r/popular, qui fonctionne de la même façon mais tire son contenu de subreddits plus diversifiés (et sans contenu pornographique) ; pour les dirigeants du site, cela s'inscrit dans une volonté de rendre le site « plus inclusif » (simbawulf, 2017).

L'absence de barrière d'entrée

Comme Reddit n'impose pas de barrière d'entrée aux internautes, encourageant même la création de comptes jetables, un utilisateur banni peut immédiatement revenir sur le site avec un nouveau profil. Ce mode de fonctionnement peut encourager les Redditors à adopter un comportement répréhensible, en écrivant par exemple des commentaires racistes et sexistes sans crainte de répercussions (Massanari, 2017).

Le mode de fonctionnement du site

Pendant plus de dix ans, Reddit n'a hébergé aucun contenu audiovisuel : les Redditors pouvaient uniquement mettre des liens vers des images (généralement hébergées sur le site imgur.com) ou des vidéos. Depuis juin 2016, il est possible d'intégrer des images et des gifs (et non des vidéos), mais uniquement dans des self-posts (et non dans les commentaires), et sur les subreddits dits « SFW » (*safe for work*), c'est-à-dire non pornographiques ou violents. Les contenus violents, racistes ou sexistes restent ainsi « en dehors » de Reddit. Ce système permet aux administrateurs du site de se défausser de toute responsabilité quant au contenu de ces images et vidéos.

La politique sur les contenus choquants

Les administrateurs de Reddit souhaitent conserver leur neutralité, et rechignent à interdire des subreddits, même lorsque leur contenu est choquant ou peut porter préjudice. Au sein des subreddits, la modération est effectuée par des Redditors bénévoles, ce qui est une façon pour les administrateurs du site d'éviter de « faire la police ». En même temps, le site fournit à ces modérateur·trices peu d'outils permettant de bannir les utilisateur·trices gênant·es ou les contenus offensants. Pour Massanari, cette neutralité « valorizes the rights of the majority while often trampling over the rights of others » (p. 339).

3.3.8 De l'Old Reddit au New Reddit : un site en quête de respectabilité

Naissance du New Reddit

Avec sa popularité grandissante, Reddit a beaucoup évolué. Les changements dans le fonctionnement et l'esthétique du site se sont accélérés en 2015, date à laquelle Alexis Ohanian et Steve Huffman, qui avaient quitté leurs postes en 2009, ont repris la direction de Reddit (Isaac & Streitfeld, 2015). En mai 2017, deux mois après le début de la construction du corpus, Reddit a introduit une nouvelle interface et la possibilité de créer de nouveaux profils. Dans un premier temps, seuls quelques Redditors trié·es sur le volet ont pu profiter de cette nouvelle fonctionnalité, puis cette phase de bêta-test s'est ouverte à tou·tes les Redditors intéressé·es (HideHideHidden, 2017).

Ces modifications avaient pour but de casser l'image geek du site, et de le rendre plus accessible et compréhensible aux « internautes moyens » (« Reddit rolls out user profiles amid site makeover », 2017). Elles s'inscrivaient dans une dynamique d'ouverture et de « nettoyage » du site : Reddit souhaitait se distancier de la culture geek qui lui a donné naissance et de son cousin le forum 4chan, au design vieillot et chaotique, autre foyer des « technocultures toxiques » (Massanari, 2017). Les nouveaux profils offrent plusieurs fonctionnalités inédites : la possibilité d'écrire des commentaires directement sur son profil, de suivre un Redditor (comme on suit quelqu'un sur Twitter), d'ajouter un avatar, une bannière et un texte descriptif, et de mettre en avant les subreddits sur lesquels les Redditors sont les plus actif·ves. L'esthétique est résolument plus moderne et épurée que celles des anciens profils. Il faut toutefois noter que Reddit n'est pas pour autant devenu un réseau social. Le nombre de « followers » d'un·e Redditor n'est indiqué nulle part, et le karma reste le seul indicateur du statut de l'internaute sur le site. Le New Reddit est aujourd'hui devenu l'interface par défaut. Les deux versions de Reddit coexistent toutefois sur le web, avec les URLs new.reddit.com et old.reddit.com.

Les administrateurs de Reddit commencent à intervenir

En 2015, la politique de Reddit sur les contenus choquants a changé avec l'arrivée d'Ellen Pao en tant que CEO intérimaire du site. En février 2015, elle a banni les subreddits consacrés au « revenge porn », une pratique qui consiste à publier sur internet des images ou vidéos à caractère sexuel explicite sans la permission des personnes qui y figurent, pour se venger, pour faire du chantage, ou pour réduire une personne au silence. En juin, Pao a mis en place des mesures pour lutter contre le harcèlement (reddit, 2012).

Ces changements ont été mal accueillis par de nombreux·ses Redditors, qui y ont vu la victoire du politiquement correct et des « social justice warriors » (un terme péjoratif qui désigne les personnes ayant des vues féministes ou antiracistes) (Massanari, 2017). Le mois suivant, Ellen Pao a démissionné après que 200 000 Redditors ont demandé son départ (« Sacked Reddit employee speaks out », 2015). Steve Huffman, le cofondateur de Reddit, a alors repris les rênes du site, maintenant les changements décidés par Pao. En août, le subreddit raciste et violent r/coontown a été banni. Huffman a annoncé que les subreddits proposant des contenus extrêmement offensants seraient désormais mis en quarantaine. Le but était de rendre ces forums moins visibles, sans toutefois les censurer (« Quarantined Subreddits », p. d.). Huffman a ensuite proposé de ne pas bannir les subreddits choquants, mais de ne plus les mettre en avant sur le site, en ne les incluant pas sur r/all, la page d'accueil de Reddit.

Reddit fait le ménage

Depuis 2017, la politique du site a été progressivement renforcée pour interdire le contenu illégal, la « pornographie involontaire », les contenus sexuels impliquant des mineurs, les messages encourageant la violence, les menaces, l'usurpation d'identité, le spam et l'utilisation de Reddit à des fins commerciales (« Reddit Content Policy », p. d.). Parmi les subreddits bannis par Reddit, il y a des forums de l'alt-right (r/alternativeright, r/altright, r/whiterights), les subreddits de The Fappening (bannis en juillet 2018), les *deepfakes*, où on pouvait voir de fausses images et vidéos pornographiques mettant en scène des célébrités, certains subreddits racistes et pornographiques, proposant du contenu sexuel concernant des mineur·es, ou encore des subreddits se moquant violemment des personnes obèses (« banned - ListOfSubreddits », p. d.).

En juin 2020, après les manifestations liées au mouvement Black Lives Matter, les administrateurs de Reddit ont publiquement reconnu que le site restait un des espaces privilégiés de l'expression du racisme et de la haine. Steve Huffman a annoncé sur le subreddit r/announcements :

> « As Reddit has grown, alongside much good, it is facing its own challenges around hate and racism. We have to acknowledge and accept responsibility for the role we have played » (Newton, 2020).

Le site a ainsi annoncé le renforcement de sa politique de modération, bannissant 2000 subreddits controversés dont r/The_Donald (un subreddit

non-officiel des supporters de Donald Trump), r/GenderCritical (un subreddit transphobe) et r/soyboy (un subreddit de l'alt-right, dont le nom vient d'une insulte désignant des hommes féministes et libéraux, « What Does soyboy Mean ? », p. d.). De plus, Alexis Ohanian, cofondateur du site, a annoncé sa démission, demandant à être remplacé par une personne afro-américaine (Newton, 2020).

tl ;dr

Depuis sa création en 2005, Reddit a connu une transformation spectaculaire. D'un site confidentiel, consacré aux centres d'intérêt geek, il est devenu un véritable phénomène de société aux États-Unis, avec l'émergence de millions de communautés dédiées à toutes sortes de sujets.

Reddit n'est pas un réseau social, mais un site web communautaire qui fonctionne sous le système du pseudonymat, et sur lequel il est très facile de créer un ou plusieurs comptes. Les diverses communautés, appelées subreddits, sont modérées par des bénévoles. Le système de votes permet à la fois de mettre en avant les contenus les plus populaires, et d'attribuer des points de karma aux Redditors, qui mesurent leur prestige et leur activité sur le site.

Malgré sa popularité grandissante, Reddit reste un espace masculin qui érige des barrières, par son fonctionnement, aux femmes et aux internautes issu·es de minorités. La liberté d'expression qui est un de ses principes fondateurs a permis l'essor de mouvements misogynes et la diffusion de contenus violents et racistes ; ceux-ci ont parfois eu des répercussions dans la vie « hors ligne », avec le harcèlement de femmes et des tueries.

Ces dernières années, Reddit a beaucoup changé : son interface a été refondue, ce qui vise à rendre le site plus accessible, et les administrateurs du site interviennent de plus en plus pour bannir les subreddits racistes, transphobes et misogynes. Ces changements se sont accélérés en 2020, avec la mise en lumière du racisme qui imprègne de nombreuses communautés, et de la nécessité d'intégrer davantage d'Afro-Américain·es dans les équipes de modérateur·trices des subreddits.

Deuxième partie

Méthodologie

Chapitre 4

Le corpus RedditGender

Ce chapitre décrit la méthode que nous avons utilisée pour créer le corpus RedditGender, de la sélection des internautes à l'annotation du corpus pour son exploitation dans le logiciel de textométrie TXM. Il présente ensuite la composition du corpus et les fonctionnalités employées pour extraire les données qui nous intéressaient depuis TXM pour leur analyse dans le logiciel de statistiques R, et termine en abordant la question de la mise à disposition du corpus à d'autres chercheur·es.

4.1 Origine du projet et construction du corpus pilote

La construction de RedditGender s'inscrit dans la continuité de notre mémoire de master 2 Mondes anglophones (Flesch, 2016), qui était également une étude de corpus quantitative du genre et de la CMC. À l'époque, nous avions créé un corpus de 2 millions de mots constitué des commentaires mis en ligne sur Reddit par 100 femmes et 100 hommes cisgenres, que nous avions analysé avec le logiciel de linguistique de corpus AntConc (Anthony, p. d.). Pour déterminer le genre des internautes, nous avions utilisé les flairs (→ p. 90) disponibles sur certains subreddits. Pour cette thèse, l'objectif était de créer un corpus plus grand, comprenant les interventions d'au moins 1000 personnes, et annoté de façon à pouvoir, par exemple, connaitre le forum sur lequel un commentaire a été publié. Nous souhaitions également réaliser des analyses multifactorielles, et donc recueillir d'autres variables sociodémographiques que le genre des internautes.

Reddit fonctionnant sous le système du pseudonymat (→ p. 90), le premier défi était de trouver une méthode qui permettrait d'obtenir des données sur les internautes. Nous avons pris le même point de départ que celui utilisé en master 2, à savoir les subreddits r/AskMen et r/AskWomen, dans lesquels la communauté pose des questions à des hommes ou à des femmes. De par notre familiarité avec Reddit, nous savions que les internautes se confient volontiers sur ces forums et livrent souvent, sous le couvert de l'anonymat, des informations très personnelles. Cela nous a donné l'idée

de puiser les données sociodémographiques directement dans leurs commentaires.

Il n'était évidemment pas envisageable de lire l'intégralité des commentaires d'une personne, car cela aurait été extrêmement chronophage. À la place, nous avons essayé d'extraire des informations sociodémographiques dans les commentaires par des recherches par mots clés. Cette technique a donné de bons résultats, et nous a permis de déterminer l'âge de nombreux·ses Redditors, ainsi que d'autres informations que nous ne nous attendions pas à pouvoir recueillir aussi aisément : la nationalité ou le lieu de résidence, l'âge, l'orientation sexuelle ou encore l'ethnicité. Cela nous a permis d'envisager la construction d'un corpus richement annoté, permettant une analyse multivariée de la langue de la CMC. Après la construction d'un corpus pilote composé des commentaires de 100 personnes (50 hommes et 50 femmes cisgenres), nous avons fait le point sur les données recueillies, en calculant le nombre de personnes dans chaque catégorie et le nombre d'observations manquantes (tableau 4.1).

Tableau 4.1 – Composition du corpus pilote

Variables	Catégories	Individus
Genre	Hommes	50
	Femmes	50
	Inconnu	0
Ethnicité	Blancs	21
	Autres	7
	Inconnue	72
Orientation sexuelle	Hétérosexuelle	68
	Gay	3
	Bisexuelle	5
	Inconnue	24
Pays	États-Unis	74
	Autres pays	26
	Inconnu	0
Âge	16-65 ans	91
	Inconnu	9
Total		100

Cela nous a permis de faire trois constats. Tout d'abord, il a paru évident que nous n'allions pas pouvoir recueillir des données complètes (genre, âge, pays, orientation sexuelle, ethnicité, catégorie socioprofessionnelle) pour chaque personne. Pour pouvoir réaliser des analyses statistiques prenant en compte plusieurs variables, il allait donc falloir intégrer un grand nombre de personnes au corpus, et se contenter d'échantillons réduits pour les analyses s'intéressant à plus de deux variables sociodémographiques.

Ensuite, malgré le nombre important d'observations manquantes dans certaines catégories, l'utilisation de mots clés pour trouver des informations démographiques nous a semblé avoir un fort potentiel. Nous avons donc décidé de continuer à employer cette méthode pour la suite de la cons-

truction du corpus.

Enfin, la composition démographique du corpus nous a paru très homogène, avec une majorité de personnes hétérosexuelles et blanches, ce qui correspond au profil démographique de Reddit (→ p. 88). Ce constat nous a fait envisager une autre possibilité : puisque, de toute façon, notre méthode d'échantillonnage n'est pas aléatoire et ne permettra pas d'obtenir un échantillon représentatif de Reddit, pourquoi ne pas envisager de créer un corpus diversifié, incluant une plus grande variété d'expériences et de profils ?

Dès lors, nous avons décidé d'adopter une perspective intersectionnelle, en surreprésentant certaines catégories de Redditors, dont, par exemple, les Afro-Américain·es, les Asiatiques et les Hispaniques, et les gays. L'élargissement de la catégorie binaire du genre (hommes et femmes cisgenres) à d'autres identités de genre (femmes et hommes transgenres et personnes non binaires) s'est alors imposé comme une nécessité. L'étape suivante a consisté à ajuster la méthode de recueil des données pour inclure davantage de diversité. Au vu de la composition du corpus pilote, il est apparu évident que se contenter de puiser des données dans les subreddits r/AskWomen et r/AskMen ne permettrait pas d'atteindre nos objectifs. Nous avons alors recherché des subreddits consacrés aux personnes transgenres, aux orientations sexuelles non-hétérosexuelles et aux différents profils ethniques.

Cette exploration nous a permis de prendre conscience de la richesse de Reddit pour l'étude sociolinguistique intersectionnelle. Il existe en effet un grand nombre de forums offrant les perspectives de Redditors issu·es de « minorités ». Certains d'entre eux utilisent, comme r/AskWomen et r/AskMen, un système de flair qui, combiné à la recherche de mots clés, nous a permis de créer un corpus diversifié, capable de répondre à nos questions de recherche.

4.2 Méthode de recueil des données

4.2.1 Description de la procédure

Voici la procédure et les critères de sélection utilisés pour recueillir les informations démographiques et les données textuelles (certains des points abordés sont décrits plus en détail par la suite) :

1. Sélection d'un subreddit sur le thème du genre ou sur un autre thème d'intérêt (ethnicité, orientation sexuelle, etc.).
2. Sélection d'un fil de discussion qui a généré de nombreux commentaires (plusieurs dizaines de commentaires pour les subreddits peu fréquentés, et plusieurs centaines pour les gros subreddits, pour avoir davantage de choix et visualiser rapidement des dizaines de profils).
3. Sélection des Redditors, en privilégiant celles et ceux qui utilisent un flair dans les subreddits proposant cette fonctionnalité (afin de gagner du temps dans le recueil des informations sociodémographiques).

4. Accès à l'historique des commentaires d'un·e Redditor, en cliquant sur son nom. En faisant défiler les commentaires, nous avons vérifié rapidement s'il y en avait assez (c'est-à-dire environ 200 au minimum) pour que la personne soit intégrée au corpus ; cela, afin d'obtenir suffisamment de données pour étudier des phénomènes rares, et également pour optimiser les chances d'intégrer des comptes « authentiques », à la longévité suffisante.
5. Recherche d'informations démographiques complémentaires avec des mots clés.
6. Si la recherche s'est avérée fructueuse (c'est-à-dire que nous avons pu déterminer au moins le genre et l'âge de la personne), le ou la Redditor a été intégré·e au corpus. Nous lui avons assigné un identifiant, qui contient des informations sur son genre et sur la date de collecte (par exemple, F_004_170510 pour la quatrième femme cisgenre du corpus, dont les commentaires ont été recueillis le 10 mai 2017). On a consigné son identifiant, l'URL de son profil, et ses données sociodémographiques dans un tableur (figure 4.1). Ses commentaires ont été copiés et collés dans un document de traitement de texte qui porte comme nom son identifiant.

	A	B	C	D	E	F	G	H	I	J	
1	id	n_words	n_comments	gender	exact_age	age_group3	sex_or	ethnicity	reddit_age	moderator	pc
2	F_001_170311	18083	196	female	31	3	heterosexual	NA	3	no	
3	F_002_170311	18562	375	female	26	2	NA	NA	4	no	
4	F_003_170311	18213	650	female	24	2	NA	NA	5	yes	
5	F_004_170314	17378	266	female	29	2	heterosexual	NA	4	no	
6	F_005_170314	20346	120	female	23	2	heterosexual	NA	1	no	
7	F_006_170314	19236	616	female	25	2	heterosexual	NA	2	no	
8	F_007_170314	17985	204	female	NA	2	heterosexual	NA	4	no	
9	F_008_170314	18316	525	female	26	2	heterosexual	white	6	no	
10	F_009_170314	18535	302	female	28	2	heterosexual	NA	2	no	
11	F_010_170314	18730	456	female	27	2	heterosexual	NA	0	no	
12	F_011_170315	18588	452	female	22	2	NA	NA	3	no	
13	F_012_170315	18273	318	female	23	2	heterosexual	NA	0	no	
14	F_013_170315	18945	368	female	18	1	heterosexual	NA	0.5	no	
15	F_014_170315	18445	117	female	30	2	heterosexual	NA	0.5	no	
16	F_015_170315	18932	283	female	31	3	heterosexual	NA	1	no	
17	F_016_170315	19259	269	female	27	2	heterosexual	white	5	no	
18	F_017_170315	17786	500	female	26	2	heterosexual	white	3	no	
19	F_018_170315	19387	736	female	29	2	heterosexual	NA	3	no	
20	F_019_170315	19109	252	female	26	2	heterosexual	NA	2	no	
21	F_020_170315	18435	266	female	31	3	heterosexual	NA	1	no	
22	F_021_170315	19799	510	female	27	2	heterosexual	NA	5	no	
23	F_022_170315	20624	176	female	25	2	heterosexual	NA	2	no	
24	F_023_170315	18780	402	female	NA	2	NA	black	1	no	
25	F_024_170315	18222	750	female	23	2	heterosexual	NA	4	no	
26	F_025_170315	18181	210	female	29	2	bisexual	NA	8	no	
27	F_026_170315	18748	415	female	26	2	heterosexual	white	1	no	
28	F_027_170315	18911	329	female	41	3	bisexual	white	3	no	
29	F_028_170315	17361	475	female	21	2	heterosexual	NA	0.5	no	
30	F_029_170315	18942	675	female	31	3	bisexual	white	3	no	

FIGURE 4.1 – Capture d'écran d'un extrait du fichier où ont été consignées les données sociodémographiques sur les Redditors

4.2.2 Subreddits utilisés pour cibler les Redditors

De nombreux Redditors ont été trouvé·es dans des subreddits consacrés au thème du genre, notamment sur r/AskMen ou r/AskWomen, mais aussi sur r/AskMenOver30 et r/AskWomenOver30. Ces deux subreddits ont un principe similaire à celui de r/AskMen et r/AskWomen, mais permettent de poser des questions à des hommes et des femmes de plus de 30 ans. Ils ont l'avantage de permettre à leurs membres d'utiliser des flairs indiquant leur genre mais aussi leur âge, ce qui s'est avéré très précieux.

Nous avons également ciblé des Redditors dans d'autres subreddits consacrés à des thèmes précis pour diversifier la composition de chaque variable sociale. Nous avons trouvé ces subreddits au gré de nos explorations sur le site, puisque Reddit n'est pas organisé de façon thématique. Voici une liste (non exhaustive) des subreddits utilisés :

— Genre : r/transgender, r/mtf, r/ftm, r/TransForTheMemories, r/transpassing, r/transpositive, r/NonBinary, r/nonbinaryUK, r/genderfluid, r/genderqueer, r/asktransgender, r/askGSM, r/TransSpace, r/agender, r/DualGender, r/ftm, r/transpassing.
— Âge : r/TeenFFA, r/MiddleSchool, r/highschool, r/OverFifty, r/OVER30REDDIT, r/AskOldPeople.
— Ethnicité : r/asianbros, r/AsABlackMan, r/blackladies, r/NativeAmerican, r/LatinoPeopleTwitter, r/AsianParentStories, r/asiantwoX, r/BlackHair, r/blackgirlgamers.
— Orientation sexuelle : r/pansexual, r/LGBTeens, r/gaypoc, r/gaylatinos, r/gaybros, r/actuallesbians, r/gay, r/lgbtqteens.

4.2.3 Sources des informations sociodémographiques

Le flair

Même s'il est difficile de savoir à quel point ils sont fiables, les flairs, ces indications que les Redditors peuvent accoler à leur nom d'utilisateur dans certains subreddits (sortes de « tags », → p. 90), ont été des sources d'information précieuses pour la constitution de l'échantillon. Grâce à cette méthode, nous avons pu identifier rapidement, par exemple, des hommes et femmes dans /AskMen et r/Askwomen et des hommes et des femmes transgenres sur r/asktransgender, connaitre l'orientation sexuelle et l'identité de genre des personnes qui échangent sur r/genderqueer, ou encore déterminer l'âge, l'identité de genre et le lieu de résidence des Redditors qui fréquentent r/ftm et r/mtf. Il faut toutefois noter que tous les subreddits n'utilisent pas de flair, et que sur les subreddits qui les utilisent, tous les Redditors n'en choisissent pas. De plus, les informations contenues dans un flair sont souvent limitées.

Le contenu des commentaires

Pour compléter les données obtenues grâce aux flairs, nous avons puisé des informations directement dans les commentaires des Redditors. Cette procédure s'est avérée simple et relativement efficace ; elle a nécessité de 5 à 10 minutes de travail par individu. Nous avons utilisé la fonctionnalité de recherche offerte par les navigateurs web (cmd+F sur Mac ou ctrl+F sur Windows et Linux) pour obtenir des informations sur chaque internaute. Plusieurs expressions ont été ciblées : « I am », « I'm », « I live », « I work », ou encore « my job ». La recherche la plus productive a été « I'm a ». Grâce à cette méthode, seules les déclarations des Redditors ont été prises en compte ; nous n'avons par exemple pas utilisé les photographies et selfies pour déduire le genre ou l'âge d'un·e Redditor.

Ce travail d'enquête n'a pas toujours abouti ; certain·es Redditors ne révèlent rien d'elles ou eux, ou presque rien. Ils ou elles ont donc été systématiquement écartés. Nous avons donc examiné bien plus d'historiques de commentaires que ceux inclus dans le corpus : pour parvenir au nombre total de 1044 Redditors retenu·es, plus de 5000 comptes ont été inspectés. Dans de nombreux cas, cependant, les recherches ont été étonnamment fructueuses. Il arrive qu'en une phrase ou deux, les Redditors indiquent leur genre, leur âge, leur orientation sexuelle, leur lieu de résidence et leur métier, comme le montre le tableau 4.2 (les déclarations des internautes ont été tronquées ou modifiées, de façon à ce qu'ils ou elles ne puissent pas être identifié·es).

Tableau 4.2 – Exemples de résultats de la recherche de *I'm a*

I'm a SAHM
I'm a black man
I'm a transwoman
I'm Hispanic
I'm a Chinese-Canadian software developer
I'm a 50s hetero male
I'm a 31 year old accountant
I'm a 23 year old trans woman
I'm a 45-yr-old female
I'm a 20s black woman in California
I'm a bisexual female
I'm Korean American
I'm a lesbian
I am a cis hetero girl
I'm a military veteran
I'm a nonbinary trans woman
I'm a straight trans guy
I'm a white girl

4.2.4 Autres critères de sélection

Nombre de tokens

La taille de chaque sous-corpus, en nombre de tokens (en linguistique de corpus, un token est défini comme étant « a single linguistic unit, most often a word », Baker et al., 2006) a été dictée par deux principes. Premièrement, nous avons décidé que les sous-corpus seraient tous d'une longueur équivalente, afin d'obtenir un corpus équilibré. Lorsque l'on crée un corpus, l'équilibre entre les échantillons est, avec la représentativité, un principe fondamental (même s'il s'agit plutôt en fait d'un « idéal », car difficile à atteindre, McEnery et Hardie, 2012). Nous souhaitions éviter à tout prix d'avoir, par exemple, des sous-corpus de 500, de 5000 et de 15 000 tokens, qui auraient rendu plus difficile la comparaison entre les individus.

Ensuite, il était nécessaire de disposer de sous-corpus relativement importants, pour deux raisons principales. La première découle de l'objectif de nos analyses linguistiques, qui était d'étudier 11 variables non standard.

Les phénomènes de la CMC étant relativement rares, obtenir des échantillons de grande taille pour chaque personne maximisait les chances d'observer l'utilisation (ou la non-utilisation) de ces variables dans chaque sous-corpus. La seconde est liée au fait que notre projet n'était pas uniquement d'ordre linguistique. Nous souhaitions également établir une cartographie des centres d'intérêt des Redditors ; cela n'était possible que si nous disposions de nombreux (plusieurs centaines) de commentaires par personne, afin d'avoir un panorama assez large de leurs parcours sur Reddit.

Nous avons choisi de récolter environ 15 000 tokens par sous-corpus. Nous nous sommes arrêtée sur ce chiffre de façon empirique ; il paraissait un objectif réaliste, au vu des premiers échantillons prélevés. Nous avons utilisé le logiciel Microsoft Word pour copier les échantillons, et les mots ont été comptés, lors de la construction du corpus, avec sa fonction « Statistiques » (les fonctionnalités avancées de Word ont permis d'isoler les métadonnées du contenu des commentaires, et de prendre uniquement celui-ci en compte). Soulignons que, une fois annoté, le corpus a été analysé avec le logiciel TXM, qui n'a pas la même conception du « mot » (ou token) que Word, ce qui explique les différences entre le nombre de mots utilisé comme étalon pour prélever l'échantillon, et les nombres de tokens rapportés plus bas. Même si beaucoup d'internautes ont été exclus du corpus parce que leur historique ne contenait pas suffisamment de messages, il n'a pas été très difficile de trouver des Redditors ayant produit au moins 15 000 tokens de commentaires.

La question de l'authenticité Nous n'avons à aucun moment contacté les Redditors de notre corpus pour vérifier si les informations qu'ils et elles ont fournies sur le site sont exactes. Nous avons pris le parti d'accepter leurs déclarations pour argent comptant, en partant du principe que ce qu'ils disent d'elles et d'eux, sur Reddit, est ce qu'elles et ils sont sur Reddit. Il est évidemment possible que certain·es aient menti sur leur identité, ou aient modifié la réalité afin de protéger leur anonymat. Nous pensons toutefois que la façon dont nous avons sélectionné les internautes limite l'effet des éventuel·les « menteurs·ses ». Tout d'abord, nous avons choisi des comptes Reddit très actifs, et/ou actifs depuis plusieurs années. La longévité d'un compte est un signe d'implication et un marqueur de prestige dans une communauté en ligne (Huffaker, 2010). Privilégier ces comptes très actifs a ainsi diminué la probabilité d'inclure des *throwaways* (des comptes jetables, à la durée de vie limitée, → p. 89) et des comptes créés dans le but de faire du *trolling* (des éléments perturbateurs, susceptibles de mentir pour créer la controverse).

Ensuite, nous avons trouvé les Redditors sur des forums consacrés à des questions personnelles, principalement liées au genre. Dans ce type de forum, une certaine authenticité est attendue par la communauté (Bergstrom, 2011). Enfin, le fait que nous ayons inclus plus d'un millier de comptes Reddit dans notre corpus limite l'impact statistique des « menteurs·ses » ; par exemple, la présence éventuelle d'un homme disant être une femme (à des fins malhonnêtes, pour infiltrer des subreddits dédiés aux femmes)

n'impacte pas beaucoup nos résultats.

Fréquence de publication

Le critère temporel a également été pris en compte. Notre étude n'est pas diachronique, mais examine un « instantané » de Reddit. La langue de la CMC évoluant très vite, nous ne voulions pas intégrer des commentaires trop anciens, qui auraient pu biaiser les résultats des analyses. Par conséquent, les Redditors qui ont produit des commentaires sur une durée trop importante (au-delà de deux ans) n'ont pas été intégré·es au corpus.

Un corpus de l'« Old Reddit »

La construction du corpus a commencé en mars 2017, deux mois avant le début de la phase de bêta-test (test de la première version) des nouveaux profils de Reddit (HideHideHidden, 2017). Seul·es les Redditors ayant conservé les « anciens » profils ont été intégré·es au corpus (Figure 4.2). Le code source d'un profil sur l'Old Reddit et sur le New Reddit est le même, mais, comme nous avons recueilli les données de façon « manuelle » (c'est-à-dire en faisant de simples copiés-collés sur un logiciel de traitement de texte), nous n'avons pas utilisé le code source des pages.

FIGURE 4.2 – Un profil sur l'Old Reddit

Pour pouvoir être traitées ensemble par le script écrit par l'ingénieur de recherche, les pages copiées devaient toutes avoir la même structure et contenir le même type d'informations. Or, les nouveaux profils diffèrent des anciens profils : on peut y ajouter un avatar, une bannière et un texte descriptif, et y poster directement des commentaires, sans passer par un subreddit (figure 4.3). La structure des pages est par ailleurs plus complexe : celles-ci ne montrent pas uniquement les commentaires écrits par les Redditors, mais aussi le ou les commentaires auxquels ils répondent.

Il y avait également un autre obstacle à l'intégration de nouveaux profils dans le corpus : le fait que l'outil « Never Ending Reddit » du plugin (module d'extension) Reddit Enhancement Suite ne fonctionnait pas (à l'époque) sur

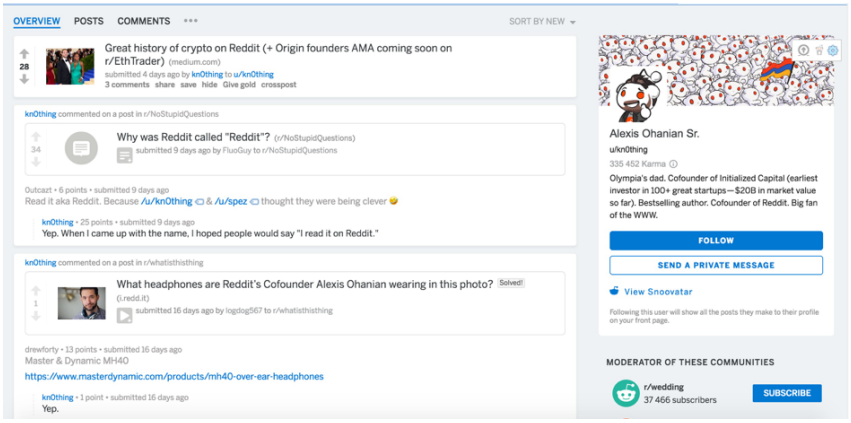

FIGURE 4.3 – Un profil sur le New Reddit

les pages des nouveaux profils. Cet outil s'est avéré très pratique pour effectuer des recherches dans les commentaires et trouver des informations personnelles, car il nous a permis de visualiser des centaines de commentaires à la fois en « scrollant » tout simplement en bas de la page au lieu de cliquer à de multiples reprises pour afficher le contenu plus ancien.

4.2.5 Difficultés rencontrées

Au fur et à mesure de la constitution du corpus, il est apparu qu'il était plus difficile de trouver certaines catégories de Redditors que d'autres. Par exemple, nous n'avons pas trouvé de subreddits dédiés aux femmes hispaniques (autres que des subreddits pornographiques). Pour disposer d'un échantillon suffisant, nous avons donc effectué des recherches dans des subreddits consacrés à la mode (r/FemaleFashionAdvice) ou à la beauté (r/MakeUpAddiction), principalement fréquentés par des femmes.

Grâce au système de flair utilisé dans plusieurs subreddits consacrés aux personnes transgenres, nous avons pu trouver relativement aisément 200 comptes Reddit à intégrer à notre corpus. La recherche a été plus difficile pour les personnes non binaires, qui représentent une minorité de la population transgenre (et non transgenre, car toutes ne s'identifient pas ainsi). Le choix était donc moins large, et, dans de nombreux cas, nous avons dû nous contenter de personnes dont nous connaissions uniquement le genre et l'âge.

L'intégration d'adolescent·es au corpus a également été problématique, principalement parce qu'il y a généralement beaucoup moins de contenu sur leurs profils que sur ceux des autres Redditors. Cela est dû à deux facteurs : premièrement, les adolescent·es fréquentent le site depuis moins longtemps (il faut être âgé de 13 ans au minimum pour créer un compte sur Reddit), et, ensuite, leurs messages sont généralement beaucoup plus courts que les messages des adultes. Comme ils sont une population souvent décrite comme fortement génératrice d'innovation langagière, nous avons décidé de dévier de notre critère de longueur des corpus (fixé à envi-

ron 15 000 tokens), de façon à pouvoir intégrer davantage d'adolescent·es dans RedditGender. Par conséquent, plusieurs sous-corpus correspondant aux productions d'adolescent·es ne dépassent pas les 10 000 tokens.

4.2.6 Un échantillon de convenance avec surreprésentation de certaines catégories

La méthode de recueil des données que nous avons adopté a entrainé la composition d'un échantillon de convenance (ou *convenience sample* en anglais, Levshina, 2015), par opposition à un échantillon aléatoire, qui consiste à prélever des individus au hasard dans une population. Le fait que nous ayons privilégié les individus qui étaient le plus facilement accessibles (c'est-à-dire ceux qui révèlent des informations sur eux), tout en surreprésentant certaines catégories, a des implications qu'il nous faut signaler. Notre échantillon ne permet pas la même généralisation à l'ensemble des utilisateur·trices de Reddit qu'offrirait un échantillon aléatoire (Levshina, 2015). Il a toutefois l'avantage de nous permettre d'explorer la diversité des internautes et de répondre à nos questions de recherche, posées dans une perspective intersectionnelle.

4.3 Composition du corpus

Le tableau 4.3 présente la composition du corpus. Il indique le nombre de tokens et de commentaires recueillis pour chaque catégorie, ainsi que la proportion que les Redditors de chaque catégorie et leurs productions représentent dans l'ensemble du corpus, en pourcentages. Ce tableau permet de faire plusieurs constats. Tout d'abord, on remarque que certaines informations sociodémographiques ont été plus faciles à recueillir que d'autres. Nous disposons ainsi de l'identité de genre, et de l'âge et du pays de tous les Redditors. L'ethnicité et l'orientation sexuelle n'ont en revanche pas pu être déterminées dans respectivement 54.79 % et 26.92 % des cas.

Les effets de nos choix d'échantillonnage sont également visibles dans ce tableau. Il montre la surreprésentation de plusieurs catégories : les personnes transgenres, qui représentent 28.74 % du corpus (alors que seuls 0.6 % des adultes américain·es se définissent comme étant transgenres Flores et al., 2016 ; des personnes non hétérosexuelles, qui représentent 57.89 % des Redditors dont l'orientation sexuelle est connue (alors que 4.5 % de la population américaine s'identifie comme LGBT Newport, 2018 ; et des personnes non blanches, qui représentent 38.67 % des Redditors dont l'ethnicité est connue (alors qu'elles représentent 27.80 % de la population américaine, « Race and ethnicity », p. d.) [1]. Ces proportions ne reflètent pas non plus la composition démographique de Reddit, un site majoritairement blanc, masculin et hétérosexuel (Barthel et al., 2016b). On constate également la prédominance des Redditors américain·es (78.83 %), même si nous

1. Notons que ces comparaisons de chiffres ont des limites, car tous les Redditors du corpus ne sont pas américain·es.

n'avons pas cherché à les privilégier au détriment des autres Redditors anglophones. Un petit nombre de Redditors (41, soit 3.93 %) ne sont pas issu·es de pays anglophones, mais écrivent sur Reddit (nous l'avons vérifié) quasi exclusivement en anglais. Nous avons supprimé, dans la mesure du possible, les commentaires écrits dans d'autres langues.

Tableau 4.3 – Composition de RedditGender

Variables	**Individus**	**Commentaires**	**Tokens**
Genre			
Hommes cisgenres	372 (35.54 %)	191 574 (41.58 %)	6 849 846 (35.44 %)
Femmes cisgenres	372 (35.54 %)	155 449 (33.74 %)	6 906 937 (35.73 %)
Hommes transgenres	100 (9.58 %)	34 776 (7.55 %)	1 867 884 (9.66 %)
Femmes transgenres	100 (9.58 %)	40 099 (8.70 %)	1 849 742 (9.57 %)
Non-binaires	100 (9.58 %)	38 809 (8.70 %)	1 854 921 (9.60 %)
Inconnu	0 (0 %)		
Âge			
14-20 ans	147 (14.08 %)	81 123 (17.61 %)	2 638 980 (13.65 %)
21-30 ans	517 (49.52 %)	219 048 (47.55 %)	9 581 767 (49.57 %)
31 ans et plus	380 (36.40 %)	160 536 (34.85 %)	7 108 583 (36.78 %)
Inconnu	0 (0 %)		
Ethnicité			
Blanc·hes	203 (19.44 %)	86 223 (18.72 %)	3 753 397 (19.42 %)
Afro-américain·es	92 (8.81 %)	44 846 (9.73 %)	1 697 823 (8.78 %)
Asiatiques	69 (6.61 %)	29 164 (6.33 %)	1 278 828 (6.62 %)
Hispaniques	68 (6.51 %)	36 471 (7.92 %)	1 207 415 (6.25 %)
Autres	50 (4.79 %)	20 970 (4.55 %)	931 720 (4.82 %)
Inconnue	562 (54.79 %)		
Orientation sexuelle			
Hétérosexuelle	468 (44.83 %)	204 805 (44.45 %)	8 704 751 (45.03 %)
Gay	189 (18.10 %)	84749 (18.40 %)	3 513 489 (18.18 %)
Bisexuelle	73 (6.99 %)	29 528 (6.41 %)	1 346 155 (6.96 %)
Autre	33 (3.16 %)	11 132 (2.42 %)	614 209 (3.18 %)
Inconnue	281 (26.92 %)		
Pays			
États-Unis	823 (78.83 %)	363 678 (78.94 %)	15 247 626 (78.88 %)
Canada	82 (7.85 %)	35 581 (7.72 %)	1 509 928 (7.81 %)
Royaume-Uni	53 (5.08 %)	22 725 (4.93 %)	997 938 (5.16 %)
Autres pays anglophones	45 (4.31 %)	18796 (4.08 %)	748 559 (3.87 %)
Pays non-anglophones	41 (3.93 %)	19 927 (4.33 %)	825 279 (4.27 %)
Inconnu	0 (0 %)		
Total	1044	460 707 (100 %)	19 329 330 (100 %)

La figure 4.4 présente deux boites à moustaches (→ p. 152) qui représentent la distribution des données du corpus en termes de nombre de tokens (à gauche) et de commentaires (à droite). Ces données ont été obtenues une fois le corpus encodé dans le logiciel TXM. En ce qui concerne le nombre de tokens, on peut voir que la médiane (18 656.5) est supérieure au nombre

de 15 000 choisi comme objectif lors du recueil des données. Cela est dû au fait que Word et TXM n'ont pas la même conception du token. Un token, pour TXM, n'est pas seulement ce que l'on considère traditionnellement comme un « mot », mais aussi un signe de ponctuation, par exemple.

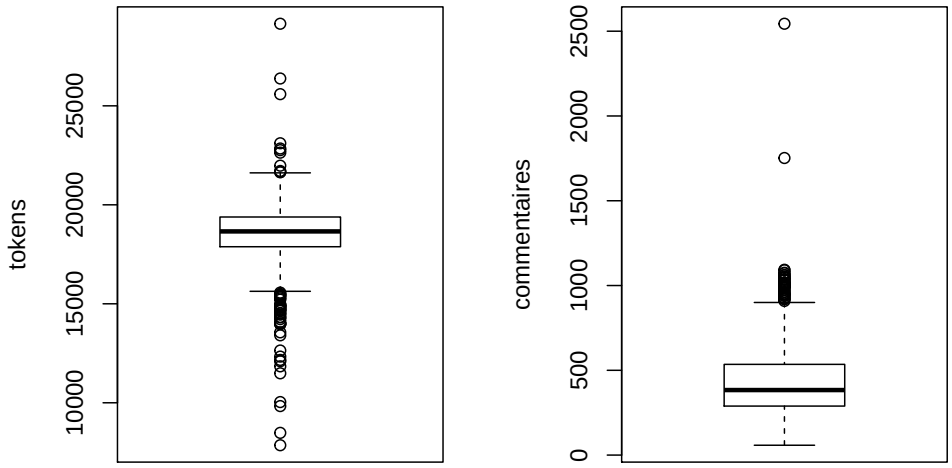

FIGURE 4.4 – Nombre de tokens et de commentaires par sous-corpus

La boite à moustaches correspondant aux tokens montre que les données sont relativement resserrées autour de la médiane. La moitié des sous-corpus (l'intérieur de la boite) ont ainsi une longueur comprise entre 17 883 et 19 384 tokens. Il y a tout de même des valeurs extrêmes. Les points les plus bas dans le graphique correspondent aux sous-corpus les plus petits, généralement ceux d'adolescent·es ; le minimum est de 7855 tokens. Ce graphique révèle que nous avons (à peu près) réussi à équilibrer les sous-corpus entre eux. La boite à moustaches représentant la distribution du nombre de commentaires par sous-corpus montre qu'il y a également de la variation, tous les commentaires n'ayant évidemment pas la même longueur. La moitié des sous-corpus comportent ainsi entre 289 et 535 commentaires ; le nombre médian de commentaires par sous-corpus est de 383.5.

4.4 Construction du corpus

4.4.1 Extraction des commentaires

En première année de thèse, il ne nous paraissait pas envisageable d'utiliser l'API (interface de programmation) de Reddit pour extraire les

données, car nous n'avions aucune connaissance en la matière. Pour notre travail de master, nous avions utilisé une méthode « artisanale », qui a consisté à faire de simples copiés-collés depuis les pages web sur un document Word. Cette méthode manuelle avait pour inconvénient d'être chronophage, et pour avantage d'être facile à mettre en œuvre. Nous avons donc décidé de l'utiliser à nouveau : le sacrifice du temps au profit de la simplicité d'utilisation nous semblait un compromis intéressant. Il nous a permis de commencer la construction du corpus assez tôt dans le travail de thèse, à savoir en mars 2017, six mois après le début du doctorat. Le processus de sélection des Redditors a de toute façon nécessité l'inspection manuelle de chaque profil ; il était ensuite simple de copier ces commentaires, accompagnés de leurs métadonnées, et de les coller dans un document Microsoft Word.

4.4.2 Annotation et encodage du corpus

Restait encore le problème de l'encodage des données, en partie conditionné par le logiciel d'exploitation du corpus. En master, nous avions utilisé AntCont (Anthony, p. d.) ; ce logiciel ne paraissait toutefois pas adapté au projet de thèse, notamment parce qu'il n'était pas assez puissant. Nous avons testé WordSmith Tools (M. Scott, 2016), que nous avons rapidement abandonné car il ne reconnaissait pas (à l'époque tout du moins) les émojis, qui font partie des phénomènes que nous souhaitions étudier. La rencontre avec Bertrand Gaiffe, ingénieur de recherche à l'ATILF, a été décisive. Il a proposé d'utiliser le logiciel TXM, qui dispose de fonctionnalités puissantes et est relativement facile à utiliser. Il s'est ensuite chargé d'encoder les fichiers Word obtenus de façon à ce qu'ils puissent être exploités par TXM. Dans un premier temps, il les a convertis en documents TEI P5 XML (un format développé par la Text Encoding Initiative qui permet de décrire la structure des documents) grâce au site Oxgarage (Initiative, p. d.), un dispositif de l'université d'Oxford. Ensuite, il a utilisé le langage informatique XSL pour annoter le corpus. Ce travail s'est étalé sur plus d'une année. Les données ont été recueillies d'avril à juillet 2017. Dès avril, les premiers tests, qui ont permis d'affiner l'annotation et d'éliminer les erreurs, ont été effectués sur des extraits du corpus. Une fois le corpus entièrement constitué, plusieurs versions ont été réalisées et testées. La dernière a été créée en janvier 2019.

Annotation

L'objectif principal du processus d'annotation était de conserver autant d'informations que possible. En effet, à ce stade de la thèse, le périmètre des analyses à effectuer n'était pas encore entièrement défini, et nous ne souhaitions pas supprimer des informations dont nous aurions pu avoir besoin par la suite.

Annotation des sous-corpus Chaque sous-corpus (correspondant à tous les commentaires d'une personne) porte comme nom l'identifiant du Red-

ditor. Il a été annoté avec toutes les informations démographiques pertinentes pour nos analyses (identité de genre, âge, orientation sexuelle, etc.).

Annotation des métadonnées des commentaires La figure 4.5 présente un commentaire mis en ligne sur Reddit (mais qui ne fait pas partie du corpus), avec les métadonnées que nous avons recueillies : titre du fil de discussion, pseudonyme de l'auteur·e du fil de discussion, pseudonyme de l'auteur·e du commentaire, nom du subreddit, date, et nombre de points de karma recueillis par le commentaire.

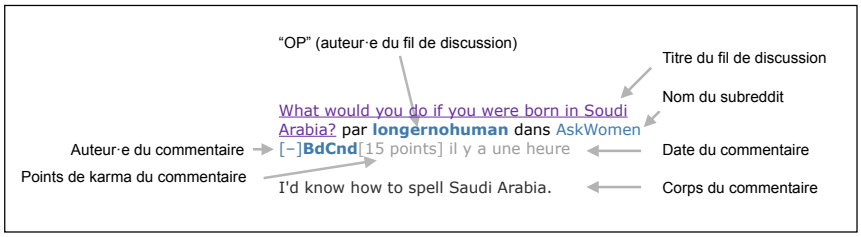

FIGURE 4.5 – Capture d'écran d'un commentaire, avec explications sur ses métadonnées

Ces données ont été intégrées au corpus et annotées (à l'exception de la date, qui le sera ultérieurement). La capture d'écran présentée dans la figure 4.6 montre les balises créées par Bertrand Gaiffe pour annoter ces informations : « forum », « fil », « auteur », « score », « nbCommentaires », « createurFil » et « type ». La balise « div » définit le début d'un commentaire.

```
<div forum="AskWomen"
fil="What_s_a_personal_topic_that_s_easier_to_talk_about_with_
friends_than_your_SO_" auteur="F_020_170315" score="1"
nbCommentaires="14" createurFil="_dans_" type="commentaire">
```

FIGURE 4.6 – Capture d'écran des métadonnées d'un commentaire du corpus

Annotation du contenu des commentaires Les commentaires ont été annotés de façon à conserver un maximum d'informations sur leur mise en forme (gras, italique, texte barré, liens hypertexte, etc.). Nous avons tenté de supprimer les URLs quand il y en avait, mais l'opération n'a pas toujours été fructueuse. Nous avons également supprimé les éventuelles citations faites par les Redditors ; il est en effet possible sur Reddit d'utiliser une option de mise en forme qui indique que l'on cite un texte dont on n'est pas l'auteur·e, qu'il s'agisse d'un commentaire d'un autre Redditor ou de tout autre texte. La figure 4.7 présente un exemple de cette option de mise en forme, souvent utilisée par les Redditors ; la citation (« I own an escape room ») apparait en gris clair.

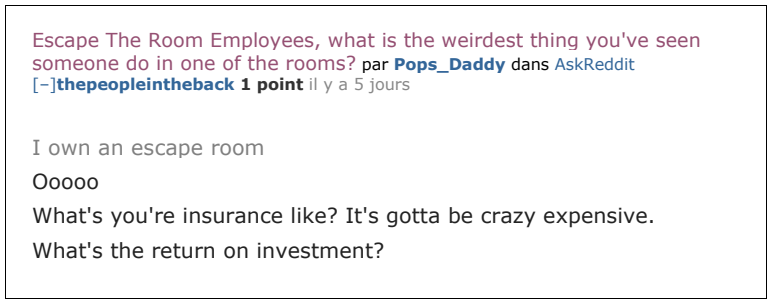

FIGURE 4.7 – Capture d'écran d'un exemple de citation dans un commentaire

Le cas des émojis Les émojis ont fait l'objet d'un traitement particulier. En effet, TXM ne prend pas en charge les caractères dont le code Unicode est supérieur à 65 735 (à la date où nous écrivons ces lignes) . À partir d'une liste des codes Unicode (« Full Emoji List, v13.0 », p. d.), Bertrand Gaiffe a créé un script qui encode les émojis de façon à ce qu'ils puissent être exploités par TXM. La plupart des émojis n'apparaissent pas dans les outils de recherche du logiciel, mais il est possible de les trouver à l'aide de leur nom anglais.

Ponctuation Nous n'avons pas, comme cela se fait parfois, supprimé la ponctuation, qui était une de nos variables d'intérêt. Cela a une conséquence importante : dans notre corpus, les signes de ponctuation sont considérés comme des « tokens », c'est-à-dire des unités lexicales (ce que l'on considère généralement comme un « mot »). C'est la raison pour laquelle nous utilisons le terme « token » et non le terme « mot » dans les analyses. Ainsi, quand nous parlons de la fréquence d'un émoji ou d'une émoticône pour « 1000 tokens », il ne faut pas oublier que ces 1000 tokens comprennent les caractères de ponctuation.

Annotation morphosyntaxique L'annotation morphosyntaxique et en lemmes a été automatiquement effectuée lors de l'import du corpus dans TXM avec TreeTagger version 3.2 (H. Schmid, p. d.), un outil qui permet de lemmatiser un texte et d'annoter les partis du discours. Chaque token est ainsi accompagné d'une balise qui indique à quelle partie du discours il appartient (par exemple, « JJ » pour les adjectifs, « NP » pour les noms propres, « VVD » pour les verbes au passé, « RB » pour les adverbes, etc.), et d'une balise qui précise son lemme (« be » pour « was », « cat » pour « cats », etc.). Nous avons toutefois très peu eu recours à ces annotations dans nos analyses.

Erreurs

Malgré les nombreux tests effectués et les modifications apportées au corpus, des erreurs subsistent. Deux sous-corpus n'ont pas pu être intégrés au corpus. Certaines URL n'ont pas été supprimées, et les dates des

commentaires n'ont pas été annotées.

4.5 Structure du corpus

Notre méthode de recueil des données sociodémographiques a dicté la structure du corpus. Contrairement à de nombreux grands corpus de Reddit (par exemple ceux de Baumgartner, p. d. ou Farrell et al., 2019), qui sont structurés autour de subreddits et de fils de discussion, RedditGender est construit autour des profils des Redditors. Il contient donc uniquement les commentaires mis en ligne sur le site par chaque personne, et non pas l'intégralité des échanges ayant eu lieu dans un fil de discussion.

Cette structure présente à la fois des inconvénients et des avantages. Elle fournit une vue tronquée des interactions se déroulant sur Reddit. On sait, pour chaque commentaire recueilli, dans quel subreddit et dans quel fil de discussion il a été écrit. On ne sait pas, en revanche, quels commentaires précèdent ou suivent les commentaires du corpus. Comme nous souhaitions étudier les procédés d'écriture, et non pas les interactions entre Redditors, cela ne nous a pas semblé être problématique. Il est de toute façon généralement possible, en cliquant sur le lien correspondant à chaque fil de discussion contenu dans la version Word de notre corpus, d'accéder à l'intégralité des échanges (à condition que le commentaire ou le profil n'ait pas été supprimé depuis la création du corpus).

Les avantages de cette méthode l'emportent, pour notre étude, sur ses inconvénients. Tout d'abord, il aurait été impossible d'obtenir des données sociodémographiques sur tou·tes les Redditors ayant écrit un message dans un fil de discussion donné, car de nombreux·ses Redditors ne dévoilent pas d'informations sur leur identité. Ensuite, en utilisant les profils des internautes, nous avons pu cibler celles et ceux qui avaient écrit un grand nombre de commentaires, ce qui nous a permis d'obtenir un aperçu assez large de la façon dont chaque personne utilise la langue du web.

Enfin grâce à cette structure, nous disposons non pas uniquement des commentaires mis en ligne par une personne dans un subreddit, mais aussi des commentaires qu'elle a publiés dans tous les subreddits auxquels elle a participé. Cela nous a permis d'étudier, en plus des phénomènes langagiers, les centres d'intérêt des Redditors.

4.6 Exploitation du corpus

4.6.1 TXM

TXM a été développé dans le cadre du projet de recherche « Fédération des recherches et développements en textométrie autour de la création d'une plateforme logicielle ouverte » financé par l'Agence Nationale de la Recherche française (ANR). Le logiciel permet de construire et d'analyser des corpus au format XML. Il a été créé afin de rendre la textométrie, ou analyse de corpus, plus accessible aux utilisateur·trices issu·es du domaine des sciences humaines et sociales (Heiden et al., 2010).

Outre sa simplicité d'utilisation, TXM a l'avantage d'être suffisamment performant pour analyser des corpus complexes et de grande taille. Cet outil open source articule plusieurs éléments : le moteur de recherche IMS Corpus Workbench, l'environnement de calcul statistique R, et un module d'importation de corpus XML-TEI (Heiden et al., 2010). Il est proposé sous deux formes : un logiciel à installer sur un poste local, compatible avec Mac, Linux et Windows, et une application en ligne. Nous avons utilisé le logiciel.

Le langage CQL

Dans TXM, les requêtes sont traitées avec le moteur de recherche CQP, ou « Corpus Query Processor », un composant logiciel développé à l'université de Stuttgart (Christ et al., 1999). Pour effectuer des recherches sur TXM, il faut donc utiliser le CQL, ou « Corpus Query Language », « un langage formel, avec un lexique et une syntaxe d'opérateurs qui forment un métalangage permettant de combiner des éléments pour la recherche de motifs structurés » (*Manuel de TXM*, 2018, p. 179). Des exemples de requêtes effectuées pour réaliser nos analyses sont présentés dans le chapitre suivant.

Limites de TXM

Malgré sa puissance et la richesse de ses fonctionnalités, TXM ne répondait pas entièrement à nos besoins. Deux problèmes principaux se posaient. Tout d'abord, le logiciel ne fournit que la fréquence brute d'un mot, c'est-à-dire le nombre de fois où il apparait dans un corpus, et non sa fréquence relative (ou normalisée), c'est-à-dire le nombre d'occurrences d'un token pour, par exemple, 1000 ou 10 000 tokens. Disposer de la fréquence relative d'un token est indispensable pour pouvoir comparer des corpus de taille différente entre eux. Il est possible de calculer manuellement la fréquence relative d'un mot à partir des informations fournies par TXM, mais cette méthode n'était pas envisageable au vu de la quantité de données que nous souhaitions analyser.

L'autre problème, c'est que, lorsque l'on effectue une requête dans un corpus avec TXM (par exemple, si on recherche toutes les occurrences de l'acronyme *lol*), le logiciel ne fournit qu'un résultat global. Il n'est pas possible d'obtenir directement le nombre d'occurrences de *lol* dans chacun des 1044 fichiers dont notre corpus est composé. La seule solution est de créer 1044 sous-corpus, puis d'effectuer 1044 requêtes, ce qui est évidemment extrêmement fastidieux. Heureusement, TXM permet d'exporter les résultats des requêtes au format .csv, afin de les exploiter dans un autre logiciel. Nous avons opté pour R, un logiciel aujourd'hui largement utilisé en linguistique quantitative.

4.6.2 R

R est un langage de programmation et un logiciel libre dédié à la statistique et à la création de graphiques. Le langage R a été développé par Ross

Ihaka et Robert Gentleman à l'université d'Auckland, en Nouvelle-Zélande, à partir de 1993. La première version libre du logiciel est sortie en 1995. Il est depuis 1997 développé par une équipe d'une vingtaine de développeurs, la R Development Core Team. R est principalement inspiré par le langage S. Son nom est à la fois un clin d'œil aux prénoms de ses deux fondateurs et à S (« R FAQ », p. d.). Le logiciel permet de réaliser de nombreuses analyses statistiques, dont, entre autres, les modèles de régression linéaires et généralisés, l'analyse factorielle, l'analyse de cluster ou encore les tests paramétriques et non paramétriques. R est également réputé pour la flexibilité et la puissance de son environnement graphique. Le logiciel s'utilise généralement à partir de l'interface utilisateur graphique gratuite RStudio.

R était en 2019 le 16ème langage de programmation le plus utilisé au monde (« TIOBE Index », p. d.). Le logiciel offre de nombreux avantages : il est gratuit et libre, et de qualité comparable aux autres logiciels de statistiques payants, comme SAS, SPSS et Stata. Sa syntaxe est relativement simple (Baayen, 2008), et il est compatible avec Mac, Windows et Linux. De nombreux packages, créés par des statisticiens et des spécialistes de diverses disciplines, permettent de réaliser des tâches spécifiques. En janvier 2020, plus de 15 000 packages étaient ainsi disponibles (« The Comprehensive R Archive Network », p. d.). Aujourd'hui, R est devenu un outil de référence dans de nombreux domaines de la linguistique, en particulier la linguistique de corpus et la linguistique computationnelle (Levshina, 2015). Nous avons utilisé la version 3.5.1 de R (R Core Team, 2018), avec la version 1.1.456 de l'environnement de développement RStudio (RStudio Team, 2016).

4.6.3 Extraction des données depuis TXM

Utilisation de la fonction « Lexique » de TXM

Pour certaines variables linguistiques, nous avons adopté une perspective « corpus-driven » (Tognini-Bonelli, 2001), qui a consisté à explorer le corpus de façon « naïve », sans liste préalable des éléments à rechercher, en inspectant le corpus. Dans la mesure du possible, nous sommes donc partie du contenu du corpus pour déterminer les variables à étudier. Cette approche a été rendue possible par la fonction « Lexique » de TXM. Celle-ci « calcule la liste des fréquences de toutes les valeurs de propriétés lexicales d'un corpus ou d'un sous-corpus », c'est-à-dire de tous les types (mots uniques) présents dans un corpus (*Manuel de TXM*, 2018, p. 97). Elle génère un tableau qui peut être exporté au format .csv. La figure 4.8 montre un extrait du « Lexique » ainsi généré. On y voit, par exemple, que le mot *some* a une fréquence de 36 436 dans le corpus ; il est suivi par les mots qui arrivent après lui dans la liste de fréquence. Cette liste a été inspectée manuellement dans le but de repérer toutes les formes lexicales non standard caractéristiques de l'anglais d'internet.

Cette méthode a toutefois des limites, imposées par le fonctionnement de TXM. Émoticônes et émojis, notamment, ne sont pas identifiés par la

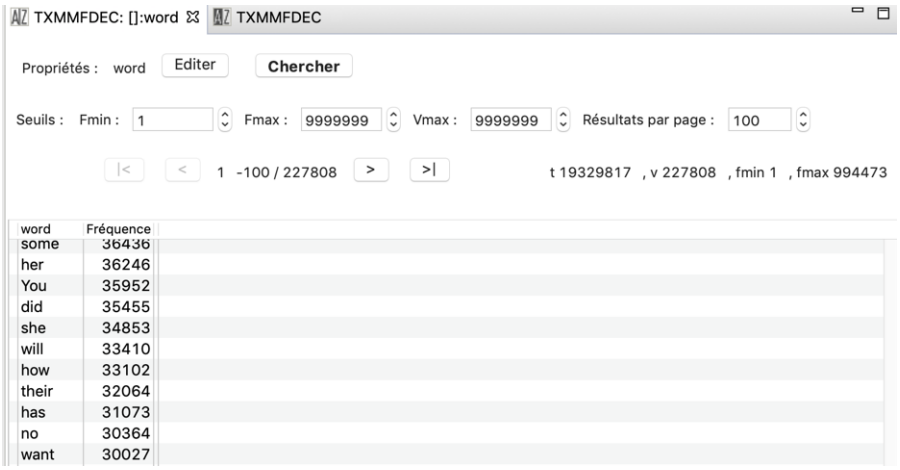

FIGURE 4.8 – Capture d'écran montrant le résultat de la fonction « Lexique » de TXM, avec RedditGender

fonction « Lexique ». Les premières sont en effet composées de plusieurs types de caractères : lettres, symboles et signes de ponctuation. Quand une émoticône intègre un ou plusieurs signes de ponctuation, elle n'est pas considérée comme une unité lexicale par le logiciel, mais comme plusieurs tokens distincts. Ses différents éléments apparaissent donc de façon séparée dans le « Lexique » du corpus. Pour repérer les émoticônes, il a été nécessaire de faire des recherches avec la fonction « Concordance », à partir de listes d'émoticônes trouvées sur internet. Les émojis, quant à eux, n'étaient pas reconnus par TXM, comme nous l'avons expliqué plus haut, et ont été annotés automatiquement par l'ingénieur de recherche. Par ailleurs, inspecter la totalité du lexique est fastidieux, à cause du nombre élevé de types différents dans le corpus (227 808). Pour certaines variables, comme les interjections, nous avons plutôt eu recours à des listes préétablies, comme nous l'expliquons dans le chapitre suivant (→ p. 144).

Génération des concordances

Chaque type (ou mot) identifié comme faisant partie du lexique de la CMC a fait l'objet d'une recherche séparée avec la fonction « Concordance » de TXM, qui « construit une concordance kwic [2] à partir des résultats de recherche correspondant à une requête CQL sur un corpus » (*Manuel de TXM*, 2018, p. 103). La concordance peut être exportée au format .csv, et être exploitée avec un autre logiciel (R, dans notre cas).

Comme le montre la figure 4.9, les résultats de la fonction « Concordance » intègrent dans notre cas quatre informations (a minima). Tout à gauche figure l'identifiant de chaque texte ou sous-corpus (nous avons masqué le début des identifiants, qui contient l'abréviation correspondant à l'identité de genre des internautes). Dans notre cas, chaque identifiant correspond à la production d'un·e Redditor. On trouve ensuite les deux co-

2. Acronyme de *Key Word in Context*

lonnes de « contexte » gauche et droit, qui montrent un extrait du texte précédant et suivant le mot recherché (nous avons réduit le contexte à 1 token à gauche et à 1 token à droite, afin de ne pas révéler le contenu des messages). Enfin, la colonne « pivot » contient la ou les unités lexicales recherchées.

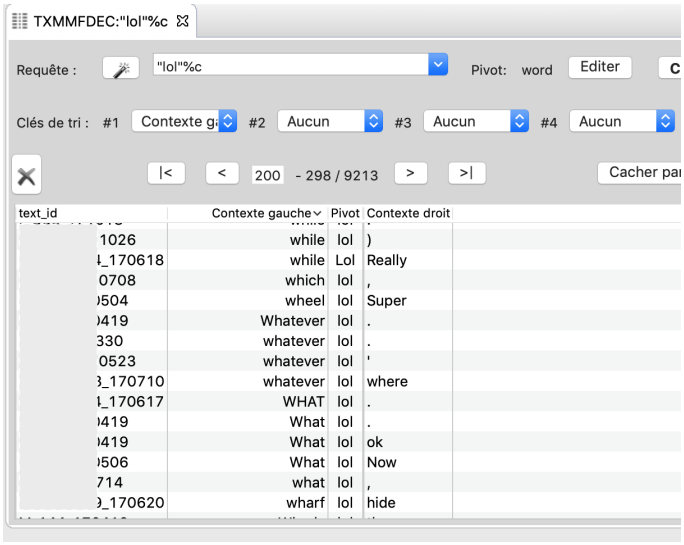

FIGURE 4.9 – Capture d'écran montrant le résultat de la requête « lol » avec la fonction « Concordance » de TXM, avec RedditGender

Les requêtes ont toujours été effectuées avec la syntaxe "unitélexicale"%c, qui permet de neutraliser la casse. Cela a permis d'obtenir les différentes variantes d'un mot, avec majuscules ou minuscules, ou avec une combinaison des deux. Le résultat de la recherche de "lol"%c inclut ainsi *lol*, *Lol*, *LOL*, *loL* ou encore *LoL*.

4.6.4 Importation des données dans R

Les fichiers obtenus, correspondant chacun aux résultats d'une recherche de concordance, ont été analysés avec R. Avant de pouvoir réaliser des statistiques, il a fallu transformer le format des données. Une fois importé dans R, le fichier .csv est converti en *data frame* (tableau de données)[3]. Ce *data frame* a la forme des données originelles de TXM : à chaque ligne correspond une occurrence d'une unité lexicale, accompagnée de son contexte et de son identifiant. Les Redditors n'ayant pas utilisé une seule fois l'unité lexicale recherchée ne sont pas inclus dans le *data frame*.

Pour réaliser nos analyses, nous avions besoin d'un format différent, avec un *data frame* composé de 1044 lignes, correspondant aux 1044 Redditors de RedditGender. Chaque ligne devait comporter l'identifiant de la personne, et le nombre de fois où elle a utilisé une unité lexicale donnée. Le contexte n'est pas nécessaire. Nous avons donc écrit un script R permettant de remanier les données. Dans un premier temps, les différentes

3. Nous utilisons le terme *data frame* car le logiciel R et sa terminologie sont en anglais.

graphies de chaque unité lexicale (avec ou sans majuscules) ont été normalisées, c'est-à-dire mises en lettres minuscules. Cela permet à R de toutes les considérer comme des éléments identiques. Les colonnes « contexte droit » et « contexte gauche » ont ensuite été supprimées. Avec la fonction `table()`, nous avons calculé le nombre d'occurrences de chaque token dans chaque sous-corpus. Nous avons ainsi obtenu un objet R `table`, qui a été transformé en *data frame*. Ce *data frame* a ensuite été « collé » au *data frame* contenant les données sociodémographiques (âge, genre, etc.) avec la fonction `merge()`. Les observations manquantes, correspondant aux internautes n'ayant pas utilisé un token, ont été remplacées par des 0. Cette procédure a dû être effectuée pour la majorité des éléments étudiés dans cette thèse. Le script R a permis de le faire de façon quasi instantanée pour chaque élément d'intérêt.

4.7 Mise à disposition du corpus et éthique

Un des objectifs de notre travail était de mettre le corpus RedditGender à disposition des chercheurs·e, afin qu'il puisse être utilisé pour d'autres études et pour rendre notre recherche reproductible. Toutefois, deux obstacles principaux se posent encore. Tout d'abord, en l'état actuel, notre corpus ne répond pas aux critères « FAIR », c'est-à-dire de « Findability, Accessibility, Interoperability, and Reusability » (Wilkinson et al., 2016), aujourd'hui largement utilisés dans la communauté scientifique et décrits, pour leur application dans les corpus de CMC, par Frey et al. (2019). Ces critères ont été mis en place pour promouvoir une recherche transparente et reproductible, et pour partager des données (ici, un corpus) dont la création est coûteuse et chronophage.

Pour que RedditGender puisse être utilisé par d'autres chercheur·es, il faudrait, notamment revoir son annotation afin de la rendre plus transparente et cohérente, et fournir une documentation qui donne une sorte de « mode d'emploi » du corpus. Le problème principal n'est toutefois pas d'ordre technique, mais éthique. Il nous faudrait trouver une solution permettant partager nos données en respectant l'anonymat des internautes qui ont contribué, sans le savoir, au corpus. Il semble que RedditGender réponde aux exigences du RGDP ou Règlement européen sur la protection des données (« Le règlement général sur la protection des données - RGPD | CNIL », p. d.), car les données qu'il contient sont hébergées aux États-Unis, librement accessibles sur internet et ne permettent pas d'identifier les personnes.

Toutefois, même si les pseudonymes des Redditors sont anonymisés, ils peuvent facilement être retrouvés en effectuant une recherche de leurs commentaires sur Google. Le contenu des commentaires pourrait également servir à connaitre leur genre, leur orientation sexuelle ou leur ethnicité. Toutes ces données pourraient être utilisées par des individus malveillants pour harceler ou *doxxer* des internautes (révéler des informations privées dans le but de leur nuire). Le problème est particulièrement aigu pour les personnes transgenres et non binaires, qui peuvent vivre leur identité de

genre en tant que *stealth* (c'est-à-dire sans révéler leur statut transgenre), ou de façon secrète. Notre corpus préserve par ailleurs des commentaires et des comptes Reddit que les Redditors peuvent décider à tout moment de supprimer, pour effacer leurs traces sur Reddit.

Des solutions pourraient être envisagées pour éviter ces problèmes : mettre le corpus à disposition sur demande (et non librement), fournir les métadonnées uniquement aux chercheur·es, ne rendre disponible qu'une partie du corpus (uniquement les personnes cisgenres, par exemple), ou demander, a posteriori, le consentement des internautes. Toutes ces questions demandent encore réflexion. Pour préserver l'anonymat des personnes dont nous avons recueilli les commentaires, nous ne citons jamais leurs commentaires tels quels ; nous paraphrasons, et (plus rarement) traduisons ce qu'ils ont écrit. De la même façon, nous ne citons jamais leurs pseudonymes. Les noms d'utilisateur·trices présentés dans notre étude des pseudonymes et cités en exemple ne sont pas ceux des Redditors de RedditGender.

tl ;dr

Nous avons utilisé, pour créer le corpus, la méthode de l'échantillon de convenance. Les Redditors n'ont pas été sélectionné·es au hasard, mais ont été choisi·es parce qu'ils et elles ont produit de nombreux commentaires sur le site, et parce qu'ils et elles ont révélé, par leurs commentaires ou leurs « flairs », des informations sociodémographiques. Nous avons diversifié l'échantillon en surreprésentant certaines catégories de façon à pouvoir réaliser des analyses statistiques dans une perspective intersectionnelle.
Nous avons construit le corpus en copiant les commentaires des Redditors sur des documents Word. Grâce à l'aide d'un ingénieur de recherche de l'ATILF, nous avons transformé ces documents en un corpus richement annoté. Celui-ci a été exploité avec le logiciel de textométrie TXM. Les résultats des concordances ont été importés dans le logiciel de statistique R. La question du partage du corpus reste encore à l'étude à l'heure où nous écrivons ces lignes, à cause du caractère sensible des données recueillies.

Chapitre 5

Les variables

Cette section présente les différentes variables que nous étudions dans cette thèse. Elle décrit tout d'abord les variables sociodémographiques recueillies, puis les variables de ce que nous appelons la « Reddidentité », c'est-à-dire les informations disponibles sur les profils des utilisateur·trices de Reddit, explorées dans le chapitre 7. Ensuite, nous expliquons comment nous avons codé les subreddits pour réaliser l'analyse des centres d'intérêt (→ chapitre 8). Enfin, nous présentons les 11 variables linguistiques, classées en procédés d'ajout et de réduction, que nous analysons dans les chapitres 10 et 11.

5.1 Les variables sociales

5.1.1 Genre

Le genre a été traité comme une variable catégorielle non binaire. Elle comprend cinq niveaux : hommes cisgenres (N = 372), femmes cisgenres (N = 372), femmes transgenres (N = 100), hommes transgenres (N = 100), et individus non binaires (N = 100). Ce dernier niveau est hétérogène, car on y trouve des personnes agenre, genderqueer, genderfluid, et bigenre. Pour les personnes non binaires, nous disposons également du genre assigné à la naissance de 98 personnes : 59 AFAN (assignées filles à la naissance) et 39 AGAN (assignés garçons à la naissance) (→ p. 17).

5.1.2 Âge

Autant que possible, nous avons essayé de déterminer l'âge exact des Redditors en nombre d'années, à partir de leurs déclarations du type « I was born in 1987 » ou « I am 42 years old », ou lorsqu'ils et elles l'indiquent dans leur flair. L'âge a été calculé au moment de la récolte des données. Par exemple, si un·e internaute écrit qu'il ou elle a 27 ans dans un commentaire écrit un an avant la construction du corpus, nous avons ajouté une année à l'âge indiqué. Parfois, nous n'avons pu déterminer qu'une tranche d'âge, lorsqu'un flair indiquait « 30-35 », ou quand un·e internaute a déclaré :

« I was born in the nineties » ou « As someone who came of age in the eighties ». Dans ce dernier cas, par exemple, l'internaute dit avoir atteint sa majorité dans les années 80, ce qui signifie qu'il ou elle est né entre 1962 et 1971, et qu'il ou elle avait donc entre 55 et 46 ans en 2017, date à laquelle ce commentaire a été mis en ligne sur Reddit et à laquelle les données ont été recueillies. L'âge exact a pu être relevé pour 861 personnes (82.47 % des Redditors), mais pas pour les 183 autres (17.52 %). Nous avons traité l'âge comme une variable catégorielle, et non pas comme une variable numérique. Nous avons utilisé le codage par décennies présenté dans le tableau 5.1.

Tableau 5.1 – Composition des catégories d'âge

Catégorie	Individus	%
14 à 20 ans	147	14.08 %
21 à 30 ans	517	49.52 %
31 ans et plus	380	36.40 %
Total	1044	100 %

Notons que, dans les cas où nous disposions d'une tranche d'âge englobant deux décennies (25 à 34 ans ou 37 à 42 ans, par exemple), nous avons placé le ou la Redditor dans la catégorie correspondant au milieu de la fourchette indiquée. Par exemple, dans le premier cas, nous avons calculé : 34 – 25 = 9 ans ; 9 ans divisés par 2 = 4.5 ; 25 + 4.5 = 29.5 ans. L'internaute a donc été placé·e dans la catégorie 2 (21 à 30 ans).

En termes de tranches d'âge, notre catégorisation n'est pas équilibrée. La catégorie 1 couvre 6 ans (14 à 20 ans), la catégorie 2 10 ans (21 à 30 ans) et la catégorie 3 35 ans (31 ans à 65 ans). Nous aurions aimé pouvoir faire une catégorisation par décennies pour les Redditors les plus âgé·es (31 à 40 ans, 41 à 50 ans, etc.). Toutefois, comme le montre la figure 5.1.2, le corpus est composé majoritairement de personnes jeunes. Établir des catégories plus fines pour les adultes de plus de 31 ans aurait donc entrainé la composition de petits groupes comportant parfois de moins de 20 individus, ce qui aurait posé des problèmes pour les analyses statistiques. De plus, nous limiter à 3 groupes facilite l'interprétation des interactions entre genre et âge, comme nous l'expliquons dans la section 6.4.5.

5.1.3 Ethnicité

Comme pour le genre, nous avons utilisé les déclarations des Redditors pour déterminer leur ethnicité. Rappelons que nous employons le terme « ethnicité » pour désigner le concept de « race » (→ p. 41), et que notre catégorisation représente des groupes sociaux ; elle ne vise en aucun cas à catégoriser la population sur des bases génétiques. La majorité des internautes de RedditGender étant américain·es, notre catégorisation est basée sur les catégories ethniques de l'US Census, le recensement américain (Bureau, p. d.). Il est à noter que « hispanique » n'est pas considéré comme

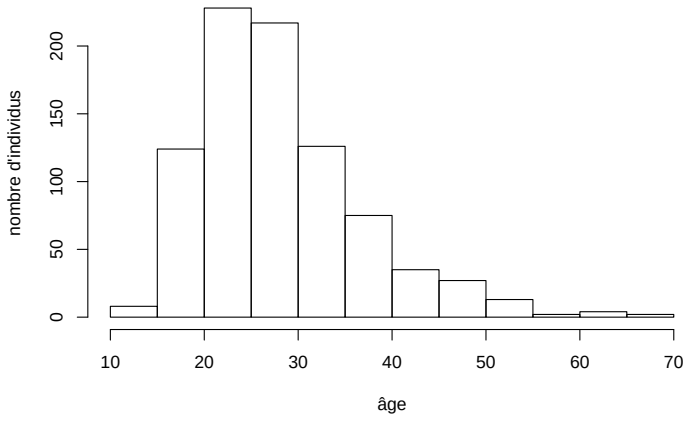

FIGURE 5.1 – Âge des Redditors de RedditGender

une « race » par l'US Census. Nous avons choisi d'inclure cette catégorie parce que, selon un sondage du Pew Research Center, deux tiers des Hispaniques considèrent leur hispanicité comme leur origine raciale (Gonzalez-Barrera & Lopez, 2015). Signalons, une fois de plus, que nous sommes bien consciente du caractère problématique de cette catégorisation. L'ethnicité est un concept complexe, et toute tentative de catégorisation est forcément réductrice. Toutefois, cette variable est une composante essentielle de notre projet intersectionnel.

Nous disposons d'informations sur l'ethnicité pour 482 Redditors (46.17 % des internautes de RedditGender). La catégorie « autre » a été constituée pour rassembler les Redditors de diverses ethnicités (amérindienne, notamment), qui ne sont pas assez nombreux pour constituer une catégorie dans notre étude. Elle regroupe également les Redditors s'identifiant comme étant « mixed race » (c'est-à-dire, ayant par exemple un parent blanc et un parent asiatique). Elle est présentée ici, mais n'a pas été incluse dans les analyses statistiques, à cause de son caractère hétérogène.

TABLEAU 5.2 – Catégories ethniques dans RedditGender

Catégorie	Individus	%
Blanc·hes	203	19.44 %
Afro-américain·es	92	8.81 %
Asiatiques	69	6.61 %
Hispaniques	68	6.51 %
Autres (non-blanc·hes)	50	4.79 %
Inconnue	562	53.83 %

5.1.4 Orientation sexuelle

L'orientation sexuelle des internautes a été déterminée soit par l'examen des commentaires des Redditors, soit par les indications fournies dans les flairs utilisés dans certains subreddits. Elle a été obtenue dans 763 cas (73.08 % des Redditors). Cette variable catégorielle comporte 6 niveaux présentés dans le tableau 5.3. Nous avons intégré cette variable à l'analyse des émoticônes (→ p. 225).

Tableau 5.3 – Orientations sexuelles dans RedditGender

Catégorie	Effectifs	%
Hétérosexuelle	468	44.83 %
Gay	104	9.96 %
Lesbienne	85	8.14 %
Bisexuelle	73	6.99 %
Asexuelle	19	1.82 %
Pansexuelle	14	1.34 %
Inconnue	281	26.92 %

5.1.5 Pays

De la même façon que l'âge, le genre ou l'orientation sexuelle, le pays des Redditors a été obtenu soit dans leurs flairs, soit dans leurs commentaires. Cette variable comporte dans certains cas une part d'ambiguïté. Il est parfois difficile de savoir si le pays indiqué par une personne est celui dans lequel elle vit, ou dont elle est originaire. Les Redditors de RedditGender proviennent de 32 pays différents : États-Unis, Canada, Royaume-Uni, Australie, Nouvelle-Zélande, Allemagne, Pays-Bas, Norvège, Afrique du Sud, Pologne, Finlande, Turquie, Inde, Autriche, Italie, Russie, Argentine, Chili, Danemark, Singapour, Égypte, Espagne, Brésil, Pakistan, Irlande, les Philippines, Suède, République tchèque, Portugal, Chine, Kenya et Mexique.

Étant donné le nombre important de pays représentés dans le corpus, et le fait que la majorité des Redditors de RedditGender proviennent d'une poignée de pays, nous avons recodé cette variable afin de réduire le nombre de catégories. Le codage retenu comporte cinq niveaux : États-Unis, Canada, Royaume-Uni, pays anglophones et pays non anglophones (présentés plus haut, dans le tableau 4.3, p. 119). Les pays anglophones sont les autres pays de l'Anglosphère, où « la langue et les valeurs culturelles anglaises prédominent », (« Anglosphere », p. d.), ici, l'Australie, la Nouvelle-Zélande, l'Irlande, et l'Afrique du Sud. Les pays non anglophones incluent des pays où l'anglais est langue officielle comme le Pakistan, le Kenya et l'Inde. Même si 78.83 % des Redditors de RedditGender habitent ou viennent des États-Unis, nous n'avons pas, lors du recueil des données, privilégié les internautes américain·es au détriment de ceux et celles issus d'autres pays anglophones ou non anglophones. Ce chiffre semble refléter le fait que les Redditors américain·es sont majoritaires sur le site et constituent sans

doute une part importante des utilisateur·trices assidu·es, qui produisent de nombreux commentaires, et que nous avons retenus dans notre corpus. Nous n'avons pas utilisé la variable « pays » dans cette thèse, mais elle pourra faire l'objet d'analyses dans l'avenir.

5.1.6 Catégories socioprofessionnelles

Les informations relevant des catégories socioprofessionnelles ont été codées, car leur forme brute ne se prête pas à une analyse statistique. Il s'agit en effet d'indications comme « middle-school teacher », « paralegal », « stay-at-home-mom », « student », « self employed artist », ou « corporate job », qui renvoient à une grande diversité d'activités et de professions. Pour coder ces données, nous avons utilisé l'ISCO, ou « International Standard Classification of Occupations », une classification internationale produite par l'Organisation internationale du travail (« Resolution Concerning Updating the International Standard Classification of Occupations », p. d.), qui divise les professions en 10 catégories. Le codage réalisé, que nous n'avons pas utilisé pour cette thèse mais qui pourra faire l'objet de futures analyses, est présenté en annexe (→ p. 389).

5.1.7 Échantillons utilisés pour les analyses statistiques

À cause des spécificités de notre corpus (→ p. 118), nous avons dû dissocier la non-conformité de genre de l'ethnicité, et réaliser nos analyses sur deux échantillons différents. En effet, nous n'avons réussi à déterminer l'ethnicité des personnes transgenres que dans de rares cas, comme le montre le tableau 5.4.

Tableau 5.4 – Ethnicité des personnes transgenres et non binaires

	FTM	MTF	NB	Total	%
Blancs	20	33	22	75	25.00 %
Afro-américain·es	3	0	0	3	1.00 %
Asiatiques	3	0	3	6	2.00 %
Hispaniques	0	1	0	1	0.33 %
« Autres »	3	2	1	6	2.00 %
Inconnus	71	64	74	209	69.67 %
Total	100	100	100	300	100 %

Pour l'analyse du genre et de la non-conformité de genre, nous avons utilisé le corpus entier, soit les contributions de 1044 personnes, dont nous connaissons pour chacune la catégorie d'âge et l'identité de genre. Pour l'analyse de l'interaction entre ethnicité et genre, nous avons utilisé un échantillon réduit de 347 personnes cisgenres, pour lesquelles nous connaissons la catégorie d'âge, l'identité de genre, et l'ethnicité. La composition de cet échantillon est présentée dans les tableaux 5.5 (pour l'ethnicité) et 5.6 (pour le genre).

Tableau 5.5 – Composition ethnique de l'échantillon réduit utilisé pour étudier l'interaction entre genre et ethnicité

	Hommes cisgenres	Femmes cisgenres	Total
Blanc·hes	72 (40.22 %)	56 (33.33 %)	128 (36.89 %)
Afro-américain·es	38 (21.23 %)	51 (30.36 %)	89 (25.65 %)
Asiatiques	33 (18.44 %)	30 (17.86 %)	63 (18.16 %)
Hispaniques	36 (20.11 %)	31 (18.45 %)	67 (19.31 %)
Total	179 (100 %)	168 (100 %)	347 (100 %)

Tableau 5.6 – Âge des Redditors de l'échantillon réduit utilisé pour étudier l'interaction du genre et de l'ethnicité

	Hommes cisgenres	Femmes cisgenres	Total
14-20 ans	27 (15.08 %)	17 (10.12 %)	44 (12.68 %)
21-30 ans	84 (46.93 %)	102 (60.71 %)	186 (53.6 %)
31 et +	68 (37.99 %)	49 (29.17 %)	117 (33.72 %)
Total	179 (100 %)	168 (100 %)	347 (100 %)

5.2 Les variables de la « Reddidentité »

Cette catégorie de variables renvoie à l'identité des internautes sur Reddit, que nous avons appelée « Reddidentité ». Les données ont été recueillies sur les pages de profil des Redditors. Même si celles-ci sont assez sommaires, on y trouve des informations qu'il nous paraissait pertinent d'étudier.

5.2.1 L'âge Reddit

« L'âge Reddit » indique le nombre d'années depuis lequel un compte Reddit a été créé. Il nous a semblé intéressant d'inclure cette variable à la thèse, car elle peut être considérée comme un indicateur de la familiarité d'un internaute avec la plateforme, et peut avoir un impact sur ses pratiques linguistiques. Dans notre corpus, cette variable a une étendue comprise entre 1 mois et 11 ans. Elle a été codée en cinq niveaux :

— Catégorie 1 : comptes qui ont été créés moins d'un an avant le recueil des données
— Catégorie 2 : âge Reddit de plus d'un 1 an et de moins de 2 ans
— Catégorie 3 : âge Reddit de 2 et 3 ans
— Catégorie 4 : âge Reddit de 4 et 5 ans
— Catégorie 5 : comptes qui ont été créés plus de 6 ans avant la récolte des données

La répartition des Redditors dans les différents niveaux est présentée dans le tableau 5.7. On y voit que, même si l'âge Reddit maximal est de 11 ans, l'immense majorité des comptes inclus dans notre corpus ont été créés cinq ans tout au plus avant la construction du corpus.

Tableau 5.7 – Âges Reddit dans le corpus

Catégories	Effectifs	%
Cat. 1 (- 1 an)	184	17.62 %
Cat. 2 (+ d'1 an, - de 2 ans)	204	19.54 %
Cat. 3 (2 et 3 ans)	302	28.93 %
Cat. 4 (4 et 5 ans)	278	26.63 %
Cat. 5 (6 ans et +)	76	7.28 %

5.2.2 Karma de post et commentaire

Les scores de karma de post et de commentaire ont été relevés sur les profils des Redditors. Ils ont été traités comme des variables numériques, sans codage. Ces scores vont de 1 à 101 687 pour le karma de post, et de 98 à 614 391 pour le karma de commentaire. Nous avons choisi d'étudier les scores de karma des Redditors parce qu'ils permettent de mesurer leur statut et leur activité sur le site (→ p. 91).

5.2.3 Modération

Nous avons noté, pour chaque Redditor, s'il ou elle modère un ou plusieurs forums (voir p. 93). La variable « Modération » comporte deux niveaux : oui/non. Le corpus compte 167 modérateur·trices, soit 16 % des Redditors de RedditGender. Nous avons également pris note du nombre de subreddits modérés par chaque modérateur·trice. Nous avons ensuite classé les modérateur·trices en trois catégories, en fonction du nombre de forums qu'ils et elles modèrent. Le classement, avec le nombre de personnes dans chaque catégorie et la proportion occupée par chaque catégorie dans l'ensemble du corpus, est présenté dans le tableau 5.8.

Tableau 5.8 – Modérateurs dans RedditGender

Catégorie	Nb de subreddits	Effectifs	%
Cat. 1	1	114	10.92 %
Cat. 2	2-3	35	3.35 %
Cat. 3	4 et +	18	1.72 %

5.2.4 Pseudonymes

Les pseudonymes (*usernames*) des Redditors ont été recueillis lors de la construction du corpus. Dans un premier temps, nous avons dû décider si le nom d'un·e internaute indique un genre, et si oui, lequel. Nous n'avons retenu que les références explicites au genre :

— Prénoms masculins ou féminins. Exemples [1] : DiggaDoug492, GeoffdeRuiter, James_Brassmoney, stephaniemac19, ThereseMercado. Les

1. Les pseudonymes donnés en exemple ont été trouvés sur Reddit, mais ne sont pas ceux des Redditors de notre corpus.

prénoms épicènes ont été classés comme des noms « agenre ».
— Noms communs genrés, termes d'adresse ou titres honorifiques. Exemples : jesuislanana, motherofdoves, thriftybabygurl, MrWhiskey1998, kingring1.
— Noms de figures historiques ou de personnages de fiction (littérature, jeux vidéo, anime/manga, mythologie, cinéma, etc.). Exemples : venus019, Zeus, jonSnow.
— Référence à une identité transgenre dans un pseudonyme, avec, par exemple, l'utilisation de termes comme *trans*, textitAFAB, textitftm, textitgenderfluid, etc. Exemples (fictifs) : transdude89, IAmgenderfluid, FabulousAMAB.

Le processus de codage a été long et compliqué, pour plusieurs raisons. Les Redditors aiment jouer avec les mots et font preuve d'une grande créativité pour trouver leurs pseudonymes. Ceux-ci comportent souvent des références, parfois obscures, à la culture populaire ou à la culture geek, ou des mots d'autres langues que l'anglais. Par ailleurs, le format des pseudonymes (→ p. 90) complique leur déchiffrage : les noms d'utilisateur·trices sont toujours composés d'une chaine de caractères, sans espace, et il est parfois difficile de savoir où un mot se termine et où un mot commence.

Pour toutes ces raisons, nous avons procédé à trois codages successifs, en recherchant minutieusement sur internet les termes qui nous semblaient obscurs, pour ne pas passer à côté de références culturelles que nous ne connaissions pas. Un quatrième codage a été effectué par un maitre de conférences de l'université de Lorraine qui connait bien Reddit. Le taux d'agrément brut avec notre codage est de 88.34 %. Les résultats du codage sont présentés dans le chapitre 7 (→ p. 176).

5.3 Thèmes des subreddits

5.3.1 Les données brutes

Une liste des 7733 forums dans lesquels les Redditors de RedditGender ont commenté, accompagnée du nombre de commentaires publiés dans chaque forum, a été générée par Bertrand Gaiffe. La figure 5.2 présente une capture d'écran des données brutes obtenues. Chaque ligne correspond à un·e Redditor, c'est-à-dire à un sous-corpus. Chacune des 7733 colonnes correspond à un forum. Les observations indiquent le nombre de commentaires mis en ligne dans chaque forum par chaque personne. Il y a de nombreux zéros, parce que chaque Redditor a commenté dans un nombre limité des 7733 forums. On peut par exemple voir que l'internaute F_004 a écrit 20 commentaires dans le subreddit r/AdviceAnimals.

5.3.2 Le codage par thèmes

La liste des forums générée a été codée manuellement de façon à indiquer les thèmes des forums. Ce codage a nécessité une exploration manuelle de chaque forum. Se fier uniquement aux noms des subreddits pour en dé-

	1	2	3	4	5	6	7	8	9
1	id	advertising	AdviceAnimal	Advice	AE86	Aerials	afcwimbledor	Affairs	afinil
2	F_001_1703	0	0	0	0	0	0	0	0
3	F_002_1703	0	0	0	0	0	0	0	0
4	F_003_1703	0	0	0	0	0	0	0	0
5	F_004_1703	0	20	0	0	0	0	0	0
6	F_005_1703	0	0	0	0	0	0	0	0
7	F_006_1703	0	0	0	0	0	0	0	0
8	F_007_1703	0	0	0	0	0	0	0	0
9	F_008_1703	0	0	0	0	0	0	0	0
10	F_009_1703	0	0	0	0	0	0	0	0
11	F_010_1703	0	0	0	0	0	0	0	0
12	F_011_1703	0	0	0	0	0	0	0	0
13	F_012_1703	0	0	0	0	0	0	0	0
14	F_013_1703	0	0	0	0	0	0	0	0
15	F_014_1703	0	0	0	0	0	0	0	0
16	F_015_1703	0	0	0	0	0	0	0	0

FIGURE 5.2 – Extrait du jeu de données généré par l'ingénieur de recherche

terminer le thème n'était pas envisageable : les noms des subreddits ne sont en effet pas toujours transparents, et sont même parfois trompeurs. Nous avons tout d'abord classé les forums de façon assez intuitive, en créant des catégories selon les thèmes découverts dans les forums de RedditGender. Ce premier codage a abouti à plus de 60 catégories, dont « humour », « jobs », « livres », « musique », « voyage », ou encore « violence ». Nous avons souhaité réduire le nombre de catégories, afin que celles-ci soient relativement équilibrées et se prêtent plus facilement aux analyses statistiques. Pour ce faire, nous avons pris comme référence la nomenclature des subreddits de Thelwall et Stuart (2018), qui se sont intéressé·es aux 100 subreddits les plus actifs, et l'avons adaptée aux forums de RedditGender.

Ce travail a été délicat, parce qu'il impliquait de faire des choix qui ont un impact sur les analyses. Certaines décisions ont été faciles à prendre : regrouper les forums ayant pour thème l'actualité et la politique, par exemple, semble naturel. D'autres catégories sont toutefois plus hétérogènes ; nous avons ainsi réuni dans la catégorie « x-rated » la pornographie, la violence, les drogues et les armes. Les jeux vidéo constituent à eux seuls une catégorie, tandis que la littérature, la musique, le cinéma, la bande dessinée et la radio ont été placés dans la catégorie « mass entertainment ». La figure 5.3 présente une capture d'écran d'un extrait du codage réalisé.

	1	2
146	adventuretime	gaming
147	advertising	edu_science
148	AdviceAnimals	memes_humor
149	Advice	general
150	AE86	hobbies
151	Aerials	sports_fitness
152	afcwimbledon	sports_fitness
153	Affairs	personal_advice
154	afinil	personal_advice
155	aflmemes	memes_humor
156	AFL	sports_fitness
157	AFOL	gaming
158	AfricanAmerican	personal_advice

FIGURE 5.3 – Capture d'écran présentant le codage du thème de chaque forum, extrait du jeu de données

Le tableau 5.9 montre les 12 catégories créées et le nombre de forums

dans chaque catégorie ainsi que quelques exemples. Nous avons également indiqué les noms des catégories en anglais, par souci de cohérence : le codage du jeu de données a été réalisé en anglais, et le graphe d'analyse factorielle des correspondances simples analysé dans le chapitre 8 (→ p. 201) présente les noms des catégories en anglais.

5.4 Les variables linguistiques

Nous avons classé les variables linguistiques en deux catégories : les procédés d'ajout et les procédés de réduction. Ces catégories sont présentées ci-dessous, avec des exemples tirés du corpus. Nous les avons paraphrasés afin de ne pas permettre l'identification des internautes, en conservant les éléments d'intérêt.

Procédés d'ajout

Les procédés d'ajout consistent à insérer plus de caractères dans les messages que n'en nécessiterait l'utilisation de l'orthographe standard de l'anglais (étirements de lettres ou de ponctuation), à utiliser des phénomènes paralinguistiques (émoticônes, émojis, interjections), ou à utiliser des majuscules de façon excessive (*all caps* ou mots en majuscules). Des exemples sont présentés dans le tableau 5.10.

Procédés de réduction

Ces phénomènes répondent au besoin d'économie qui caractérise souvent la CMC, et donnent une impression de minimalisme. Ce sont les abréviations (acronymes et réductions), les omissions d'apostrophe, les graphies phonétiques, les g-droppings (→ p. 82) et les omissions de majuscules. Dans ce dernier cas, pour faciliter le recueil des données, nous nous sommes limitée à l'omission de la majuscule de la première personne du singulier *I*. Des exemples de chacun de ces procédés sont présentés dans le tableau 5.11 ; les éléments d'intérêt ont été mis en gras. Acronymes et réductions sont ici présentés séparément, mais nous les avons analysés ensemble.

5.4.1 Méthodes d'extraction des données

Pour recueillir les données linguistiques, nous avons fait appel à trois méthodes différentes.

Utilisation d'expressions régulières

Le logiciel TXM utilise le langage de requête CQL (→ p. 125). Celui-ci est basé sur des expressions régulières, qui sont utilisées en traitement automatique des langues et permettent, par exemple, de rechercher tous les mots commençant par une lettre donnée, faisant un certain nombre de caractères, ou appartenant à une catégorie grammaticale donnée. Nous avons utilisé les expressions régulières quand cela était judicieux et en fonction de

Tableau 5.9 – Catégories thématiques des subreddits

Thème	Nombre de subreddits	Exemples
Forums généraux (*general interest*) : thèmes généraux, images, animaux, régions	1275	r/Advice, r/AskReddit, r/AskNYC, r/CasualConversation, r/ImagesofCalifornia, r/motivation
Humour (*humor*) : mèmes, humour, et subreddits « circlejerks »	476	r/oldpeoplefacebook, r/shittyadvice, r/EbayWTF, r/WastedGifs, r/Hiphopcirclejerk
Jeux (*gaming*) : jeux de société et jeux vidéos	869	r/FreeGamesOnSteam, r/gaming, r/gamedev, r/gtaonline, r/hearthstone
Actualité, politique et religion (*news, politics and religion*)	383	r/hillaryclinton, r/immigration, r/MarchAgainstTrump, r/NewPatriotism, r/syriancivilwar, r/PhilosophyofReligion
Éducation et science (*education and science*) : université, emploi, sciences, finance, business	631	r/TranslationStudies, r/college, r/Archeology, r/math, r/Harvard, r/AskLiteraryStudies
Technologie (*technology*) : internet, informatique, programmation	419	r/microsoft, r/Nexus7, r/Python, r/smarthome, r/TechnologyProTips, r/SQL
Divertissement (*mass entertainment*) : musique, radio, télévision, cinéma, manga	1124	r/Concerts, r/KingCrimsons, r/NPR, r/visualnovels, r/disney, r/criminalminds
Loisirs (*hobbies*) : hobbies, cuisine, shopping	663	r/DIY, r/Fiat, r/FoodPorn, r/Gin, r/gardening
Thèmes personnels (*personal advice*) : mode, beauté, santé, relations familiales et amoureuses	793	r/interracialdating, r/relationships, r/LatinaBeauties, r/lonely, r/LushCosmetics, r/malehairadvice
Sport (*sports and fitness*) :	328	r/FitToFat, r/Gymnastics, r/olympics, r/SFGiants, r/xxketo
Classés « X » (*x-rated*) : pornographie, violence, armes et drogues	479	r/Drugs, r/MorbidInterests, r/progun, r/SexyHalfAsians, r/malegonewild
Divers (*others*) : subreddits qui ne peuvent pas être classés dans les catégories ci-dessus	236	r/FuckChuck, r/GTAorRussia, r/ZombiesSurvivalTactics, r/ButtonAftermath

Tableau 5.10 – Exemples de procédés d'ajout

Émoticônes	Thanks :) I work close to it. I will miss going there :(Yeah, being tall is definitely a struggle for me. :\ That d be pretty cool. :-) < 3
Émojis	I've been thinking about burgers all day i'm like 🍔 🍔 🍔 🍔 🍔 not funny 😝 What is it? 🍔
Étirements de lettres	buuuuut I LOLed. Saaameeeee Daaaaaamn guuuuuurl Idk what you mean you look awesome. New York style pizza with a side of balsamic dressing. Yaaaassss.
Étirements de ponctuation	It is so good!!!!!!! Holy fuck! 120%?!?! Congrats!!! :D !!!!!!!!!!!!!!!!!!!!!!!!!!!!
Mots en majuscules	Omg you don't fucking have a clue. YOU DON'T KNOW WHAT I DID TO SUPPORT HIM!!!!! SHE IS MY SPIRIT ANIMAL! I always get excited when I meet people like that. < 3 What REALLY helped me, is cutting calories at breakfast. DISCUSSION IS OVER! GO HOME!!! ALL OF YOU.
Interjections	**Duh**. I guess I didn't pay attention during history class. **Shh, shh**, someone who has never been on hrt knows better. **ugh**, yes. that sucks. **Oh wow**, I didn't know what clafoutis was. Now wanna make it!

TABLEAU 5.11 – Exemples de procédés de réduction

ACRONYMES	**Idk** about you, but this is on point for me Dunno if I'll be there soon, but I'll **PM** you. I'm 28 **btw**. I don't know what you're on about **OP** was being respectful. I thought you had a strong southern accent since you're from **TN**. My **SO** is from Chattanooga and her accent is heeeavy !
RÉDUCTIONS	My **gf** is a teacher. Am seriously considering it**bc** it gets good gas mileage. But also **bc** cute :P As a native new yorker, I call **bs**. **srs** me too I can't stand him
OMISSIONS D'APOSTROPHE	**im** going to tell them **dont** worry, I **cant** :-(**its** not fair
I MINUSCULE	Hey dude, **i**'m just giving my 2 cents. **i**'ll have 2-5 beers when **i** go out with my friends. **it**'s fine, since **i** don't have a car anyway
GRAPHIES PHONÉTIQUES	My SO is **gonna** freak. Actually I am European myself (UK, so **kinda**) but there are still issues. You **gotta** try. Hugs. I feel **ya**.
G-DROPPINGS	How you **doin** ? I **effin** hate lice. she's the only player I like, she's a **freakin** legend. I'm real **fuckin'** excited ! ! !

nos compétences en la matière : par exemple, pour rechercher tous les étirements de lettres, les mots en majuscules ou les g-droppings présents dans le corpus. Cette méthode a l'avantage d'être rapide : une fois l'expression régulière adéquate identifiée, il suffit d'une requête dans TXM pour obtenir la liste des tokens d'intérêt. Le problème, c'est qu'elle peut dans certains cas générer des faux positifs, notamment des homographes des tokens recherchés. Dans ce cas, les données obtenues ont été nettoyées manuellement ou automatiquement, à l'aide de listes.

Utilisation de listes

Les éléments typiques du Netspeak sont caractérisés par une forte variation orthographique, que les expressions régulières ne peuvent pas toujours capturer. Il est donc commun de rechercher dans un corpus toutes les occurrences d'éléments figurant sur des listes préétablies. Nous avons par exemple adopté cette approche pour rechercher les émoticônes et les interjections présentes dans le corpus. Nous l'avons également utilisée pour « nettoyer » les résultats obtenus avec les expressions régulières, notamment dans le cas des mots en majuscules. L'inconvénient majeur de cette méthode, dans notre cas, est que TXM ne permet de rechercher simultanément que 4 éléments au maximum. Plus une liste est longue, plus il faut multiplier les requêtes, ce qui rend l'opération fastidieuse et augmente le risque d'erreur.

Examen manuel du corpus

Pour rechercher certains phénomènes, il nous a paru judicieux de partir du contenu du corpus plutôt que de listes préexistantes, comme nous l'avons expliqué plus haut (→ p. 126). La variation orthographique de la CMC est très importante, et dépend du type de plateforme utilisé. Il nous a donc semblé que les listes établies à partir d'autres corpus ne pourraient pas forcément capturer les graphies utilisées sur Reddit. Nous avons, à la place, utilisé la fonction « Lexique » de TXM pour générer la liste de toutes les unités lexicales contenues dans le corpus, puis avons parcouru cette liste pour trouver les éléments qui nous intéressaient. Nous avons utilisé cette méthode pour rechercher les omissions d'apostrophe, les graphies phonétiques, les réductions et les acronymes. Cette méthode a pour inconvénient d'être fastidieuse, à cause du nombre de « types » (mots différents) présents dans le corpus (227 808). Nous avons donc limité notre recherche aux types les plus fréquents, en fixant un seuil de fréquence (habituellement 100).

Avantages et inconvénients de notre méthodologie

Cette méthodologie mixte a offert un bon compromis entre facilité d'utilisation, rapidité de recherche et exigences techniques. Il faut toutefois noter que les données recueillies à l'aide de listes ou d'un examen manuel du corpus n'ont pas la prétention à l'exhaustivité, contrairement à celles obtenues avec les expressions régulières.

5.4.2 Procédés d'ajout

Émoticônes

Pour identifier les émoticônes présentes dans le corpus, il a été nécessaire de procéder à une recherche manuelle. En effet, TXM ne considère pas les suites de signes de ponctuation comme des unités lexicales. Les émoticônes n'apparaissent dans le « Lexique » du corpus que sous forme de fragments (un point, une virgule, un chiffre, etc.). Nous avons donc dû utiliser une liste d'émoticônes pour les repérer ; nous avons choisi la liste d'émoticônes de Wikipedia (« List of emoticons », 2020), présentée en annexe (→ p. 390. Pour des raisons pratiques, nous nous sommes limitée à la partie « Western », c'est-à-dire aux émoticônes occidentales, qui se lisent de gauche à droite, et n'avons pas inclus les émoticônes dites « orientales » (*eastern)*, qui se lisent de bas en haut. Voici deux exemples de requêtes utilisées pour trouver les émoticônes à l'aide de TXM :

— **:-)** → `[word=":"] [word="-"] [word="\)"]`
— **:'(** → `[word=":"] [word="'"] [word="\("]`

Dans plusieurs cas, il a fallu trier les résultats manuellement afin d'enlever les faux positifs. Par exemple, la suite de caractères **:3** peut représenter l'émoticône **:3** (un visage souriant) ou le deux-points suivi d'un chiffre 3. **XP** peut représenter un visage tirant la langue, ou être la réduction de « Experience » utilisée par les joueurs de Runescape. Soixante-dix émoticônes sur les 129 figurant dans la liste de Wikipédia ont été identifiées dans le corpus.

Émojis

TXM ne reconnait pas les émojis, qui ont disparu lorsque nous avons testé la première version du corpus. Pour pallier ce problème, l'ingénieur de recherche a dû les encoder. Il les a identifiés en utilisant leur code Unicode (par exemple, U+1F602 pour l'émoji 😂). Nous avons assigné aux émojis le caractère §, qui n'était pas présent dans le corpus, de façon à pouvoir facilement effectuer une recherche globale à l'aide de TXM : en entrant simplement « § » dans la barre de recherche, TXM génère les concordances correspondant à tous les acronymes du corpus. Bertrand Gaiffe a également ajouté à l'encodage les noms des émojis en anglais (« Face With Tears of Joy » pour 😂 , « Loudly Crying Face » pour 😭 , etc.), afin d'identifier les différents types d'émojis présents dans le corpus. Pour connaitre, par exemple, le nombre d'émojis « Angry Face », ou visage en colère, il a donc fallu utiliser l'expression `<c_unicode=".*ANGRY.*">"§"<c_unicode=".*FACE.*">"§"` dans l'outil « Concordance » de TXM.

Étirements de lettres

Grâce au moteur de recherche CQP de TXM, nous avons recherché tous les cas où une lettre était répétée au moins 3 fois et jusqu'à 20 fois dans un même token. La syntaxe CQL utilisée est la suivante : `".*a{3,20}.*"%c`. Nous avons décliné cette recherche pour chaque lettre de l'alphabet. Pour

identifier les types les plus fréquents, nous avons traité les résultats des concordances avec R, en mettant tous les tokens en lettres minuscules avec la fonction `tolower()`, puis en retirant les lettres répétées avec la fonction `gsub()`.

Étirements de signes de ponctuation

Pour repérer les répétitions de signes de ponctuation, nous avons utilisé la requête CQL suivante : `[word="\?"] [word="\?"] |[word ="\!"] [word="\!"] |[word="\..."][word="\."]`. Elle permet de rechercher dans le corpus toute succession d'au moins 2 points d'interrogation, ou de 2 points d'exclamation, ou de 4 points. Cette liste a ensuite été nettoyée manuellement, pour éviter les répétitions. En effet, par exemple, quand un point d'exclamation est répété 7 fois de suite, TXM a généré 7 concordances. Dans ce cas, nous avons retiré les 6 concordances superflues.

Mots en majuscules

Pour trouver les mots écrits en majuscules, nous avons utilisé l'expression régulière suivante dans TXM : `[word="[A-Z]{2,}"]`. Elle a permis de trouver tous les tokens d'au moins deux lettres, et composés uniquement de lettres en majuscules. Cette requête a généré 138 327 résultats. La liste obtenue contenait de nombreux acronymes et réductions (*AA*, *LGBT*, *AC*, *XL*, *WWII*, *USB*, etc.). Elle a donc dû être « nettoyée » dans R.

Pour ce faire, nous avons constitué une liste d'acronymes, à partir de sites web et de fichiers trouvés sur internet : acronymes du Netspeak (« Acronym definition & 3000+ acronyms list from a-z », 2019 ; « pi's Yet Another Bloody List of Acronyms », p. d.), liste d'acronymes établie par Wikipedia (« Crawl Wiki for Acronyms », p. d.), acronymes « transgenres » (« Transgender acronyms », 2019), et acronymes du web et de la technologie (« The Hackers Acronym Chart », 2000 ; « RFC Editor Abbreviations List », 2020). La liste obtenue contient environ 17 000 acronymes. Nous avons utilisé la fonction `setdiff()` de R pour supprimer automatiquement les acronymes figurant dans les concordances extraites de TXM. Après cette opération, de nombreux acronymes subsistaient dans notre liste de tokens. Nous avons parcouru manuellement les 3000 tokens les plus fréquents, et avons retiré les acronymes identifiés. Faire cette opération pour l'ensemble des types en majuscules (environ 10 000) aurait été extrêmement fastidieux. La plupart n'apparaissant qu'une seule fois, nous nous sommes contentée de la liste ainsi obtenue : la majorité des tokens sont des mots « de dictionnaire » ou des graphies non standard, et une minorité sont des acronymes, ce qui nous a paru satisfaisant.

Interjections

Le corpus a été annoté avec Tree Tagger, qui propose une catégorie « interjections ». Nous n'avons toutefois pas utilisé cette annotation pour

rechercher les interjections dans le corpus. En effet, la catégorie « interjections » de Tree Tagger comporte des interjections secondaires comme *yes*, *please* ou *indeed*, et peu d'interjections primaires (p. → 77), comme *ah*, *ugh* ou *aw*, que nous souhaitions étudier. Nous avons, à la place, effectué des recherches ciblées dans le corpus à partir de la liste d'interjections `pos_interjections` du package R `lexicon` (Rinker, 2018), qui contient des listes de mots permettant de réaliser des analyses textuelles. Cette liste, qui comprend 140 éléments, est présentée en annexe A.2.

Nous avons écarté certains éléments de cette liste, à cause des nombreux cas d'ambiguïté observés : *boo*, qui, dans RedditGender, est surtout utilisé dans le sens de « partenaire » (« Definition of BOO », p. d.), *mm*, qui est notamment une abréviation de *millimeter*, ou encore *sis*, clipping de *sister*. Nous avons également retiré les étirements graphiques, comportant une succession de trois lettres identiques, car nous les avons analysés dans une catégorie distincte. La liste proposée par *lexicon* n'est pas exhaustive, évidemment, mais elle nous a paru relativement complète et variée.

5.4.3 Procédés de réduction

Acronymes

La fonction « Index » de TXM a permis d'afficher une liste de tous les « types » présents dans le corpus. Ensuite, un relevé manuel de tous les acronymes apparaissant au moins 100 fois dans le corpus a été effectué, sans distinction. Cette étape a été réalisée deux fois, afin de limiter le risque d'oubli. Dans un second temps, il a fallu décider quels acronymes relèvent du Netspeak, et quels acronymes ne sont pas spécifiques à la langue du web. La distinction entre ces deux types d'acronymes n'était pas évidente. Afin d'y voir plus clair, nous avons classé les acronymes en plusieurs catégories. Voici les grands thèmes qui se sont dégagés, avec des exemples issus du corpus :

- Acronymes médicaux : *AHDH* (*attention deficit hyperactivity disorder*), *GP* (*general practitioner*, *ASD* (*autism spectrum disorder*), *IUD* (*intrauterine device*).
- Acronymes politiques : *GOP* (*Grand Old Party*, le parti républicain), *BLM* (*Black Lives Matter*), *DNS* (*Democratic National Committee*).
- Lieux : *SF* (*San Francisco*), *NY* (New York), *NZ* (*New Zealand*)
- Acronymes de la vie de tous les jours : *BBQ* (*barbecue*)
- Acronymes faisant référence à des personnes : *MIL* (*mother in law*), *SO* (*significant other*)
- Acronymes de la vie des personnes transgenres et non binaires : *afab* (*assigned female at birth*), *NB* (*non binary*), *HRT* (*hormone replacement therapy*)
- Acronymes techniques : *USB* (*Universal Serial Bus*), *DVD* (*Digital Versatile Disc*)
- Acronymes du Netspeak : *OMG* (*oh my god*), *idk* (*I don't know*), *lol* (*laughing out loud*)

— Acronymes de Reddit, généralement dérivés de noms de subreddits : *TIL* (*Today I Learned*)

Le tri des acronymes a été compliqué par l'ambiguïté qui leur est propre. Un acronyme peut avoir plusieurs sources (Mattiello, 2013), ou peut avoir comme homographes un ou plusieurs mots «de dictionnaire. Il a été possible de traiter les cas d'homographie par un tri manuel des concordances, en s'appuyant sur le contexte pour inférer le sens de l'acronyme. Voici quelques exemples d'homographie notables :

— *PM* : *private message*, *Prime Minister*, *post meridiem*, ou encore *project manager*
— *rn* : *registered nurse*, *right now*
— *ATM* : *at the moment* ; *automated teller machine*
— *ED* : *eating disorder* ; *education*
— *TIL* : acronyme de *Today I Learned* (nom d'un subreddit) ; réduction de *until*
— *SO* : acronyme de *significant other* ; adverbe ou conjonction *so*

Le cas de *SO* était particulièrement problématique, étant donné la fréquence élevée du token dans le corpus (plus de 91 000 occurrences). Pour éviter les faux positifs, nous avons uniquement recherché les cas où *SO* apparait en lettres capitales, puis nous avons inspecté les concordances manuellement pour enlever celles qui contenaient l'adverbe.

L'objectif de cette thèse étant d'étudier les éléments lexicaux typiques de l'anglais d'internet, nous avons uniquement analysé les acronymes référant à des expressions ou *phrases*, à des personnes, ainsi que le seul acronyme spécifique à Reddit ayant une fréquence supérieure à 100 (*TIL*). La plupart des acronymes cités dans la liste ci-dessus, qui font référence à des lieux, des termes techniques, médicaux, ou qui sont spécifiques à une population, comme les termes « transgenres » ont ainsi été exclus de l'analyse. Nous avons également pris en compte les différentes variantes typographiques des acronymes (par exemple *LOL*, *lol*, *Lol*, etc). Nous ne l'avons pas fait pour l'acronyme *SO*, pour les raisons pratiques expliquées ci-dessus.

Réductions

Les réductions ont été identifiées de la même façon que les acronymes, par l'examen de la liste de tous les types du corpus. Seules les réductions ayant une fréquence supérieure à 100 ont été incluses dans l'étude, afin de faciliter le recueil des données. Les formes complètes de ces abréviations ont également été extraites du corpus, pour comparaison.

Graphies phonétiques

Les graphies phonétiques ont été identifiées par l'examen de la liste des types contenus dans le corpus grâce à la fonction « Lexique » de TXM. L'opération étant chronophage, nous avons inclus uniquement les éléments ayant une fréquence supérieure à 100. Nous avons également généré des concordances des formes standard correspondantes (*want to* pour *wanna*,

par exemple). Les graphies standard correspondant aux graphies phonétiques ont également été extraites du corpus.

G-droppings

Les g-droppings du corpus ont été identifiés à l'aide de l'expression régulière `[word="..*in"%c]`, qui a permis de rechercher tous les tokens composés de trois lettres au minimum et se terminant par *-in*. Les concordances ont été classées par ordre alphabétique, puis triées manuellement afin d'éliminer les tokens qui n'étaient pas des g-droppings. Les g-droppings ont tous été mis en minuscules afin de faciliter leur analyse. Nous avons ensuite recherché les formes standard des g-droppings identifiés.

Omissions d'apostrophe

Les omissions d'apostrophe ont été identifiées de la même façon que les réductions et les graphies phonétiques. Plusieurs cas d'homographie ont nécessité un examen manuel des concordances, pour éliminer les faux positifs. C'était notamment le cas de *its*, qui peut être le résultat de l'omission de l'apostrophe de *it's* ou le pronom possessif *its*. Des 10 518 occurrences de *its*, la majorité (7457) a été identifiée comme étant des omissions d'apostrophes.

Les résultats des concordances de *lets* (omission de l'apostrophe de *let's* ou verbe *let* conjugué à la troisième personne du singulier, de *Id* (*I would/I had* ou *ID*, *identity*), *Im* (*I am* et acronyme d'*instant message* ou d'*intra muscular*) et *Ill* (*I will* et *ill*, malade) ont également été nettoyés manuellement. Nous n'avons en revanche pas trié les résultats des concordances d'autres possibles omissions d'apostrophe, à cause du nombre important d'occurrences de leurs homographes. Cela a été notamment le cas de *well* (fréquence dans le corpus : 22 889), qui peut être le résultat de l'omission de l'apostrophe de *we will*, mais aussi un nom, un adjectif, un adverbe, une interjection ou un verbe. *Were* (fréquence dans le corpus : 26 187) posait un problème similaire, pouvant à la fois être l'omission de l'apostrophe dans *we are*, ou la forme passée du verbe *be* à la deuxième personne. Certaines contractions, qui semblent très rares et qui présentaient également des cas d'ambiguïté importante, n'ont pas été incluses. C'est le cas de *hell* (*he will* ou *enfer*), et *shell* (*she will* ou *carapace*). Nous avons écarté tous ces tokens.

Pour des raisons pratiques, ce relevé ne comprend pas l'omission de l'apostrophe du possessif, comme dans *my friends car* pour *my friend's car*. L'omission de l'apostrophe possessive est en effet difficile à identifier, pouvant être confondue avec le pluriel. Les données analysées ne sont donc pas exhaustives. Une fois la liste des tokens à étudier établie, nous avons généré des concordances pour chacun de leurs équivalents avec apostrophe (*I am* pour *Im*, *it's* pour *its*, *will not* pour *wont*, etc.), afin de pouvoir déterminer si l'omission d'apostrophe se produit à la même fréquence pour chaque forme standard.

Omission de la majuscule du pronom *I*

Toutes les occurrences de *i* minuscule ont été extraites du corpus, ainsi que les occurrences de *I* majuscule, à l'aide d'une simple recherche dans TXM.

tl ;dr

Ce chapitre a présenté trois types de variables :

- Les variables sociales : le genre, l'âge, l'ethnicité et l'orientation sexuelle, que nous étudions dans cette thèse, mais aussi les pays et catégories socioprofessionnelles, qui pourront servir de base à de futures analyses.
- Les variables de la Reddidentité : l'âge Reddit, le karma de post et de commentaire, la modération de forums et les pseudonymes, étudiés dans le chapitre 7.
- Les thèmes des subreddits, étudiés dans le chapitre 8.
- Les variables linguistiques : les procédés « d'ajout » (émoticônes, émojis, étirements de lettres, étirements de ponctuation, mots en majuscules et interjections), étudiés dans le chapitre 10, et les procédés « de réduction » (acronymes, réductions, graphies phonétiques, g-droppings, omissions d'apostrophe et omission de la majuscule du pronom personnel *I*), analysés dans le chapitre 11.

Nous avons montré que, dans certains cas, le recueil des données linguistiques a été simple et direct, mais que dans d'autres, il a nécessité un travail de réflexion et de tri important, dû à l'importante variation orthographique typique de la CMC. Même si nos données ne prétendent pas à l'exhaustivité, elles nous donnent suffisamment de matière pour tenter de répondre à nos questions de recherche.

Chapitre 6

Les méthodes statistiques

Étudier un grand corpus entraine nécessairement l'utilisation de techniques statistiques. Nous avons choisi les techniques les plus adaptées à la démarche intersectionnelle ; il fallait d'une part, dans un souci de rigueur, qu'elles soient adaptées à nos données (de comptage, donc asymétriques), et, de l'autre, qu'elles puissent permettre l'exploration des interactions entre les variables sociodémographiques. Comme certaines de ces méthodes sont encore peu utilisées dans le domaine de la sociolinguistique en France, nous les décrivons en détail dans ce chapitre.

6.1 Statistiques descriptives présentées dans la thèse

Les techniques statistiques peuvent être divisées en deux familles : les méthodes descriptives et les méthodes inférentielles. Les premières servent à décrire les caractéristiques d'un échantillon (ici, tous les Redditors de RedditGender). Les techniques inférentielles permettent de généraliser les résultats obtenus à partir de l'échantillon. Grâce à elles, on peut savoir si une différence constatée par une méthode descriptive (entre deux moyennes, par exemple) est significative, ou si elle est simplement due au hasard (Levshina, 2015). Dans notre thèse, nous présentons généralement des analyses inférentielles, précédées par une exploration descriptive des données, soit à l'aide de mesures, soit à l'aide de boites à moustaches.

6.1.1 Les mesures

Fréquences

Le concept de fréquence est essentiel en linguistique de corpus. Les fréquences peuvent être données sous forme brute : il y a par exemple 9215 occurrences de l'acronyme *lol* dans RedditGender. Quand on veut comparer plusieurs corpus de tailles différentes, il faut normaliser les fréquences, en calculant leur proportion ou leur pourcentage (Baker et al., 2006). Par exemple, dans les sous-corpus correspondant aux productions des hommes

transgenres (1 867 884 tokens) et des hommes cisgenres (6 849 846 tokens), *lol* apparait respectivement 1096 et 3159 fois. La normalisation de ces fréquences permet de savoir que *lol* est plus fréquent chez les hommes transgenres (0.59 occurrence par 1000 tokens) que chez les hommes cisgenres (0.46 occurrence par 1000 tokens).

Mesures de tendance centrale

Les mesures de tendance centrale nous informent sur la valeur la plus typique d'une distribution (Levshina, 2015). Dans nos analyses, nous indiquons deux mesures de tendance centrale : la moyenne et la médiane. La moyenne est la mesure la plus populaire, mais elle a comme défaut d'être très sensible aux valeurs aberrantes. Quand les données ne suivent pas une distribution normale, ce qui est le cas de toutes les analyses réalisées ici, il vaut donc mieux indiquer la médiane, car elle est plus robuste (Gries, 2013). La moyenne et la médiane sont calculées dans R avec les fonctions `mean()` et `median()`.

Mesures de dispersion

Les mesures de tendance centrale ne fournissent qu'une vision partielle des données. Sur la recommandation de Levshina (2015), nous les accompagnons toujours de mesures de dispersion, qui informent sur la variation présente dans les données. Ces mesures permettent de savoir si les internautes utilisent une variable de façon homogène, ou s'il existe des écarts importants. Avec la moyenne, nous indiquons l'écart type, qui est la racine carrée de la variance, et qui est plus fréquemment utilisé que celle-ci (Levshina, 2015). Avec la médiane, nous précisons l'écart interquartile, qui est la différence entre le 1er quartile et le 3ème quartile. C'est une mesure plus robuste que l'écart type, car moins sensible aux valeurs aberrantes. Plus l'écart type ou l'écart interquartile est faible, moins il y a de variation dans les données. Dans R, l'écart type est calculé avec la fonction `sd()`, et l'écart interquartile avec la fonction `IQR()`.

6.1.2 Boites à moustaches

La boite à moustaches, ou *boxplot* en anglais, fournit un résumé graphique des données. C'est un graphique très informatif, qui, dans certains cas, peut se suffire à lui-même (Gries, 2013). La figure 6.1 présente une boite à moustaches créée avec la fonction `boxplot()` de R. Elle présente la distribution de la fréquence du pronom personnel *I* (sous sa forme en majuscule) dans l'ensemble du corpus. La boite centrale contient 50 % des observations. Ici, la moitié des valeurs sont donc comprises entre (environ) 2.2 et 3.9 *I* par 1000 tokens. Cette boite est coupée en deux (pas forcément en son milieu) par la médiane, qui sépare les 50 % inférieurs des données des 50 % supérieurs.

La limite inférieure de la boite représente le 1er quartile, c'est-à-dire la valeur en dessous de laquelle se trouvent 25 % des observations. La

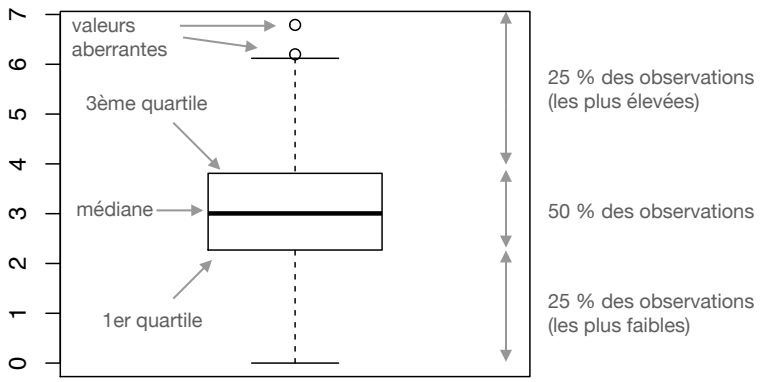

FIGURE 6.1 – Fréquence du *I* majuscule dans le corpus, boite à moustaches créée avec la fonction `boxplot()`

limite supérieure correspond au 3ème quartile, à savoir la valeur en dessous de laquelle se trouvent 75 % des observations. La différence entre le 3ème quartile et le 1er (ici, environ 1.7) indique l'écart interquartile. Les « moustaches », c'est-à-dire les lignes verticales qui dépassent de chaque côté de la boite, ne sont jamais plus longues que l'écart interquartile multiplié par 1.5. Les observations qui excèdent cette valeur sont représentées sous forme de points et peuvent être considérées comme des valeurs aberrantes, c'est-à-dire extrêmement basses ou élevées (Levshina, 2015).

Les boites à moustaches sont particulièrement utiles parce qu'elles permettent de visualiser la dispersion des données et de comparer facilement des groupes. Par exemple, dans la figure 6.2, qui représente la fréquence des émoticônes pour 1000 tokens chez les hommes et les femmes cisgenres, la boite correspondant aux femmes est plus haute que celle des hommes. Cela indique qu'il y a davantage de variations individuelles chez les femmes, alors que, chez les hommes, les données sont plus resserrées autour de la médiane, qui est par ailleurs plus faible que celle des femmes. Cette boite à moustaches a été générée avec le package `ggplot2` (Wickham, 2016). Nous avons utilisé la fonction `geom_jitter()` pour ajouter des points qui correspondent à chacune des observations (donc, ici, à la fréquence relative des émoticônes dans chacun des 1044 sous-corpus de RedditGender). C'est ce type de boite à moustaches qui est généralement présenté dans nos analyses ; dans d'autres cas, nous avons simplement utilisé la fonction `boxplot()`.

6.2 Analyse des corrélations

6.2.1 Diagramme en mosaïque

Le diagramme en mosaïque permet de visualiser la distribution de deux variables catégorielles, et donc d'explorer les corrélations entre ces deux variables. Il est basé sur un tableau de contingence, un tableau à double

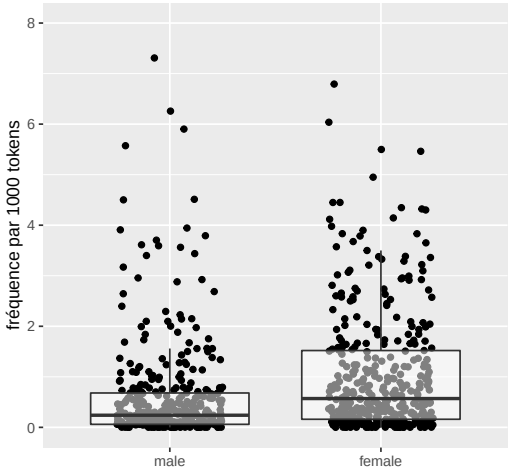

FIGURE 6.2 – Fréquence des émoticônes, par sous-corpus

entrée qui place des données issues de comptage dans deux catégories. Ce type de tableau est très souvent utilisé pour analyser des variables catégorielles, notamment dans le domaine de la linguistique de corpus (Glynn, 2014). Par exemple, le tableau 6.1 détaille la composition du corpus, en présentant le nombre d'individus de chaque groupe de genre dans chaque groupe d'âge (les noms des groupes sont en anglais, par souci de cohérence avec le diagramme en mosaïque présenté plus bas).

TABLEAU 6.1 – Composition de RedditGender, âge et genre

	Male	Female	MTF	FTM	NB
14-20	61	32	13	26	15
21-30	139	202	54	59	63
31 et +	172	138	33	15	22

Le diagramme en mosaïque (figure 6.3) représente les mêmes informations de façon visuelle. La taille de chaque rectangle est proportionnelle aux effectifs présents dans chaque groupe. On voit par exemple que la largeur des colonnes correspondant aux femmes et aux hommes cisgenres est plus importante que celle des autres groupes, parce qu'il y a plus d'hommes et de femmes cisgenres dans le corpus que de personnes transgenres et non binaires. Le rectangle correspondant aux 14 à 20 ans est moins haut chez les femmes cisgenres que chez les hommes cisgenres, parce qu'il y a moins de femmes que d'hommes dans cette catégorie (32 femmes et 61 hommes). Ce diagramme en mosaïque, réalisé avec la fonction `mosaic` du package `vcd` (Meyer et al., 2020), contient également des couleurs. Celles-ci indiquent les résidus de Pearson, calculés en divisant la différence entre les effectifs observés et les effectifs théoriques par la racine carrée des effectifs théoriques (Levshina, 2015). Les gros résidus négatifs sont en rouge, et les gros résidus positifs en bleu. Plus la couleur est foncée, plus la distance entre

les résidus de Pearson et les effectifs théoriques est grande. Ici, dans les rectangles gris, il n'y a pas de déviation significative par rapport aux effectifs théoriques (c'est-à-dire aux nombres d'individus que ces catégories comporteraient s'il y avait, proportionnellement, autant de personnes de chaque catégorie de genre dans chaque catégorie d'âge). Les rectangles colorés indiquent des déviations. On voit ainsi qu'il y a moins de Redditors de 14 à 20 ans chez les femmes cisgenres que dans les autres catégories. Il y a en revanche plus d'hommes transgenres dans cette tranche d'âge, par rapport à l'ensemble des groupes de genre.

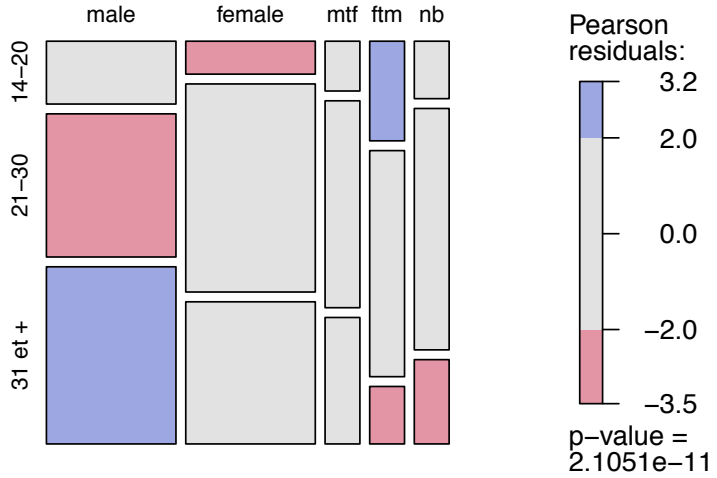

FIGURE 6.3 – Diagramme en mosaïque représentant la composition du corpus par effectifs de genre et d'âge, avec résidus de Pearson

6.2.2 Analyse factorielle des correspondances simples

Comme le diagramme en mosaïque, l'analyse des factorielle des correspondances simples (AFC, ou *correspondence analysis* en anglais) met en lumière les corrélations entre deux variables catégorielles de façon graphique (Salles, 2009). Elle établit donc des liens entre les catégories représentées par les lignes d'un tableau et celles qui sont représentées par ses colonnes (Husson et al., 2011). L'AFC a été mise au point par le mathématicien et statisticien français Jean-Paul Benzécri dans les années 1960 (Yelland, 2010). Initialement développée pour analyser des données linguistiques (Murtagh, 2005), elle est aujourd'hui utilisée dans de nombreuses disciplines.

Une méthode qui permet de résumer des données complexes

Les tableaux de contingence qui sont composés d'un nombre réduit de colonnes et de lignes peuvent souvent être interprétés en réalisant des pourcentages, ou en créant un diagramme en barres. Par exemple, les effectifs des catégories du tableau 6.2 sont représentés par un diagramme en barres

dans la figure 6.4. Toutefois, quand un tableau de contingence contient de nombreuses lignes et colonnes, il devient difficile à interpréter.

TABLEAU 6.2 – Tableau de contingence représentant les fréquences d'usages métaphoriques et non-métaphoriques du verbe anglais *see* dans quatre registres du VU Amsterdam Metaphor Corpus (Levshina, 2015, p. 215)

	Academic	Conversations	Fiction	News
Metaphoric	44	48	27	17
Non-metaphoric	26	135	98	19

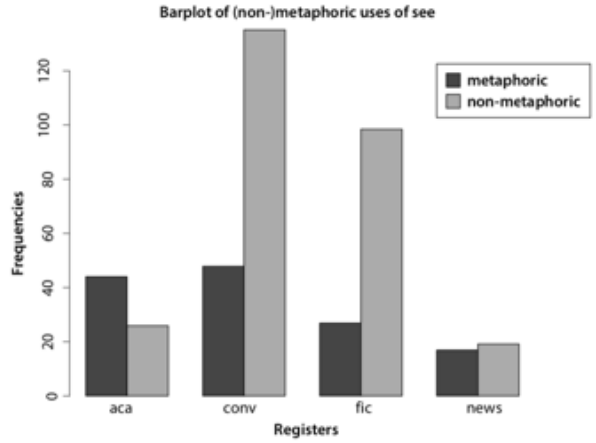

Figure 9.3. Bar plot of metaphoric and non-metaphoric uses of *see* in four registers with grouped bars

FIGURE 6.4 – Diagramme en barres représentant les données du tableau de contingence ci-dessus (Levshina, 2015, p. 216)

L'analyse factorielle des correspondances réduit la dimension d'un tableau de contingence en un nombre restreint de facteurs tout en conservant un maximum d'informations (Salles, 2009). Elle permet de visualiser les associations (ou l'absence d'association) entre les lignes et les colonnes d'un tableau par une représentation graphique sur les plans factoriels. Cette méthode répond notamment aux problématiques rencontrées par les linguistes de corpus qui souhaitent analyser des tableaux de fréquences complexes pour mettre en lumière des liens entre des formes linguistiques et les contextes dans lesquels elles sont utilisées : la projection graphique des données facilite l'identification des associations (Glynn, 2014). Elle peut être considérée comme une extension du test du χ^2, et ne doit être « mise en œuvre si et seulement si l'hypothèse d'indépendance entre les variables, par le test duχ^2, est rejetée » (Salles, 2009, p. 135). Comme le test du χ^2 peut uniquement répondre par oui ou par non à la question de l'indépendance (il n'indique pas où se trouvent les différences), l'analyse factorielle des correspondances est utile : elle met en valeur des relations complexes de similarités et de différences entre les variables (Brezina, 2018).

Une méthode exploratoire

L'analyse factorielle des correspondances simples est une méthode purement exploratoire. Elle n'offre pas d'information sur la significativité des associations constatées (Glynn, 2014). Elle ne décrit que l'échantillon concerné, et ne permet pas de savoir si ses résultats peuvent être généralisés à une population. Elle permet toutefois de révéler des tendances dans les données.

L'inertie

L'inertie est le terme utilisé pour désigner le degré de variation en analyse factorielle des correspondances (Glynn, 2014). Elle est élevée quand les valeurs des lignes et des colonnes sont éloignées du profil moyen. Elle mesure la qualité de la représentation graphique du tableau de contingence (Yelland, 2010). Plus l'inertie est élevée, mieux c'est. Une inertie de 55 % signifie qu'un graphique explique 55 % de la variation contenue dans les données. Ce niveau d'inertie peut paraitre faible, mais il est assez commun dans l'analyse factorielle des correspondances simples. Cela ne signifie pas que le graphique est une mauvaise représentation des données, mais tout simplement qu'il faut l'interpréter avec précaution (Yelland, 2010).

La représentation graphique

Les données du tableau de contingence sont projetées sur un espace bidimensionnel (*biplot* en anglais), sous forme d'un nuage de points représentant les différentes catégories. Cette représentation graphique utilise les distances du χ^2 pour représenter la proximité ou l'éloignement entre les catégories du tableau. Le nombre de dimensions possible, en analyse factorielle des correspondances simples, est égal au nombre de lignes ou de colonnes d'un tableau (en prenant le nombre le plus petit), moins 1. En général, les deux premières dimensions capturent un pourcentage élevé de la variation, et sont donc suffisantes (Glynn, 2014). Pour réaliser les graphiques, nous avons utilisé la fonction `CA` du package `FactoMineR` (Lê et al., 2008).

Interprétation des graphiques en nuage de points

La distance entre les différentes colonnes peut être interprétée directement dans un graphique d'AFC : les catégories les plus proches sur le nuage de points sont celles qui ont des profils similaires, et il en va de même pour les lignes. En revanche, il n'est pas possible d'interpréter directement la distance entre les lignes et les colonnes. Pour cette raison, l'interprétation d'un nuage de points d'AFC est délicate (Levshina, 2015) : ce n'est pas parce qu'une colonne est située près d'une ligne dans le nuage de points qu'elles sont associées. Yelland (2010) décrit une procédure permettant d'interpréter les associations entre les colonnes et les lignes. Elle consiste à tracer des lignes reliant l'origine du graphique aux différents points. Si l'angle formé

par deux lignes reliant une colonne et une ligne est aigu, il y a une association entre la ligne et la rangée. Si l'angle fait 90 degrés, il n'y a pas de relation. Si l'angle est obtus, il y a une association négative entre la ligne et la colonne.

La longueur des lignes doit également être prise en compte. Quand une ligne reliant une catégorie (ou colonne d'un tableau de contingence) à l'origine du graphique est longue, cela signifie que cette catégorie est fortement associée à une ou plusieurs rangées du tableau. Quand elle est courte, l'association est faible. Il en va de même pour les rangées : une ligne longue indique une association forte à une ou plusieurs colonnes et une ligne courte une association faible (Bock, 2017). Dans l'exemple tiré de Yelland (2010) (figure 6.5), la lettre « Y » apparait plus fréquemment dans l'échantillon MT2 (composé de textes de Mark Twain) que la moyenne dans l'ensemble des échantillons figurant sur le graphique. L'association semble forte, car les lignes sont longues. La lettre « P » apparait en revanche moins souvent dans MT2 que dans la moyenne de l'ensemble des échantillons. Notons que l'on ne sait pas si « Y » est plus fréquent que « P » dans MT2.

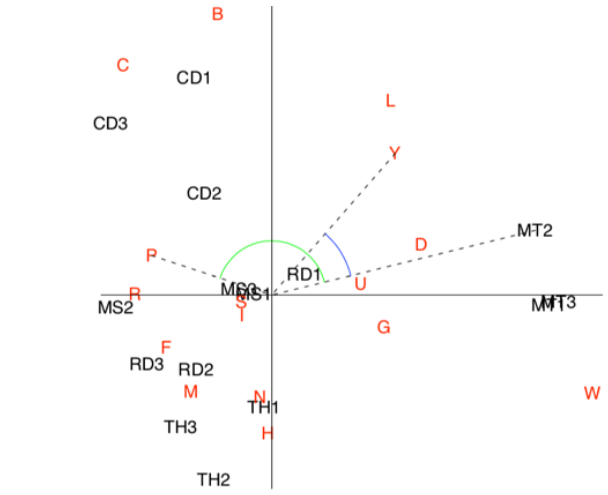

FIGURE 6.5 – Interprétation d'un graphique d'AFC par l'examen des angles (Yelland, 2010, p. 17)

On peut également tracer sur le graphique une ligne qui passe par l'origine et par le point que l'on souhaite étudier, puis tracer des lignes perpendiculaires à cette ligne passant par les points correspondants aux catégories des colonnes du tableau de contingence. Dans la figure 6.6, nous avons tracé des perpendiculaires bleu clair, qui rencontrent la ligne reliant MT2 à l'origine du graphe, du côté de l'origine du graphe. Cela signifie que les lettres « W » à « B », auxquelles elles sont reliées, apparaissent plus souvent dans MT2 qu'en moyenne dans l'ensemble des échantillons. Les perpendiculaires vertes qui rencontrent la ligne de l'autre côté de l'origine sont reliées à des points qui sont moins fréquents dans MT2 qu'en moyenne dans l'ensemble des échantillons. La position des intersections nous ren-

seigne également sur la fréquence d'une lettre au sein d'un échantillon. Plus l'intersection est éloignée de l'origine (mais du même côté que le point de l'échantillon), plus la lettre est fréquente dans l'échantillon. Plus elle est éloignée de l'origine, de l'autre côté du point de l'échantillon, moins la lettre est fréquente. Ici, « W » est plus fréquente dans MT2 que U ou G, par exemple.

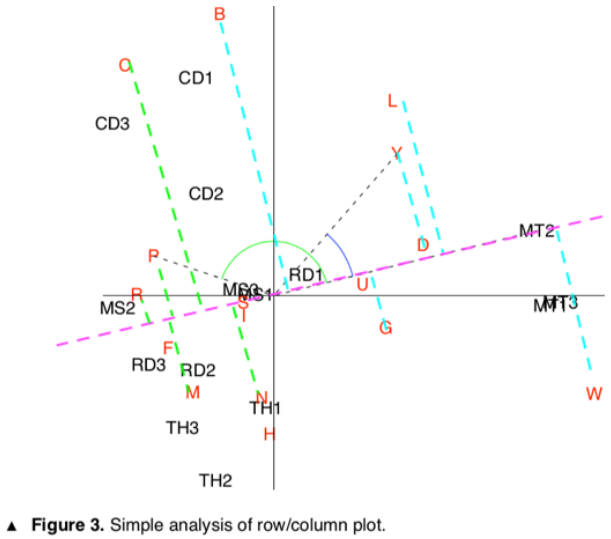

FIGURE 6.6 – Interprétation d'un graphique d'AFC par l'examen des angles (nous avons ajouté les lignes colorées) (Yelland, 2010, p. 17)

6.3 Tests statistiques

6.3.1 Le test du χ^2 d'indépendance

Le test du χ^2 (ou khi deux) est un des tests les plus utilisés en statistique et en linguistique de corpus (Desagulier, 2017). Il permet de savoir si la relation entre deux variables catégorielles représentées dans un tableau de contingence est significative. Il est basé sur une idée simple : comparer les fréquences observées dans certaines catégories aux fréquences que l'on obtiendrait par hasard (Field et al., 2012). Il ne requiert pas que les données suivent la distribution normale : les données catégorielles n'étant pas continues, elles ne peuvent de toute façon pas être normalement distribuées (Field et al., 2012).

Le test du χ^2 peut être réalisé si deux conditions sont réunies. Tout d'abord, les observations doivent être indépendantes les unes des autres. Dans le cas de la linguistique de corpus, cette condition est difficile à remplir, puisque la langue n'est jamais aléatoire, et que, dans un corpus, les mots sont toujours connectés les uns aux autres (Kilgarriff, 2005). Il faut donc être prudent, car le test peut renvoyer des résultats faussement significatifs (Brezina, 2018). Pour faire un test du χ^2, il faut également que

les fréquences figurant dans le tableau de contingence soient toutes supérieures à 5. Dans les tableaux comportant plus de 4 cellules, il suffit que 80 % des fréquences soient supérieures à 5.

Le test du χ^2 permet uniquement de savoir s'il existe une association significative entre deux variables, mais ne donne pas d'informations sur la nature de cette association. Celle-ci peut être parfois déduite de l'observation des fréquences ou du calcul des pourcentages correspondant à chaque cellule du tableau. Dans les tableaux comportant plus de deux lignes ou deux colonnes, il est conseillé de calculer les résidus standardisés (la différence entre les effectifs théoriques et les effectifs observés divisée par la racine carrée des effectifs théoriques) pour en savoir plus sur la relation entre les variables. En complément, plusieurs mesures de tailles d'effet peuvent être calculées, dont le test V de Cramer pour les tableaux comportant plus de 2 lignes ou 2 colonnes, que nous avons utilisé dans notre analyse des g-droppings (→ p. 278). Dans R, le test du χ^2 se calcule avec la fonction `chisq.test()` et le V de Cramer avec la fonction `assocstats()` du package `vcd`.

6.4 La régression

La régression est la méthode que nous utilisons le plus fréquemment dans cette thèse. Elle est en effet particulièrement adaptée à nos données : sous sa forme linéaire généralisée, elle est capable de prendre en charge l'asymétrie des données de corpus. Elle peut également être utilisée dans une perspective intersectionnelle, en intégrant des interactions (→ p. 35) entre plusieurs variables indépendantes, c'est-à-dire les variables dont nous étudions l'impact sur une variable dépendante. Dans notre thèse, par exemple, les variables indépendantes sont le genre, l'âge et l'ethnicité (entre autres), et les variables dépendantes la fréquence des émojis ou des abréviations (entre autres).

6.4.1 La régression linéaire : une méthode pas adaptée à nos données

Dans sa forme la plus simple, la régression linéaire consiste à essayer de tracer une ligne droite de façon à représenter au mieux un nuage de points (Levshina, 2015). Le modèle de régression linéaire dit « général » est basé sur la loi de probabilité normale, ou loi de Gauss. L'ajustement de la droite est réalisé avec la méthode des moindres carrés ordinaire. Son objectif est de minimiser la somme des carrés des résidus, c'est-à-dire la différence entre les valeurs observées et les valeurs prédites par la ligne de régression (Baayen, 2008). L'hypothèse nulle d'un modèle de régression est que les variables indépendantes n'ont pas d'effet sur la variable dépendante. Pour qu'un modèle de régression linéaire produise des résultats fiables, il faut que les données répondent à plusieurs conditions. Il faut notamment que la variable dépendante soit numérique et continue (avec des valeurs comme 2,84 ou -1,3201).

Dans cette thèse, nous étudions principalement des fréquences, c'est-à-dire des données de comptage. Ces données ne se prêtent pas à la régression linéaire à cause de leurs deux caractéristiques principales : elles sont numériques, mais pas continues, puisque les fréquences sont toujours des nombres entiers (1, 492, 10 012), et elles ne peuvent pas être négatives (la fréquence d'un mot dans un corpus ne peut pas être -13). D'autres données qui nous intéressent ici, comme le genre d'un pseudonyme ou le fait de modérer ou non un forum, ne sont pas numériques et ne peuvent pas non plus être analysées avec la régression linéaire. Dans ces deux cas (variables catégorielles et données de comptage), nous avons utilisé des modèles appartenant à la famille des modèles linéaires généralisés (Gries, 2013).

6.4.2 Les modèles linéaires généralisés

Le modèle linéaire généralisé (*generalized linear model* ou *GLM*) est une extension du modèle général. Il utilise une fonction de lien (*link function*) qui transforme l'étendue des valeurs prédites par le modèle linéaire (de moins l'infini à plus l'infini) à une étude plus appropriée à la variable dépendante. Pour les données de comptage, l'étendue des valeurs prédites par un modèle linéaire généralisé va ainsi de 0 à l'infini. Pour les variables catégorielles binaires, les valeurs prédites vont de 0 à 1, et peuvent être interprétées comme des probabilités (Gries, 2013).

Il existe plusieurs types de modèles linéaires généralisés. Le choix d'un modèle s'effectue en fonction de la nature de la variable dépendante : de Poisson ou binomial négatif pour les variables numériques, logistique binaire pour les variables catégorielles à deux niveaux, et multinomial pour les variables catégorielles à plus de deux niveaux. Pour les variables numériques, la dispersion doit également être prise en compte : modèle *zero-inflated* pour les données contenant beaucoup de zéros, modèle binomial négatif pour les données surdispersées, ou modèle Poisson généralisé pour gérer la sous-dispersion.

Modèle de Poisson

La régression de Poisson est une des distributions les plus communément utilisées pour modéliser les fréquences et données de comptage. La distribution de Poisson a notamment pour caractéristiques le fait que les valeurs ne peuvent pas être inférieures à 0, une moyenne égale à la variance, et une asymétrie quand sa moyenne est faible (Zuur et al., 2015). Dans R, le modèle de Poisson est mis en œuvre de la façon suivante, avec la fonction `glm` :

```
glm(emoticons ~ gender + age, family = ‘‘poisson’’, data =
RedditGender)
```

Modèle binomial négatif

La surdispersion est très courante dans les données de comptage. Elle est présente quand la variation des données est supérieure à la variation

attendue avec la distribution de Poisson. Le modèle négatif binomial est la solution la plus communément utilisée pour régler ce problème (Zuur et al., 2015). Il utilise un paramètre de dispersion qui permet de gérer la surdispersion. Il existe plusieurs types de modèles binomiaux négatifs ; le plus courant utilise une distribution de mélange Poisson-gamma (Hilbe, 2011). La majorité des données de ce thèse étant surdispersées, nous avons principalement eu recours à des modèles binomiaux négatifs (sauf si précisé). Nous les avons créés avec la fonction `glm.nb` du package `MASS` (Venables & Ripley, 2002), comme ci-dessous :

```
glm.nb(emoticons ~ gender * age, data = RedditGender) [1]
```

Modèle logistique binaire

La régression logistique binaire est utilisée pour analyser des variables catégorielles ayant deux niveaux (comme oui/non, graphie standard/graphie non standard, etc.). Cette méthode a une longue histoire en sociolinguistique, qui a été la première discipline linguistique à l'employer avec le programme « Variable Rule », mis au point dans les années 1970 et implementé dans plusieurs logiciels dont le célèbre Varbrul (Tagliamonte & Baayen, 2012). La régression logistique permet en effet d'étudier un type de question courante en sociolinguistique : les situations où les locuteur·trices ont le choix entre deux réalisations d'une forme qui a le même sens, et où ce choix est conditionné par des variables sociales ou contextuelles (Tagliamonte & Baayen, 2012).

Nous avons utilisé la régression logistique binaire sous sa forme classique dans notre analyse des pseudonymes (→ p. 176), et sous forme de modèle à effets mixtes dans notre analyse de la variante de l'omission de l'apostrophe de *it's* (→ p. 285). Nous avons utilisé la `glm` pour créer les modèles logistiques binaires, avec ce type de formule :

```
glm(i ~ gender * age, family = binomial, data = RedditGender)
```

Modèles de régression logistique mixtes Les modèles mixtes sont une méthode relativement récente, qui a vu le jour grâce aux progrès informatiques rapides de la fin du 20ème siècle (Galwey, 2014). Un modèle mixte fait la distinction entre deux types d'effets : les effets fixes et les effets aléatoires. Les effets fixes correspondent aux variables indépendantes que l'on souhaite étudier (par exemple, dans notre cas, le genre ou l'âge). Les effets aléatoires correspondent à une autre source de variation présente dans les données mais qui ne fait pas partie des questions de recherche que l'on a posées, et que l'on souhaite donc neutraliser (dans notre cas, les préférences individuelles des Redditors) (Brezina, 2018). Lorsque l'on a recueilli plusieurs observations par personne, les modèles mixtes représentent une amélioration considérable par rapport aux modèles classiques. Ils permettent de neutraliser la variation individuelle, et d'isoler l'effet des variables sociolinguistiques d'intérêt (Brezina, 2018).

1. L'astérisque permet ici de prendre en compte l'interaction entre le genre et l'âge

Nous avons utilisé des modèles logistiques mixtes dans les analyses où nous souhaitions comparer le choix des Redditors entre deux variantes d'une même forme, comme l'utilisation du *I* majuscule ou minuscule, ou l'omission de l'apostrophe de *it's*. Pour ce faire, nous avons employé la fonction `glmer()` du package `lme4` (Bates et al., 2015), en utilisant les identifiants des Redditors comme un effet aléatoire. Nous avons également utilisé un effet aléatoire pour chaque observation (OLRE) dans un cas où nous avons rencontré une importante surdispersion (Harrison, 2014).

Modèles *zero-inflated*

Dans les cas où les données contiennent beaucoup plus de zéros que le nombre de zéros auquel on pourrait s'attendre étant donné la moyenne de la distribution, il est conseillé d'utiliser un modèle Poisson ou binomial négatif *zero-inflated* (Hilbe, 2014). Ce type de modèle fait la différence entre les « bons » et les « mauvais » zéros ; cette distinction vient de la recherche écologique, pour laquelle il est important de savoir d'où proviennent les observations zéro (Hilbe, 2014, p. 198). En effet, certains zéros peuvent être dus à des erreurs. Hilbe donne l'exemple des chants d'oiseaux. Quand un·e scientifique compte le nombre de fois où une espèce d'oiseau chante, les observations zéros peuvent se produire pour deux raisons. Tout d'abord, il se peut que les oiseaux n'aient pas chanté au moment où l'enregistrement a été réalisé ; c'est ce que Hilbe appelle les « bons zéros ». Ensuite, il est possible que les chants n'aient pas été enregistrés parce que le ou la scientifique était au mauvais endroit ou au mauvais moment. Ce sont les « mauvais zéros » : les chants d'oiseaux auraient pu être enregistrés, car les oiseaux chantaient. La distinction entre les « bons » et les « mauvais » zéros est une « fiction mathématique » (Hilbe, 2014, p. 197), mais elle aide à l'interprétation des modèles *zero-inflated*. Dans notre cas, on pourrait dire que les « mauvais » zéros correspondent à des zéros obtenus non pas parce qu'une personne n'utilise pas tel ou tel procédé, mais parce que la façon dont les échantillons ont été constitués n'a pas permis de capter le ou les moments où elle les utilise.

Les modèles *zero-inflated* sont principalement utilisés dans le domaine de l'écologie, mais aussi de la santé et des transports (Hilbe, 2014). Ils commencent également à être utilisés en linguistique ; par exemple, Burch et Egbert (2020) ont utilisé des modèles *zero inflated* pour étudier la fréquence de certains mots dans le British National Corpus. Les modèles *zero-inflated* sont composés de deux parties. La première est binaire, et modélise d'un côté les zéros, qui deviennent des « 1 », et de l'autre les observations supérieures à zéro, qui deviennent des « 0 ». La seconde, ou *count component*, modélise toutes les observations (Hilbe, 2014). Dans nos analyses, nous commentons les résultats de la première partie, mais ne présentons que les résultats de la seconde partie du modèle, pour des raisons pratiques.

6.4.3 Sélection des modèles

Nous avons choisi les modèles les plus adaptés aux types de données dont nous disposions en comparant les modèles entre eux par l'utilisation de tests. Nous avons par exemple utilisé le *boundary likelihood ratio test* pour comparer un modèle binomial négatif *zero-inflated* à un modèle de régression de Poisson *zero-inflated*, ou un modèle binomial négatif à un modèle de régression de Poisson (c'est-à-dire, des modèles qui sont *nested*, ou imbriqués) (Hilbe, 2014). Nous avons pour ce faire utilisé la fonction `lrtest` du package `lmtest` (Zeileis & Hothorn, 2002).

Nous avons également eu recours au *Vuong test*, qui permet de comparer un modèle *zero-inflated* à sa version *non-zero-inflated* (Hilbe, 2014) à l'aide de la fonction `vuong` du package `pscl` (Jackman, 2020). Dans une vaste majorité des cas, ces tests nous ont permis de déterminer que les modèles de régression binomiaux négatifs étaient les plus adaptés à nos données, car ils peuvent prendre en charge des données qui comportent une forte dispersion (Hilbe, 2014).

6.4.4 Processus de sélection des variables

Quand un modèle contient plus d'une variable indépendante, il faut savoir quelle(s) variable(s) on conserve, ou pas. Selon le principe de parcimonie, également appelé « rasoir d'Ockham », il convient de privilégier les modèles simples aux modèles complexes, à pouvoir explicatif équivalent (Gries, 2013). En d'autres termes, le modèle optimal est celui qui ne contient que les variables indépendantes qui permettent de comprendre la variance de la variable dépendante (Levshina, 2015).

Pour sélectionner les variables, nous avons utilisé la méthode appelée *stepwise selection* et implémentée dans R par la fonction `step`. Elle permet d'ajouter ou de retirer des variables pour trouver le modèle optimal. Nous avons choisi la deuxième option, en partant d'un modèle maximal qui contient toutes les variables que l'on souhaitait étudier, et en retirant ensuite (ou pas) une ou plusieurs variables indépendantes. La fonction `step()` utilise comme critère de sélection l'Akaike Information Criterion, qui pénalise les modèles qui ont trop de variables. Plus l'AIC est faible, mieux le modèle explique la variance. La fonction renvoie le modèle à l'AIC le plus faible. Comme cette fonction ne prend pas en charge (à notre connaissance) les modèles linéaires généralisés à effets mixtes, nous avons utilisé à la place, dans ce cas, la fonction `drop1()`. Celle-ci renvoie la liste des variables, avec, sur la même ligne, l'AIC d'un modèle qui ne contiendrait pas cette variable.

Toutefois, dans certains cas, nous avons fait des exceptions à ce processus de sélection pour conserver toutes les variables dans un modèle. En effet, parfois, l'algorithme de la fonction `step()` a supprimé le genre du modèle ; or, comme nos groupes ne sont pas de taille égale (il y a par exemple plus d'hommes cisgenres de moins de 20 ans que de femmes cisgenres du même âge), nous ne souhaitions pas que l'effet de la variable qui subsistait dans le modèle (l'âge, en général) soit influencé par la taille des groupes.

6.4.5 Les interactions

Lorsque l'on intègre plusieurs variables indépendantes dans un modèle, on doit considérer la question de leur possible interaction (Gries, 2013). Dans un modèle de régression, les effets de deux variables indépendantes peuvent être additifs : la combinaison des deux variables produit un effet similaire à celui que l'on s'attendrait à obtenir en additionnant les deux effets (Gries, 2013). Mais, parfois, deux variables (ou plus) interagissent : l'effet d'une variable indépendante dépend de celui d'une autre variable indépendante. On ne peut donc pas prédire l'effet de l'interaction en examinant les effets principaux.

Pour Gries (2013), les interactions sont souvent « often misunderstood and/or underutilized » (p. 252-253). Leur intégration dans un modèle apporte des informations supplémentaires qui peuvent complètement changer la donne. Quand on ne les utilise pas, on ne peut pas se fier aux effets principaux. Parfois, un effet principal peut être seulement significatif pour un groupe et pas pour les autres. Par ailleurs, sans utiliser d'interactions, ou en utilisant uniquement les interactions auxquelles on s'attend, il devient plus difficile d'obtenir des résultats inattendus. Intégrer les interactions dans un modèle de régression est donc, selon Gries (2013), absolument indispensable.

L'interprétation d'une interaction est fastidieuse car elle change le niveau de référence du modèle. En intégrant l'interaction de l'âge et du genre, par exemple, le niveau de référence devient par exemple les femmes cisgenres de 14 à 20 ans, qui peuvent uniquement être comparés avec les hommes cisgenres, les femmes transgenres, les hommes transgenres et les personnes non binaires du même âge. Il faut donc changer les niveaux de référence des variables « âge » et « genre » plusieurs fois pour comparer tous les groupes entre eux. La façon la plus simple d'interpréter une interaction dans un modèle de régression est d'en réaliser une représentation graphique (Levshina, 2015). Pour cela, nous avons utilisé le package `visreg` (Breheny & Burchett, 2017), qui permet de représenter l'effet de deux variables dans un même graphe.

6.4.6 L'offset

Nous avons utilisé dans les modèles de régression la fréquence brute des phénomènes lexicaux étudiés (→ p. 151). Pour compenser le fait que chaque sous-corpus, qui correspond à la production d'un·e internaute, a une taille différente, nous avons intégré un offset aux modèles de régression. L'offset est le logarithme naturel (également appelé népérien) du nombre de tokens de chaque sous-corpus (Zuur et al., 2015). Il permet de modéliser des taux (*rate*) au lieu de fréquences brutes ; c'est-à-dire, ici, le nombre de phénomènes linguistiques en fonction de la taille d'un sous-corpus. Il est ajouté dans les modèles comme toute autre variable indépendante ; par exemple, ici, dans le cas d'un modèle binomial négatif :

```
glm.nb(emoticons ~ gender + age + offset(log(n_tokens)),
data = RedditGender)
```

6.4.7 Interprétation des modèles linéaires généralisés

Pour savoir comment il faut interpréter un modèle linéaire généralisé, nous présentons ici un exemple (sans interaction pour privilégier la simplicité, tableau 6.3), tiré du chapitre 2 de la partie 4 (→ p. 221). L'objectif n'est pas de donner un aperçu des résultats, sur lesquels nous reviendrons, mais d'expliquer comment les nombreux tableaux de régression présentés dans cette thèse doivent être lus.

Tableau 6.3 – Régression binomiale négative, exemple

	Variable dépendante :
	Émoticônes
Intercept	0.001^{**} (0.001, 0.001)
Femmes cisgenres	1.811^{**} (1.502, 2.183)
Femmes transgenres	3.294^{**} (2.501, 4.398)
Hommes transgenres	2.085^{**} (1.583, 2.784)
Non-binaires	2.076^{**} (1.575, 2.776)
21-30 ans	0.635^{**} (0.498, 0.801)
31 ans et +	0.477^{**} (0.372, 0.607)
Observations	1,044
Log Likelihood	−4,059.691
θ	0.640^{**} (0.028)
Akaike Inf. Crit.	8,133.381
Note :	*p<0.05 ; **p<0.01

La constante ou *intercept*

Dans une régression linéaire, l'*intercept* (également appelé « constante ») représente la valeur prédite de la variable dépendante (ici la fréquence des émoticônes) à l'endroit où la ligne de régression traverse l'axe des ordonnées y. En d'autres termes, il s'agit de la valeur prédite de la variable dépendante quand toutes les variables sont à leur niveau de référence. Pour les variables numériques, il s'agit de la valeur prédite de la variable dépendante quand les valeurs des variables numériques sont égales à 0. Dans le cas des variables catégorielles, comme ici (et généralement dans cette thèse), le niveau de référence est déterminé par la personne qui réalise l'analyse. En général, on choisit le niveau de la variable où il y a le plus d'observations ou individus, ou le niveau qui fait le plus sens sur le plan théorique. Pour l'âge, la catégorie 1 (14 à 20 ans) est donc notre niveau de référence, même si elle contient moins d'individus que les autres niveaux (21 à 30 ans, et 31 ans et +). La variable « genre » comporte autant d'hommes que de femmes cisgenres (372). Dans le modèle présenté ici, les hommes cisgenres sont le niveau de référence, mais, dans d'autres modèles présentés dans la thèse, ce sont les femmes cisgenres. Le choix du niveau de référence n'a pas d'impact sur le modèle, mais uniquement sur la façon dont sont présentés et doivent être interprétés les résultats.

Dans une régression linéaire généralisée, le principe est le même, mais

l'*intercept* indique le logarithme de la valeur prédite (ici, le logarithme de la fréquence des émoticônes). Il faut lui appliquer la fonction exponentielle (dans R, avec la fonction `exp()`) pour connaitre la valeur prédite (Faraway, 2016). L'*intercept* originel du modèle présenté dans le tableau 6.3 est de -0.6880 ; sa valeur exponentielle est de 0.001. Dans ce modèle, comme dans les autres modèles réalisés dans cette thèse (sauf exceptions, toujours précisées), les coefficients et les *intercepts* sont présentés sous leur forme exponentielle. Ils peuvent donc être interprétés directement. Notons une autre spécificité de ce modèle, et de la majorité des modèles réalisés dans cette thèse ; par l'utilisation d'un offset, qui correspond au logarithme de nombre de tokens par sous-corpus, le modèle prédit non pas le nombre d'émoticônes dans un sous-corpus, mais le nombre d'émoticônes par token. La valeur exponentielle du coefficient, 0.01, indique donc qu'un homme cisgenre produit 0.001 émoticône par token, ou 1 émoticône par 1000 tokens.

Significativité des effets

Dans la partie supérieure du tableau de régression, chaque ligne correspond à une catégorie, et à sa différence (ici dans la fréquence des émoticônes) par rapport au niveau de référence (ici, les hommes cisgenres). Les valeurs p, représentées sous forme d'astérisques, indiquent s'il y a une différence significative entre le niveau de référence et les autres niveaux. On constate ici une différence significative entre les hommes cisgenres et tous les autres groupes.

Coefficients

Dans les modèles de régression généralisée, les coefficients sont représentés sous forme de *log odds ratios*. Leur valeur exponentielle (présentée dans le tableau 6.3) correspond à des *odds ratio*. Ils peuvent être interprétés ainsi : un *odds ratio* de 1 signifie qu'il n'y a pas d'effet. Un *odds ratio* supérieur à 1 indique une augmentation du phénomène étudié, et un *odds ratio* inférieur une diminution (Larmarange, p. d.). Dans le modèle présenté plus haut (tableau 6.3), si on regarde les groupes de genre, on voit que tous les coefficients sont supérieurs à 1. Cela signifie que femmes cisgenres, femmes transgenres, hommes transgenres et personnes non binaires ont tous utilisé davantage d'émoticônes que les hommes cisgenres. La taille d'effet la plus forte est constatée quand on compare les hommes cisgenres aux femmes transgenres, qui produisent 3.3 fois plus d'émoticônes qu'eux. Les coefficients des groupes d'âges sont quant à eux inférieurs à 1, ce qui signifie que les Redditors de 21 à 30 ans et les Redditors de 31 ans et plus utilisent moins d'émoticônes que le niveau de référence de la variable « âge », c'est-à-dire les Redditors de 14 à 20 ans.

Intervalles de confiance

Les nombres indiqués entre parenthèses correspondent aux intervalles de confiance des coefficients. Ils sont également présentés sous leur va-

leur exponentielle ; cela signifie que les intervalles de confiance qui comprennent le 1 (qui indique l'absence de différence entre deux groupes) ne sont pas significatifs.

Explorer d'autres comparaisons

Dans le modèle 6.3, les différentes catégories d'une variable (genre et âge) peuvent uniquement être comparées au niveau de référence de cette variable (hommes cisgenres et 14-20 ans). On ne peut donc pas savoir s'il y a une différence significative entre, par exemple, les femmes cisgenres et les femmes transgenres. Pour explorer d'autres contrastes, nous avons changé plusieurs fois le niveau de référence du modèle en utilisant la fonction `relevel()`.

6.5 Organisation des analyses linguistiques

Dans un premier temps, nous présentons une analyse descriptive de chaque variable dans le corpus entier (fréquence et différents types observés pour les variables linguistiques, par exemple). Dans un second temps, nous proposons une analyse axée sur le genre et son interaction avec l'âge. Cette analyse commence avec la présentation des statistiques descriptives pour chaque groupe d'âge et de genre. Nous présentons ensuite un modèle de régression qui met en lumière les possibles effets de l'âge, du genre et de leur interaction, quand elle est significative.

Pour les variables linguistiques, nous proposons également une analyse qui intègre l'ethnicité, et qui est basée sur l'échantillon réduit décrit dans la section 5.1.7. Nous présentons des statistiques descriptives sous forme de boites à moustaches ou de mesures de tendance centrale et de dispersion. Vient ensuite un modèle de régression, créé en intégrant l'effet principal de l'âge et l'interaction du genre et de l'ethnicité. Nous n'avons pas intégré d'interaction à trois niveaux (genre, âge et ethnicité) aux modèles, à cause de la grande complexité de son interprétation. Cela signifie en effet comparer, par exemple, les femmes afro-américaines de 14 à 20 ans aux femmes afro-américaines de 21 à 30, puis aux femmes afro-américaines de 31 ans et plus, et ainsi de suite pour chaque groupe ethnique et chaque groupe de genre. Même si les interactions entre âge, genre et ethnicité peuvent exister, il nous a semblé plus judicieux de nous cantonner à une interaction à deux niveaux, et de privilégier l'interaction du genre et de l'ethnicité (après nous être rendue compte, lors d'essais, que l'algorithme de sélection des variables préférait généralement cette interaction à celle de l'âge et du genre, sauf exception).

Pour certaines variables linguistiques, il nous a semblé pertinent de réaliser des analyses supplémentaires pour explorer les possibles effets d'autres variables. Nous avons ainsi intégré l'orientation sexuelle à notre étude des émoticônes (→ p. 225), et le genre assigné à la naissance des personnes non binaires dans l'analyse de plusieurs variables, dont les émoticônes (→ p. 223), les étirements de lettres (→ p. 236), les étirements

de ponctuation (→ p. 241), et les g-droppings (→ p. 284). Ces analyses sont basées sur des échantillons réduits du corpus : 275 femmes cisgenres, 48 femmes transgenres, 256 hommes cisgenres et 68 hommes transgenres pour l'analyse de l'orientation sexuelle, et 98 personnes non binaires dont nous connaissons le genre assigné à la naissance.

6.5.1 Note sur les graphiques et les tableaux

Pour des raisons pratiques, les étiquettes des graphiques sont généralement en anglais (et parfois en français). Les variables et leurs niveaux ont en effet été codés en anglais lors du recueil des données, et nous ne les avons pas traduits avant l'analyse dans R. Autant il est facile de personnaliser la légende des axes des abscisses et des ordonnées, autant changer les étiquettes correspondant aux noms des groupes peut être fastidieux. Cela nécessite en général plusieurs lignes de code, avec une syntaxe qui varie selon les packages utilisés. Nous avons donc décidé de privilégier la simplicité.

Les tableaux de régression présentés dans la thèse ont généralement été créés avec la fonction `stargazer` du package `stargazer` (Hlavac, 2018). À cause des limites de ce package, nous avons dû personnaliser le code afin d'obtenir le format que nous souhaitions (valeur exponentielle des coefficients, valeur exponentielle des intervalles de confiance et significativité des coefficients, pour permettre une interprétation directe des résultats). Nous avons utilisé une solution proposée sur le site Stack Overflow (MC808, 2013). Le code utilisé est présenté en annexe A (p. 391). Pour certains modèles, non pris en charge par `stargazer`, nous avons créé les tableaux de régression manuellement.

Les tableaux autres que les tableaux de régression (comme les tableaux de contingence) ont été générés avec le package `xtable` (Dahl et al., 2019), ce qui a permis de les copier facilement dans LaTeX, le système de composition utilisé pour composer cette thèse.

6.6 Tableau récapitulatif des méthodes utilisées

Le tableau 6.4 présente les méthodes statistiques utilisées dans la thèse, avec les variables concernées, et le chapitre, section (et parfois sous-section) où se trouvent les analyses.

TABLEAU 6.4 – Tableau synthétique des méthodes statistiques utilisées

Méthode	Variable	Ch.
AFC	Centres d'intérêt	8.4
Diagramme en mosaïque	Âge Reddit	7.3
	Étirements de ponctuation	10.5.1
Régression logistique binaire	Pseudonymes	7.2.3
	Profils supprimés	7.4
	Modération	7.5
Régression logistique binaire avec effets mixtes	Variante *its*/*it's*	11.5.4
	Variante *i*/*I*	11.6.2
Test du χ^2 d'indépendance	G-droppings	11.4.1
Régression binomiale négative	Karma de post et de commentaire	7.6
	Mobilité des Redditors	8.2
	Longueur des commentaires	8.3
	Émoticônes	10.2
	Étirements de lettres	10.4
	Étirements de ponctuation	10.5
	Interjections	10.7
	Mots en majuscules	10.6.3
	Abréviations	11.2
	G-droppings	11.4
Régression *zero-inflated*	Émojis	10.3.3
Poisson avec OLRE	Mots en majuscules	10.6.2

tl ;dr

Nos analyses statistiques sont organisées en deux parties. Tout d'abord, nous présentons des mesures ou graphiques descriptifs, qui ne s'appliquent qu'à notre échantillon, avant d'utiliser des techniques inférentielles qui visent à généraliser nos résultats. La régression binomiale négative avec interactions, méthode que nous avons la plus fréquemment utilisée, permet à la fois de répondre aux contraintes posées par nos données (de comptage, typiques de la linguistique de corpus) et d'inscrire nos analyses dans une perspective intersectionnelle.

En fonction des spécifiques de chaque jeu de données, nous avons utilisé d'autres types de régression (logistique binaire pour les données catégorielles, ou *zero-inflated* pour les données comportant de nombreux zéros, par exemple).

L'intégration des interactions complexifie l'interprétation des modèles. Nous avons privilégié l'interaction à deux niveaux du genre et de l'ethnicité. Nous avons également eu recours, pour l'analyse des centres d'intérêt, à l'analyse factorielle des correspondances, une méthode exploratoire qui génère une représentation graphique des associations entre variables.

Troisième partie

Identités et itinéraires

Chapitre 7

La Reddidentité

Ce chapitre est consacré à ce que nous avons baptisé la « Redditentité », c'est-à-dire aux marques visibles, en dehors de leurs commentaires, des Redditors sur le site. Il explore plusieurs aspects essentiels de cette identité : les pseudonymes choisis par les internautes, leur âge Reddit, la durée de vie de leurs comptes Reddit, leur choix de devenir modérateur·trice bénévole d'un ou de plusieurs subreddits, et leurs scores de karma, symboles de statut sur le site.

7.1 Hypothèses

Nos hypothèses, pour l'étude des variables de la Reddidentité, sont basées sur le fait que les hommes cisgenres sont, historiquement, les premiers utilisateurs de Reddit et qu'ils sont plus nombreux sur le site que les autres groupes. Au vu du caractère geek et misogyne du site, nous pensons qu'ils sont plus à l'aise et moins exposés au *doxxing* (fait de révéler des informations personnelles sur un·e internaute à des fins malveillantes) ou au harcèlement que les autres Redditors. Par conséquent, ils sont peut-être plus susceptibles de choisir des pseudonymes transparents, c'est-à-dire qui indiquent leur genre, par le choix d'un prénom masculin par exemple. Les femmes et les personnes transgenres et non binaires, par contraste, auraient plus intérêt à masquer leur identité de genre. Nous émettons également l'hypothèse que les hommes cisgenres ont créé leurs comptes avant les autres Redditors et qu'ils les conservent plus longtemps ; les femmes cisgenre et les personnes non binaires et transgenres pourraient davantage être amenées à les supprimer à cause du harcèlement dont elles font l'objet, ou pour effacer des commentaires qui pourraient leur porter préjudice.

Il est par ailleurs possible que les hommes cisgenres soient plus nombreux à modérer des subreddits et qu'ils obtiennent davantage de karma que les autres groupes de genre, ce qui leur conférerait davantage de pouvoir et de statut sur le site. Nous pensons également qu'il est possible que les scores de karma reflètent l'implication des Redditors dans la communauté, que nous mesurons ici par l'âge Reddit et la modération de subreddits ; il se peut donc que ces deux variables soient corrélées positivement

avec les scores de karma.

7.2 Pseudonymes

7.2.1 Recodage de la variable

À partir du codage des pseudonymes décrit page 137, nous avons procédé à un second codage dans R. Nous avons ainsi ajouté une nouvelle variable qui indique si le genre d'un internaute est reflété de façon explicite par son pseudonyme (c'est le cas si, par exemple, un homme cisgenre choisit un pseudonyme comportant un prénom masculin ou un terme qui indexe explicitement la masculinité). Cette variable comporte deux niveaux : les noms non genrés et les noms qui reflètent l'identité de genre de la personne. Dans le cas des personnes transgenres et non binaires, nous avons inclus dans cette catégorie les noms qui indiquent une identité transgenre ou non binaire (par exemple, *transperson1987*).

7.2.2 Statistiques descriptives

Le tableau 7.1 présente le nombre et le pourcentage de personnes ayant choisi un pseudonyme indiquant explicitement leur genre, ou pas, dans le corpus. La proportion de noms genrés est faible : elle est de 18.49 % (193 pseudonymes), soit moins de 1 pseudonyme sur 5. Les proportions sont relativement similaires dans la plupart des groupes ; elles vont de 11 % pour les hommes transgenres à 18.82 % pour les femmes cisgenres. Le groupe des femmes transgenres semble se démarquer, avec une proportion environ deux fois plus importante (36 %) que les femmes cisgenres.

Tableau 7.1 – Types de pseudonymes, par sous-corpus, en effectifs et en pourcentages

Sous-corpus	Nom non genrés ou $\neq$ du genre de la personne	Nom qui reflète le genre de la personne
Hommes cisgenres	310 (83.33 %)	62 (16.67 %)
Femmes cisgenres	302 (81.18 %)	70 (18.82 %)
Femmes transgenres	64 (64.00 %)	36 (36.00 %)
Hommes transgenres	89 (89.00 %)	11 (11.00 %)
Non-binaires	86 (86.00 %)	14 (14.00 %)
Tous	851 (81.51 %)	193 (18.49 %)

7.2.3 Effet de l'identité de genre sur le choix d'un pseudonyme : régression logistique binaire

Pour vérifier si ces résultats peuvent être généralisés, nous avons réalisé un modèle de régression logistique binaire avec le genre, l'âge, leur interaction, et l'âge Reddit comme variables indépendantes. Nous avons utilisé la

fonction `step()` pour sélectionner les variables qui contribuent au modèle. Seul le genre a été retenu. Le modèle est présenté dans le tableau 7.2, avec les hommes cisgenres comme niveau de référence.

Tableau 7.2 – Noms d'utilisateur·trices

	Variable dépendante :
	Pseudonyme genré
Intercept	0.200^{**} (0.151, 0.261)
Femmes cisgenres	1.159 (0.795, 1.692)
Femmes transgenres	2.813^{**} (1.714, 4.588)
Hommes transgenres	0.618 (0.297, 1.180)
Non-binaires	0.814 (0.420, 1.486)
Observations	1,044
Log Likelihood	−487.983
Akaike Inf. Crit.	985.967
Note :	*p<0.05 ; **p<0.01

Le modèle montre que la probabilité qu'un·e Redditor choisisse un nom d'utilisateur qui reflète son identité de genre est significativement plus importante chez les femmes transgenres que dans tous les autres groupes : elle est 2.81 fois plus élevée pour les femmes transgenres que chez les hommes cisgenres, 2.43 fois plus élevée que chez les femmes cisgenres, 4.55 fois plus élevée que chez les hommes transgenres, et 3.45 fois plus élevée que chez les personnes non binaires.

7.3 Âge Reddit

Nous avons réalisé une analyse des corrélations à l'aide de diagrammes en mosaïque pour savoir si le genre et l'ethnicité ont un rapport significatif avec l'âge Reddit. En d'autres termes, nous souhaitions savoir si les internautes d'un groupe de genre ou d'un groupe ethnique sont plus susceptibles que les autres d'avoir un âge Reddit plus important, c'est-à-dire de fréquenter le site depuis plus longtemps. Pour réaliser cette analyse, nous avons réduit le nombre de niveaux de la variable « Âge Reddit » décrite page 136 en groupant les catégories 1 et 2, et 4 et 5. Les trois niveaux sont donc les suivants : catégorie 1 (de 0 à 2 ans), catégorie 2 (de 2 à 3 ans) et catégorie 3 (4 ans et plus).

7.3.1 Âge Reddit et genre

Le diagramme 7.1, basé sur le tableau de contingence présenté en annexe B, montre plusieurs corrélations entre l'âge Reddit et les groupes de genre. Les hommes cisgenres sont surreprésentés parmi les Redditors ayant les comptes les plus anciens (4 ans et plus), et sous-représentés parmi ceux ayant les comptes les plus jeunes (moins de 2 ans). Chez les femmes transgenres et les hommes transgenres, c'est l'inverse : leurs comptes ont

plus fréquemment été créés il y a peu de temps. Les hommes transgenres sont sous-représentés chez les comptes de plus de 2 ans.

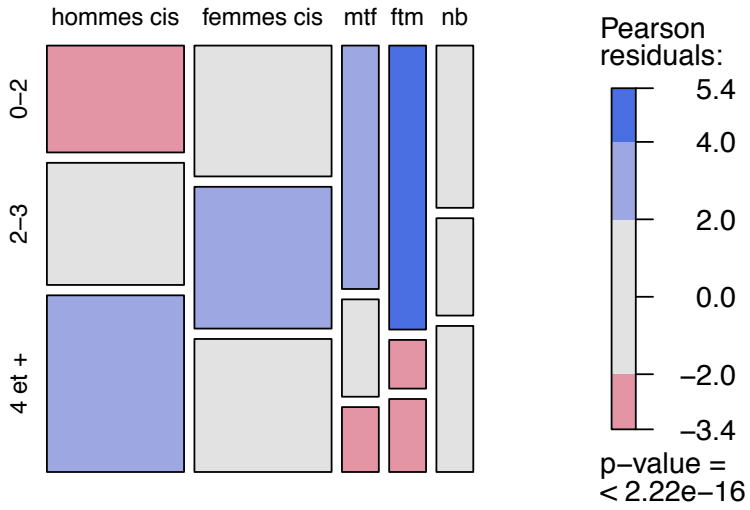

FIGURE 7.1 – Diagramme en mosaïque présentant les corrélations entre l'âge Reddit et les groupes de genre

7.3.2 Âge Reddit et ethnicité

Le diagramme en mosaïque 7.2 est basé sur le tableau de contingence présenté en annexe B, qui présente l'âge Reddit des Redditors de chaque groupe ethnique. Les rectangles ont approximativement tous la même hauteur : dans RedditGender, il ne semble y avoir aucune corrélation entre l'ethnicité d'un Redditor et son âge Reddit.

7.4 Profils supprimés : étude longitudinale

7.4.1 Données

En mai 2020, 3 ans après la fin de la construction du corpus, nous avons à nouveau accédé aux profils de tou·tes les Redditors de RedditGender. L'objectif était de savoir si leur compte Reddit était encore actif ou non, et de connaitre la cause de l'inactivité d'un profil : suppression du profil par l'utilisateur·trice (que Reddit indique comme étant *deleted*), ou suspension du profil (marquée comme *suspended* par Reddit) (→ p. 89). Nous avons obtenu une nouvelle variable, qui comporte trois niveaux : profil actif, profil supprimé, et profil suspendu.

7.4.2 Effet du genre sur la suppression de profils

Des 1044 profils de RedditGender, 928 étaient encore actifs en mai 2020. Nous avons également relevé 2 cas de profils toujours existants, mais com-

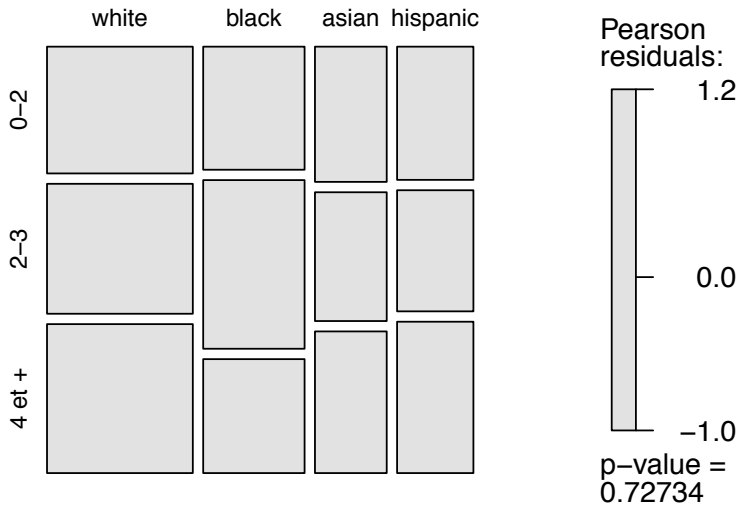

FIGURE 7.2 – Diagramme en mosaïque présentant les corrélations entre l'âge Reddit et les groupes ethniques

plètement vides et inactifs, le ou la Redditor ayant supprimé toutes ses contributions. Dans ces deux cas, nous avons considéré ces comptes comme ayant été supprimés par leurs propriétaires. En tout, 116 personnes ont supprimé leur compte entre 2017 et 2020. Seuls quatre cas de suspension ont été trouvés (trois hommes cisgenres et une femme cisgenre). En 2020, l'âge médian des comptes de RedditGender était de 6 ans, avec un maximum de 14 ans. Le tableau 7.3 présente le nombre de profils actifs, supprimés ou suspendus pour chaque groupe. Le nombre de profils supprimés dans chaque groupe est assez similaire : 9.95 % pour les hommes cisgenres, 9.10 % pour les femmes cisgenres, 8 % pour les hommes transgenres, 8 % pour les personnes non binaires, et 7 % pour les femmes transgenres. Le nombre de profils suspendus est si faible (N = 4) qu'il est difficile de comparer les groupes. Les statistiques descriptives ne révélant pas de différence entre les groupes, nous n'avons pas réalisé d'analyse inférentielle pour cette variable.

TABLEAU 7.3 – Nombre de profils actifs, supprimés ou suspendus au 4 mai 2020

	Actifs	Supprimés	Suspendus
Hommes cisgenres	332 (89.25 %)	37 (9.95 %)	3 (0.8 %)
Femmes cisgenres	337 (90.6 %)	34 (9.1 %)	1 (0.3 %)
Hommes transgenres	92 (92 %)	8 (8 %)	0 (0 %)
Femmes transgenres	93 (93 %)	7 (7 %)	0 (0 %)
Non binaires	92 (92 %)	8 (8 %)	0 (0 %)
Tous	946 (90.61 %)	94 (9.01 %)	4 (0.38 %)

7.5 Modération

7.5.1 Statistiques descriptives

Nous avons réalisé cette analyse avec l'ensemble du corpus, soit 1044 Redditors. Le tableau 7.4 montre la proportion de modérateur·trices dans chaque groupe. Les femmes cisgenres sont les moins nombreuses à modérer des subreddits (11.29 %). Le pourcentage de modérateurs est le plus élevé chez les hommes cisgenres (20.97 %) et les personnes non binaires (20 %). Les Redditors les plus jeunes sont plus nombreux à modérer des forums que les autres catégories. L'âge Reddit semble quant à lui corrélé positivement avec la modération de forums.

Tableau 7.4 – Pourcentages de modérateurs dans les différents groupes

Variable	Sous-corpus	Non modérateurs	Modérateurs
Genre	Hommes cisgenres	294 (79.03 %)	78 (20.97 %)
	Femmes cisgenres	330 (88.71 %)	42 (11.29 %)
	Femmes transgenres	86 (86.00 %)	14 (14.00 %)
	Hommes transgenres	87 (87.00 %)	13 (13.00 %)
	Non-binaires	80 (80.00 %)	20 (20.00 %)
Âge	14-20 ans	111 (75.51 %)	36 (24.49 %)
	21-30 ans	445 (86.07 %)	72 (13.93 %)
	31 ans et +	321 (84.47 %)	59 (15.53 %)
Âge Reddit	Cat. 1 (1 an maxi)	174 (94.57 %)	10 (5.43 %)
	Cat. 2 (1-2 ans)	177 (86.76 %)	27 (13.24 %)
	Cat. 3 (2-3 ans)	247 (81.79 %)	55 (18.21 %)
	Cat. 4 (4-5 ans)	225 (80.94 %)	53 (19.06 %)
	Cat. 5 (6 ans et +	54 (71.05 %)	22 (28.95 %)
	Tous	877 (84 %)	167 (16 %)

7.5.2 Modèle

Nous avons créé un modèle de régression logistique avec la variable binaire dépendante « modération » ; elle comporte les niveaux « non » et « oui », et prédit donc la probabilité qu'une personne modère un forum. Les variables indépendantes sont le genre, l'âge, leur interaction, et l'âge Reddit. Nous avons utilisé la fonction `step()` pour savoir si toutes les variables contribuaient au modèle. La fonction a retiré l'interaction du genre et de l'âge. Le modèle est présenté dans le tableau 7.5. Les hommes cisgenres sont le niveau de référence.

L'âge Reddit et l'âge sont tous deux corrélés avec le fait d'être modérateur·trice ou non. Les comptes créés depuis moins d'un an sont moins susceptibles que les autres d'être modérateurs. La catégorie 5 (6 ans et plus) compte davantage de modérateur·trices que les catégories 1, 2 et 3. Il n'y a pas de différence avec la catégorie 4. La proportion de modérateur·trices augmente donc globalement avec l'âge Reddit : pour 1 modérateur·trice de

Tableau 7.5 – Modération de forums, régression logistique binaire

	Variable dépendante :
	Modération
Intercept	0.127* (0.057, 0.256)
Femmes cisgenres	0.548* (0.357, 0.830)
Femmes transgenres	0.920 (0.466, 1.719)
Hommes transgenres	0.802 (0.394, 1.540)
Non-binaires	1.070 (0.591, 1.878)
21-30 ans	0.425* (0.265, 0.689)
31 ans et +	0.388* (0.233, 0.650)
Âge Reddit cat. 2	2.880* (1.381, 6.472)
Âge Reddit cat. 3	4.600* (2.320, 10.025)
Âge Reddit cat. 4	5.153* (2.570, 11.331)
Âge Reddit cat. 5	8.932* (3.884, 21.814)
Observations	1,044
Log Likelihood	−429.993
Akaike Inf. Crit.	881.986
Note :	*p<0.05 ; **p<0.01

la catégorie 5, il y en a seulement 0.11 dans la catégorie 1, 0.32 dans la catégorie 2 et 0.51 dans la catégorie 3. On constate l'effet inverse pour l'âge. Il y a davantage de modérateur·trices chez les Redditors les plus jeunes que chez les Redditors des autres catégories d'âge. Il n'y a en revanche pas de différence significative entre les Redditors de 21 à 30 ans et les Redditors de 31 ans et plus.

Les femmes cisgenres occupent moins souvent la fonction de modératrice que les hommes cisgenres et les personnes non binaires. La probabilité d'être modérateur·trice est ainsi deux fois plus importante pour les hommes cisgenres et les personnes non binaires que pour les femmes cisgenres. Il n'y a pas de différence significative entre les femmes cisgenres, les femmes transgenres et les hommes transgenres.

7.5.3 Nombre de subreddits par modérateur·trice

Le diagramme en barres 7.3 montre le nombre de modérateur·trices dans chaque catégorie. Les femmes transgenres, les hommes transgenres et les personnes non binaires ont été rassemblé·es dans une même catégorie, pour que le graphique soit plus lisible. La majorité des modérateur·trices (115 sur 165) modèrent seulement 1 subreddit. Dans cette catégorie (1 subreddit), les trois groupes sont présents dans des proportions similaires. En revanche, les hommes cisgenres semblent dominer dans la catégorie 2 ; 23 hommes cisgenres (sur 372) modèrent 2 à 3 subreddits, contre 4 femmes cisgenres (sur 372). C'est également le cas dans la catégorie 3 (9 hommes cisgenres contre 3 femmes cisgenres). Les effectifs très réduits dans cette catégorie (18 modérateur·trices au total) incitent toutefois à la prudence quant à l'interprétation de ce résultat.

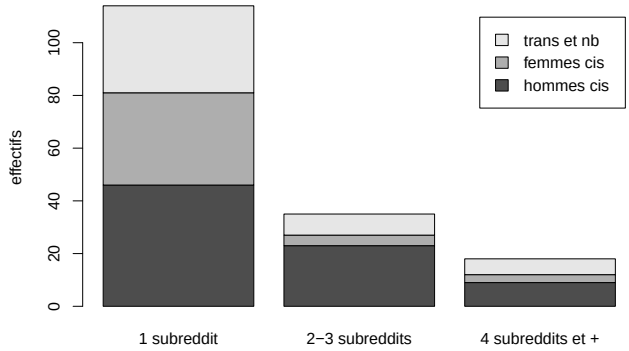

FIGURE 7.3 – Nombre de modérateur·trices modérant 1, 2-3 et plus de 4 subreddits

7.6 Analyse du karma : étude longitudinale

7.6.1 Données

Lors de la construction du corpus, nous avons pris note du nombre de points de karma de commentaire (points obtenus lorsque l'on commente dans un fil de discussion) et de karma de post (points que l'on obtient quand on lance un fil de discussion, → p. 91) que les Redditors de RedditGender avaient accumulés. Mi-août 2018 et mi-septembre 2019, nous avons accédé à nouveau aux profils des Redditors de RedditGender et nous avons procédé une nouvelle fois à ce relevé. Nous avons ensuite soustrait les scores de 2018 aux scores de 2019, ce qui nous a permis de savoir combien de points chaque personne a obtenus entre ces deux dates, c'est-à-dire en l'espace de 13 mois environ. Nous n'avons pas utilisé les scores de 2017, car ils ont été relevés sur une période de plusieurs mois, ne permettant donc pas de comparer les Redditors entre eux. Ce relevé n'a pas pu être effectué pour les Redditors ayant supprimé leur compte. Dans certains cas également, il y a eu des erreurs ou des absences de relevé en 2018, ce qui a réduit le nombre d'observations à 941. Les données utilisées pour faire les analyses statistiques sont décrites en annexe (→ p. 393).

7.6.2 Statistiques descriptives

Une première exploration graphique des données à l'aide de boites à moustaches (figure 7.4) révèle quatre valeurs de karma de post extrêmes, supérieures à 90 000, alors que le score médian, pour tous les Redditors, est de 158. Ces observations correspondent à une femme cisgenre, qui a un karma de post de 479 577, à deux hommes cisgenres (scores : 188 271 et 93 876), et à un homme transgenre (score : 554 375). Nous avons donc retiré ces quatre observations des données, y compris pour l'analyse du karma

de commentaire (pour des raisons pratiques). D'autres valeurs très élevées subsistent, tout comme pour le karma de commentaire, mais elles bien sont plus proches de la médiane. Nous avons donc choisi de les conserver.

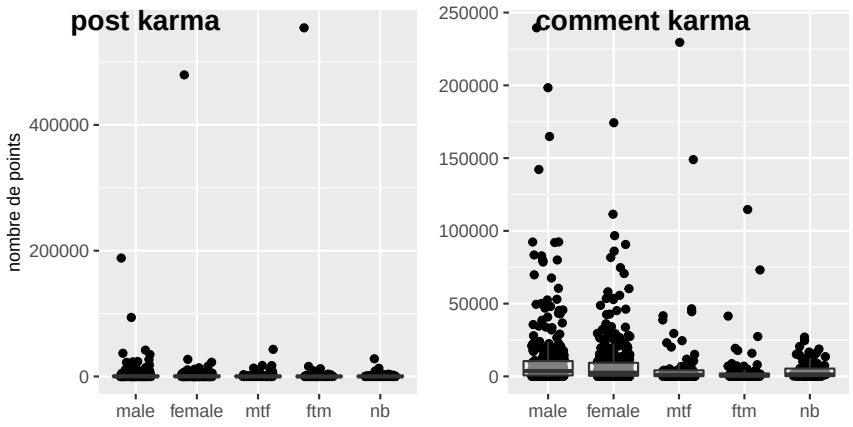

FIGURE 7.4 – Karma de post et de commentaire, boites à moustaches

Le tableau 7.6 montre que, pour le karma de post, les trois quarts des valeurs sont en dessous de 864 (le troisième quartile), avec un maximum de 43 115. Les scores de karma de commentaire sont plus élevés, avec une médiane approximativement 13 fois plus importante que celle du karma de post (2101). Le maximum est de 239 515, mais 75 % des valeurs sont en dessous de 8029. À cause de cette dispersion importante, nous avons choisi de rapporter dans le tableau 7.7 les médianes et non les moyennes de chaque groupe.

TABLEAU 7.6 – Statistiques descriptives, karma de post et de commentaire

	Karma de post	Karma de commentaire
Min.	-18	-49
1st Qu.	6	304
Médiane	154	2101
Moyenne	1382.92	9017.92
3rd Qu.	864	8029
Max.	43115	239515

Ce tableau montre que les femmes cisgenres ont le karma de post médian le plus élevé (190). Viennent ensuite les hommes cisgenres, les femmes transgenres, puis les personnes non binaires et enfin, assez loin derrière, les hommes transgenres avec un karma de post de 48. L'écart entre les groupes est plus important pour le karma de commentaire : les hommes cisgenres sont au premier rang, avec une médiane de 3550, suivis par les femmes cisgenres (2478). Notons que ces deux groupes présentent également la dispersion la plus importante, avec des écarts interquartiles respectifs de 9759.75 et 8897, suggérant des différences individuelles plus prononcées que dans les autres groupes. Les groupes transgenres ont des

scores de karma de commentaire beaucoup plus faibles : entre 1332 pour les non binaires, et 525 pour les hommes transgenres.

Le karma de post est le plus élevé chez le groupe le plus jeune, avec une médiane de 222, contre 119 pour les Redditors de 21 à 30 ans et de 161 pour ceux de 31 ans et plus. Pour le karma de commentaire, il semble que le phénomène inverse se produise : les plus jeunes ont la médiane la plus faible, et les plus âgés la médiane la plus élevée. Enfin, les modérateur·trices ont un score de karma de post 4.44 fois plus élevé que les Redditors qui ne modèrent pas de subreddit, et un score de karma de commentaire 1.69 fois plus élevé.

Tableau 7.7 – Karma de post et de commentaire médians par groupes de genre et d'âge

	Karma de post		Karma de commentaire	
	Médiane	EI	Médiane	EI
Hommes cisgenres	161.50	950.75	3550.00	9759.75
Femmes cisgenres	190.00	1035.00	2478.00	8897.00
Femmes transgenres	112.00	732.00	1109.00	4064.00
Hommes transgenres	48.00	423.00	525.00	1939.00
Non-binaires	100.00	681.50	1332.00	5159.00
14-20 ans	222.00	1478.00	1350.00	7898.00
21-30 ans	119.00	789.00	1736.00	6119.00
31 ans et +	161.00	697.00	2950.00	9152.00
Non-modérateur	114.00	693.75	1957.00	7133.00
Modérateur	507.00	2528.00	3301.00	14581.00

7.6.3 Effets de l'âge et du genre sur l'obtention de points de karma

Les données à analyser étant issues de comptage et étant fortement dispersées, nous avons opté pour des modèles binomiaux négatifs. Comme ces modèles ne peuvent pas prendre en compte de valeurs négatives (Hilbe, 2011), nous les avons remplacées par des zéros. Il y en avait 5 pour le karma de post et 5 pour le karma de commentaire (indiquant que les contributions de certain·es Redditors ont reçu plus de votes négatifs que de votes positifs). Nous avons intégré aux modèles plusieurs variables, toutes catégorielles : le genre, l'âge, et leur interaction, la catégorie d'âge Reddit, et le fait qu'un Redditor soit modérateur·trice ou non. Nous avons utilisé la fonction `step()` pour savoir si ces variables contribuaient toutes au modèle. L'âge Reddit n'a pas été conservé. Les niveaux de référence des modèles sont les hommes cisgenres de 14 à 20 ans.

Karma de post

Les coefficients du modèle, avec les valeurs p et les intervalles de confiance de 95 %, sont présentés dans le tableau 7.8. Le modèle montre que

le fait de modérer un ou plusieurs subreddits a un effet significatif sur le nombre de points de karma de post obtenus. Quand un Redditor non modérateur·trice obtient 1 point de karma, un modérateur en accumule 2.49.

Tableau 7.8 – Karma de post, modèle de régression binomial négatif

	Variable dépendante :
	Karma de post
Intercept	2,227.104** (1,280.709, 4,365.295)
Femmes cisgenres	0.351* (0.133, 1.009)
Femmes transgenres	1.965 (0.543, 11.493)
Hommes transgenres	0.567 (0.200, 1.889)
Non-binaires	0.246* (0.073, 1.229)
21-30 ans	0.415* (0.193, 0.833)
31 ans et +	0.723 (0.344, 1.405)
Modérateur·trice	2.495** (1.703, 3.776)
Femmes cisgenres :21-30 ans	3.187* (0.986, 9.574)
Femmes transgenres :21-30 ans	0.474 (0.072, 2.135)
Hommes transgenres :21-30 ans	0.816 (0.206, 2.940)
Non-binaires :21-30 ans	5.332* (0.945, 22.284)
Femmes cisgenres :31 ans et +	1.228 (0.381, 3.704)
Femmes transgenres :31 ans et +	0.271 (0.040, 1.340)
Hommes transgenres :31 ans et +	0.203 (0.041, 1.201)
Non-binaires :31 ans et +	1.887 (0.297, 10.349)
Observations	937
Log Likelihood	−6,296.696
θ	0.201** (0.008)
Akaike Inf. Crit.	12,625.390
Note :	*p<0.05 ; **p<0.01

L'âge a également un effet significatif sur l'obtention de points de karma de post, mais, l'interaction avec le genre étant significative, l'effet n'est pas uniforme. Les hommes cisgenres, les hommes transgenres et les femmes transgenres les plus âgé·es obtiennent moins de karma de post que les plus jeunes ; l'effet de l'âge n'est pas significatif quand on compare le groupe 2 (21 à 30 ans) avec les autres groupes. Pour ces groupes, il semble donc y avoir une corrélation négative entre âge et score de karma de post, mais qui n'atteint pas toujours le niveau de significativité. Il n'y a en revanche pas de corrélation entre âge et karma de post pour les femmes cisgenres et les personnes non binaires.

Chez les Redditors les plus jeunes, les femmes cisgenres et les personnes non binaires accumulent significativement moins de points de karma de post que les hommes cisgenres et les femmes transgenres. L'effet est particulièrement important chez les non-binaires : pour chaque point de karma accumulé par les hommes cisgenres, les non-binaires en obtiennent seulement 0.25. Dans le groupe d'âge 2 (21 à 30 ans), les hommes transgenres ont obtenu moins de points de karma de post que les hommes cisgenres, les femmes cisgenres et les non binaires. Dans le groupe d'âge 3 (31 ans et plus), les hommes transgenres obtiennent moins de points que tous les autres groupes. La différence est particulièrement marquée avec les hom-

mes cisgenres : selon le modèle, pour chaque point obtenu par un homme transgenre, un homme cisgenre en obtient 8.69. Il y a également une différence significative entre hommes cisgenres et femmes cisgenres : quand une femme cisgenre accumule 1 point de karma de post, un homme en obtient 2.32. Les interactions de l'âge et du genre sont présentées dans la figure 7.5. On peut y voir la corrélation négative de l'âge et du score de karma de post chez les hommes transgenres, les femmes transgenres et les hommes cisgenres.

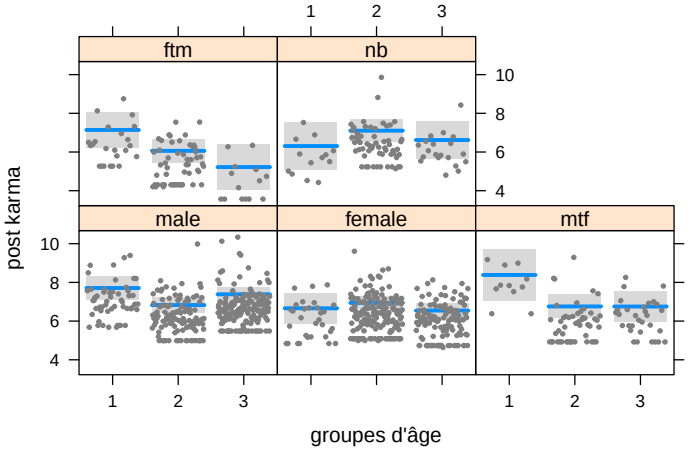

FIGURE 7.5 – Interaction du genre et de l'âge, karma de post

Karma de commentaire

La variable indépendante « âge Reddit » a, comme pour le karma de post, été retirée du modèle par la fonction `step()`. Ce modèle ne montre pas de corrélation positive entre le fait de modérer des forums et le score de karma de commentaire.

L'interaction de l'âge et du genre, présentée dans la figure 7.6, est significative. La corrélation entre âge et obtention de points de karma de commentaire n'est pas clairement définie, et varie selon les groupes. Elle est inexistante chez les personnes non binaires et les hommes cisgenres. Les femmes transgenres les plus jeunes obtiennent significativement plus de points de karma que les deux groupes de femmes transgenres plus âgées. Les femmes cisgenres et les hommes transgenres les plus âgé·es, en revanche en récoltent plus que les groupes de 21 à 30 ans.

Chez les Redditors les plus jeunes, les femmes transgenres obtiennent significativement plus de points que toutes les autres catégories : 5.26 plus que les femmes cisgenres, 4.46 que les hommes cisgenres, 9.88 fois plus que les personnes non binaires, et 11.22 fois plus que les hommes transgenres. Dans la catégorie des Redditors de 21 à 30 ans, les hommes cisgenres obtiennent plus de points de karma de commentaire que les groupes transgenres et non binaire. L'écart est particulièrement fort avec les hommes

TABLEAU 7.9 – Karma de commentaire, modèle de régression binomial négatif

	Variable dépendante :
	Karma de commentaire
Intercept	8,596.310** (5,383.345, 14,908.510)
Femmes transgenres	0.847 (0.378, 2.015)
Femmes transgenres	4.459* (1.507, 18.341)
Hommes transgenres	0.397* (0.166, 1.063)
Non-binaires	0.451 (0.161, 1.645)
21-30 ans	1.108 (0.591, 1.985)
31 ans et +	1.509 (0.819, 2.630)
Modérateur·trice	1.299 (0.940, 1.836)
Femmes transgenres :21-30 ans	0.784 (0.297, 1.965)
Femmes transgenres :21-30 ans	0.086** (0.019, 0.305)
Hommes transgenres :21-30 ans	0.569 (0.182, 1.670)
Non-binaires :21-30 ans	1.035 (0.254, 3.437)
Femmes transgenres :31 ans et +	1.206 (0.458, 3.033)
Femmes transgenres :31 ans et +	0.093** (0.020, 0.353)
Hommes transgenres :31 ans et +	1.280 (0.334, 5.499)
Non-binaires :31 ans et +	0.391 (0.086, 1.602)
Observations	937
Log Likelihood	−8,590.890
θ	0.289** (0.011)
Akaike Inf. Crit.	17,213.780
Note :	*p<0.05 ; **p<0.01

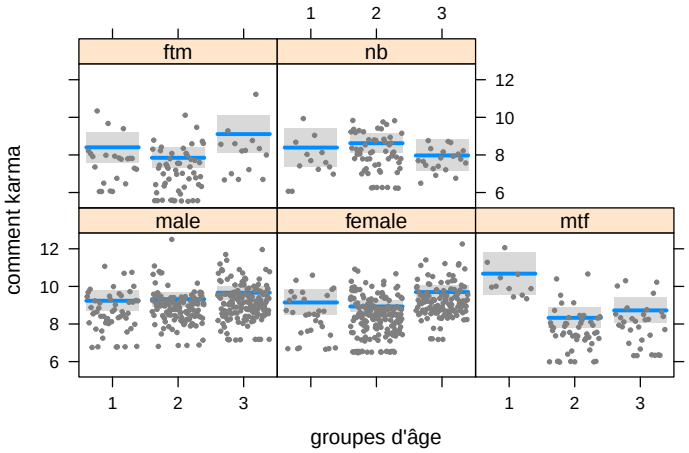

FIGURE 7.6 – Interaction du genre et de l'âge, karma de commentaire

transgenres, qui récoltent 0.23 point quand les hommes cisgenres en ont 1. Les femmes cisgenres touchent également davantage de points que les hommes transgenres et les femmes transgenres. Les hommes transgenres accumulent moins de points que tous les groupes à l'exception des femmes transgenres. Enfin, chez les Redditors plus âgé·es, les hommes et les femmes cisgenres accumulent davantage de points de karma de commentaire que les personnes non binaires et les femmes transgenres. Il n'y a pas de différence entre les groupes transgenres.

Scores de karma et ethnicité

Nous avons réalisé ces mêmes analyses sur l'échantillon réduit décrit page 135, qui ne contient que les internautes cisgenres, afin d'examiner l'effet éventuel de l'ethnicité. Par souci de concision, ces modèles sont présentés en annexe (→ p. 394). Aucun effet significatif n'a été révélé par les modèles de régression. Nous pouvons en conclure qu'il n'y a pas de corrélation significative entre ethnicité et obtention de points de karma de post et de commentaire.

7.7 Discussion

7.7.1 Pseudonymes

Notre analyse montre que, sur Reddit, les pseudonymes sont peu genrés : moins d'1 pseudonyme sur 5, dans notre corpus, indique un genre de façon explicite et sans ambiguïté. Ce phénomène avait déjà été remarqué par Thelwall et Stuart (2018), qui, en utilisant une liste des 10 000 prénoms les plus courants aux États-Unis, avaient seulement pu déterminer le genre de 4.9 % des auteur·es des messages de leur corpus. Bucholtz (2002) avait également noté, dans son étude du site geek Slashdot, que les pseudonymes choisis par les utilisateur·trices avaient peu fréquemment un lien avec le genre.

Contrairement à ce que nous pensions, les hommes cisgenres ne sont pas ceux qui choisissent le plus fréquemment des pseudonymes reflétant leur genre, et il n'y a pas de différence significative entre les hommes cisgenres et les femmes cisgenres. Étonnamment, les femmes transgenres sont les plus susceptibles de choisir un pseudonyme reflétant leur identité de genre. Quand on les compare aux différents groupes, c'est avec le groupe des hommes transgenres que la différence est la plus forte : les femmes transgenres sont 4.55 fois plus susceptibles de choisir un pseudonyme genré que les hommes transgenres.

Il est peut-être possible de trouver une explication à ce résultat dans l'expérience particulière des femmes transgenres. Parce qu'il efface le corps, internet offre aux personnes transgenres la possibilité de vivre leur véritable identité de genre, là où le « monde réel » est souvent inhospitalier pour elles et où elles sont parfois obligées de cacher qui elles sont (Whittle, 1998). Marciano écrit ainsi que :

> « Some transgender users employ cyberspace as an alternative sphere that constitutes a parallel world that provides its inhabitants with different and sometimes contradictory experiences from those available in the offline world » (Marciano, 2014, p. 830).

Dans cet espace désincarné et anonyme, les femmes transgenres peuvent vivre comme de « vraies » femmes. Cet espace est sans doute plus précieux pour les femmes transgenres que pour les hommes transgenres. Ceux-ci ont moins de problèmes à « passer pour » (être considérés comme) des hommes, grâce aux effets puissants de la testostérone, qui modifie rapidement leur voix et leur apparence physique. Les femmes transgenres sont ainsi plus souvent isolées et harcelées que les hommes transgenres, et ont davantage tendance à cacher leur identité de genre dans la vie réelle (Rankin & Beemyn, 2012). Dans ce contexte, on peut imaginer qu'il est plus essentiel pour les femmes transgenres d'affirmer leur identité de genre par le choix de leur pseudonyme sur un site comme Reddit. Il est possible que davantage de femmes transgenres de notre corpus vivent « cachées » dans la vie « hors ligne » pour diverses raisons, et qu'elles ne puissent pas se faire appeler par un prénom féminin. Choisir un pseudonyme explicitement féminin serait donc une façon pour elles d'effacer leur *dead name* (le nom masculin qu'on leur a donné à la naissance), et d'affirmer haut et fort leur identité de genre dans la communauté de Reddit.

7.7.2 Âge Reddit

L'analyse des corrélations avec les diagrammes en mosaïque semble indiquer, comme nous nous y attendions, que les hommes cisgenres sont « installés » sur le site depuis plus longtemps que les autres Redditors, puisqu'ils ont les comptes les plus anciens. Les femmes cisgenres de RedditGender semblent être arrivées sur le site plus récemment. Il semble par ailleurs y avoir eu un afflux récent de femmes et d'hommes transgenres sur Reddit, peut-être dû à l'éclosion ou à la montée en popularité des communautés transgenres sur le site. En revanche, il n'existe pas de corrélation entre âge Reddit et ethnicité. Les personnes non blanches de notre corpus ne sont donc pas arrivées plus tard sur le site que les internautes blanc·hes. Cela suggère que les Redditors afro-américain·es, hispaniques et asiatiques de RedditGender sont aussi familier·es avec le fonctionnement du site et ses usages linguistiques que les Redditors blanc·hes, alors qu'il est possible que ce ne soit pas le cas pour les hommes et les femmes transgenres, comparé·es aux hommes cisgenres.

7.7.3 Longévité des profils

En l'espace de trois ans, seuls 10.73 % des comptes Reddit du corpus ont été supprimés par leurs propriétaires. Il semble donc que les internautes de notre corpus, que nous avions choisis en 2017 parce qu'ils étaient des Redditors assidu·es et écrivaient beaucoup de commentaires, soient pour la plupart attaché·es à leurs comptes. Notre analyse a révélé qu'il n'y a pas de différence entre les différents groupes de genre dans la longévité des

comptes Reddit. Le fait que Reddit soit un site geek et majoritairement masculin ne signifie donc pas que les hommes cisgenres conservent leurs comptes plus longtemps. Malgré les informations sensibles que contiennent leurs historiques de commentaires, les personnes transgenres ne sont pas non plus plus susceptibles que les hommes cisgenres de supprimer leurs comptes. Cela peut paraitre étonnant, au vu de l'extrême facilité avec laquelle on peut supprimer un compte et en créer un autre. Il semble donc que, tout au moins chez les Redditors prolifiques comme celles et ceux qui composent notre corpus, il est important de maintenir la même identité en ligne. En effet, la longévité d'un compte dans une communauté en ligne est un des facteurs qui déterminent l'influence et le statut d'un·e utilisateur·trice dans la communauté (Huffaker, 2010).

7.7.4 Modération

Notre analyse a montré que les comptes Reddit les plus récents (existant depuis moins d'un an) étaient les moins susceptibles de modérer un ou plusieurs subreddits, alors que les plus anciens (6 ans et plus) étaient plus fréquemment modérateurs. Le fait que les comptes les plus récents soient rarement modérateurs peut être dû à la courbe d'apprentissage de la culture Reddit (Kilgo et al., 2016). Le fonctionnement du site, tout comme son langage, n'est pas transparent et doit être appris par les internautes, ce qui prend du temps.

En même temps, les Redditors les plus jeunes (14 à 20 ans) sont plus fréquemment modérateur·trices que les plus âgé·es (les comptes Reddit les plus âgés ne sont pas forcément ceux des internautes les plus âgés). Même si Reddit est un site jeune (64 % des utilisateur·trices adultes ont moins de 29 ans, Barthel et al., 2016b), il peut paraitre étonnant que les adolescent·es et les très jeunes adultes soient plus susceptibles de modérer des forums que les adultes plus âgés. Il est possible que les jeunes Redditors aient plus de facilité à comprendre le fonctionnement du site, ou qu'ils soient plus enthousiastes et souhaitent s'impliquer dans la communauté en devenant modérateur·trices volontaires. Il se peut aussi que les modérateur·trices les plus jeunes modèrent des subreddits moins prestigieux et moins fréquentés que les modérateur·trices plus âgé·es, c'est-à-dire des subreddits qui requièrent moins de travail. Tout comme l'âge Reddit ou le karma, le fait d'être modérateur·trice peut être considéré comme une marque de prestige dans la communauté, notamment parce que ce statut permet plus facilement de créer des liens avec d'autres Redditors (Del Valle et al., 2020). Le prestige n'est évidemment pas le même quand on modère un subreddit fréquenté par 100 personnes ou un subreddit qui figure régulièrement sur la première page de Reddit et compte des centaines de milliers de membres.

En ce qui concerne le genre, il y a, comme nous nous y attendions, une différence significative entre femmes et hommes cisgenres. Les hommes cisgenres dominent le site par leur présence massive (Barthel et al., 2016b), mais aussi par le fait qu'ils sont plus souvent modérateurs que les femmes cisgenres. De plus, il semble également que, quand ils sont modérateurs,

les hommes cisgenres modèrent davantage de subreddits que les femmes cisgenres et les personnes transgenres et non binaires. Ils ont davantage de pouvoir sur le site que les autres modérateur·trices : ils contrôlent plusieurs forums, où ils peuvent bannir des utilisateur·trices, supprimer des messages, édicter des règles de bonne conduite et choisir le type de sujets acceptés (ou, au contraire, décider de ne rien censurer).

Ce déséquilibre entre femmes et hommes existe dans d'autres communautés en ligne, comme celle des éditeur·trices de Wikipedia (Lam et al., 2011). Menking et Erickson (2015) ont montré que, sur Wikipedia, la faible présence des femmes était liée à la charge émotionnelle auxquelles elles sont confrontées. Il semble que cela soit la même chose sur Reddit, comme le suggère l'étude de Dosono et Semaan (2019) : les femmes modératrices sont plus sujettes au « burn-out » que les hommes, notamment à cause des commentaires misogynes et du harcèlement qu'elles subissent. Sur Reddit, il semble donc que la modération soit encore, en grande partie, une affaire d'hommes. Cela est est problématique, parce que les modérateur·trices jouent un rôle de premier plan dans le contrôle du contenu publié sur le site (même si, aujourd'hui, les administrateurs de Reddit essaient d'intervenir davantage, comme nous l'avons expliqué page 104). Il faudrait cependant, pour avoir une image plus précise du phénomène, réaliser une étude plus large de la modération sur Reddit, car le nombre de modérateur·trices dans RedditGender est réduit. Notre analyse montre par ailleurs que les personnes non binaires sont aussi plus fréquemment modératrices que les femmes cisgenres. Il serait intéressant, pour mieux comprendre ce résultat, de savoir quels subreddits modèrent les hommes cisgenres et les personnes non binaires car, encore une fois, tous les subreddits n'ont pas le même poids sur le site.

7.7.5 Karma

Les *karma whores*

L'analyse du karma révèle des valeurs extrêmement élevées. Quatre Redditors (une femme cisgenre, deux hommes cisgenres dont un modérateur, et un homme transgenre) ont ainsi obtenu plus de 90 000 points de karma de post en 13 mois, soit 569 fois plus que le score médian de 158. Il semble que ces Redditors soient ce que l'on appelle des « karma whores », c'est-à-dire des Redditors qui sont souvent prêts à tout pour obtenir les scores les plus élevés possible, et qui passent un temps considérable sur le site (Burlage, 2019).

Karma et modération

La corrélation significative et positive entre le fait d'être modérateur·trice et le score de karma de post peut avoir plusieurs explications. Tout d'abord, il est très probable que les modérateur·trices écrivent davantage de *self-posts* que les autres Redditors. De nombreux subreddits ont des fils de discussion récurrents ou prévisibles, qui sont créés par les modé-

rateur·trices ; il peut s'agir d'« episode discussion threads » sur les subreddits consacrés à des séries télévisées (par exemple, sur r/gameofthrones), ou de « daily discussion threads » et de « weekend discussion threads » (par exemple, sur r/personalfinance ou r/wallstreetbets).

Les modérateur·trices créent également des *self-posts* pour informer leur communauté de décisions ou de nouveautés. Il semble par ailleurs logique que ces Redditors soient très impliqué·es dans des subreddits auxquels elles et ils ont choisi de consacrer du temps, toujours de façon bénévole, et lancent donc davantage de fils de discussion que la plupart des Redditors. Enfin, il est tout à fait possible que les *self-posts* écrits par les modérateur·trices soient bien accueillis dans ces communautés où elles et ils sont reconnus (les subreddits indiquent la liste de leurs modérateur·trices) et où elles et ils jouissent d'un certain prestige. Le fait que les modérateur·trices ne gagnent pas davantage de karma de commentaire suggère que, à l'intérieur des fils de discussion, leur comportement n'est pas différent de celui des autres Redditors.

Karma et âge Reddit

S'il n'y a pas de corrélation entre l'âge Reddit et le karma de commentaire, l'âge Reddit semble avoir un impact significatif sur le karma de post. Les comptes Reddit les plus récents (moins de 1 an) obtiennent significativement plus de points de karma que ceux qui ont entre 3 et 6 ans. Il est possible que ce résultat illustre la différence d'état d'esprit entre les Redditors expérimentés, pour qui accumuler du karma n'est pas essentiel, et les Redditors arrivés récemment sur le site, qui sont séduits par le système du karma et en tirent profit pour se faire remarquer par la communauté (Richterich, 2014). Notons toutefois que la corrélation est partielle, et qu'il n'y a par exemple pas de différence entre les comptes très anciens (de plus de 6 ans) et les comptes les plus récents.

Karma et interaction de l'âge et du genre

Nos analyses n'ont pas montré de corrélation claire entre l'obtention de points de karma de post et de commentaire et l'âge. La corrélation est soit inexistante, soit partielle, soit négative pour certains groupes de genre et positive pour un autre. La réalité semble donc très contrastée, et les résultats sont difficiles à interpréter. On note toutefois que, pour certains groupes, l'âge est généralement corrélé négativement avec l'obtention de points de karma de post (pour les hommes cisgenres et les groupes transgenres) et le karma de commentaire (pour les femmes transgenres). On peut peut-être avancer la même explication que pour l'âge Reddit ; les utilisateur·trices les plus jeunes sont plus préoccupé·es par leur score de karma que les contributeur·trices plus âgés. Il est également possible qu'ils fréquentent de plus gros subreddits, où il est plus facile d'obtenir des scores de karma importants que sur des petits subreddits. Notons que, dans le cas du karma de post, l'interaction de l'âge et du genre a un effet inverse

chez les femmes cisgenres et les hommes transgenres ; le groupe plus âgé obtient davantage de karma que le plus jeune.

Comme nous en avions fait l'hypothèse, les hommes cisgenres sont, en général, ceux qui obtiennent le plus de points de karma de post. Cela peut être lié au fait qu'ils sont plus souvent modérateurs que les femmes cisgenres, comme nous l'avons montré. Il se peut également que les hommes cisgenres recherchent davantage le karma que les femmes cisgenres, ou qu'ils sont tout simplement plus à l'aise sur le site. Créer un *self-post* donne en effet davantage de visibilité à un·e Redditor que d'écrire un commentaire dans un fil de discussion ; les hommes cisgenres seraient donc moins « timides » que les autres groupes, et plus enclins à commencer des discussions. Il est également possible que les hommes cisgenres commentent sur des subreddits plus fréquentés que les autres Redditors, et engrangent donc davantage de points : les femmes cisgenres pourraient tout à fait commencer autant de discussions que les hommes cisgenres, mais le faire dans des communautés plus réduites, où il n'est pas possible de récolter autant de points. Les hommes transgenres de plus de 21 ans obtiennent quant à eux moins de karma de post que les autres groupes. Cela peut indiquer que le karma n'a pas d'importance pour eux, qu'ils lancent peu de discussions, ou qu'ils fréquentent des communautés de taille limitée.

Pour le karma de commentaire, il n'y a pas de différence entre femmes et hommes cisgenres. Les plus jeunes femmes transgenres sont celles qui obtiennent le plus de points de karma de commentaire ; les femmes transgenres plus âgées ne font en revanche pas partie des groupes qui récoltent le plus de points. Les jeunes femmes transgenres semblent donc avoir un comportement distinct, propice à l'obtention de karma de post et de commentaire, soit par la publication plus fréquente de commentaires, soit par la publication de commentaires jugés particulièrement pertinents, soit par la fréquentation de gros subreddits.

tl ;dr

Ce chapitre consacré aux variables de la « Reddidentité » donne un aperçu de la façon dont les Redditors créent leur identité en ligne et habitent l'espace de Reddit. Les interactions de l'âge et du genre étant significatives dans plusieurs cas, ces dynamiques sont parfois complexes et difficiles à résumer.

De nos analyses, nous dégageons toutefois plusieurs résultats marquants. Il y a, tout d'abord, le fait que les Redditors semblent attaché·es à leurs identités en ligne : en l'espace de trois ans, 10.73 % seulement les ont supprimées, malgré la facilité avec laquelle on peut créer de nouveaux comptes le site. Ensuite, si on considère que le pseudonyme est une composante essentielle de la Reddidentité, ce sont les femmes transgenres qui s'en saisissent le plus volontiers pour exprimer leur identité de genre.

Nos analyses de l'âge Reddit, de la modération, et, dans une certaine mesure, du karma, confirment le fait que les hommes cisgenres sont les plus installés dans la communauté, et qu'ils jouissent de davantage de prestige que les autres Redditors, à l'exception des jeunes femmes transgenres, qui accumulent beaucoup de karma et qui, nous le verrons dans le chapitre suivant, ont d'autres comportements inattendus.

Enfin, contrairement au genre, l'ethnicité n'est dans RedditGender pas corrélée au prestige des internautes (mesuré par le karma), ni à leur familiarité avec le site (mesurée par leur âge Reddit).

Chapitre 8

Mobilité et centres d'intérêt des Redditors

Dans ce chapitre, nous poursuivons notre exploration de Reddit par l'examen de la façon dont les Redditors investissent l'espace virtuel. Pour ce faire, nous avons cartographié leurs centres d'intérêt, mesuré leur mobilité sur le site, et étudié comment ils occupent l'espace textuel en analysant la longueur de leurs commentaires.

8.1 Hypothèses et questions de recherche

Étant donné que les hommes sont majoritaires sur Reddit, et qu'une grande partie du site reflète leurs intérêts (Massanari, 2017), nous émettons l'hypothèse qu'ils s'y sentent « chez eux » et qu'ils fréquentent un nombre de forums plus important que les autres groupes. Nous nous attendons à ce que les femmes cisgenres, les internautes transgenres et les personnes non binaires soient moins mobiles, et donc commentent sur un nombre plus réduit de subreddits. Il est également possible, comme Thelwall et Stuart (2018) l'ont trouvé dans leur étude d'un corpus de commentaires de Reddit, que les hommes cisgenres écrivent des commentaires plus courts que les femmes cisgenres. Nous n'avons pas d'hypothèse particulière quant à la longueur des commentaires rédigés par les personnes transgenres et non binaires.

Nous pensons par ailleurs que les centres d'intérêt étant traditionnellement considérés comme féminins ou masculins (→ p. 59) seront reflétés par l'analyse des catégories thématiques dans lesquelles femmes et hommes commentent le plus fréquemment. Selon cette hypothèse, les femmes fréquenteraient davantage les subreddits dédiés aux relations personnelles que les hommes, qui graviteraient plus vers les forums consacrés aux jeux vidéo, aux sports, à la politique ou encore à la pornographie. En ce qui concerne les Redditors transgenres et non binaires, nous nous demandons si leur comportement est plus proche du groupe cisgenre auquel elles et ils s'identifient, ou si leurs centres d'intérêt sont spécifiques et communs aux

autres groupes transgenres.

8.2 Étude de la mobilité des Redditors

Cette section explore le degré auquel les internautes de RedditGender se « déplacent » sur le site en se basant sur le nombre de subreddits sur lesquels ils ont écrit des commentaires.

8.2.1 Statistiques descriptives

Dans un premier temps, nous avons calculé le nombre de forums dans lesquels chaque personne a commenté, pour 100 commentaires. Nous avons donc multiplié le nombre de forums fréquentés par chaque personne dans RedditGender par 100, puis avons divisé le résultat par le nombre total de ses commentaires. La moyenne, pour l'ensemble des Redditors, est de 10.21 (écart type = 6.47). La médiane est de 9.17 (écart interquartile = 8.27). Les moyennes et les médianes, par groupes de genre et d'âge, sont présentées dans le tableau 8.1. Les femmes transgenres ont la moyenne la plus élevée, avec 13.26 forums fréquentés pour 100 commentaires. Il est par conséquent possible qu'elles se déplacent plus sur Reddit que les autres groupes. Les hommes transgenres et les personnes non binaires ont en revanche écrit des commentaires dans le moins de forums, avec respectivement 7.09 et 6.44 forums fréquentés pour 100 commentaires, soit environ deux fois moins que les femmes transgenres. Les boites à moustaches de la figure 8.1 permettent de visualiser l'écart entre les femmes transgenres d'un côté, et les hommes transgenres et les personnes non binaires de l'autre.

Tableau 8.1 – Nombres moyens et médians de forums dans chaque sous-corpus, pour 100 commentaires

Sous-corpus	Moyenne (et ET)	Médiane
Hommes cisgenres	11.31 (6.84)	9.93
Femmes cisgenres	10.13 (6.07)	9.46
Hommes transgenres	7.09 (4.94)	5.39
Femmes transgenres	13.26 (7.05)	11.60
Non-binaires	6.44 (3.95)	5.60
13 à 20 ans (N = 147)	9.32 (5.88)	8.63
21 à 30 ans (N = 517)	9.72 (6.20)	8.65
31 ans et + (N = 380)	11.22 (6.92)	9.82

8.2.2 Effet du genre sur la mobilité des Redditors

Choix du modèle de régression

Un modèle de régression a été créé avec, comme variable dépendante, le nombre de forums dans lequel chaque personne a écrit au moins un commentaire. Le genre et l'âge et leur interaction ont été inclus au modèle

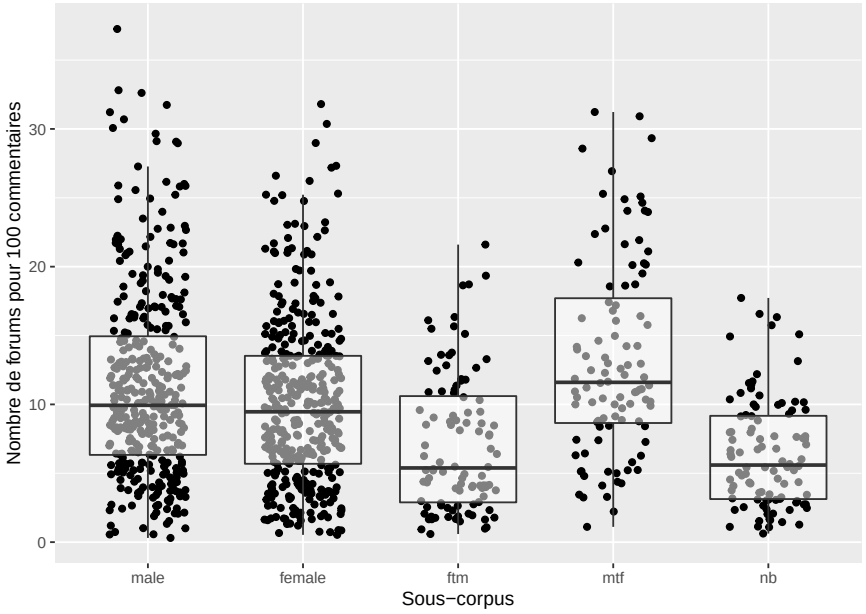

FIGURE 8.1 – Nombre de forums pour 100 commentaires, par groupe de genre

comme variables indépendantes. Nous avons ajouté dans le modèle un offset (→ p. 165), afin de prendre en compte le fait que chaque sous-corpus contient un nombre de commentaires différent (Zuur et al., 2015). Comme nous traitons ici des données de comptage, nous avons créé des modèles de régression de Poisson et binomial-négatif, et les avons comparé à l'aide du *boundary likelihood ratio test* (→ p. 164). Les résultats ont montré que le modèle négatif binomial était plus adapté ici. La fonction `step()` a éliminé l'interaction de l'âge et du genre. Les hommes cisgenres sont le niveau de référence du modèle.

Résultats

Le tableau 8.2 présente le modèle de régression créé. Il ne révèle pas de résultat significatif pour l'âge, ce qui semble indiquer que l'âge des Redditors n'a pas d'effet sur le nombre de forums qu'ils et elles fréquentent. En revanche, le genre des Redditors semble avoir un impact sur leur mobilité sur le site. Les coefficients exponentialisés sont inférieurs à 1 pour les femmes cisgenres (0.90), les hommes transgenres et les personnes non binaires. Cela signifie que, selon le modèle, ces internautes commentent sur un nombre de forums plus réduit que les hommes cisgenres, qui sont le niveau de référence du modèle. La taille d'effet est particulièrement importante pour les hommes transgenres (0.65) et les personnes non binaires (0.58). Ces dernières commentent dans moitié moins de forums que les hommes cisgenres. Le coefficient des femmes transgenres est en revanche supérieur à 1 (1.19), ce qui signifie qu'elles fréquentent un nombre de forums plus important que les hommes cisgenres. Elles se déplacent sur le

site davantage que tous les autres groupes, et particulièrement ceux des hommes transgenres et des personnes non binaires. Les femmes transgenres fréquentent ainsi deux fois plus de forums différents que ces deux groupes.

Tableau 8.2 – Nombre de forums fréquentés pour 100 commentaires, effets de l'âge et du genre

	Variable dépendante :
	Nombre de forums par 100 commentaires
Intercept	0.105^{**} (0.094, 0.118)
Femmes cisgenres	0.903^{*} (0.824, 0.989)
Hommes transgenres	0.650^{**} (0.564, 0.752)
Femmes transgenres	1.195^{*} (1.042, 1.375)
Non-binaires	0.585^{**} (0.508, 0.675)
21-30 ans	1.032 (0.917, 1.158)
31 ans et +	1.126 (0.995, 1.271)
Observations	1,044
Log Likelihood	−4,648.407
θ	2.755^{**} (0.126)
Akaike Inf. Crit.	9,310.814
Note :	*p<0.05 ; **p<0.01

8.3 Longueur des commentaires

8.3.1 Statistiques descriptives

La longueur moyenne des commentaires a été calculée pour chaque Redditor en divisant le nombre de tokens par le nombre de commentaires de chaque sous-corpus. Le nombre moyen de tokens par commentaire, dans l'ensemble du corpus, est de 53.32 tokens (écart type = 31.31) ; la médiane est 48.07 tokens par commentaire. Les moyennes et écarts types par groupe de genre et classes d'âge sont présentés dans le tableau 8.3.

Tableau 8.3 – Longueur des commentaires en nombre de tokens

Sous-corpus	Moyenne	Écart type
Hommes cisgenres	46.22	27.04
Femmes cisgenres	54.11	26.33
Hommes transgenres	68.24	38.69
Femmes transgenres	55.16	28.42
Non-binaires	60.07	29.66
14-20 ans	43.94	31.05
31-30 ans	54.83	29.79
31 ans et +	54.91	27.00
Tous	53.32	29.21

Ces statistiques descriptives révèlent plusieurs tendances. Les hommes cisgenres semblent écrire les commentaires les moins longs (46.22 tokens en moyenne), et les hommes transgenres les commentaires les plus longs (68.24 tokens en moyenne). Les internautes les plus jeunes produisent quant à eux des messages plus courts que les deux autres groupes d'âge (43.94 tokens en moyenne).

8.3.2 Modèle de régression

Les données présentent une forte asymétrie, comme on peut le voir dans l'histogramme présenté dans la figure 8.2. Peu de Redditors ont mis en ligne des commentaires très longs. Une majorité de Redditors a écrit des commentaires d'une longueur moyenne comprise entre 25 et 75 mots environ. Nous avons donc testé des modèles linéaires généralisés de Poisson et binomial négatif. Ce dernier s'est avéré plus efficace pour prendre en compte la forte dispersion des données. Pour créer le modèle, nous avons intégré comme variables indépendantes le genre et l'âge (leur interaction a été éliminée par le processus de sélection des variables). La variable dépendante est le nombre de tokens par commentaires, arrondi de façon à avoir des nombres entiers, condition nécessaire pour créer un modèle binomial négatif.

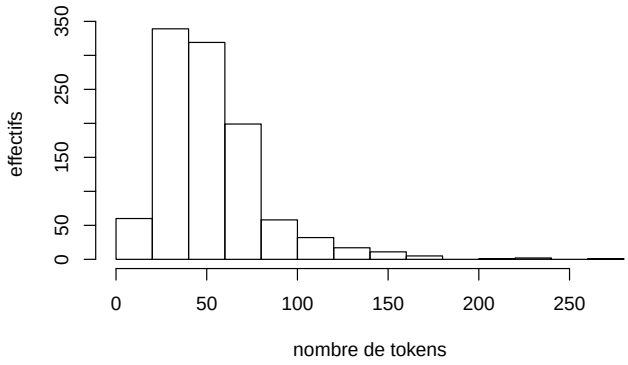

FIGURE 8.2 – Longueur des commentaires dans RedditGender

8.3.3 Résultats

Les coefficients, leurs valeurs p et les intervalles de confiance de 95 % sont présentés dans le tableau 8.4. Le niveau de référence du modèle est les femmes cisgenres. Le modèle montre que les hommes cisgenres écrivent des commentaires plus courts que tous les autres Redditors. Les femmes cisgenres écrivent des commentaires 1.16 fois plus longs qu'eux. C'est toutefois quand on compare les hommes transgenres aux hommes cisgenres que l'on remarque la différence la plus importante : les commentaires des hommes

transgenres contiennent 1.57 fois plus de tokens que les commentaires des hommes cisgenres. Les hommes transgenres écrivent par ailleurs des commentaires plus longs que tous les autres Redditors, à l'exception des non binaires. Il n'y a pas de différence entre les femmes cisgenres et les femmes transgenres.

Tableau 8.4 – Longueur des commentaires, effets de l'âge et du genre

	Variable dépendante :
	Nombre de tokens/commentaire
Intercept	42.653** (38.917, 46.814)
Hommes cisgenres	0.858** (0.800, 0.920)
Femmes transgenres	1.026 (0.924, 1.142)
Hommes transgenres	1.345** (1.209, 1.498)
Non-binaires	1.184** (1.066, 1.318)
21-30 ans	1.252** (1.143, 1.369)
31 ans et +	1.361** (1.238, 1.494)
Observations	1,044
Log Likelihood	−4,779.102
θ	4.720** (0.219)
Akaike Inf. Crit.	9,572.204
Note :	*p<0.05 ; **p<0.01

Les résultats révèlent par ailleurs une corrélation positive et significative entre âge et longueur des commentaires : les Redditors de 14 à 20 ans écrivent des commentaires plus courts que les Redditors de 21 à 30 ans, qui écrivent des commentaires plus courts que les plus de 31 ans. Les commentaires des plus âgé·es comportent ainsi 1.36 fois plus de tokens que ceux des plus jeunes.

8.4 Centres d'intérêt

Pour répondre à nos questions de recherche, nous avons utilisé l'analyse factorielle des correspondances simples, une technique présentée dans la section 6.2.2, qui permet d'étudier les corrélations entre deux variables catégorielles (ici, le genre et les thèmes des forums).

8.4.1 Moyennes par groupes

Le tableau 8.5 présente le pourcentage de commentaires mis en ligne par chaque groupe dans les 12 catégories recensées et décrites dans le chapitre 5.3.2. Les noms des catégories sont en anglais, par souci de cohérence avec le graphique présenté plus bas. Le tableau montre que les forums « personal advice » (thèmes personnels) (40.2 %) et « general interest » (forums généraux) (21.6 %) sont les plus fréquentés par les internautes de RedditGender : à elles deux, ces catégories représentent 61.8 % des commentaires contenus dans le corpus. D'autres catégories sont beaucoup moins

fréquentes dans RedditGender ; c'est notamment le cas des forums consacrés à la technologie (1.8 %), des forums « x-rated » (1.4 %), et des forums « divers » (1.2 %). Quand on compare les groupes de genre entre eux, on voit quelques tendances se dessiner. Les hommes cisgenres, par exemple, semblent moins fréquenter les forums dédiés aux thèmes personnels que les autres internautes, et les personnes transgenres semblent moins attiré·es par les forums généraux que les autres groupes.

Tableau 8.5 – Pourcentage de commentaires mis en ligne dans chaque catégorie, par sous-corpus

Catégories	**M**	**F**	**MTF**	**FTM**	**NB**	**Tous**
General interest	26.7	22.2	21.5	13.1	8.6	21.6
Humor	5.3	5.7	7.8	4.1	8.7	5.9
Gaming	7.9	3.5	6.1	3.2	6.5	5.6
News & politics	7.4	3.6	6.3	2.6	3.7	5.1
Science & education	5.0	4.1	3.6	2.6	2.4	4.0
Technology	3.3	0.8	2.2	0.5	1.1	1.8
Mass entertainment	6.6	6.5	7.8	2.6	1.9	5.8
Hobbies	4.8	4.6	3.3	0.8	1.2	3.9
Personal advice	23.4	44.3	37.0	67.1	63.9	40.2
Sports & fitness	5.8	3.0	1.6	1.8	0.7	3.5
X-rated	2.4	0.5	1.6	0.6	1.0	1.4
Others	1.4	1.2	1.2	1.1	0.3	1.2
Total	100	100	100	100	100	100

8.4.2 Analyse factorielle des correspondances simples

L'analyse factorielle des correspondances simples a été effectuée avec un tableau de contingence contenant les fréquences absolues (brutes) des commentaires mis en ligne dans chaque catégorie de forums par chaque groupe de genre. Ce tableau est présenté en annexe B. Avant de procéder à l'analyse factorielle des correspondances simples, un test du χ^2 a été effectué. La valeur p étant inférieure à 0.001, nous avons pu rejeter l'hypothèse nulle du test du χ^2 : il y a bien une ou plusieurs relations de dépendance entre les variables. Nous avons ensuite réalisé le graphe d'AFC présenté dans la figure 8.3.

Ce graphe montre que les deux premières dimensions de l'analyse capturent 94.71 % de la variation contenue dans les données, ce qui est très satisfaisant. Nous avons interprété les relations entre les groupes de genre et les thèmes des forums en nous basant sur la méthode présentée dans la section 6.2.2. Des lignes reliant l'origine du graphe à chaque point ont été tracées ; quand les lignes reliant un sous-corpus à l'origine du graphe sont longues et quand l'angle qu'elles forment est aigu, cela signifie que l'association entre les variables est positive et forte (Yelland, 2010). Le graphe suggère que les comportements des femmes et des hommes cisgenres sont différents, car la distance entre les deux points est importante. Les hommes transgenres et les individus non binaires sont relativement proches les uns

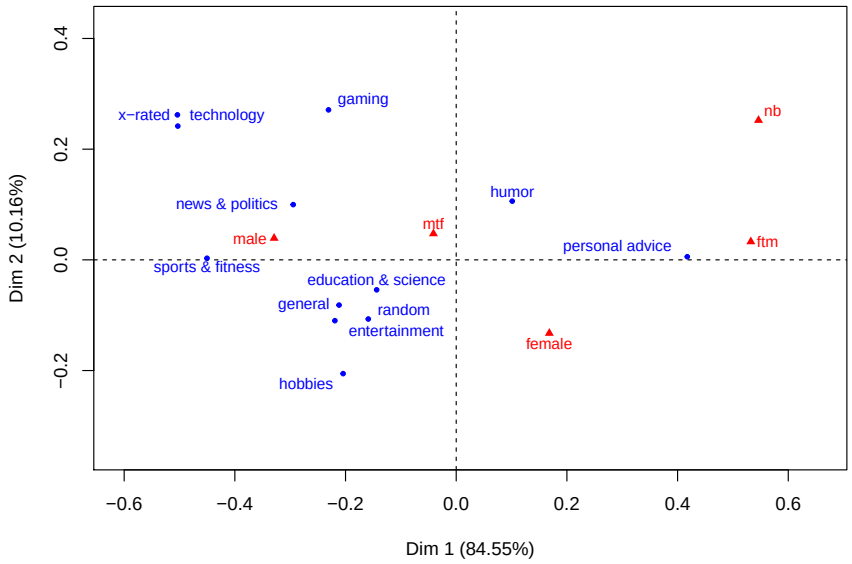

FIGURE 8.3 – Graphe d'analyse des correspondances simples : centres d'intérêt des Redditors

des autres, ce qui suggère un comportement similaire. Les femmes transgenres sont situées près de l'origine du graphe, ce qui indique qu'elles sont le groupe le plus indistinct.

Il y a une association positive entre les hommes cisgenres et 10 catégories de forums sur 12. Les hommes cisgenres sont particulièrement associés aux forums traitant de technologie, de thèmes « x-rated », de jeux vidéo et d'actualité et politique. Par contraste, ils commentent moins que l'ensemble des Redditors de RedditGender sur les forums traitant de sujets personnels. Les femmes transgenres commentent quant à elles plus fréquemment que l'ensemble des Redditors sur 11 thèmes sur 12. Cette association est toutefois faible, car la ligne reliant les femmes transgenres à l'origine du graphe est courte. Les femmes cisgenres sont associées positivement avec une seule catégorie de forums : les forums traitant de sujets personnels. Les hommes transgenres et les personnes non binaires sont également associé·es avec cette catégorie, ainsi qu'avec les subreddits humoristiques. L'association avec les subreddits dédiés aux sujets personnels est particulièrement forte pour les hommes transgenres.

8.5 Discussion

8.5.1 Femmes et hommes cisgenres

Les résultats présentés dans ce chapitre semblent confirmer le statut d'espace masculin de Reddit. Dans notre analyse de la mobilité, le modèle

de régression montre que les femmes cisgenres fréquentent moins de subreddits que les hommes cisgenres, et donc se déplacent moins sur le site ; elles écrivent des commentaires plus longs, sur un nombre plus réduit de forums. Leur « territoire » semble donc plus petit que celui des hommes, ce que l'on pourrait expliquer par le caractère misogyne d'une grande partie du site, qui est souvent inhospitalier pour les femmes (Massanari, 2017).

L'analyse des correspondances simples a montré que les hommes cisgenres écrivent plus fréquemment des commentaires que l'ensemble des Redditors de RedditGender sur 10 types de forums, tandis que les femmes cisgenres sont associées avec seulement un seul type de subreddits. Les hommes cisgenres sont donc mobiles non seulement en termes de nombre de subreddits fréquentés, mais aussi en termes du nombre de types de forums qu'ils visitent.

Comme nous en avions fait l'hypothèse et comme Thelwall et Stuart (2018) l'ont montré, les hommes cisgenres sont fortement associés avec les subreddits traitant de sujets typiquement considérés comme « masculins » : le sport, les jeux vidéo, l'actualité et la politique, la technologie et la pornographie. Les femmes cisgenres, quant à elles, sont associées positivement et fortement avec les subreddits regroupés dans le thème « personal advice », qui traitent de relations, de famille, de mode et de beauté, des centres d'intérêt traditionnellement vus comme « féminins ».

Notons également que les femmes sont plus susceptibles que les hommes de créer des comptes jetables (→ p. 89) que les hommes (Leavitt, 2015). Il est donc possible qu'elles créent des comptes jetables pour écrire des commentaires sur des subreddits sur lesquels elles ne veulent pas être identifiées en tant que femmes, par peur du harcèlement ou du doxxing. Dans ce cas, certaines femmes pourraient avoir des « territoires » plus larges que ce que notre étude a révélé, mais fragmentés par l'utilisation de plusieurs comptes.

8.5.2 Groupes transgenres et non binaire

Pour les internautes transgenres et non binaires, cette analyse révèle des tendances intéressantes, et peut-être inattendues. Nous nous demandions si les personnes transgenres avaient un comportement similaire à celui du groupe cisgenre auquel elles s'identifient, ou similaire à celui des autres groupes non cisgenres. La réponse est complexe. Les hommes transgenres et les personnes non binaires semblent se comporter d'une façon parallèle, et significativement différente des femmes transgenres. Ils écrivent les commentaires les plus longs, fréquentent un nombre limité de subreddits et sont associés positivement à seulement deux types de forums, qui traitent des relations personnelles et de l'humour. Les femmes transgenres, d'un autre côté, semblent occuper l'espace de Reddit d'une façon qui leur est propre. Elles commentent sur davantage de forums que tous les autres groupes de genre, et écrivent des commentaires significativement plus courts que ceux des hommes transgenres et, pour les plus de 31 ans, que les hommes transgenres et les personnes non binaires.

L'analyse factorielle des correspondances simples montre que les femmes transgenres sont le groupe le plus indistinct ; elles sont associées positivement avec 10 des 12 thèmes, comme les hommes cisgenres, mais cette association est faible. Il semble donc que, sur Reddit tout du moins, les femmes transgenres ont une expérience très différente de celles des hommes transgenres et des personnes non binaires.

On pourrait trouver plusieurs explications possibles à ce résultat. Tout d'abord, il se peut que, ayant été socialisées en tant que garçons et parfois hommes, les femmes transgenres soient plus à l'aise que les femmes cisgenres et les autres personnes transgenres sur les subreddits traitant de sujets perçus comme étant typiquement masculins ; pour elles, la barrière d'entrée serait ainsi plus facile à franchir. Les parcours différents, sur Reddit, des femmes et des hommes transgenres pourraient également s'expliquer par le fait que leurs expériences dans la vie « hors ligne » sont très différentes, avant, pendant, et après leur transition (Schilt, 2010). Les hommes transgenres sont en effet dans une position unique. Ils rejoignent une catégorie de genre qui leur offre plus d'avantages socioéconomiques que celle qu'ils quittent. Pour les femmes transgenres, c'est l'inverse. De plus, l'utilisation de testostérone permet aux hommes transgenres de développer une apparence masculine en l'espace de quelques mois. L'œstrogène n'est pas aussi efficace, et ne peut pas effacer des caractéristiques physiques perçues comme étant masculines, comme la pilosité ou la pomme d'Adam. Ainsi, après la transition, de nombreuses femmes transgenres ne « passent » pas aussi bien que les hommes transgenres (c'est-à-dire qu'elles ne sont pas vues comme des femmes), ce qui a un impact profond sur leur vie sociale et professionnelle. Les hommes transgenres, par contraste, trouvent plus facilement leur place dans les espaces masculins, y compris dans le monde du travail (Schilt, 2010). Par conséquent, les femmes transgenres sont souvent plus isolées que les hommes transgenres, et cachent davantage leur identité de genre (Rankin & Beemyn, 2012). Dans ce contexte, Reddit, et le cyberespace de manière générale, pourrait leur offrir un espace où elles se sentent bien, et où elles peuvent interagir avec les autres sans souffrir de la marginalisation et du harcèlement qu'elles vivent « hors ligne » (Marciano, 2014 ; Whittle, 1998). Les hommes transgenres, d'un autre côté, pourraient avoir un besoin moindre d'utiliser Reddit pour vivre pleinement leur identité d'hommes, puisque leur statut transgenre est souvent invisible ou aisément accepté dans le monde réel.

Nous ne pouvons par ailleurs pas exclure la possibilité que les personnes non binaires et les hommes transgenres sont plus susceptibles que les autres Redditors de créer plusieurs comptes sur Reddit, pour pouvoir commenter sur certains forums de façon réellement anonyme (sans que l'on puisse savoir qu'ils sont transgenres). Il se peut également qu'ils fréquentent des subreddits sans y commenter plus que les femmes transgenres (ce que l'on appelle *lurk* en anglais). Comme les hommes transgenres peuvent souvent vivre de façon *stealth*, c'est-à-dire sans révéler leur identité transgenre, ils pourraient créer un compte pour échanger avec d'autres personnes transgenres, et un autre compte pour échanger sur d'autres sujets, sans que

ces deux comptes puissent être reliés.

En ce qui concerne les Redditors non binaires, il est difficile d'interpréter nos résultats, pour plusieurs raisons : tout d'abord, parce qu'il existe très peu d'études sur leurs centres d'intérêt ou leurs comportements en ligne, et ensuite parce qu'il s'agit d'un groupe hétérogène.

Rappelons, enfin, que l'analyse factorielle des correspondances simples est une méthode purement descriptive et exploratoire (→ p. 155). Les résultats présentés ici ne peuvent donc pas être généralisés à l'ensemble des Redditors : ils ne concernent que les internautes de RedditGender.

tl ;dr

Les analyses présentées dans ce chapitre ont révélé que le genre a un impact sur la mobilité, la longueur des commentaires et les centres d'intérêt des Redditors. Les femmes cisgenres ont un territoire plus réduit sur le site que les hommes cisgenres : elles commentent sur moins de forums, et fréquentent une gamme de thèmes limitée.
Comme c'était déjà le cas dans le chapitre précédent, les femmes transgenres semblent se démarquer par leur comportement des hommes transgenres et des personnes non binaires, sans pour autant s'aligner sur les femmes cisgenres. Là où les hommes transgenres et les personnes non binaires écrivent des commentaires longs dans un nombre réduit de forums, les femmes transgenres se déplacent beaucoup sur Reddit, écrivant des commentaires plus courts.
Nous verrons dans les chapitres suivants que le parallèle entre hommes transgenres et personnes non binaires et le comportement singulier des femmes transgenres se manifestent également souvent dans leurs pratiques linguistiques.

Quatrième partie

Analyses linguistiques

Chapitre 9

Production de Netspeak

Ce chapitre présente une analyse synthétique des 11 variables linguistiques étudiées dans cette thèse : les émoticônes, émojis, étirements de lettres, étirements de ponctuation, mots en majuscules, interjections, abréviations (acronymes et réductions), g-droppings, graphies phonétiques, omissions d'apostrophe et omissions de la majuscule du pronom *I*. Nous avons réalisé deux analyses : la première examine les effets de l'âge et du genre, et la seconde prend également en compte l'ethnicité des Redditors.

9.1 Hypothèses et questions de recherche

Nous faisons les hypothèses suivantes :

— Les Redditors les plus jeunes utilisent davantage d'éléments non standard que les plus âgés, de nombreuses études ayant montré que l'âge est corrélé de façon négative avec la production de graphies non standard (→ p. 56).

— Il est possible que l'« âge Reddit », c'est-à-dire le nombre d'années depuis lequel un compte Reddit a été créé, soit corrélé positivement avec la production d'éléments du Netspeak. En effet, on peut imaginer qu'un·e Redditor qui fréquente le site depuis plusieurs années soit plus familier·e avec la langue utilisée sur Reddit, et donc utilise davantage ses éléments non standard.

— Il est également possible que le genre et l'ethnicité aient un impact sur la production de Netspeak. On sait, par exemple, que les Afro-américain·es sont considérés comme des innovateur·trices en la matière (→ p. 56).

9.2 Données

Les fréquences brutes des 11 variables linguistiques sont présentées dans le diagramme en barres 9.1. En tout, nous avons recensé 221 096 occurrences de ces 11 procédés. Les plus fréquents sont les mots en majuscules (*all caps*) et les abréviations, avec respectivement 39 478 et 37 357 oc-

currences. Les moins fréquents sont les g-droppings, avec 1662 occurrences seulement.

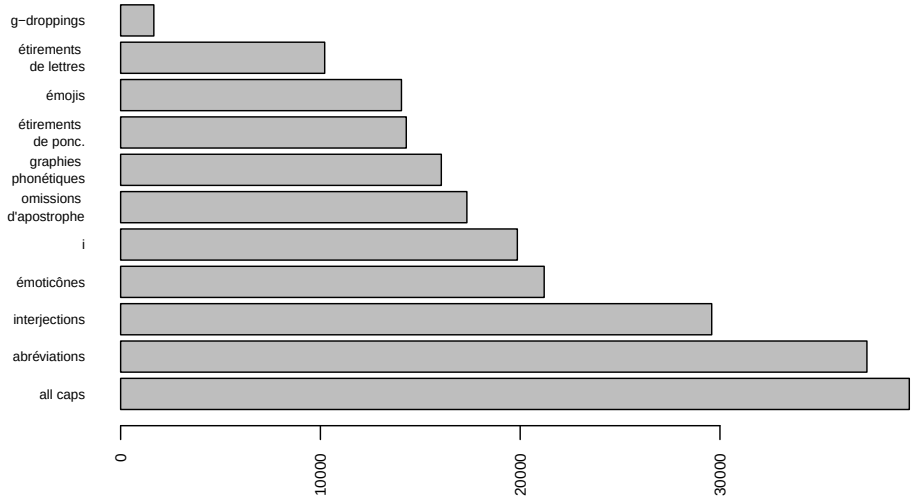

FIGURE 9.1 – Fréquence des 11 variables linguistiques dans RedditGender

Pour analyser les effets du genre, de l'âge et de l'ethnicité sur la production de Netspeak, nous avons additionné les fréquences brutes de ces 11 variables. Une première visualisation de la fréquence relative du Netspeak à l'aide de boites à moustaches (figure 9.2) a révélé la présence de trois valeurs aberrantes, qui correspondent à des personnes qui ont utilisé respectivement environ 92.79, 144.02 et 360.82 éléments non standard par 1000 tokens. Nous avons décidé d'écarter ces trois individus (deux hommes de 14 et 15 ans et une femme de 29 ans) pour cette analyse, afin que leurs productions extrêmes de graphies non standard, d'émoticônes et d'émojis n'influencent pas nos résultats. L'analyse présentée dans la section suivante est ainsi basée sur les productions de 1041 individus. La boite à moustaches montre également qu'il y a une majorité de valeurs faibles, et un nombre réduit de valeurs élevées : la plupart des Redditors utilise le Netspeak rarement, ou de façon modérée, et une minorité emploie beaucoup plus fréquemment ses procédés non standard.

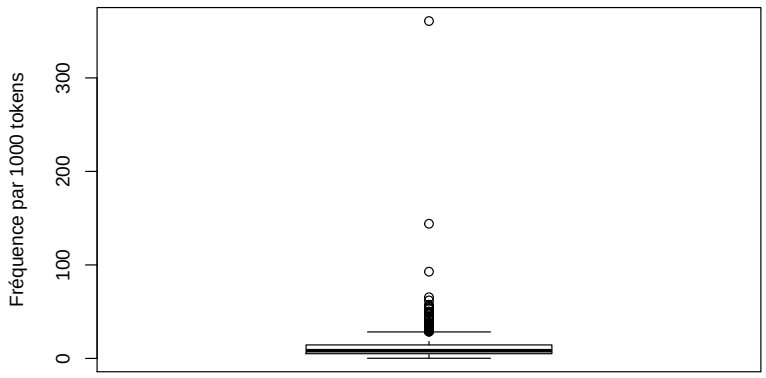

FIGURE 9.2 – Fréquence du Netspeak dans RedditGender, par 1000 tokens

9.3 Effets du genre et de l'âge sur la production de Netspeak

9.3.1 Statistiques descriptives

Le tableau 9.1 indique que les individus non binaires utilisent un nombre médian d'éléments non standard plus important que les autres groupes (médiane de 9.44 par 1000 tokens). Viennent ensuite les femmes cisgenres, avec une médiane de 8.76 éléments non standard par 1000 tokens, puis les trois groupes transgenres, qui ont des médianes comprises entre 7.34 pour les femmes transgenres et 7.82 pour les hommes transgenres. Nous constatons également que la fréquence de ces éléments décline avec l'âge des internautes : la fréquence médiane est de 12.77 éléments non standard par 1000 tokens pour les Redditors âgé·es de 14 à 20 ans, de 8.86 pour celles et ceux qui ont de 21 à 30 ans, et de 6.54 pour les plus âgé·es. Les moyennes sont plus élevées, ce qui est normal étant donné le nombre de valeurs très élevées dans chaque groupe. On remarque également une dispersion importante, en particulier chez les plus jeunes (écart interquartile = 12.92, contre 6.78 chez les plus de 30 ans), ce qui suggère qu'il y a davantage de variation chez les adolescent·es et les jeunes adultes que chez les Redditors les plus âgé·es.

9.3.2 Modèle de régression : effet du genre et de l'âge sur la production de Netspeak

La variable dépendante du modèle est la fréquence brute des éléments du Netspeak utilisés par chaque personne, soit la somme des omissions d'apostrophe, des émoticônes, des émojis, des graphies phonétiques, etc.

Tableau 9.1 – Fréquence des éléments non standard dans le corpus

	Moyenne	Écart-type	Médiane	EI
Hommes cisgenres	10.63	9.11	7.68	9.66
Femmes cisgenres	11.41	9.54	8.76	8.58
Femmes transgenres	10.40	9.06	7.34	10.04
Hommes transgenres	12.08	11.66	7.82	10.26
Non-binaires	12.27	9.14	9.44	9.44
14-20 ans	15.74	12.11	12.77	12.92
21-30 ans	11.62	9.35	8.86	9.04
31 ans et +	8.85	7.82	6.54	6.78
Tous	11.18	9.53	8.24	9.21

Les variables indépendantes sont l'âge, le genre, et l'âge Reddit (une variable catégorielle qui indique le nombre d'années depuis lequel un compte Reddit a été créé, → p. 136). L'interaction du genre et de l'âge a également été intégrée au modèle. Les hommes cisgenres de 14 à 20 ans sont le niveau de référence du modèle. Nous avons commencé par créer des modèles de régression logistique Poisson et binomial-négatif, qui sont conseillés pour analyser les données de comptage (Hilbe, 2014). Toutefois, les résultats montraient une surdispersion importante pour ces deux modèles ; dans le cas du modèle binomial négatif, la déviance résiduelle était de 11 048 pour 1022 degrés de liberté, alors qu'elle devrait être d'une valeur à peu près équivalente au nombre de degrés de liberté (Harrison, 2014).

Cette surdispersion importante est sans doute due au fait que, pour chaque personne, notre variable indépendante est le résultat de l'addition de données de comptage, ou « aggregated count data » (Harrison, 2014). Cela crée une dépendance dans les données : un individu qui utilise beaucoup d'émoticônes, par exemple, peut avoir tendance à utiliser beaucoup d'émojis, d'acronymes, etc. À l'inverse, d'autres Redditors peuvent employer peu fréquemment toutes les catégories linguistiques. En additionnant toutes nos variables, nous avons donc amplifié la variation idiosyncrasique. Pour remédier à ce problème, nous avons ajouté au modèle des « observation-level random effects », ou OLRE, qui sont communément utilisés dans ce cas (Harrison, 2014). Cette méthode assigne à chaque observation un effet aléatoire qui permet de prendre en compte la surdispersion des données. Nous l'avons mise en oeuvre en utilisant la fonction `glmer` du package `lme4` (Bates et al., 2015) pour créer un modèle généralisé à effets mixtes de type Poisson. Nous avons ajouté au modèle un effet aléatoire correspondant à l'identifiant de chaque observation. Un offset a été intégré pour prendre en compte les différences de longueur des sous-corpus.

9.3.3 Résultats

Les coefficients, les intervalles de confiance de 95 % et les valeurs p du modèle sont présentés dans le tableau 9.2. L'« âge Reddit » des internautes ne semble pas être corrélé avec la fréquence d'éléments non standard. L'in-

teraction de l'âge et du genre est significative, ce qui veut dire que l'effet de l'âge n'est pas le même dans tous les groupes. La fréquence des éléments du Netspeak est corrélée négativement avec l'âge chez les hommes cisgenres : plus un Redditor est âgé, moins il utilise ces éléments. On remarque le même phénomène chez les femmes cisgenres, à ceci près que la différence entre les femmes de 14 à 20 ans et les femmes de 21 à 30 ans n'est pas significative, mais approche du niveau de significativité (p = 0.07). L'âge n'a en revanche pas d'effet chez les groupes transgenres et non binaires, où on ne remarque pas de diminution significative de la fréquence du Netspeak chez les Redditors les plus âgé·es.

Tableau 9.2 – Production de Netspeak, effets de l'âge Reddit et l'interaction de l'âge et du genre

	Variable dépendante :
	Tout le Netspeak
Intercept	0.013** (0.0107, 0.0165)
Femmes cisgenres	0.866 (0.625, 1.198)
Femmes transgenres	0.574* (0.364, 0.904)
Hommes transgenres	−0.349 (0.705, 1.003)
Non-binaires	0.679 (0.442, 1.042)
21-30 ans	0.610** (0.483, 0.769)
31 ans et +	0.408** (0.324, 0.513)
Reddit cat. 2	1.079 (0.926, 1.256)
Reddit cat. 3	1.118 (0.969, 1.292)
Reddit cat. 4	1.054 (0.909, 1.221)
Reddit cat. 5	1.047 (0.846, 1.296)
Femmes cisgenres :21-30 ans	1.264 (0.877, 1.818)
Femmes transgenres :21-30 ans	1.523 (0.912, 2.545)
Hommes transgenres :21-30 ans	1.240 (0.815, 1.885)
Non-binaires :21-30 ans	1.721* (1.060, 2.793)
Femmes cisgenres :31 ans et +	1.399 (−0.967, 2.024)
Femmes transgenres :31 ans et +	0 2.230** (1.305, 3.808)
Hommes transgenres :31 ans et +	2.203** (1.293, 3.751)
Non-binaires :31 ans et +	2.302** (1.335, 3.975)
Observations	1,041
Log Likelihood	−6,425.139
Akaike Inf. Crit.	12,890.280
Bayesian Inf. Crit.	12,989.240
Note :	*p<0.05 ; **p<0.01

Quand on compare les groupes de genre dans chaque catégorie d'âge, les différences sont peu nombreuses. Chez les plus jeunes, la seule différence significative se trouve entre les femmes transgenres et les hommes cisgenres ; ces derniers produisent plus de Netspeak que les femmes transgenres. Chez les internautes de 21 à 30 ans, il n'y a pas de différence entre les hommes cisgenres et les autres groupes. Les hommes transgenres et les femmes transgenres utilisent moins de Netspeak que les non-binaires ; les femmes cisgenres en emploient plus que les hommes transgenres. Dans le groupe le plus âgé, les hommes cisgenres utilisent moins de Netspeak que les hommes transgenres, les femmes cisgenres et les non-binaires.

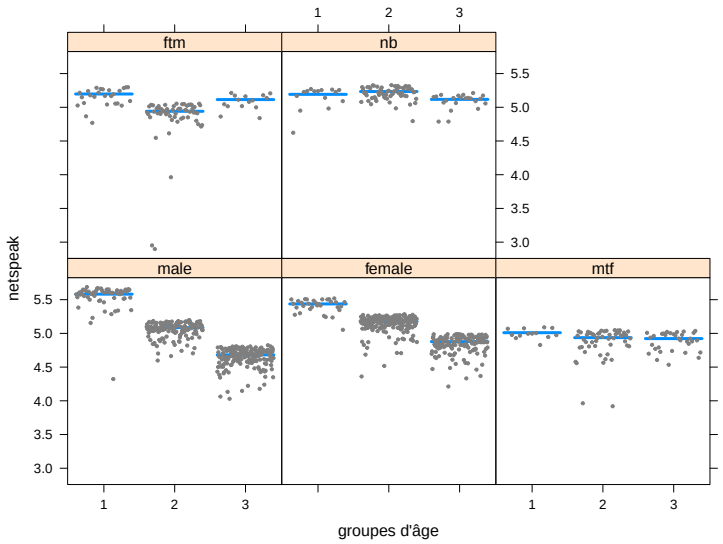

FIGURE 9.3 – Interaction âge et genre dans la production de Netspeak

Ces interactions entre genre et âge sont présentées dans la figure 9.3. On y voit la corrélation négative entre âge et production d'éléments du Netspeak pour les hommes cisgenres et les femmes cisgenres. Chez les personnes transgenres et les individus non binaires, l'interaction du genre et de l'âge ne semble pas produire les mêmes effets. La fréquence décroit très faiblement avec l'âge, et augmente même entre le groupe d'âge 2 (21 à 30 ans) et 3 (31 ans et plus) chez les hommes transgenres. Les individus transgenres et non binaires les plus âgés utilisent donc davantage d'éléments non standard que les personnes cisgenres du même âge.

9.4 Effets du genre, de l'âge et de l'ethnicité sur la production de Netspeak

9.4.1 Statistiques descriptives

Pour réaliser cette étude, nous avons utilisé l'échantillon réduit présenté dans page 135, qui comporte 346 individus cisgenres (puisqu'une valeur aberrante a été enlevée). La figure 9.4 présente la production de Netspeak chez les femmes et les hommes de chaque sous-corpus. Elle montre que les femmes de chaque groupe, à l'exception du groupe afro-américain, ont utilisé davantage de Netspeak que les hommes de leur groupe. Le nombre médian d'éléments du Netspeak est plus élevé dans les groupes afro-américain et hispanique que chez les blanc·hes et les Asiatiques.

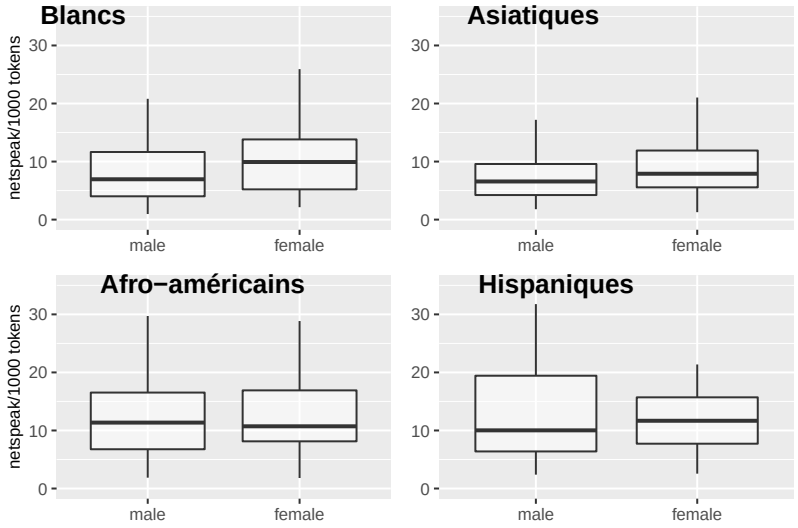

FIGURE 9.4 – Production de Netspeak par 1000 tokens, par sous-corpus

9.4.2 Effets du genre, de l'âge et de l'ethnicité sur la production de Netspeak : modèle de régression

Nous n'avons pas rencontré le même problème de surdispersion que lors de notre analyse du corpus entier. Il a donc été possible de créer un modèle de régression logistique binomial négatif. L'interaction du genre et de l'ethnicité a été intégrée au modèle, ainsi que l'âge Reddit ; ils ont été retirés lors du processus de sélection des variables à l'aide de la fonction `step()` de R. Les effets de l'âge, du genre et de l'ethnicité ont été conservés. Le niveau de référence du modèle est les hommes cisgenres.

TABLEAU 9.3 – Effets du genre, de l'âge et de l'ethnicité sur la production de Netspeak

	Variable dépendante :
	Tout le Netspeak
Intercept	0.015^{**} (0.012, 0.019)
Femmes	1.138 (0.988, 1.312)
Afro-américain·es	1.191 (0.994, 1.430)
Asiatiques	0.897 (0.732, 1.102)
Hispaniques	1.281^{*} (1.048, 1.571)
21-30 ans	0.701^{**} (0.558, 0.873)
31 ans et +	0.496^{**} (0.391, 0.624)
Observations	346
Log Likelihood	−2,139.940
θ	2.318^{**} (0.167)
Akaike Inf. Crit.	4,293.880
Note :	*p<0.05 ; **p<0.01

Le modèle ne révèle pas de différence significative entre femmes et hom-

mes cisgenres. Il montre en revanche une corrélation négative et significative entre âge et production de Netspeak : les internautes de 14 à 20 ans produisent davantage de Netspeak que les 21 à 30 ans, qui en produisent davantage que les 31 ans et plus. L'ethnicité a également un effet significatif sur la fréquence du Netspeak : les internautes hispaniques produisent davantage d'éléments non standard que les internautes blanc·hes et asiatiques. Ainsi, quand un·e internaute asiatique emploie 1 procédé du Netspeak, un·e blanc·he en utilise 0.78 et un Asiatique 0.70. Les Afro-américain·es emploient également davantage de Netspeak que les Asiatiques, mais pas plus que les blanc·hes.

9.5 Discussion

9.5.1 Effet de l'interaction entre genre et âge

Cette analyse a montré que, contrairement à une des hypothèses que nous avions émises, le nombre d'années depuis lequel une personne est sur Reddit n'a pas d'impact sur sa production d'éléments non standard. Une personne qui est active sur le site depuis 5 ans ou plus ne produit pas davantage d'éléments non standard qu'une personne qui n'utilise le site que depuis un an. Cela suggère que les éléments non standard étudiés ici ne sont pas propres à Reddit, et qu'ils font partie du répertoire des internautes avant leur arrivée sur le site. Comme nous en avions fait l'hypothèse, l'âge est corrélé de façon négative avec la fréquence des éléments du Netspeak. Cette corrélation semble toutefois très faible, voire inexistante, chez les personnes transgenres et non binaires, ce qui est étonnant, parce que de nombreuses études ont mis en évidence une corrélation négative entre âge et utilisation de graphies non standard (→ p. 56). L'étude séparée de chaque variable (émoticônes, acronymes, etc.) permettra sans doute de mieux comprendre ce phénomène. Nous n'avons pas mis en évidence de différence significative entre les hommes cisgenres et les femmes cisgenres : dans l'ensemble, ils utilisent les éléments non standard et graphiques du Netspeak aussi fréquemment.

Nos résultats révèlent qu'il est pertinent, et même indispensable, de ne pas étudier le genre isolément : l'âge, et sans doute d'autres variables sociales, ont également un impact sur la façon dont les internautes s'expriment en ligne. Ils suggèrent également que l'intégration des interactions est particulièrement utile, l'interaction de l'âge et du genre ne semblant pas avoir le même effet dans les groupes transgenres que dans les groupes cisgenres.

9.5.2 Effets de l'âge, du genre et de l'ethnicité

L'analyse du groupe cisgenre montre que, quand on étudie les procédés du Netspeak dans leur ensemble, le genre n'a pas un effet significatif, mais l'âge et l'ethnicité en ont un. Ici, l'âge est corrélé négativement avec la production de Netspeak. On remarque également une démarcation ethnique

entre, d'un côté, les internautes hispaniques et afro-américain·es, et, de l'autre, les internautes asiatiques, qui emploient significativement moins de Netspeak. Nous apportons, dans notre synthèse générale, des éléments d'explication sur ce résultat (→ p. 329). Évidemment, cette analyse fournit uniquement une vue d'ensemble de l'utilisation des procédés du Netspeak. Comme nous le verrons dans les deux chapitres qui suivent, il y a des différences significatives entre femmes et hommes pour certaines variables, ainsi que des interactions significatives entre genre et ethnicité. Cette exploration confirme, en tout cas, l'intérêt de l'approche intersectionnelle : prendre uniquement en compte le genre des internautes ne suffit pas pour comprendre leurs pratiques linguistiques en ligne.

tl ;dr

Ce chapitre introductif à nos analyses linguistiques a présenté une vue d'ensemble de la fréquence des différentes variables dans le corpus. Il a également esquissé des tendances que nous remarquerons à plusieurs reprises par la suite : l'absence de corrélation entre âge et production de graphies non standard chez les personnes transgenres, et la distanciation entre les groupes hispanique et afro-américain d'un côté, et le groupe asiatique de l'autre. Dans le corpus entier, les différences significatives entre groupes de genre varient selon les catégories d'âge. Dans l'échantillon composé uniquement de personnes cisgenres, le genre n'a pas d'effet. Nous verrons par la suite que la variation liée au genre, ainsi que son interaction avec l'ethnicité, a un effet significatif pour certaines variables.

Chapitre 10

Procédés d'ajout

Ce chapitre est consacré aux phénomènes que nous avons appelés « procédés d'ajout » : émoticônes, émojis, étirements de lettres, étirements de ponctuations, mots en majuscules et interjections. Il présente, pour chaque procédé, un aperçu des types les plus fréquents dans le corpus, suivi d'une analyse sociolinguistique basée sur la technique de la régression qui explore dans un premier temps les effets du genre et de l'âge, puis examine, en plus, celui de l'ethnicité.

10.1 Hypothèses et questions de recherche

Nous émettons l'hypothèse que plusieurs procédés d'ajout sont utilisés plus fréquemment par les femmes cisgenres que par les hommes cisgenres, en nous appuyant sur les études réalisées par d'autres chercheur·es. Ce serait le cas des mots majuscules (Parkins, 2012 ; Rosen et al., 2010), des émoticônes, et sans doute également des émojis et des étirements de lettres et de ponctuation (Bamman et al., 2014 ; Coats, 2017b ; Rao et al., 2010). Il est également très probable que les Redditors les plus jeunes aient tendance à produire davantage ces procédés typographiques, à cause de l'affinité des adolescent·es pour les pratiques langagières non standard.

10.2 Émoticônes

10.2.1 Résultats sur l'ensemble du corpus

Émoticônes les plus fréquentes

Nous avons recensé 21 201 émoticônes dans le corpus, dont 50 types différents. La liste complète des émoticônes recensées est présentée en annexe (→ p. 397) ; la liste des 20 émoticônes les plus fréquentes est présentée dans le tableau 10.1, avec leur proportion dans l'ensemble des émoticônes étudiées. Les 5 émoticônes les plus fréquentes représentent 73.20 % de toutes les émoticônes du corpus, et les 10 émoticônes les plus fréquentes 87.77 % de l'ensemble des émoticônes. Les émoticônes positives (comme **:)**, **:p**, **:>** ou

encore **XD**) représentent 71.87 % de toutes les émoticônes relevées dans le corpus.

Les émoticônes abrégées ou « sans nez », comme **:)** au lieu de **:-)**, sont plus fréquentes que les émoticônes « avec nez ». **:)** est ainsi 11.72 fois plus fréquent que **:-)**, et **:(** est 22.08 fois plus fréquent que **:-(**.

Tableau 10.1 – Vingt émoticônes les plus fréquentes dans RedditGender

Émoticône	Fréquence	Proportion
:)	8382	39.54 %
:(	3113	14.68 %
:/	1409	6.65 %
<3	1307	6.16 %
;)	1299	6.13 %
:P	1046	4.93 %
:-)	715	3.37 %
:p	497	2.34 %
XD	462	2.18 %
xD	367	1.73 %
:3	361	1.70 %
D :	218	1.03 %
:-(	141	0.67 %
;-)	139	0.66 %
:\	136	0.64 %
:'(	109	0.51 %
:=)	105	0.50 %
:-\	100	0.47%
:O	90	0.42 %
=/	87	0.41 %

Fréquence par utilisateur

935 personnes, soit 89.56 % des Redditors, ont utilisé au moins une émoticône. La fréquence moyenne des émoticônes, par 1000 tokens, est de 1.10 (écart type = 1.62). La médiane est de 0.5 émoticône par 1000 tokens (écart interquartile = 1.3). Cent-vingt personnes n'ont pas utilisé d'émoticône du tout. Le maximum est de 13.5 émoticônes par 1000 tokens. La boite à moustaches présentée dans la figure 10.1 montre qu'il y a une dispersion importante, ce qui suggère de fortes différences individuelles dans l'utilisation des émoticônes. La forte dispersion des données est également visible dans l'histogramme présenté dans la figure 10.1. Les données présentent une asymétrie positive, avec une masse d'observations positionnées sur la gauche. Cela signifie que l'immense majorité des internautes a utilisé très peu d'émoticônes, et qu'une minorité en a utilisé beaucoup. Notons que tous les phénomènes étudiés par la suite suivent ce type de distribution asymétrique, dite de Zipf.

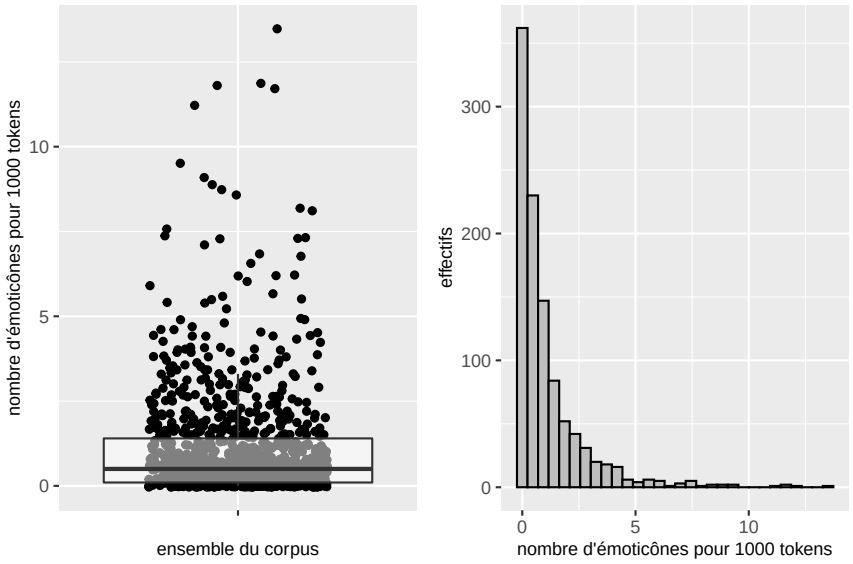

FIGURE 10.1 – Fréquence des émoticônes dans RedditGender : boite à moustaches et histogramme

10.2.2 Effets du genre et de l'âge sur la fréquence des émoticônes

Statistiques descriptives

Cette analyse a été réalisée avec l'ensemble du corpus, soit 1044 individus. Le tableau 10.2 présente le nombre moyen et médian d'émoticônes par 1000 mots dans chaque sous-corpus, avec l'écart type (ET) et l'écart interquartile (EI). Elle montre que les femmes transgenres ont utilisé le plus d'émoticônes (moyenne = 2.04, médiane = 1.2). Les hommes cisgenres ont utilisé en moyenne 2 fois moins d'émoticônes que les hommes transgenres et les personnes non binaires, et 3 fois moins que les femmes transgenres (moyenne = 0.67 ; médiane = 0.2). On remarque également que la dispersion est la moins importante chez les hommes cisgenres, avec un écart interquartile 2 fois moins élevé que les femmes cisgenres et les hommes transgenres, et 3 fois moins élevé que les femmes transgenres et personnes non binaires. Cela suggère que la variation individuelle est moins forte chez les hommes cisgenres que dans les autres groupes. Par ailleurs, la fréquence d'utilisation d'émoticônes décline avec l'âge : les Redditors les plus jeunes ont utilisé le plus d'émoticônes, et les Redditors les plus âgé·es le moins d'émoticônes.

Modèle de régression : effets de l'âge et du genre

Des modèles de régression de Poisson et binomial négatif ont été créés et comparés à l'aide du *likelihood ratio test* (→ p. 164). Celui-ci préfère le modèle binomial négatif, qui gère mieux l'importante dispersion des données ($p < 0.001$) (sauf exception, c'est le type de modèle que nous avons utilisé

Tableau 10.2 – Fréquence des émoticônes par 1000 tokens

Sous-corpus	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.67	1.23	0.24	0.62
Femmes cisgenres	1.12	1.47	0.57	1.37
Femmes transgenres	2.06	2.29	1.18	2.10
Hommes transgenres	1.41	1.82	0.78	1.54
Non-binaires	1.41	2.01	0.60	1.81
14-20 ans	1.57	2.02	0.68	1.76
21-30 ans	1.18	1.68	0.55	1.40
31 ans et +	0.82	1.34	0.33	0.84
Tous	1.10	1.63	0.49	1.25

pour analyser les variables linguistiques dans ce chapitre et le suivant). La variable dépendante est la fréquence brute des émoticônes produites par chaque personne. Nous avons inclus l'âge, le genre et leur interaction comme variables indépendantes, puis avons utilisé la fonction `step()` pour sélectionner les variables. Seuls les effets principaux de l'âge et du genre ont été retenus par l'algorithme. Les coefficients, intervalles de confiance de 95 % et valeurs p du modèle sont présentés dans le tableau 10.3 ; le niveau de référence est les hommes cisgenres.

Tableau 10.3 – Émoticônes, effets du genre et de l'âge

	Variable dépendante :
	Émoticônes
Intercept	0.001^{**} (0.001, 0.001)
Femmes cisgenres	1.811^{**} (1.502, 2.183)
Femmes transgenres	3.294^{**} (2.501, 4.398)
Hommes transgenres	2.085^{**} (1.583, 2.784)
Non-binaires	2.076^{**} (1.575, 2.776)
21-30 ans	0.635^{**} (0.498, 0.801)
31 ans et +	0.477^{**} (0.372, 0.607)
Observations	1,044
Log Likelihood	−4,059.691
θ	0.640^{*} (0.028)
Akaike Inf. Crit.	8,133.381
Note :	*p<0.05 ; **p<0.01

Les effets de l'âge et du genre sont significatifs. Pour le genre, les coefficients sont positifs pour tous les groupes, ce qui signifie que les hommes cisgenres (le niveau de référence) utilisent les émoticônes le moins fréquemment. Les femmes cisgenres les emploient quant à elles 1.81 fois plus que les hommes cisgenres. Les femmes transgenres sont celles qui utilisent le plus fréquemment les émoticônes : 1.81 fois plus que les femmes cisgenres, 1.57 fois plus que les hommes transgenres et 1.58 fois plus que les personnes non binaires. Il n'y a pas de différence significative entre les femmes cisgenres, les hommes transgenres et les personnes non binaires. L'âge est

corrélé significativement et négativement avec la production d'émoticônes. Les internautes de 14 à 20 ans utilisent plus d'émoticônes que les internautes de 21 à 30 ans, qui en utilisent davantage que les Redditors les plus âgé·es.

Analyse de l'échantillon non-binaire

L'effet principal du genre étant significatif dans le premier modèle réalisé, nous avons décidé de savoir si le genre assigné à la naissance (AFAN, ou assignée fille à la naissance et AGAN, assigné garçon à la naissance, → p. 17) des personnes non binaires avait un impact sur la production d'émoticônes. S'il l'est, nous nous attendons à ce que les personnes AFAN utilisent davantage d'émoticônes que les personnes AGAN. Nous avons donc réalisé un modèle à partir d'un échantillon réduit, qui ne contient que les personnes non binaires ; il est composé de 98 personnes, car nous ne disposons pas du genre assigné à la naissance pour 2 internautes. Nous avons intégré comme variables indépendantes l'âge, le genre assigné à la naissance, et leur interaction, et avons sélectionné les variables à l'aide de la fonction `step()`. La fonction a uniquement conservé l'âge, mais nous avons ajouté le genre comme variable de contrôle. Le modèle ainsi réalisé est présenté dans le tableau 10.4. L'âge est la seule variable retenue par le processus de sélection des variables. Son effet n'est pas significatif. Chez les personnes non binaires, ni le genre assigné à la naissance, ni l'âge ne semblent donc avoir d'impact sur la fréquence des émoticônes.

Tableau 10.4 – Émoticônes, groupe non-binaire

	Variable dépendante :
	Émoticônes
Intercept	0.002^{**} (0.001, 0.005)
21-30 ans	0.594 (0.262, 1.211)
31 ans et +	0.456 (0.183, 1.086)
AGAN	1.235 (0.721, 2.153)
Observations	98
Log Likelihood	−407.978
θ	0.585^{**} (0.079)
Akaike Inf. Crit.	823.956
Note :	*p<0.05 ; **p<0.01

10.2.3 Effets du genre, de l'âge et de l'ethnicité sur la production d'émoticônes

Les analyses qui suivent ont été réalisées avec l'échantillon réduit décrit page 135, qui contient 347 personnes.

Statistiques descriptives

Le tableau 10.5 présente le nombre moyen et médian d'émoticônes pour 1000 tokens dans les sous-corpus. Comme dans le corpus entier, il semble que, dans cet échantillon, les hommes cisgenres utilisent moins d'émoticônes que les femmes, avec une médiane deux fois plus faible. L'écart interquartile est 2.26 fois plus élevé pour les femmes, ce qui indique une dispersion plus importante et des différences individuelles plus marquées. Le tableau montre également une possible corrélation négative de l'utilisation d'émoticônes avec l'âge, avec une fréquence moins élevée chez les Redditors les plus âgé·es que chez les plus jeunes. L'observation des médianes, plus robustes que la moyenne, montre que les Redditors blanc·hes et hispaniques ont utilisé à peu près autant d'émoticônes (respectivement 0.48 et 0.47 par 1000 tokens). Les Redditors afro-américain·es et asiatiques en ont utilisé moins, avec des médianes respectives de 0.27 et 0.32 émoticônes par 1000 tokens.

Tableau 10.5 – Fréquence des émoticônes par 1000 tokens dans l'échantillon réduit

	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.65	1.23	0.23	0.61
Femmes cisgenres	1.15	1.51	0.63	1.38
14-20 ans	1.40	2.02	0.66	1.31
21-30 ans	0.99	1.46	0.42	1.20
31 ans et +	0.65	1.02	0.30	0.72
Blanc·hes	0.85	1.27	0.48	1.01
Afro-Américain·s	0.88	1.37	0.27	1.14
Asiatiques	0.60	0.73	0.32	0.77
Hispaniques	1.03	1.24	0.47	1.30
Tous	0.89	1.40	0.38	0.97

Modèle avec interactions

Nous avons intégré à un modèle de régression binomial négatif l'interaction du genre et de l'ethnicité et l'interaction de l'âge et de l'ethnicité. La fonction `step()` a ensuite été utilisée pour enlever les variables qui ne contribuent pas au modèle. L'interaction de l'âge et de l'ethnicité a été supprimée. Les femmes cisgenres blanches sont le niveau de référence du modèle. Le modèle obtenu (tableau 10.6) montre que l'effet de l'âge est limité ; les Redditors les plus âgé·es emploient significativement moins d'émoticônes que les groupes 1 (14-20 ans) et 2 (21-30 ans).

La figure 10.2 représente l'interaction du genre et de l'ethnicité. On voit que dans les groupes blancs et asiatiques, il n'y a pas de différence significative dans la fréquence des émoticônes entre femmes et hommes. Il y en a en revanche dans les deux autres groupes : les femmes afro-américaines utilisent 2.94 fois plus d'émoticônes que les hommes afro-américains, et

Tableau 10.6 – Émoticônes, effets de l'âge et de l'interaction entre genre et ethnicité

	Dependent variable :
	Émoticônes
Intercept	0.002** (0.001, 0.003)
Hommes	0.753 (0.474, 1.188)
Hispaniques	1.408 (0.798, 2.551)
Afro-américaines	1.314 (0.807, 2.150)
Asiatiques	0.655 (0.369, 1.196)
21-30 ans	0.669 (0.423, 1.024)
31 ans et +	0.428** (0.266, 0.670)
Hommes :Hispaniques	0.494 (0.228, 1.065)
Hommes :Afro-Américains	0.452* (0.222, 0.928)
Hommes :Asiatiques	1.024 (0.463, 2.264)
Observations	347
Log Likelihood	−1,257.367
θ	0.626** (0.049)
Akaike Inf. Crit.	2,534.733
Note :	*p<0.05 ; **p<0.01

les femmes hispaniques en emploient 2.69 fois plus que les hommes hispaniques. Quand on compare les femmes entre elles, on remarque que les femmes asiatiques emploient environ deux fois moins d'émoticônes que les femmes hispaniques et afro-américaines. Chez les hommes, il n'y a pas de différence significative entre les Asiatiques et les autres. Les hommes afro-américains emploient en revanche moins d'émoticônes que les blancs : ils en utilisent 0.59 quand les hommes blancs en produisent 1.

10.2.4 Effet de l'orientation sexuelle sur la production d'émoticônes

Dans la section précédente, nous avons vu que l'ethnicité a un effet sur l'utilisation des émoticônes. Ici, nous nous intéressons à l'effet potentiel de l'orientation sexuelle sur la fréquence des émoticônes. C'est la seule variable linguistique pour laquelle nous prenons en compte cette variable ; cette décision a été motivée par le fait que les émoticônes ont été abondamment étudiées (→ p. 67), mais qu'aucune étude, à notre connaissance, n'a examiné un lien éventuel avec l'orientation sexuelle des internautes. Nous avons créé deux modèles séparés : le premier pour les femmes (cisgenres et transgenres) et le second pour les hommes (cisgenres et transgenres) pour lesquels nous connaissons l'orientation sexuelle.

Pour les femmes, l'échantillon contient 275 femmes cisgenres (227 hétérosexuelles et 48 gays) et 48 femmes transgenres (15 femmes hétérosexuelles et 33 gays). Pour les hommes, l'échantillon comprend 256 hommes cisgenres (194 hétérosexuels, 62 gays) et 68 hommes transgenres (30 hétérosexuels, 38 gays). Les modèles ont été réalisés en intégrant comme variables indépendantes l'âge, le genre, et l'orientation sexuelle, ainsi que

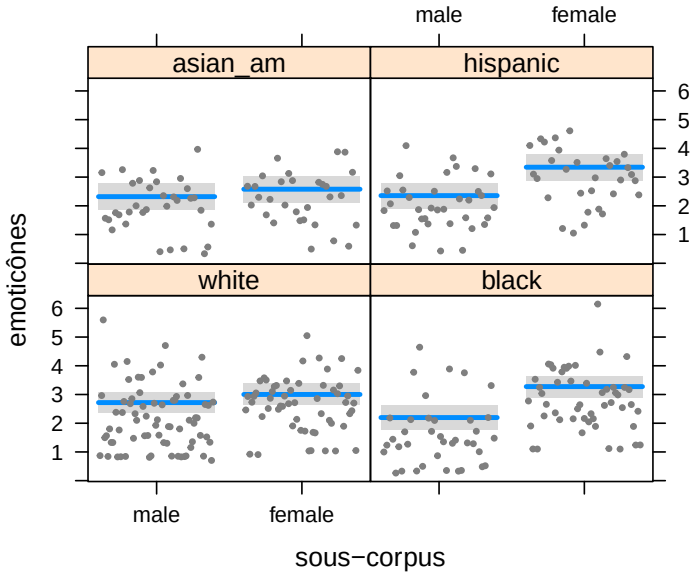

FIGURE 10.2 – Interaction du genre et de l'ethnicité dans la production d'émoticônes

l'interaction du genre et de l'âge, de l'orientation sexuelle et de l'âge, et du genre et de l'orientation sexuelle. La fonction `step()` a été utilisée pour sélectionner les variables.

TABLEAU 10.7 – Émoticônes, effet de l'orientation sexuelle : femmes

	Variable dépendante :
	Émoticônes
Intercept	0.001** (0.001, 0.002)
Femmes transgenres	1.463 (0.996, 2.203)
Gays	1.271 (0.918, 1.780)
21-30 ans	0.990 (0.585, 1.586)
31 ans et +	0.804 (0.467, 1.319)
Observations	323
Log Likelihood	−1,316.132
θ	0.732** (0.057)
Akaike Inf. Crit.	2,642.264
Note :	*p<0.05 ; **p<0.01

Pour les femmes, l'algorithme a uniquement retenu les effets principaux du genre et de l'orientation sexuelle. Nous avons toutefois conservé l'âge, pour en faire une variable de contrôle. Le modèle ne révèle aucun effet significatif de l'âge, du genre ou de l'orientation sexuelle sur l'utilisation d'émoticônes.

La fonction `step()` a conservé l'interaction du genre et de l'âge pour les hommes. Le niveau de référence du modèle (10.8) est les hommes cis-

Tableau 10.8 – Émoticônes, effet de l'orientation sexuelle : hommes

	Variable dépendante :
	Émoticônes
Intercept	0.001** (0.001, 0.002)
Hommes transgenres	0.805 (0.376, 1.853)
Gays	1.672** (1.212, 2.328)
21-30 ans	0.458** (0.269, 0.749)
31 ans et +	0.299** (0.178, 0.481)
Hommes transgenres :21-30 ans	1.671 (0.650, 4.099)
Hommes transgenres :31 ans et +	5.229** (1.711, 17.022)
Observations	324
Log Likelihood	−1,145.709
θ	0.629** (0.051)
Akaike Inf. Crit.	2,305.418
Note :	*p<0.05 ; **p<0.01

genres de 14 à 20 ans. L'âge est corrélé négativement avec la production d'émoticônes pour les hommes cisgenres, mais pas chez les hommes transgenres. Les hommes cisgenres de 14 à 20 ans produisent ainsi 3.34 fois plus d'émoticônes que les hommes les plus âgés. Il n'y a pas de différence dans la fréquence des émoticônes entre cisgenres et transgenres pour les deux groupes le plus jeunes. Chez le groupe le plus âgé, la différence est significative, avec une taille d'effet importante : les hommes transgenres emploient 4.21 plus d'émoticônes que les hommes cisgenres. L'effet principal de l'orientation sexuelle est significatif. Les hommes gays, qu'ils soient cisgenres ou transgenres, emploient davantage d'émoticônes que les hétérosexuels : ils en utilisent 1.67 quand les hétérosexuels en utilisent 1.

10.3 Émojis

10.3.1 Statistiques générales

Il y a 14 057 émojis dans le corpus. Notons la présence d'une valeur particulièrement extrême : un Redditor, un garçon de 14 ans, a utilisé 6486 émojis dans ses commentaires, soit une fréquence relative de 347.37 émoji par 1000 tokens. Nous avons supprimé ce Redditor des données utilisées pour effectuer les analyses ci-dessous, pour qu'il n'influence pas les résultats. Sans cette valeur aberrante, il y a 7571 émojis dans le corpus, dont 314 types différents. La moyenne est de 0.41 émoji par 1000 mots (écart type = 1.49). La médiane est de 0 (écart interquartile = 0.2) : 614 Redditors, soit 58.87 %, n'ont pas utilisé d'émoji du tout. Les données sont donc très dispersées et contiennent un nombre important de zéros.

Le tableau 10.9 présente les 20 types d'émojis les plus fréquents dans le corpus. 😂 est l'émoji le plus fréquent (1389 occurrences, soit 18.35 % de tous les émojis recensés). On remarque également, comme pour les émoticônes, la prédominance des émojis positifs, avec des visages souriants ou riants,

des cœurs et d'autres symboles qui indiquent l'enthousiasme, comme 🔥, 💯 et les signes 👍 et 👌.

Tableau 10.9 – 20 émojis les plus fréquents dans RedditGender

Rang	Émoji	Fréq.	%
1	😂	1389	18.35
2	😭	362	4.78
3	😊	338	4.46
4	😍	338	4.46
5	👏	277	3.66
6	❤️	274	3.62
7	😉	206	2.72
8	💯	192	2.54
9	👌	171	2.26
10	👍	151	1.99
11	🤔	146	1.93
12	🔥	131	1.73
13	😘	114	1.51
14	🖤	111	1.47
15	😅	102	1.35
16	😁	100	1.32
17	™	96	1.27
18	🅱	75	0.99
19	😜	70	0.92
20	😩	61	0.81

10.3.2 Comparaison entre émojis et émoticônes

Le corpus contient plus d'émoticônes (21 189) que d'émojis (14 057). Pour vérifier si les émoticônes sont plus fréquentes que les émojis, nous avons utilisé le test de Wilcoxon-Mann-Whitney, qui ne présuppose pas une distribution normale et permet de comparer deux médianes (Brezina, 2018). Il confirme que cette différence est significative ($W = 846\,3440$, $p < 0.001$).

10.3.3 Effet de l'âge et du genre sur la production d'émojis

Statistiques descriptives

Le tableau 10.10 présente le pourcentage d'internautes qui ont utilisé au moins 1 émoji, ainsi que le nombre moyen et médian d'émojis pour 1000 tokens, par sous-corpus, et l'écart type et interquartile. Dans l'ensemble du corpus, 613 personnes (58.72 %) n'ont pas utilisé un seul émoji.

On peut y voir que la proportion d'internautes qui ont utilisé des émojis au moins une fois est la plus faible chez les hommes cisgenres et les personnes non binaires. Elle est la plus forte chez les femmes transgenres, suivies par les femmes cisgenres. La moyenne des femmes cisgenres et des hommes transgenres est la plus élevée (0.47 émojis par 1000 tokens). Les

Tableau 10.10 – Fréquence des émojis par sous-corpus

	1 émoji min.	Moyenne	Écart type	Médiane	EI
Hommes cisgenres	38.81 %	0.41	1.62	0.00	0.20
Femmes cisgenres	42.74 %	0.46	1.60	0.00	0.30
Femmes transgenres	46.00 %	0.30	0.61	0.00	0.20
Hommes transgenres	41.00 %	0.47	1.80	0.00	0.20
Non-binaires	38.00 %	0.25	0.69	0.00	0.20
14-20 ans	60.96 %	0.93	2.63	0.15	0.60
21-30 ans	45.26 %	0.43	1.47	0.00	0.30
31 ans et plus	27.89 %	0.17	0.62	0.00	0.10
Tous	41.18 %	0.41	1.49	0.00	0.2

moyennes doivent toutefois être interprétées avec précaution, car la dispersion est très importante, comme le montrent les écarts types. La fréquence des émojis semble décroitre avec l'âge. Seule la médiane des Redditors de 14 à 20 ans est supérieure à 0 (0.15). Deux fois plus de Redditors de 14 à 20 ans que de Redditors âgé·es de 31 ans et plus ont utilisé au moins un émoji dans leurs commentaires. La moyenne baisse également fortement d'un groupe à l'autre, passant de 0.93 émojis par 1000 tokens pour les plus jeunes à 0.17 pour les plus âgé·es. Encore une fois, la prudence est de mise, car la dispersion est importante, surtout chez le groupe le plus jeune (écart type = 2.63).

Effets de l'interaction de l'âge et du genre : régression *zero-inflated*

Étant donné la prédominance des zéros dans les données, nous avons opté pour un modèle *zero-inflated*. Nous avons créé deux modèles *zero-inflated* : un modèle de Poisson et un modèle binomial négatif avec, comme variables indépendantes, les effets principaux de l'âge et du genre ainsi que l'interaction de l'âge et du genre. La variable indépendante est la fréquence brute des émojis dans chaque sous-corpus. Nous avons ajouté un offset au modèle pour prendre en compte la différence de taille de chaque sous-corpus. Toutes les variables intégrées au modèle ont été retenues. Les deux modèles obtenus ont été comparés avec le *log likelihood test*. Le test indique que le modèle *zero-inflated* binomial négatif est plus performant ($p < 0.001$). Les hommes âgés de 14 à 20 ans sont le niveau de référence du modèle.

La partie à *zero-inflation* (qui n'est pas présentée ici pour des raisons pratiques) ne renvoie aucun résultat significatif. Tous les groupes sont donc aussi susceptibles que les autres de ne pas utiliser d'émojis. L'autre partie du modèle (tableau 10.11) , qui indique la probabilité d'utiliser 1 émoji, montre une corrélation négative entre âge et fréquence des émojis chez les hommes cisgenres. La taille d'effet est importante : un homme de plus de 31 ans utilise seulement 0.08 émoji quand un homme de 14 à 20 ans en utilise 1. Cette corrélation est seulement partielle chez les femmes cisgenres (les femmes de 31 ans et plus en utilisent plus que les femmes de 21 à 30

ans) et les hommes transgenres (les hommes transgenres de 21 à 30 ans en utilisent plus que les hommes de 14 à 20 ans). Elle est inexistante chez les personnes non binaires et les femmes transgenres.

Dans le groupe d'âge le plus jeune, les hommes cisgenres utilisent davantage d'émojis que les femmes transgenres et les personnes non binaires, avec une taille d'effet importante : un homme cisgenre produit respectivement 4.02 et 7 émojis quand une femme transgenre et une personne non binaire en produisent 1. Les hommes transgenres utilisent davantage d'émojis que les femmes transgenres et les personnes non binaires. Il n'y a pas de différence entre les femmes cisgenres et les hommes cisgenres. Chez les internautes de 21 à 30 ans, les femmes cisgenres et les hommes cisgenres produisent plus d'émojis que les hommes transgenres. Les femmes transgenres les utilisent également moins souvent que les femmes cisgenres. Dans le groupe le plus âgé, une seule différence significative est relevée : les hommes cisgenres utilisent moins d'émojis que tous les autres groupes.

Tableau 10.11 – Émojis, effets de l'âge et du genre

	Variable dépendante :
	Émojis
Intercept	0.001** (0.001, 0.0022)
Femmes cisgenres	0.445 (0.165, 1.197)
Femmes transgenres	0.248* (0.082, 0.752)
Hommes transgenres	1.299 (0.414, 4.075)
Non-binaires	0.140** (0.0426, 2.172)
21-30 ans	0.437* (0.224, 0.850)
31 ans et +	0.079** (0.038, 0.164)
Femmes cisgenres :21-30 ans	2.834 (0.917, 8.758)
Femmes transgenres :21-30 ans	2.371 (0.613, 9.161)
Hommes transgenres :21-30 ans	0.314 (0.077, 1.278)
Non-binaires :21-30 ans	4.697* (1.088, 20.267)
Femmes cisgenres :31 ans et +	9.924** (2.866, 34.329)
Femmes transgenres :31 ans et +	17.115** (3.532, 82.269)
Hommes transgenres :31 ans et +	1.635 (5.129, 28.876)
Non-binaires :31 ans et +	32.395** (6.347, 165.504)
Observations	1,043
Log Likelihood	−2,180.771
Note :	*p<0.05 ; **p<0.01

10.3.4 Effet de l'ethnicité sur la production d'émojis

Statistiques descriptives

Le tableau 10.12 présente le pourcentage de Redditors ayant utilisé au moins un émoji, la fréquence moyenne des émojis pour chaque groupe, par 1000 tokens, et l'écart type. On voit que, comme dans la section précédente, plus de femmes que d'hommes ont utilisé au moins un émoji. La fréquence relative est également, en moyenne, plus élevée chez les femmes ; la dispersion est également plus forte chez elles.

Tableau 10.12 – Fréquence des émojis, échantillon réduit

	1 émoji	Moyenne	ET	Médiane	EI
Hommes	40.78 %	0.47	1.39	0.00	0.30
Femmes	51.50 %	0.65	1.76	0.10	0.40
14-20 ans	61.36 %	0.92	1.87	0.10	0.60
21-30 ans	54.05 %	0.71	1.87	0.10	0.60
31 ans et +	27.35 %	0.17	0.60	0.00	0.00
Blancs	38.28 %	0.39	1.40	0.00	0.20
Afro-Américain·es	55.06 %	0.53	1.09	0.10	0.60
Asiatiques	33.33 %	0.32	1.24	0.00	0.10
Hispaniques	60.60 %	1.12	2.44	0.10	0.95
Tous	45.95 %	0.55	1.58	0.00	0.40

Dans le groupe asiatique, il y a moins de Redditors qui ont utilisé au moins 1 émoji que dans les autres groupes (33.33 %). Les groupes où le plus de Redditors ont employé au moins 1 émoji sont le groupe afro-américain (55.06 %) et le groupe hispanique (60.60 %). La moyenne est la plus élevée chez les Hispaniques (1.12 émoji par 1000 tokens), mais l'écart type est important, ce qui indique une dispersion forte et la présence possible de valeurs aberrantes. La fréquence moyenne est la plus faible dans le groupe asiatique (0.32 émojis par 1000 tokens). Seuls les Afro-Américain·es et les Hispaniques affichent une médiane supérieure à 0. L'utilisation des émojis semble par ailleurs décroitre avec l'âge. Les Redditors les plus jeunes sont ceux qui sont utilisé le plus d'émojis : ils et elles en ont produit 0.92 par 1000 tokens en moyenne, contre 0.17 pour les Redditors les plus âgé·es. Un peu plus d'un quart (27.35 %) des Redditors âgé·es de 31 ans et plus ont utilisé au moins un émoji.

Effet du genre et de l'ethnicité : régression binomiale négative

Nous avons retiré une observation, qui était bien plus élevée que les autres dans cet échantillon réduit ; elle correspond à une femme hispanique âgée de 21 à 30 ans. Nous avons ensuite tenté de créer un modèle de régression *zero-inflated* en intégrant l'ethnicité. Toutefois, l'échantillon étant beaucoup plus réduit, et le nombre de zéros étant très important, le modèle n'a pas pu être créé par R, sans doute à cause de données insuffisantes. Nous avons à la place créé un modèle binomial négatif (tableau 10.13) qui inclut, après le processus de sélection des variables, l'âge et l'interaction du genre et de l'ethnicité. Les hommes blancs sont le niveau de référence du modèle.

Le modèle montre qu'il n'y a pas de différence entre femmes et hommes dans la fréquence d'utilisation des émojis dans le groupe blanc et dans le groupe hispanique. Dans les groupes afro-américain et asiatique, les femmes emploient davantage d'émojis que les hommes, avec des tailles d'effet importantes : 3.76 fois plus pour les femmes afro-américaines, et 10.37 fois plus pour les femmes asiatiques. L'interaction du genre et de l'ethni-

Tableau 10.13 – Effets de l'âge et de l'interaction entre genre et ethnicité, fréquence des émojis

	Variable dépendante :
	Émojis
Intercept	0.001** (0.0003, 0.002)
Femmes	0.751 (0.306, 1.884)
Asiatiques	0.122** (0.042, 0.389)
Afro-Américains	0.755 (0.287, 2.142)
Hispaniques	1.598 (0.624, 4.510)
21-30 ans	0.763 (0.301, 1.765)
31 ans et +	0.166** (0.059, 0.436)
Femmes :Asiatiques	13.824** (2.856, 67.470)
Femmes :Afro-américaines	5.017* (1.269, 19.129)
Femmes :Hispaniques	2.133 (0.492, 9.396)
Observations	345
Log Likelihood	−814.042
θ	0.177** (0.017)
Akaike Inf. Crit.	1,648.084
Note :	*p<0.05 ; **p<0.01

cité est présentée dans la figure 10.3. Chez les hommes, les Asiatiques utilisent moins d'émojis que tous les autres, avec encore une fois des tailles d'effet fortes : ils en emploient par exemple 13.12 fois moins que les Hispaniques et 8.23 moins que les blancs. Chez les femmes, en revanche, il n'a pas de différence significative entre les Asiatiques et les autres. Les femmes blanches emploient moins d'émojis que les femmes afro-américaines et que les femmes hispaniques. Quand une femme blanche produit 1 émoji, une femme afro-américaine en utilise 3.78 et une femme hispanique 3.41.

L'âge a un effet significatif, mais uniquement quand on compare les Redditors de plus de 31 ans aux deux autres groupes. Les internautes les plus âgé·es emploient 0.16 émoji quand les 14-20 ans en emploient 1, et 0.22 émoji quand les 21-30 ans en emploient 1.

10.4 Étirements de lettres

10.4.1 Types d'étirements de lettres

10 214 étirements de lettres ont été identifiés dans le corpus, dont 3626 types différents sous leur forme brute (avec majuscules et minuscules). Une fois tous les tokens mis en minuscules (à l'aide de la procédure décrite p. 145), nous comptons 955 étirements de lettres différents. Les 100 types les plus fréquents sont présentés dans l'annexe C.2. Nous avons ensuite retiré les lettres répétées des tokens afin d'obtenir une liste des mots les plus souvent allongés. Les 25 types les plus fréquents, avec leur graphie la plus fréquente dans le corpus, sont présentés dans la figure 10.4. Dans cette liste, les interjections sont les mots les plus fréquemment étirés (13 types sur 25, ou 17 sur 25 si on prend en compte les variantes de *yes* et *no*). L'ad-

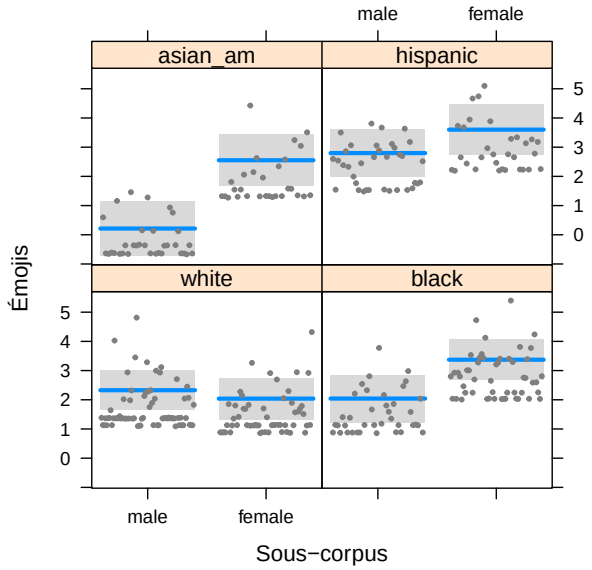

FIGURE 10.3 – Émojis, effets de l'âge et de l'interaction du genre et de l'ethnicité

verbe et conjonction *so* est, de loin, le mot le plus fréquemment étiré, avec plus de 700 occurrences dans le corpus. La lettre *o* est la plus souvent répétée dans le corpus (2427 répétitions), suivie par le *a* (1408 répétitions), le *h* (1123 répétitions), le *e* (833 répétitions), le *m* (728 répétitions), le *w* (525 répétitions) et le *s* (486 répétitions). Les lettres *z*, *b*, *v*, *c*, *n*, *q*, et *j* sont les moins fréquemment répétées.

10.4.2 Effets du genre et de l'âge sur la production d'étirements de lettres

Statistiques descriptives

Neuf-cent-cinquante-trois Redditors (91.28 %) ont utilisé au moins 1 étirement de lettres dans le corpus. Une inspection visuelle des données à l'aide d'une boite à moustaches révèle la présence de 3 valeurs extrêmes, qui correspondent aux productions d'une femme cisgenre de 33 ans, d'un homme cisgenre de 18 ans et d'une femme transgenre de 30 ans. Ils ont utilisé entre 4.70 et 5.15 étirements de lettres par 1000 tokens, alors que la médiane est de 0.33. Nous avons retiré ces observations des données pour les analyses qui suivent. Le tableau 10.14 présente les fréquences relatives moyennes et médianes pour chaque groupe, pour 1000 tokens.

L'observation des médianes, plus fiables que les moyennes étant donné l'importante dispersion des données, montre que les femmes cisgenres ont utilisé le plus d'étirements de lettres. Elles sont aussi le groupe où il y a le plus de dispersion (écart interquartile = 0.70), ce qui indique des variations individuelles plus prononcées que dans les autres groupes. Les hommes

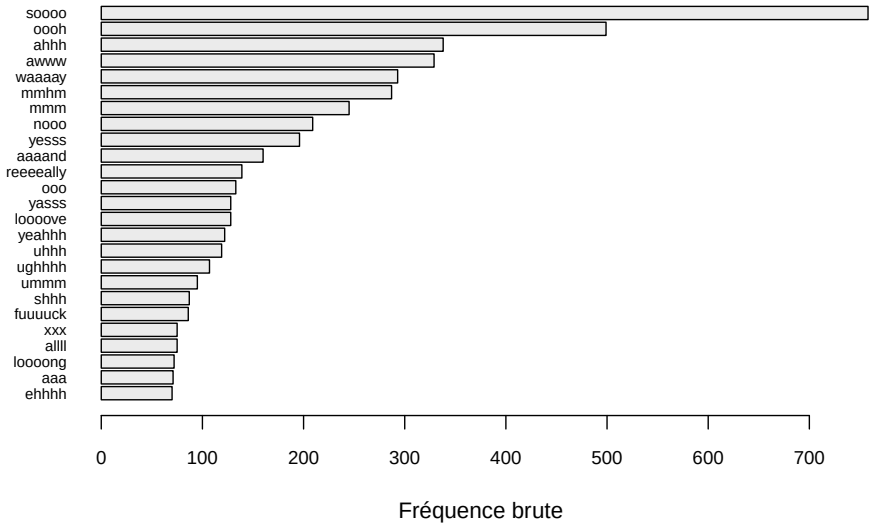

FIGURE 10.4 – 25 mots les plus fréquemment allongés dans RedditGender

TABLEAU 10.14 – Fréquence des étirements de lettres, pour 1000 tokens

Sous-corpus	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.47	0.59	0.27	0.50
Femmes cisgenres	0.61	0.61	0.44	0.70
Femmes transgenres	0.50	0.49	0.37	0.54
Hommes transgenres	0.46	0.61	0.22	0.51
Non-binaires	0.46	0.52	0.26	0.45
14-20 ans	0.72	0.72	0.52	0.73
21-30 ans	0.54	0.58	0.36	0.62
31 et +	0.41	0.51	0.22	0.44
Tous	0.52	0.59	0.33	0.59

transgenres ont utilisé le moins d'étirements de lettres. La fréquence des étirements de lettres semble diminuer avec l'âge, avec une médiane de 0.52 pour les plus jeunes, et de 0.22 pour les âgé·es.

Modèle de régression : effets de l'âge et du genre

Nous avons créé un modèle de régression linéaire généralisé binomial négatif. La variable dépendante est la fréquence brute des étirements de lettres ; les variables indépendantes sont l'âge, le genre et leur interaction. Un offset a été ajouté au modèle pour prendre en compte les différentes tailles des sous-corpus. La fonction `step()` a montré que toutes les variables intégrées contribuent au modèle. Le modèle ainsi créé est présenté dans le tableau 10.15.

Tableau 10.15 – Étirements de lettres, effet de l'interaction du genre et de l'âge

	Variable dépendante :
	Étirements de lettres
Constante	0.001^{**} (0.001, 0.001)
Femmes cisgenres	0.815 (0.531, 1.275)
Femmes transgenres	0.877 (0.491, 1.679)
Hommes transgenres	0.460^{**} (0.289, 0.752)
Non-binaires	0.616 (0.353, 1.139)
21-30 ans	0.534^{**} (0.390, 0.725)
31 ans et +	0.370^{**} (0.272, 0.498)
Femmes cisgenres :21-30 ans	1.670^{*} (1.015, 2.711)
Femmes transgenres :21-30 ans	1.271 (0.618, 2.486)
Hommes transgenres :21-30 ans	2.180^{**} (1.220, 3.844)
Non-binaires :21-30 ans	1.589 (0.802, 3.019)
Femmes cisgenres :31 ans et +	1.988^{**} (1.202, 3.246)
Femmes transgenres :31 ans et +	1.239 (0.584, 2.520)
Hommes transgenres :31 ans et +	3.310^{**} (1.618, 6.961)
Non-binaires :31 ans et +	2.027 (0.950, 4.248)
Observations	1,041
Log Likelihood	−3,415.139
θ	1.039^{**} (0.050)
Akaike Inf. Crit.	6,860.277
Note :	*p<0.05 ; **p<0.01

L'effet de l'âge est significatif pour les hommes cisgenres ; plus ils sont âgés, moins ils produisent d'étirements de lettres. Pour les autres groupes, il ne l'est pas ; on constate uniquement une différence significative entre la production des femmes transgenres les plus jeunes et les plus âgées, avec également une diminution de la fréquence des étirements de lettres avec l'âge. La figure 10.5 présente l'interaction du genre et de l'âge ; notons que, même si la fréquence des étirements graphiques diminue chez les femmes cisgenres, la différence n'est pas significative (ce qui indique qu'elle est due au hasard).

Dans le groupe le plus jeune, les femmes et les hommes cisgenres pro-

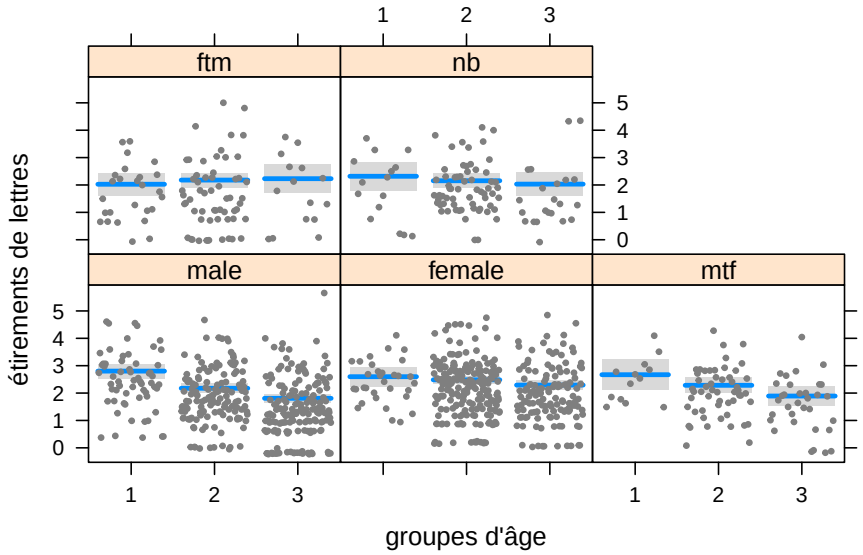

FIGURE 10.5 – Interaction du genre et de l'âge, étirements de lettres

duisent plus de répétitions de lettres que les hommes et les femmes transgenres. Les femmes cisgenres de 21 à 30 ans utilisent plus ce procédé que les hommes cisgenres et transgenres et les personnes non binaires. Il n'y a pas de différence significative entre femmes cisgenres et transgenres. Chez les Redditors les plus âgé·es, les femmes cisgenres utilisent plus de répétitions de lettres que les femmes transgenres et les hommes cisgenres.

Étude du groupe non-binaire

Les étirements graphiques de lettres semblant être un marqueur de féminité, nous avons pensé qu'il pourrait être intéressant de savoir s'il existe des différences dans leur fréquence dans le groupe non binaire. Comme pour les émoticônes (→ p. 223), nous avons donc créé un modèle séparé à partir du groupe non binaire, avec l'âge, le genre assigné à la naissance et leur interaction comme variables indépendantes. Le niveau de référence du modèle (tableau 10.16) est les personnes non binaires AFAN.

Les variables ont été sélectionnées de la façon habituelle. L'algorithme a conservé l'interaction du genre assigné à la naissance et de l'âge. Le modèle montre qu'il n'y a pas de différence significative entre les personnes AFAN et les personnes AGAN dans le groupe le plus jeune (14-20 ans) et le groupe le plus âgé (31 ans et plus). Chez les Redditors non binaires de 21 à 30 ans, les personnes AFAN produisent 2.22 fois plus d'étirements de lettres que les personnes AGAN. Chez les personnes AFAN, l'âge n'a pas d'effet significatif sur la production d'étirements de lettres. Il existe en revanche une différence significative entre le groupe 1 (14-20 ans) et le groupe 2 (21-30 ans) : les plus jeunes produisent davantage d'étirements de lettres que les Redditors qui ont de 21 à 30 ans. Il n'y a pas de différence entre les deux

Tableau 10.16 – Étirements de lettres, effet du genre assigné à la naissance

	Variable dépendante :
	Étirements de lettres
Intercept	0.0004** (0.0002, 0.001)
AGAN	1.504 (0.543, 4.092)
21-30 ans	1.296 (0.549, 2.738)
31 ans et +	0.888 (0.338, 2.203)
AGAN :21-30 ans	0.301* (0.097, 0.949)
AGAN :31 ans et +	0.771 (0.211, 2.866)
Observations	98
Log Likelihood	−309.079
θ	1.174** (0.188)
Akaike Inf. Crit.	630.157
Note :	*p<0.05 ; **p<0.01

groupes les plus âgés.

10.4.3 Effet de l'ethnicité sur la fréquence des étirements de lettres

Statistiques descriptives

Le tableau 10.17 présente les fréquences moyennes et médianes des étirements de lettres dans l'échantillon réduit décrit page 135.

Tableau 10.17 – Fréquence des étirements de lettres par 1000 tokens

Sous-corpus	Moyenne	ET	Médiane	EI
Hommes	0.48	0.55	0.28	0.47
Femmes	0.65	0.63	0.47	0.69
14-20 ans	0.70	0.70	0.46	0.63
21-30 ans	0.62	0.64	0.42	0.68
31 ans et +	0.42	0.45	0.21	0.45
Blancs	0.55	0.59	0.33	0.69
Afro-Américain·es	0.60	0.60	0.43	0.69
Asiatiques	0.42	0.50	0.26	0.32
Hispaniques	0.67	0.68	0.45	0.71

On remarque la même utilisation plus fréquente des étirements de lettres par les femmes que dans l'ensemble du corpus, ainsi que l'apparente corrélation négative avec l'âge. Il semble par ailleurs que les groupes afro-américain et hispanique emploient davantage d'étirements de lettres que les autres groupes, avec des fréquences médianes respectives de 0.43 et 0.45, contre 0.33 et 0.26 pour les blanc·hes et les Asiatiques.

Effet de l'ethnicité, du genre et l'âge : modèle de régression

Nous avons inclus le genre, l'âge, et la race comme variables dépendantes et leur interaction dans le modèle binomial négatif. La fonction `step()` a conservé l'interaction du genre et de l'âge ainsi que les effets principaux du genre, de l'âge et de l'ethnicité. Le modèle ainsi réalisé est présenté dans le tableau 10.18. Les hommes cisgenres de 14 à 20 ans sont le niveau de référence.

TABLEAU 10.18 – Étirements de lettres, effets principaux de l'ethnicité, du genre et de l'âge

	Variable dépendante :
	Étirements de lettres
Intercept	0.001^{**} (0.001, 0.001)
Femmes	0.743 (0.413, 1.366)
21-30 ans	0.681 (0.438, 1.032)
31 ans et +	0.432^{**} (0.274, 0.666)
Afro-américaines	1.052 (0.805, 1.380)
Asiatiques	0.722^{*} (0.532, 0.985)
Hispaniques	1.138 (0.849, 1.535)
Femmes :21-30 ans	1.904 (0.976, 3.650)
Femmes :31 ans et +	2.298^{*} (1.133, 4.598)
Observations	347
Log Likelihood	−1,154.425
θ	1.162^{**} (0.098)
Akaike Inf. Crit.	2,326.849
Note :	*p<0.05 ; **p<0.01

On remarque la même corrélation négative constatée précédemment entre âge et fréquence des étirements de lettres chez les hommes. Elle est toutefois partielle, car il n'y a pas de différence significative entre le groupe 1 et le groupe 2 ; la valeur p approche toutefois du seuil de significativité (p = 0.08). Cette corrélation est absente chez les femmes, comme le montre la figure 10.6. Dans le groupe le plus jeune, il n'y a pas de différence entre femmes et hommes dans la fréquence des étirements graphiques, mais il y en a une dans les deux groupes plus âgés, les femmes produisant davantage d'étirements graphiques que les hommes. L'effet de l'ethnicité est lui aussi significatif ; les Asiatiques emploient moins d'étirements graphiques que tous les autres groupes. La différence la plus marquée est avec le groupe hispanique, qui emploie 1.58 étirement graphique quand le groupe asiatique en produit 1.

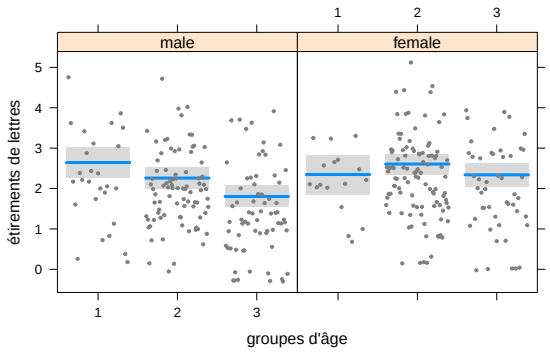

FIGURE 10.6 – Interaction âge et genre dans la production d'étirements de lettres, groupes cisgenres

10.5 Étirements de ponctuation

10.5.1 Effets de l'âge et du genre sur la production d'étirements de ponctuation

Statistiques descriptives

Le corpus contient 14 298 étirements de signes de ponctuation. Cent-trente-huit personnes (13.22 %) n'ont pas utilisé d'étirements de ponctuation du tout. L'exploration des données avec des boites à moustaches (figure 10.7) révèle la présence deux valeurs extrêmes, qui correspondent à la production d'une femme cisgenre et d'une femme transgenre. Elles ont utilisé respectivement 23.28 et 35.62 répétitions de ponctuation par 1000 tokens, alors que la médiane est de 0.32 pour l'ensemble du corpus. Nous avons retiré ces deux observations des données pour les analyses qui suivent.

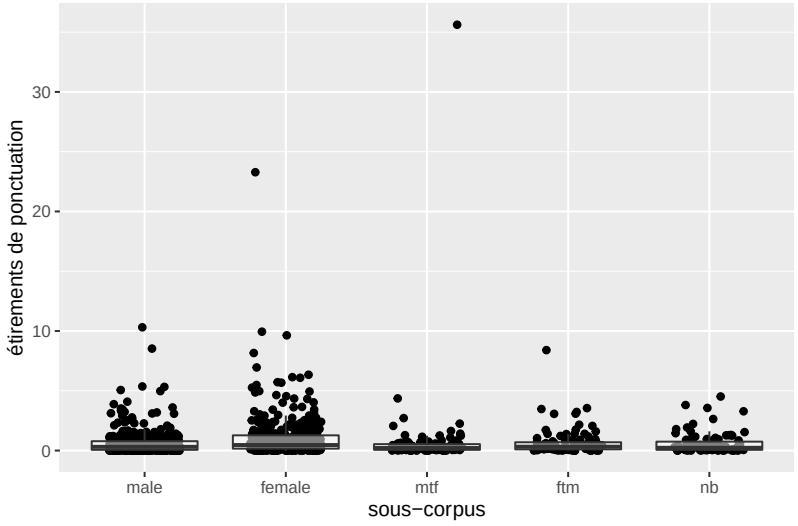

FIGURE 10.7 – Fréquence des étirements de ponctuation par 1000 tokens

Le tableau 10.19 présente les fréquences relatives moyennes et médianes des étirements de ponctuation, par 1000 tokens, par groupes de genre et d'âge. Les femmes cisgenres ont utilisé le plus de répétitions de signes de ponctuation, avec une moyenne de 1 répétition par 1000 tokens, et une médiane de 0.45. Elles sont aussi le groupe où la variation est la plus importante, comme le montrent l'écart type (1.44) et l'écart interquartile (1.10). Elles sont suivies par les hommes transgenres et les cisgenres. Les femmes transgenres ont quant à elles utilisé le moins d'étirements de ponctuation. La fréquence médiane des répétitions des signes de ponctuation semble décroitre légèrement avec l'âge. La différence entre les groupes est faible, avec 0.32 étirements de ponctuation pour les Redditors de 21 à 30 ans, et 0.30 pour les plus âgé·es.

Tableau 10.19 – Fréquence relative des étirements de ponctuation, par 1000 tokens

	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.66	1.09	0.31	0.72
Femmes cisgenres	1.00	1.44	0.45	1.10
Femmes transgenres	0.42	0.64	0.20	0.47
Hommes transgenres	0.67	1.11	0.32	0.59
Non-binaires	0.56	0.86	0.23	0.68
14 à 20 ans	0.86	1.43	0.41	0.78
21 à 30 ans	0.71	1.08	0.32	0.82
31 ans et plus	0.75	1.25	0.30	0.72
Tous	0.75	1.2	0.32	0.79

Effets du genre et de l'âge : modèle de régression

Nous avons créé un modèle de régression binomial négatif pour explorer les effets de l'âge et du genre sur la production d'étirements de ponctuation. Il intègre l'âge, le genre et leur interaction comme variables indépendantes, la fréquence brute des étirements de ponctuation comme variable dépendante, et un offset pour compenser les différentes tailles des sous-corpus. Nous avons utilisé la fonction `step()` pour sélectionner les variables indépendantes. Le modèle, dont les hommes cisgenres est le niveau de référence, est présenté dans le tableau 10.20. Il montre que les femmes cisgenres utilisent davantage d'étirements graphiques que les hommes cisgenres (coefficient exponentialisé = 1.57) et que tous les autres groupes. Les femmes transgenres utilisent moins de répétitions de signes de ponctuation que tous les autres groupes à l'exception des personnes non binaires. Elles en produisent 2 fois moins que les femmes cisgenres (coefficient exponentialisé = 0.42). L'effet de l'âge est significatif uniquement lorsque l'on compare le groupe 1 (14-20 ans) au groupe 2 (21-30 ans). Les Redditors du groupe 2 utilisent moins d'étirements graphiques que les Redditors plus jeunes.

TABLEAU 10.20 – Étirements de ponctuation, effets de l'âge et du genre

	Variable dépendante :
	Étirements de ponctuation
Intercept	0.001** (0.001, 0.001)
Femmes cisgenres	1.566** (1.291, 1.900)
Femmes transgenres	0.662** (0.494, 0.902)
Hommes transgenres	0.987 (0.737, 1.341)
Non-binaires	0.876 (0.653, 1.193)
21-30 ans	0.777* (0.603, 0.992)
31 ans et +	0.784 (0.603, 1.011)
Observations	1,042
Log Likelihood	−3,697.560
θ	0.590** (0.026)
Akaike Inf. Crit.	7,409.119
Note :	*p<0.05 ; **p<0.01

Groupe non binaire

Les étirements de ponctuation semblant être clairement privilégiés par les femmes cisgenres, nous avons réalisé une analyse de cette variable avec le groupe non binaire, pour voir si les internautes AFAN l'utilisent plus fréquemment que les AGAN. Le modèle est présenté dans le tableau 10.21, avec les personnes non binaires AFAN comme niveau de référence. L'interaction entre genre assigné à la naissance et âge a été conservée dans le processus de sélection des variables.

TABLEAU 10.21 – Étirements de ponctuation, effets de l'âge et du genre assigné à la naissance

	Variable dépendante :
	Étirements de ponctuation
Intercept	0.0004** (0.0002, 0.001)
AGAN	1.071 (0.279, 3.970)
21-30 ans	1.650 (0.523, 4.208)
31 ans et +	0.533 (0.148, 1.724)
AGAN :21-30 ans	0.390 (0.090, 1.748)
AGAN :31 ans et +	2.641 (0.490, 14.691)
Observations	98
Log Likelihood	−316.399
θ	0.671** (0.099)
Akaike Inf. Crit.	644.799
Note :	*p<0.05 ; **p<0.01

Le modèle révèle une seule différence entre les personnes AGAN et AFAN. Les personnes AFAN de 21 à 30 ans utilisent 2.39 fois plus d'étirements de ponctuation que les personnes AGAN du même âge. Les personnes AFAN de 21 à 30 ans emploient par ailleurs davantage d'étirements de ponctuation que les personnes AFAN les plus âgées.

Corrélations entre le genre et les différents types d'étirements de ponctuation

Dans les 14 298 répétitions de signes de ponctuation identifiées dans le corpus, les répétitions de points d'exclamation sont majoritaires, avec 6970 occurrences (soit 48.75 % de tous les étirements de ponctuation), contre 2627 (18.37 %) pour les points d'interrogation, et 4701 (32.88 %) pour les points de suspension. Nous avons souhaité savoir s'il existe une corrélation entre genre et types d'étirements de ponctuation. Nous avons donc réalisé un diagramme en mosaïque, basé sur le tableau de contingence 10.5.1, qui présente les fréquences relatives des trois types d'étirements de ponctuation, par million de tokens, pour chaque groupe de genre.

Tableau 10.22 – Fréquence des étirements de ponctuation par million de tokens

	Exclamation	Interrogation	Suspension
Hommes cisgenres	257.09	102.19	285.70
Femmes cisgenres	539.49	197.77	261.04
Femmes transgenres	181.45	62.85	178.71
Hommes transgenres	350.13	132.24	165.43
Non-binaires	272.25	108.90	166.58

Le diagramme en mosaïque de la figure 10.8 présente les résidus de Pearson. Il montre que la fréquence de tous les étirements graphiques est la plus importante chez les femmes cisgenres (f) : les tailles de leurs rectangles sont les plus importantes. Les femmes transgenres en ont utilisé le moins. L'examen des résidus montre que, chez les femmes cisgenres, les valeurs observées des étirements de points d'exclamation sont supérieures aux effectifs théoriques (rectangle bleu, et résidus compris entre 2 et 4). Pour les points de suspension, avec des effectifs observés inférieurs aux effectifs théoriques. On constate le phénomène inverse chez les hommes (m), avec une fréquence observée des points d'exclamation inférieure à la fréquence théorique (rectangle rose), et une fréquence des points de suspension bien supérieure à la fréquence théorique (rectangle bleu foncé). Il semble y avoir une association positive entre l'utilisation de répétitions de points de suspension et les femmes transgenres (mtf), et une association négative avec les hommes transgenres (ftm). Aucune relation entre les différents types d'étirements graphiques n'est constatée chez le groupe non binaire (nb).

10.5.2 Effets de l'interaction entre genre et ethnicité sur la production d'étirements de ponctuation

Statistiques descriptives

Les boites à moustaches de la figure 10.9 présentent la fréquence des étirements de ponctuation par 1000 tokens dans chaque sous-corpus. On y voit que le nombre médian d'étirements de ponctuation est plus élevé chez

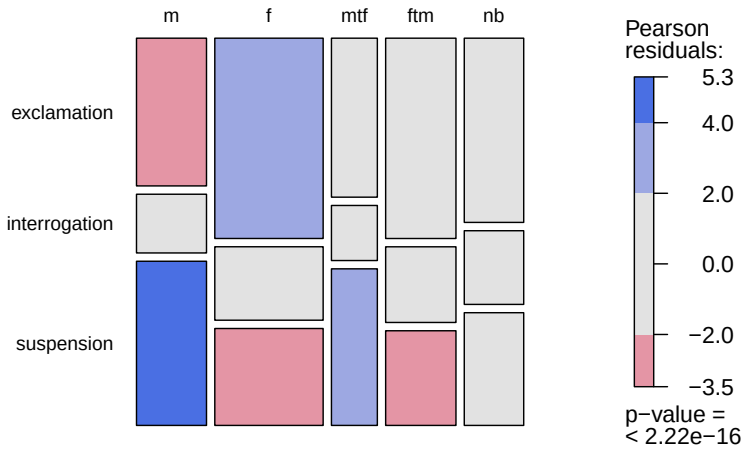

FIGURE 10.8 – Fréquence des différents étirements de ponctuation par groupe de genre

les femmes que chez les hommes, à l'exception du groupe blanc. Les données sont également plus dispersées chez les femmes que chez les hommes, ce qui indique qu'il y a davantage de variation entre les individus. Ainsi, même si la médiane des femmes et des hommes blancs est similaire, il y a davantage de valeurs élevées chez les femmes blanches.

Régression : effets de l'ethnicité, de l'âge et du genre

Le modèle de régression négative binomiale (tableau 10.23) a comme variable dépendante la fréquence brute des étirements de ponctuation pour chaque personne, et comme variables indépendantes l'âge et l'interaction du genre et de l'ethnicité.

Selon le modèle, l'âge n'a pas d'effet significatif sur la production d'étirements de ponctuation. Les femmes produisent davantage d'étirements de ponctuation que les hommes dans tous les groupes à l'exception du groupe hispanique. L'effet est le plus fort dans le groupe asiatique, où les femmes produisent 2.43 fois plus d'étirements de ponctuation que les hommes, contre 1.80 fois dans le groupe blanc et 2.31 dans le groupe afro-américain.

Quand on compare les hommes entre eux, on voit que les hommes asiatiques emploient moins d'étirements de ponctuation que tous les autres groupes, et surtout que les hommes hispaniques, qui en utilisent 2.63 fois plus. Chez les femmes, le groupe asiatique emploie 2 fois moins d'étirements de ponctuation que les femmes afro-américaines, et près de 3 fois moins que les femmes hispaniques. L'interaction du genre et de l'ethnicité

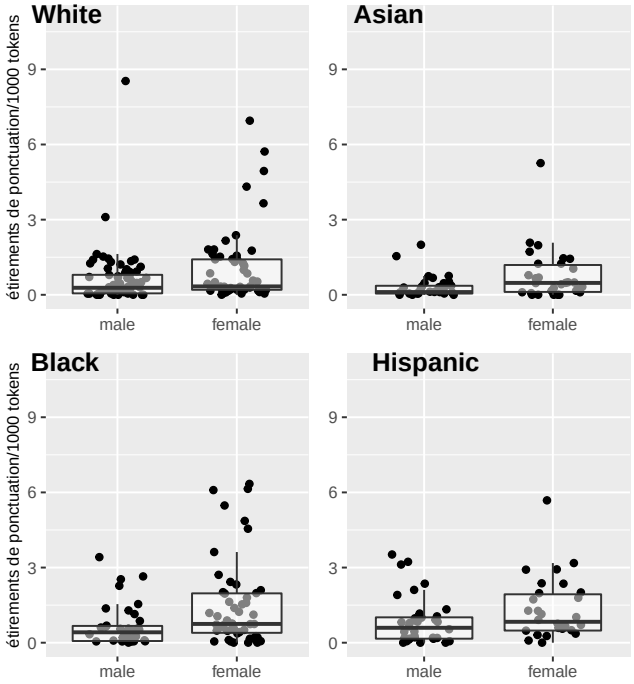

FIGURE 10.9 – Fréquence des étirements de ponctuation par 1000 tokens, groupes ethniques

TABLEAU 10.23 – Étirements de ponctuation, effets de l'âge, du genre et de l'ethnicité

	Variable dépendante :
	Étirements de ponctuation
Intercept	0.001** (0.0005, 0.001)
Femmes	1.806** (1.152, 2.849)
Afro-Américains	1.135 (0.684, 1.924)
Asiatiques	0.539* (0.314, 0.949)
Hispaniques	1.418 (0.851, 2.424)
21-30 ans	0.877 (0.555, 1.344)
31 ans et +	0.798 (0.497, 1.250)
Femmes :Afro-Américaines	1.280 (0.633, 2.559)
Femmes :Asiatiques	1.351 (0.615, 2.975)
Femmes :Hispaniques	0.835 (0.391, 1.793)
Observations	346
Log Likelihood	−1,272.654
θ	0.653** (0.051)
Akaike Inf. Crit.	2,565.307
Note :	*p<0.05 ; **p<0.01

est présentée dans la figure 10.10.

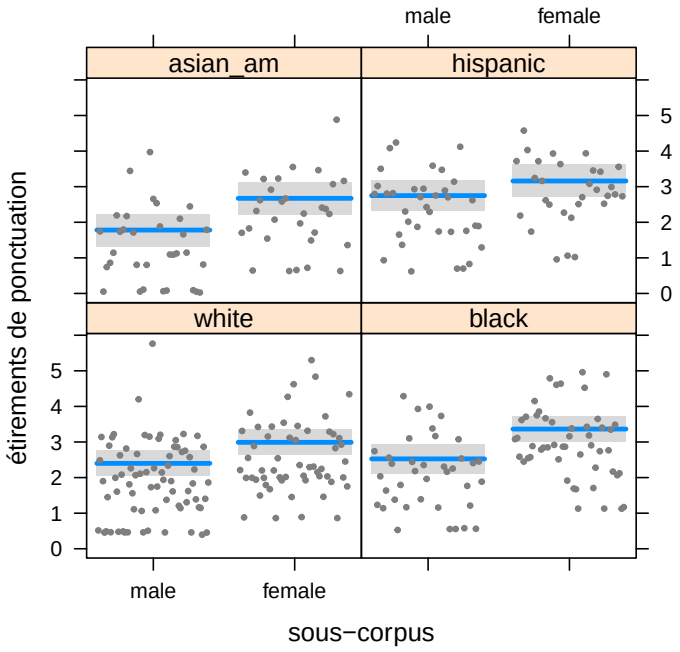

FIGURE 10.10 – Fréquence des étirements de ponctuation par 1000 tokens, groupes ethniques

10.6 Mots en majuscules (*all caps*)

10.6.1 Statistiques générales

Une première exploration visuelle des données révèle la présence d'une valeur extrême. Il s'agit d'un garçon de 15 ans qui a utilisé 1380 mots en majuscules. L'inspection de ses commentaires montre qu'il a répété plus d'un millier de fois le nom d'un constructeur de moto. Nous l'avons donc supprimé du jeu de données pour cette analyse. En moyenne, les Redditors de RedditGender ont utilisé 2.16 mots en majuscules par 1000 tokens (écart type = 3). La médiane est de 1.17 (écart interquartile = 1.86). Une fois la valeur aberrante retirée des données, il y a dans le corpus 39 478 tokens de trois lettres et plus en majuscules, dont 9579 types différents. Le tableau 10.24 présente les 20 mots en majuscules les plus fréquents dans le corpus, avec leur rang, leur fréquence brute dans le corpus et la proportion qu'ils occupent dans l'ensemble des mots en majuscules.

La majorité des mots en majuscules sont des mots grammaticaux : les déterminants *the*, *this*, *that*, *one*, les pronoms *you*, *your*, *what*, la conjonction de coordination *but*, l'auxiliaire *have*, le modal *will*, ou l'adverbe *yes*. On trouve également dans la liste des adverbes ou adjectifs d'intensité comme *(a) lot*, *really*, *very*, *much*, *just*, *fucking*, ou *never*, ainsi que *love*. Le mot

Tableau 10.24 – 20 mots en majuscules les plus fréquents dans le corpus

Rang	Mots	Fréq. maj.	%
1	THE	1265	3.31 %
2	YOU	1141	2.99 %
3	EDIT	841	2.20 %
4	THIS	540	1.41 %
5	THAT	527	1.38 %
6	LOVE	471	1.23 %
7	LOT	398	1.04 %
8	REALLY	373	0.98 %
9	BUT	347	0.91 %
10	YES	339	0.89 %
11	YOUR	322	0.84 %
12	VERY	315	0.82 %
13	WHAT	307	0.80 %
14	HAVE	273	0.71 %
15	FUCKING	261	0.68 %
16	NEVER	252	0.66 %
17	ONE	250	0.65 %
18	JUST	228	0.60 %
19	MUCH	196	0.51 %
20	WILL	190	0.50 %

edit arrive en troisième position. Cela s'explique par le fait que le code de conduite de Reddit, la Reddiquette, conseille aux internautes d'expliquer pourquoi ils ont modifié le contenu d'un message (« reddit », p. d.), comme dans les deux exemples suivants tirés du corpus (paraphrasés) :

> My comment may sound stupid, but I think it's a fair question.
> EDIT : Spelling

> I've seen neck tattoos in NYC. Not a face one though. EDIT : TIL very few artists are willing to do a face tattoo.

10.6.2 Effets de l'âge et genre sur la fréquence des mots en majuscules

Statistiques descriptives

Le tableau 10.25 montre que les hommes et les femmes cisgenres ont utilisé plus de mots en majuscules que les groupes transgenres. Les personnes non binaires ont utilisé le moins de mots en majuscules. La fréquence des mots en majuscules semble diminuer avec l'âge, avec une médiane de 1.62 pour les personnes de 14 à 20, de 1.20 pour les personnes de 21 à 30 ans, et de 0.99 pour les Redditors les plus âgé·es. La dispersion est importante ; elle est la plus élevée chez les personnes cisgenres et les Redditors les plus jeunes, ce qui pointe vers des différences individuelles plus marquées.

Tableau 10.25 – Fréquence des mots en majuscules dans RedditGender

	Moyenne	ET	Médiane	EI
Hommes cisgenres	2.37	3.28	1.26	1.92
Femmes cisgenres	2.13	3.20	1.34	1.98
Femmes transgenres	1.64	2.27	1.02	1.15
Hommes transgenres	1.65	2.50	1.00	1.48
Non-binaires	1.48	1.92	0.71	1.74
14-20 ans	3.16	4.05	1.62	2.87
21-30 ans	1.89	2.31	1.20	1.83
31 ans et +	1.87	3.27	0.99	1.65
Tous	2.16	3.00	1.17	1.86

Choix du modèle

Au vu de la dispersion des données, nous avons créé un modèle binomial négatif. La déviance résiduelle était importante (11 747 pour 1028 degrés de liberté), indiquant une surdispersion. Nous avons donc créé à la place un modèle linéaire généralisé mixte Poisson, en ajoutant un effet aléatoire pour chaque observation (OLRE), technique conseillée en cas de surdispersion (Harrison, 2014). Nous avons également dû résoudre l'échec de convergence du modèle en changeant son *optimizer* (« lme4 convergence warnings : troubleshooting », p. d.). Le modèle créé est présenté dans le tableau 10.26, avec les coefficients, l'intervalle de confiance de 95 % et les valeurs p. Le modèle contient comme variable dépendante la fréquence brute des mots en majuscules, et comme variables indépendantes le genre, l'âge et leur interaction. Nous avons utilisé la fonction `step()` pour vérifier qu'elles contribuaient bien au modèle. Les hommes cisgenres de 14 à 20 ans sont le niveau de référence du modèle. Les coefficients et les intervalles de confiance n'ont pas été exponentialisés parce que ce type de modèle n'est pas pris en charge par le script que nous avons utilisé pour les exponentialiser de façon automatique (→ p. 169).

Résultats

Dans le groupe le plus jeune, il n'y a pas de différence significative entre les femmes et les hommes cisgenres. Les groupes transgenres utilisent en revanche moins de mots en majuscules que les hommes cisgenres et les femmes cisgenres. Les personnes non binaires sont celles qui utilisent le moins de mots en majuscules comparées aux hommes cisgenres du même âge, avec une fréquence équivalente à un tiers (coefficient exponentialisé = 0.34) de celle du groupe des hommes cisgenres. On constate le même phénomène chez les Redditors de 21 à 30 ans : il n'y a pas de différence entre les groupes cisgenres, tandis que les groupes transgenres et non binaire emploient significativement moins de mots en majuscules que les hommes cisgenres. Chez les Redditors les plus âgé·es, les différences s'atténuent. Seules les personnes non binaires utilisent moins de mots en majuscules

TABLEAU 10.26 – Mots en majuscules, interaction de l'âge et du genre

	Variable dépendante :
	Mots en majuscules
Intercept	-6.069^{**} (-6.352, -5.787)
Femmes cisgenres	-0.121 (-0.600, 0.359)
Femmes transgenres	-0.819^{*} (-1.491, -0.147)
Hommes transgenres	-0.947^{**} (-1.468, -0.426)
Non-binaires	-1.098^{**} (-1.745, -0.452)
21-30 ans	-0.555^{**} (-0.894, -0.217)
31 ans et +	-0.885^{**} (-1.215, -0.556)
Femmes cisgenres :21-30 ans	0.073 (-0.464, 0.610)
Femmes transgenres :21-30 ans	0.216 (-0.546, 0.977)
Hommes transgenres :21-30 ans	0.453 (-0.172, 1.078)
Non-binaires :21-30 ans	0.650 (-0.079, 1.379)
Femmes cisgenres :31 ans et +	0.221 (-0.321, 0.763)
Femmes transgenres :31 ans et +	1.011^{*} (0.218, 1.804)
Hommes transgenres :31 ans et +	1.012^{*} (0.220, 1.804)
Non-binaires :31 ans et +	0.515 (-0.310, 1.341)
Observations	1,043
Log Likelihood	$-4,759.447$
Akaike Inf. Crit.	9,550.895
Bayesian Inf. Crit.	9,630.092
Note :	*p<0.05 ; **p<0.01

que les hommes cisgenres, les femmes cisgenres et les femmes transgenres. Aucune autre différence significative n'est constatée.

L'interaction de l'âge et du genre ne produit pas les mêmes effets chez les groupes cisgenres et les groupes transgenres. Chez les hommes et les femmes cisgenres, on constate une corrélation négative entre l'âge et la fréquence des mots en majuscules, visible dans la figure 10.11. Les Redditors les plus jeunes utilisent davantage de mots en majuscules que les Redditors des deux autres groupes d'âge, et les Redditors les plus âgé·es en utilisent moins que les autres. Il faut noter, toutefois, que la différence entre le groupe de femmes de 21 à 30 ans et le groupe de femmes de 31 ans et plus n'est pas significative. On ne retrouve pas cette corrélation dans les groupes transgenres, où aucune différence significative n'est constatée entre les catégories d'âge.

10.6.3 Effet de l'ethnicité et de son interaction avec le genre sur la production des mots en majuscules

Statistiques descriptives

La figure 10.12 présente la fréquence relative des mots en majuscules dans chaque sous-corpus. On remarque que ceux-ci sont plus fréquents chez les femmes, à l'exception du groupe hispanique, où la différence semble très faible. Ce groupe est celui où on trouve le plus de dispersion, ce qui montre que les différences individuelles y sont les plus fortes. Il semble également

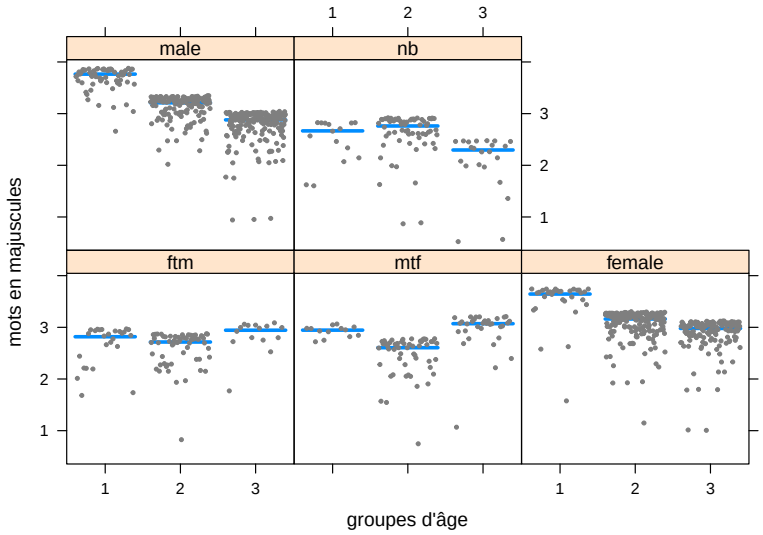

FIGURE 10.11 – Interaction genre et âge, fréquence des mots en majuscules

que les femmes et les hommes de ce groupe utilisent davantage de mots en majuscules que les internautes des autres groupes.

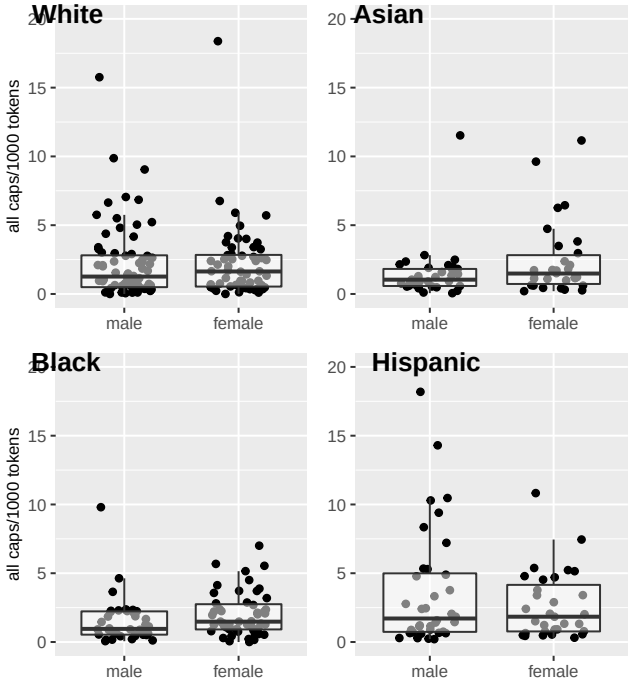

FIGURE 10.12 – Fréquence des mots en majuscules par sous-corpus

Effet de l'ethnicité : modèle de régression

Le modèle de régression binomiale négative présenté dans le tableau 10.27 intègre, comme variables indépendantes, le genre, l'ethnicité et leur interaction. Le niveau de référence est les hommes blancs. Toutes les variables ont été conservées par la fonction `step()`. Le modèle indique une corrélation négative entre âge et fréquence des mots en majuscules. Il n'y a pas de différence significative entre femmes et hommes dans aucun groupe. Aucune différence significative entre les groupes de femmes n'est également révélée. En revanche, chez les hommes, les Hispaniques utilisent davantage de mots en majuscules que tous les autres groupes : 1.57 fois plus que les hommes blancs, et 2.17 fois plus que les hommes asiatiques et afro-américains.

Tableau 10.27 – Effets de l'âge et de l'interaction genre et ethnicité sur la fréquence des mots en majuscules

	Variable dépendante :
	Mots en majuscules
Intercept	0.003** (0.002, 0.005)
Femmes cisgenres	1.045 (0.750, 1.463)
Afro-Américains	0.707 (0.487, 1.039)
Asiatiques	0.706 (0.478, 1.060)
Hispaniques	1.535* (1.059, 2.260)
21-30 ans	0.706* (0.508, 0.964)
31 ans et +	0.505** (0.357, 0.706)
Femmes :Afro-américaines	1.329 (0.787, 2.233)
Femmes :Asiatiques	1.392 (0.780, 2.486)
Femmes :Hispaniques	0.648 (0.369, 1.139)
Observations	347
Log Likelihood	−1,622.178
θ	1.156** (0.083)
Akaike Inf. Crit.	3,264.355
Note :	*p<0.05 ; **p<0.01

10.7 Interjections

10.7.1 Fréquence des interjections dans le corpus

Statistiques générales

29 587 interjections, dont 87 types différents, ont été identifiées dans le corpus. Une première exploration des données à l'aide d'une boite à moustaches révèle la présence de 2 valeurs aberrantes : un homme cisgenre de 25 ans et une personne non binaire de 28 ans qui ont utilisé respectivement 10.73 et 10.75 interjections par 1000 tokens, alors que la médiane est de 1.27. Nous avons enlevé ces deux observations des données pour les analyses qui suivent.

Types d'interjections identifiés

Le tableau 10.28 présente les 15 interjections les plus fréquentes dans le corpus, avec leur fréquence brute et la proportion qu'elles occupent dans l'ensemble des interjections recensées. La liste complète des 87 interjections identifiées est présentée en annexe C.

Tableau 10.28 – 15 interjections les plus fréquentes dans RedditGender

	Interjections	Fréq.	%
1	oh	9225	31.55
2	haha	3814	13.05
3	hey	3134	10.72
4	ah	1396	4.78
5	ugh	1280	4.38
6	ha	784	2.68
7	um	778	2.66
8	eh	747	2.56
9	yay	718	2.46
10	huh	694	2.37
11	hmm	631	2.16
12	uh	630	2.15
13	aww	452	1.55
14	meh	434	1.48
15	blah	380	1.30
1 à 15	-	25097	85.85

Les 15 interjections les plus fréquentes du corpus représentent 85.85 % des interjections identifiées dans le corpus. *oh*, la plus fréquente, représente près d'un tiers (31.55 %) de toutes les interjections recensées.

10.7.2 Effets du genre et de l'âge sur la fréquence des interjections

Statistiques descriptives

Le tableau 10.29 présente les fréquences relatives moyennes et médianes des interjections par sous-corpus, pour 1000 tokens, ainsi que les écarts types (ET) et les écarts interquartiles (EI) de chaque groupe. Il semble indiquer que la fréquence des interjections est corrélée négativement avec l'âge : moyenne et médiane baissent d'un groupe à l'autre. Les interjections sont donc plus nombreuses chez les plus jeunes ; il faut toutefois noter que c'est dans ce groupe que la dispersion est la plus importante. Les femmes cisgenres et les hommes transgenres semblent employer plus d'interjections que les autres groupes. Les hommes cisgenres semblent en utiliser le moins. Notons par ailleurs que seuls 3 Redditors (0.29 %) n'ont pas utilisé une seule des interjections étudiées ici.

Tableau 10.29 – Fréquence relative des interjections, pour 1000 tokens

Groupes	Moyenne	ET	Médiane	EI
Hommes cisgenres	1.35	1.15	1.03	1.12
Femmes cisgenres	1.68	1.12	1.46	1.35
Femmes transgenres	1.54	1.12	1.34	1.30
Hommes transgenres	1.67	1.21	1.40	1.60
Non-binaires	1.43	0.98	1.14	1.21
14-20 ans	2.16	1.60	1.90	2.06
21-30 ans	1.57	1.04	1.37	1.25
31 et +	1.21	0.92	1.02	1.05
Tous	1.54	1.20	1.25	1.33

Effet de l'interaction du genre et de l'âge sur la fréquence des interjections : modèle de régression

Pour étudier les corrélations possibles entre âge, genre et leur interaction et l'utilisation d'interjection, nous avons créé un modèle de régression linéaire généralisé binomial négatif (tableau 10.30). Nous avons utilisé la fonction `step()` pour procéder à la sélection des variables. Toutes les variables ont été conservées. Les femmes cisgenres de 14 à 20 ans sont le niveau de référence du modèle.

Tableau 10.30 – Interjections, effet de l'interaction de l'âge et du genre

	Variable dépendante :
	Interjections
Intercept	0.002^{**} (0.002, 0.003)
Hommes cisgenres	1.229 (0.921, 1.627)
Femmes transgenres	1.044 (0.687, 1.623)
Hommes transgenres	0.975 (0.692, 1.379)
Non-binaires	0.773 (0.516, 1.176)
21-30 ans	0.878 (0.680, 1.118)
31 ans et +	0.732^{*} (0.563, 0.940)
Hommes cisgenres :21-30 ans	0.613^{**} (0.446, 0.846)
Femmes transgenres :21-30 ans	0.845 (0.522, 1.346)
Hommes transgenres :21-30 ans	0.909 (0.612, 1.350)
Non-binaires :21-30 ans	1.083 (0.684, 1.693)
Hommes cisgenres :30 ans et +	0.534^{**} (0.388, 0.739)
Femmes transgenres :30 ans et +	0.830 (0.500, 1.360)
Hommes transgenres :30 ans et +	1.089 (0.667, 1.801)
Non-binaires :30 ans et +	1.051 (0.628, 1.749)
Observations	1,042
Log Likelihood	−4,348.567
θ	2.416^{**} (0.112)
Akaike Inf. Crit.	8,727.135
Note :	*p<0.05 ; **p<0.01

L'âge est corrélé négativement avec la fréquence des interjections chez les hommes cisgenres. Quand les hommes de 14 à 20 ans utilisent 1 inter-

jection, les hommes de 21 à 30 ans en utilisent 0.53 et les hommes de 31 ans et plus 0.39. La différence est également significative entre le groupe 2 et 3 : les Redditors qui ont entre 21 et 30 ans produisent des interjections à un rythme 1.37 fois plus élevé que les plus âgés. Chez les femmes cisgenres, il y a une différence significative entre les femmes cisgenres les plus âgées et les deux autres groupes, mais pas entre les femmes de 14 à 20 ans et les femmes de 21 à 30 ans. Quand une femme du groupe 3 produit 1 interjection, les femmes de 14 à 20 ans en produisent 1.37 et les femmes de 21 à 30 ans en utilisent 1.19. Chez les femmes transgenres, l'effet de l'âge est plus limité ; il y a une différence significative entre les femmes de 14 à 20 ans et les femmes de 31 ans et plus. L'âge n'a pas d'impact significatif dans la production d'interjection chez les hommes transgenres et les personnes non binaires. La figure 10.13 présente l'interaction de l'âge et du genre.

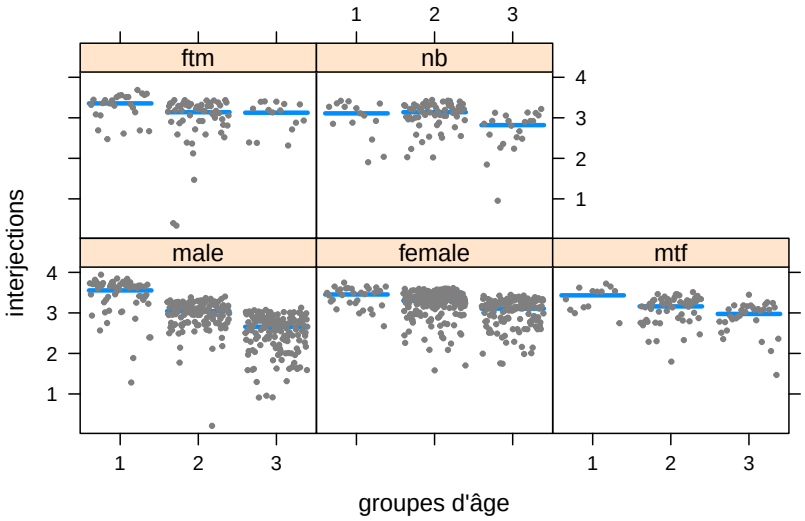

FIGURE 10.13 – Interaction genre et âge, fréquence des interjections

Chez les Redditors les plus jeunes, le modèle indique une seule différence significative entre les groupes de genre : les non-binaires utilisent moins d'interjections que les hommes cisgenres. Chez les Redditors de 21 à 30 ans, les femmes cisgenres utilisent plus d'interjections que les hommes cisgenres ; elles en produisent 1.32 quand un homme cisgenre en utilise 1. C'est chez les Redditors les plus âgé·es que l'on trouve le plus de différences : les hommes cisgenres emploient moins d'interjections que les femmes cisgenres, les femmes transgenres, et les hommes transgenres. La taille d'effet est la plus importante quand on compare les hommes cisgenres aux hommes transgenres, qui utilisent 61 % d'interjections de plus. Il n'y a pas de différence significative entre les autres groupes.

Groupe non binaire

Les interjections étant davantage utilisées par les femmes cisgenres dans le corpus, nous avons souhaité savoir s'il y avait une différence, dans le groupe non binaire, entre les personnes assignées femmes à la naissance et les personnes assignées hommes. Le modèle réalisé (tableau 10.31) ne comprend pas l'interaction du genre et de l'âge, qui selon la fonction `step()` ne contribuait pas au modèle. Les AFAN sont le niveau de référence. Le modèle indique qu'il n'y a pas de différence significative entre les groupes.

Tableau 10.31 – Interjections, effet de l'âge assigné à la naissance

	Variable dépendante :
	Interjections
Intercept	0.002^{*} (0.001, 0.002)
AGAN	0.934 (0.706, 1.243)
21-30 ans	0.930 (0.621, 1.360)
30 ans et +	0.761 (0.482, 1.190)
Observations	97
Log Likelihood	−402.738
θ	2.334^{**} (0.353)
Akaike Inf. Crit.	813.477
Note :	*p<0.05 ; **p<0.01

10.7.3 Effet de l'ethnicité et de son interaction avec le genre sur la production d'interjections

Statistiques descriptives

La figure 10.14 présente la distribution des données dans les différents sous-corpus. On remarque que, dans chaque groupe ethnique, les hommes ont utilisé moins d'interjections que les femmes. La différence semble la plus marquée dans le groupe blanc, où la médiane des femmes est près de 2 fois plus élevée que celle des hommes. La différence entre femmes et hommes semble la moins forte dans le groupe hispanique, qui contient également davantage de valeurs extrêmes que les autres groupes.

Régression : effets de l'âge, du genre et de l'ethnicité

Nous avons créé un modèle de régression binomial négatif avec, comme variables dépendantes, l'âge, le genre, l'ethnicité, et l'interaction du genre et de l'ethnicité. Les hommes blancs sont le niveau de référence du modèle. Le modèle de régression montre que les femmes blanches et asiatiques utilisent davantage d'interjections que les hommes de leurs groupes. Une femme blanche produit ainsi 1.52 interjection quand un homme blanc en produit 1. Une femme asiatique produit 1.40 interjection quand un homme asiatique en produit 1. Il n'y a pas de différence significative entre femmes

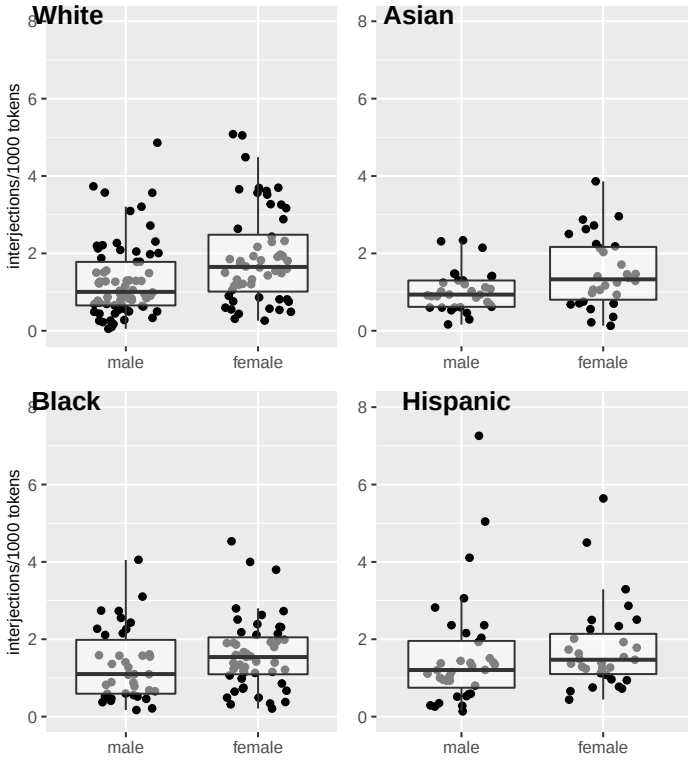

FIGURE 10.14 – Fréquence des interjections par 1000 tokens

et hommes dans les groupes hispanique et afro-américain. Quand on compare les hommes entre eux, et les femmes entre elles, une seule différence significative est révélée par le modèle : les hommes hispaniques emploient davantage d'interjections que les hommes asiatiques. L'effet de l'âge est limité : les Redditors de 31 ans et plus utilisent moins d'interjections que les deux groupes plus jeunes.

10.8 Discussion

10.8.1 Émoticônes

Sans surprise, l'émoticône la plus fréquente du corpus est **:)**, ce qui est le cas dans l'immense majorité des corpus de CMC (→ p. 68). On remarque des tendances notées par les études précédentes : le fait qu'un petit nombre d'émoticônes est utilisé fréquemment, le fait que les émoticônes abrégées (ou sans nez) soient plus fréquentes que les émoticônes « longues », ou encore la prédominance des émoticônes positives. Notre liste des émoticônes les plus fréquentes ressemble ainsi beaucoup à celles que l'on peut trouver dans d'autres études, comme celle de T. Schnoebelen (2012). La fréquence des émoticônes par millier de tokens dans tout le corpus (1.10) est également similaire à ce qu'avait trouvé Coats (2017b) dans son corpus de tweets américains (1.22 émoticône par 1000 mots).

Tableau 10.32 – Interjections, effets de l'âge et de l'interaction du genre et de l'ethnicité

	Variable dépendante :
	Interjections
Intercept	0.002** (0.001, 0.002)
Femmes cisgenres	1.524** (1.220, 1.906)
Afro-Américains	1.060 (0.826, 1.368)
Asiatiques	0.833 (0.636, 1.098)
Hispaniques	1.166 (0.905, 1.510)
21-30 ans	0.831 (0.666, 1.028)
31 ans et +	0.653** (0.519, 0.816)
Femmes :Afro-américaines	0.796 (0.563, 1.124)
Femmes :Asiatiques	0.922 (0.625, 1.360)
Femmes :Hispaniques	0.732 (0.504, 1.065)
Observations	346
Log Likelihood	−1,426.098
θ	2.774** (0.227)
Akaike Inf. Crit.	2,872.196
Note :	*p<0.05 ; **p<0.01

Les modèles de régression réalisés avec différents échantillons et différentes variables permettent de nuancer les résultats des études précédentes sur les émoticônes, qui ont généralement trouvé que les hommes les utilisent moins fréquemment que les femmes. Le premier modèle semble aller dans le sens des résultats des nombreuses études des émoticônes sur les réseaux sociaux, dans le chat ou les SMS. Il révèle une corrélation négative, attendue, de l'âge et de la fréquence des émoticônes ; c'est par exemple ce qu'ont trouvé Oleszkiewicz et al. (2017), qui ont étudié plus de 70 000 utilisateur·trices de Facebook de 16 à 60 ans. Ce premier modèle montre également que les hommes cisgenres emploient significativement moins fréquemment les émoticônes que les femmes cisgenres et que les trois autres groupes de genre. Les hommes transgenres ne s'alignent pas sur les hommes cisgenres, et ne semblent pas avoir un comportement différent des femmes cisgenres et des personnes non binaires. Les femmes transgenres, quant à elles, sont les premières utilisatrices d'émoticônes, loin devant les autres groupes : elles en emploient notamment 3.29 fois plus que les hommes cisgenres, et 1.81 fois plus que les femmes cisgenres.

On pourrait voir ces résultats comme une confirmation du statut de marqueur de féminité des émoticônes : les hommes cisgenres construisent leur masculinité en les évitant, et les femmes transgenres construisent leur féminité en les utilisant fréquemment. L'intégration de l'orientation sexuelle à l'analyse apporte une précision intéressante. Les hommes gays, qu'ils soient cisgenres ou transgenres, utilisent plus fréquemment les émoticônes que les hommes hétérosexuels ; il est possible que les émoticônes entrent dans leur stratégie d'expression de leur identité de genre et de leur identité sexuelle. En d'autres termes, les hommes hétérosexuels n'utiliseraient pas d'émoticônes pour se dissocier des femmes et des hommes gays,

et les hommes gays les utiliseraient pour se dissocier des hommes hétérosexuels. Cela montre que la non-utilisation d'émoticônes n'est pas uniquemen un marqueur de masculinité : elle indexe un certain type de masculinité, c'est-à-dire une masculinité hétéronormative (des hommes cisgenres, et des hommes transgenres hétérosexuels).

Chez les femmes, on ne remarque pas cette dissociation : les femmes gays et les femmes hétérosexuelles ne semblent pas avoir le même besoin de se distancer les unes des autres. Hazenberg (2015) a remarqué le même phénomène dans son étude des adverbes intensifs *so* (marqué « féminin ») et *pretty* (marqué « masculin ») dans un corpus oral d'anglais du Canada. Les hommes gays, les femmes gays et les femmes hétérosexuelles utilisent tous *so* bien plus fréquemment que *pretty*. Les hommes hétérosexuels se distancient de ces groupes, et utilisent 3 fois plus fréquemment *pretty* que *so*.

L'interaction du genre avec l'ethnicité apporte un autre éclairage, et nuance les différences entre les femmes et les hommes. Les femmes blanches et asiatiques n'utilisent pas significativement plus d'émoticônes que, respectivement, les hommes blancs et asiatiques. Chez les femmes, on remarque une démarcation entre les femmes hispaniques et afro-américaines d'un côté, et les femmes asiatiques de l'autre ; ces dernières semblent bouder les émoticônes. Chez les hommes, il y a une différence significative entre les internautes blancs et les internautes afro-américains, qui, comme les femmes asiatiques, utilisent peu ce procédé. Nos résultats apportent ici une précision à ceux de Eisenstein et al. (2011), qui avait trouvé que les internautes blanc·hes utilisaient plus fréquemment les émoticônes que les Afro-Américain·es et les Hispaniques. Tout comme la non-utilisation d'émoticônes participe à la construction d'une identité hétéronormative, on pourrait émettre l'hypothèse que le rejet des émoticônes participe à la construction de l'identité des femmes asiatiques et des hommes afro-américains.

Notre étude a évidemment des limites : parce que nous ne disposons pas d'informations sociodémographiques complètes pour les Redditors du corpus, nous avons dû créer des modèles séparés, et toutes les interactions ne sont donc pas explorées. Toutefois, nos résultats apportent un éclairage nouveau et nuancé sur l'utilisation des émoticônes, ces véritables « stars » des études sociolinguistiques de la CMC.

10.8.2 Comparaison entre émoticônes et émojis

La comparaison entre la fréquence des émojis et des émoticônes montre que ces dernières sont plus fréquentes dans RedditGender. En outre, les différences individuelles semblent être plus prononcées dans l'utilisation des émojis : seules 120 personnes n'ont pas utilisé d'émoticônes, tandis que plus de la moitié du corpus (614 personnes) n'ont pas employé un seul émoji. La présence d'une valeur vraiment extrême pourrait également indiquer que les émojis se prêtent davantage à une utilisation excessive, peut-être parce qu'ils sont, dans certaines conditions, plus faciles à produire que les émoticônes : sur un smartphone, ils sont créés avec une seule touche (en sortant

du clavier proposé par défaut), alors qu'il faut taper au moins deux caractères pour obtenir une émoticône. La question du terminal utilisé pour accéder à Reddit semble particulièrement pertinente quand on veut comparer émojis et émoticônes. Il se peut qu'elle influence les choix des internautes, puisqu'il est plus simple d'écrire un émoji sur un smartphone que sur un ordinateur. Si les Redditors de RedditGender écrivent davantage leurs commentaires depuis un ordinateur que depuis un smartphone, produire des émoticônes est pour eux plus simple. Nous n'avons malheureusement pas d'information sur ce sujet.

Il est également possible que la prédominance des émoticônes sur le site soit liée à la culture du site. Une recherche rapide sur Google montre que les Redditors eux-mêmes s'interrogent sur l'absence relative des émojis, avec des fils de discussion intitulés « Why is there an unspoken agreement to never use emojis on Reddit ? » (Sulfruous, 2017), « Why don't people use emojis on Reddit ? » (dsamanthas, 2018) ou encore « Why doesn't anyone use emojis on reddit ? » (sballens, 2018). Les réponses apportées par la communauté tournent autour de deux explications. La première est liée à la mauvaise qualité de l'application mobile Reddit, et au fait que le site soit principalement consulté par les internautes depuis leur ordinateur (ce qui n'est pas forcément vrai, → p. 96). Les autres pointent vers la différence entre le « dialecte » utilisé sur Reddit et celui des autres plateformes, et la mauvaise réputation des émojis (et des émoticônes) sur le site :

> « I think the reddit language is just more clinical than elsewhere. Emojis would be out of place in this dialect (though I throw a smiley in every once in a while when I'm being passive aggressive because it just works, man) » (dandeeo, 2017)

> « Also, a lot of people think they look dumb or immature. In a lot of circles on the web, emojis or emoticons (like XD) are frowned upon for being less serious. Personally, I use them sometimes, but they have no place in most discussions. They're an IM thing for me. » (ThirdEyeTrippyShit, 2014)

Ou encore, plus succinctement :

> « because reddit tries its best to not be stupid and full of degenerate fucks that only speak in emojis. » (Novaraa, 2017)

Il serait intéressant de voir si la popularité grandissante du site, alliée à l'utilisation massive des smartphones, a conduit ou conduira à une hausse dans la fréquence des émojis au détriment des émoticônes, comme l'ont montré Pavalanathan et Eisenstein (2016) sur Twitter. Notre étude n'est pas diachronique et ne nous permet pas d'examiner cette question.

10.8.3 Émojis

Comme pour les émoticônes, les émojis positifs sont plus fréquents que les émojis négatifs. L'émoji le plus fréquent du corpus, 😂, semble également être un des émojis les plus fréquents sur d'autres plateformes. Par exemple, le 19 juin 2020, c'était l'émoji le plus fréquent sur le site emojitracker, qui présente en temps réel les émojis les plus populaires sur Twitter

(« emojitracker : realtime emoji use on twitter », p. d.). En 2015, il a par ailleurs également été choisi comme « mot » de l'année par Oxford Dictionaries (Steinmetz, 2015).

Le modèle de régression indique l'existence d'une corrélation négative forte entre utilisation d'émojis et âge dans le cas des hommes cisgenres, mais pas chez les autres groupes. Il ne révèle pas d'autre tendance marquée ; dans chaque groupe d'âge, les résultats sont différents. Les hommes cisgenres et transgenres utilisent plus d'émojis que les personnes non binaires et les femmes transgenres dans le groupe le plus jeune, mais cela change dans les autres groupes. Le nombre important de zéros dans les données complique encore la tâche d'interprétation. Il semble que les différences individuelles sont particulièrement marquées, avec quelques utilisateur·trices très prolifiques, et beaucoup d'internautes qui n'utilisent pas du tout ou très peu d'émojis.

Intégrer l'interaction du genre et de l'ethnicité aux analyses des émojis est peut-être une solution pour mieux comprendre leur utilisation sur Reddit. On s'aperçoit que les hommes asiatiques les utilisent moins que les autres hommes. Chez les femmes, on remarque un phénomène assez similaire à celui que l'on avait noté avec les émoticônes : les femmes hispaniques et afro-américaines emploient plus d'émojis que les femmes blanches (pour les émoticônes, elles les utilisent plus que les femmes asiatiques). On pourrait peut-être interpréter ce résultat à la lumière de ce que l'on sait sur la culture geek dont Reddit est un des sites privilégiés sur internet. Les femmes hispaniques et afro-américaines sont sans doute les internautes les plus éloignées de cette culture à la fois racialisée (blanche, et dans une moindre mesure asiatique) et genrée (hétéro-masculine). À la marge de la culture geek, elles apportent sans doute sur Reddit des pratiques linguistiques venues d'autres plateformes, et pourraient avoir moins conscience de (ou choisir d'ignorer) la « mauvaise réputation » des émojis et des émoticônes sur le site, ce qui pourrait être vu comme une forme de transgression.

10.8.4 Étirements graphiques

Types d'étirements graphiques

Les étirements de lettres sont moins fréquents (10 214) dans notre corpus que les étirements de ponctuation (14 298). D'un autre côté, ces derniers ont été utilisés par une proportion légèrement plus faible de Redditors : 86.78 % contre 91.28 % pour les étirements de lettres. Il semble, enfin, que les points d'exclamation soient plus fréquemment étirés que les points d'interrogation et de suspension. Dans notre corpus, le mot qui fait le plus souvent l'objet d'étirements de lettres est *sooo*, comme c'est le cas dans le corpus de tweets américains de Coats (2017b). Comme chez Coats (2017b) et Kalman et Gergle (2014), les étirements de lettres concernent essentiellement des interjections (*oooh*, *ahhh*, *mmm*) et certains mots lexicaux et grammaticaux (*waaay*, *aaaand*, *reeeeally*). En revanche, dans la liste des étirements de lettres les plus fréquents, on ne trouve pas d'étirements d'acronymes de type *lmaoooo*, qui étaient fréquents chez Coats

(2017b). Les étirements de lettres les plus fréquents dans RedditGender sont tous « prononçables » (par opposition à, par exemple, l'étirement de *lmaoooo*), à l'exception de *xxx*, qui pourrait de toute façon être considéré comme une abréviation de « kisses », et qui est par exemple utilisé ainsi « hugs xxx ». Cela suggère, comme l'avaient noté Kalman et Gergle (2014), que les étirements de lettres ont pour fonction d'imiter l'intonation.

Fonctions des étirements graphiques

Les chercheur·es qui se sont intéressé·es aux étirements graphiques de la CMC ont généralement, comme nous, adopté une méthodologie quantitative qui ne permet pas de cerner précisément leurs fonctions. Ils ont généralement établi un lien entre l'imitation de la langue parlée et l'expression de l'émotion (Coats, 2017a ; Kalman & Gergle, 2014 ; Riordan & Kreuz, 2010). Toutefois, il est possible que, comme les émoticônes (→ 2.4.4) ce procédé ait d'autres fonctions que d'indiquer des émotions. Une journaliste américaine fournit quelques pistes : les étirements graphiques pourraient exprimer la gentillesse (*sooooo sorry*), le sarcasme (*riiiight*), l'emphase (*thaaaaanks!* signifiant *thank you very much!*), ou être tout simplement un procédé satisfaisant :

> « Holding down shift+1 and watching your screen fill with !!!!!!!!!!!!! is simply satisfying. It's the repetitive joy of popping bubble wrap, the glee of shouting, and the spatial luxury of lying across two seats on an airplane just because you can. » (O'Connor, 2013)

Certain·es internautes soulignent le caractère ludique, amical ou affectueux du procédé :

> « So that way we don't seem short. yeahhhhhh seems more playful than yeah. » (larbarr, 2014)
>
> « Idk I do just cause "hiii" sounds friendlier to me than "hi" » (Sexyoldmann, 2014)
>
> « Hey for me is just a greeting with no emotion. Hey! is "I'm excited to talk to you, I like you." Heyyyy is "thank god somebody messaged me I'm bored so I'm going to draw out all of my words." » (zombreness, 2015)

Effet de l'âge

L'âge n'est pas forcément corrélé négativement avec la production d'étirements de lettres ; il l'est chez les hommes cisgenres et, partiellement, chez les femmes transgenres, mais pas chez les femmes cisgenres, les hommes transgenres et les personnes non binaires. Pour l'ensemble des étirements de ponctuation, l'interaction de l'âge et du genre n'est pas significative, et la corrélation est partielle : les Redditors de moins de 20 ans produisent davantage d'étirements de ponctuation que les Redditors de 21 à 30 ans, mais il n'y a pas de différence significative entre les Redditors de 21 à 30 ans et les plus Redditors les plus âgé·es, ou entre les plus jeunes et les plus âgé·es. Nos résultats montrent ainsi, une fois de plus, que l'effet de l'âge n'est pas

uniforme, et que les internautes plus âgés n'utilisent pas forcément moins de procédés du Netspeak que les plus jeunes.

Les étirements graphiques, marqueurs de féminité ?

Notre hypothèse, au vu des études réalisées sur les étirements graphiques, était que ces procédés seraient plus fréquents chez les femmes cisgenres que chez les hommes cisgenres. Elle semble confirmée par nos analyses, même si elle est parfois nuancée par l'interaction avec l'âge. Ainsi, pour les étirements de lettres, il n'y a pas de différence entre les femmes et les hommes cisgenres de moins de 20 ans. En revanche, les femmes cisgenres de plus de 21 ans utilisent significativement davantage d'étirements de lettres que les hommes cisgenres du même âge. Les étirements de signes de ponctuation sont également privilégiés par les femmes cisgenres, comme l'ont montré d'autres études (Rubin & Greene, 1992). Elles en sont les utilisatrices les plus prolifiques, non seulement devant les hommes cisgenres, mais devant tous les autres groupes. L'analyse des corrélations, avec le diagramme en mosaïque, suggère que ce sont surtout les étirements de signes d'exclamation qu'elles semblent privilégier, tandis que les hommes cisgenres semblent préférer les multiplications de points de suspension.

Notons également que de nombreux internautes semblent associer les étirements graphiques aux femmes et aux filles :

> « Why [sic] does it mean when a girl message or texts you with extra letters like Hiii, thanksss and bossss ? » ([inconnu], 2020)
>
> « Why do some girls repeat multiple same letters in a word when typing ? » ([inconnu], 2017).
>
> « I find it girly sort of like heyyy. Wish I could speak irl, but you get what I mean. I think it's a subtle like "I know you !" » (JustALivingHuman, 2019)

Leurs commentaires révèlent parfois un agacement :

> « Some people do it to act cute. Others use it as a sign of affection. I thinks its annoying. » (SenpaiThrowMeAway, 2019)
>
> « DAE [1] hate it when people add extra letters to the end of a word to be cute like thisss ? » (ginganinja2507, 2010)
>
> « Reddit really really haaaatttesss thissss... And I kinda like itt. » ([deleted], 2010)

L'agacement et l'association à la féminité qui semblent dominer les attitudes des internautes nous font penser un autre phénomène, issu non pas du domaine de la CMC, mais de la phonologie : l'*uptalk*, ou intonation montante. Tout comme la *creaky voice* (ou friture vocale), ce phénomène est aujourd'hui stigmatisé en anglais américain : il est associé à l'image stéréotypée de la *Valley Girl* à la Kim Kardashian (Tyler, 2015). Évidemment, les étirements graphiques ne sont pas des transcriptions graphiques de l'intonation montante, mais ils ont en commun le fait d'être perçus comme étant typiquement féminins et agaçants. On leur a également attribué sans doute

1. Does anyone else

trop rapidement des fonctions : exprimer l'émotion pour les étirements graphiques, et l'incertitude pour l'intonation montante (Lakoff, 1973). Les très nombreuses études sur l'*uptalk* ont, depuis, mis en évidence ses fonctions multiples (Linneman, 2013), et son interaction avec l'ethnicité : les femmes afro-américaines l'utiliseraient moins fréquemment que les blanches (Pratt-Johnson, 2005, cité par Warren, 2016). Toutefois, la question subsiste : même si toutes les femmes n'utilisent pas les étirements graphiques et même si certains hommes les utilisent abondamment, nos résultats indiquent que ce procédé est privilégié par les femmes, comme c'est le cas pour l'intonation montante (Warren, 2016). Pourquoi ? Il se peut que les femmes innovent davantage que les hommes, quand il en vient aux jeux sur la typographie (comme en matière de phonologie), ou que les hommes boudent ce procédé à cause de ses connotations « girly ».

Les analyses des effets de l'ethnicité et des groupes transgenres apportent toutefois de précieuses nuances, qui indiquent peut-être que les étirements de ponctuation et de lettres ne sont pas forcément utilisés pour indexer la féminité. À notre connaissance, notre analyse est la seule à avoir intégré la variable « ethnicité » à l'étude des étirements graphiques. Nos résultats montrent la pertinence de ce choix, parce qu'il tempère les différences entre femmes et hommes constatées par d'autres chercheur·es. Chez les Hispaniques, il n'y a ainsi pas d'effet de genre pour les étirements de ponctuation. De plus, quand l'ethnicité est intégrée au modèle, les différences entre femmes et hommes, pour les étirements de lettres (et non de ponctuation), sont limitées à un groupe d'âge (21-30 ans). L'autre résultat marquant de cette analyse est le fait que les internautes asiatiques, hommes comme femmes, semblent éviter les étirements de lettres et de ponctuation. Non seulement la différence est significative, mais la taille d'effet est importante : les femmes asiatiques produisent par exemple deux fois moins d'étirements de ponctuation que les femmes afro-américaines et blanches, et trois fois moins que les femmes hispaniques.

Les comportements des femmes transgenres font naitre d'autres questions. Nous avons vu qu'elles utilisent fréquemment les émoticônes, autre marqueur féminin de la CMC ; nous pensions donc qu'elles s'aligneraient également avec les femmes cisgenres dans leur utilisation des étirements graphiques. Or, ce n'est pas ce que nous avons constaté. Pour les étirements de lettres, l'alignement est très partiel : les femmes transgenres de 21 à 30 ans utilisent autant d'étirements de lettres que les femmes cisgenres. Dans les deux autres groupes, les femmes cisgenres utilisent davantage ce procédé que les femmes transgenres. Pour les étirements de ponctuation, la différence est encore plus nette : non seulement les femmes transgenres utilisent moins ce procédé que les femmes cisgenres, mais elles les utilisent également moins que les hommes cisgenres et transgenres. On ne constate par ailleurs pas de prédilection des personnes non binaires AFAN pour les étirements graphiques (à l'exception des AFAN de 21 à 30 ans, et uniquement pour les étirements de lettres).

Ces résultats peuvent être interprétés de deux manières différentes. D'un côté, il est possible que les étirements graphiques n'indexent pas uni-

quement la féminité, ou qu'ils indexent un certain type de féminité. De l'autre, s'ils indexent la féminité, il est possible que les femmes transgenres choisissent de s'en écarter, construisant leur identité de genre en ligne par d'autres ressources que les femmes cisgenres.

10.8.5 Mots en majuscules

Notre étude des mots en majuscules montre que les mots grammaticaux sont les plus susceptibles de faire l'objet de ce procédé. Ce n'est pas étonnant, car les mots grammaticaux sont les plus nombreux dans le corpus (et dans tous les corpus en général). On retrouve également parmi les mots en majuscules les plus fréquents plusieurs adverbes ou adjectifs d'intensité, ce qui confirme leur statut de marqueur d'emphase et/ou d'émotion de ce procédé.

L'âge semble être lié à l'utilisation de mots en majuscules, mais pas pour tous les groupes. La corrélation négative est claire chez les hommes cisgenres, seulement partielle chez les femmes cisgenres, et absente chez les groupes transgenres et non-binaire. Contrairement à ce que Rosen et al. (2010) et Parkins (2012) suggèrent, l'utilisation de mots en majuscules n'est pas plus fréquente chez les femmes que chez les hommes dans notre corpus. C'est, à la place, entre les groupes cisgenres d'un côté et les groupes transgenres et non binaires de l'autre qu'il y a une différence significative. Elle est présente chez les internautes de 14 à 30 ans, mais s'estompe chez les internautes de 31 ans et plus. Cette démarcation entre personnes cisgenres et transgenres est liée à l'absence d'effet de l'âge chez ces dernières, pour la production de mots en majuscules.

10.8.6 Interjections

Oh est, de loin, l'interjection la plus fréquente dans RedditGender ; c'est également le cas dans le Bergen Corpus of London Teenage Language (COLT) (Aijmer, 2009), ce qui suggère un possible alignement de la langue de la CMC sur la langue orale dans le cas des interjections. La corrélation entre âge et fréquence des interjections n'est, comme c'était le cas pour les émojis et les étirements de lettres, pas présente dans tous les groupes. Chez les hommes cisgenres, la fréquence des interjections décroit avec l'âge. Cette corrélation négative n'est significative que partiellement chez les femmes cisgenres et transgenres, et elle est absente chez les personnes transgenres. Une fois de plus, l'intégration de l'interaction entre genre et âge se révèle pertinente, et fournit une image nuancée des usages des internautes. Verheijen (2017) a montré que les adolescents utilisent plus d'interjections que les jeunes adultes sur plusieurs plateformes de CMC ; dans notre cas, c'est uniquement vrai pour les hommes cisgenres.

Nos résultats semblent aller, globalement, dans le sens de ceux de Coats (2017a), qui a mis en évidence le fait que les femmes utilisent davantage d'interjections que les hommes sur Twitter, mais avec quelques nuances importantes. Dans RedditGender, il y a uniquement une différence significative entre femmes cisgenres et hommes cisgenres chez les plus de 21

ans. La démarcation entre les hommes cisgenres et les autres internautes se creuse chez les plus de 31 ans. L'analyse de l'échantillon cisgenre révèle également une interaction avec l'ethnicité : les différences entre femmes et hommes sont uniquement présentes dans les groupes blanc et asiatique. Chez les non-binaires, le genre assigné à la naissance n'a pas d'impact sur la production d'interjections.

Enfin, comparées à d'autres phénomènes du Netspeak, comme les émoticônes, les interjections semblent entrer assez peu en jeu dans la construction des identités ethniques en ligne. Il y a un seul contraste significatif : les hommes asiatiques utilisent moins fréquemment les interjections que les hommes hispaniques. Sur le plan ethnique, les interjections semblent être donc relativement « neutres ». Il est possible que cela soit lié au fait qu'elles ne sont pas réellement des innovations langagières, au même titre que les émoticônes (ou que les acronymes, comme nous le verrons plus bas, → p. 300), qui sont fortement associées à certains groupes ethniques.

tl ;dr

Ce chapitre a montré que les procédés d'ajout sont tous, à l'exception des mots en majuscules, plus fréquemment utilisés par les femmes cisgenres que par les hommes cisgenres.
Les hommes transgenres et les personnes non binaires ont souvent un comportement similaire et peu marqué ; ils s'alignent dans de rares cas sur les hommes cisgenres. Dans deux cas au moins, les femmes transgenres se démarquent à la fois des hommes et des femmes cisgenres, avec une utilisation très importante des émoticônes, et très faible des étirements de ponctuation.
Les analyses qui prennent en compte l'ethnicité nuancent les résultats constatés sur l'ensemble du corpus. Les différences entre femmes et hommes ne sont pas présentes dans tous les groupes ethniques.
Dans trois cas sur six (émoticônes, émojis et étirements de ponctuation), on remarque une distanciation entre femmes afro-américaines et hispaniques d'un côté, et femmes asiatiques de l'autre, qui semblent bouder ces procédés. Chez les hommes, la démarcation est moins claire.

Chapitre 11

Procédés de réduction

Ce chapitre est consacré aux procédés de réduction qui, à première vue, demandent moins « d'effort » aux internautes que les procédés d'ajout. Il s'agit des abréviations (acronymes et réductions), des graphies phonétiques, des g-droppings (omission du *g* dans les mots normalement terminés par la graphie *-ing*), des omissions d'apostrophe et des omissions de la majuscule du pronom personnel *I*. L'organisation de ce chapitre est la même que celui du chapitre précédent. Chaque section est consacrée à un procédé, et contient une présentation de son utilisation dans le corpus (nombre de tokens, de types, etc.), suivie de plusieurs analyses examinant les effets du genre, de l'âge et de l'ethnicité sur sa fréquence.

11.1 Hypothèses et questions de recherche

Nous avons abordé l'analyse de ces procédés avec plusieurs hypothèses. Tout d'abord, il est possible que les internautes Afro-Américain·es utilisent davantage d'abréviations que les autres internautes, puisque plusieurs études semblent indiquer qu'ils sont pionniers en la matière (→ p. 74). Nous pensons que l'âge Reddit, c'est-à-dire la durée depuis laquelle les internautes ont un compte sur le site, peut avoir un impact sur l'utilisation des abréviations, qui font partie du code langagier de Reddit (→ p. 95). Il est également probable que les femmes produisent davantage d'abréviations que les hommes, comme l'ont montré Baron (2004), Cougnon et François (2010) et Herring et Zelenkauskaite (2009). Il est possible que le genre ait un effet sur la fréquence de plusieurs autres de ces variables linguistiques, à commencer par les omissions d'apostrophe. Squires (2012) a par exemple montré que les hommes omettent plus fréquemment l'apostrophe que les femmes dans la messagerie instantanée. Comme pour les variables présentées dans le chapitre précédent, il est probable que les jeunes Redditors se démarquent des plus âgé·es par une utilisation plus forte de ces variables. Enfin, il est possible que le genre et l'ethnicité aient tous les deux un effet sur la production de g-droppings, un phénomène qui, à l'oral, est associé (entre autres) aux hommes et aux Afro-Américain·es (→ p. 82).

11.2 Abréviations

11.2.1 Comparaison entre les acronymes et leurs formes standard

Le tableau 11.1 contient les 33 types d'acronymes étudiés ici, qui représentent 31 529 tokens. La variante graphique indiquée (tout en majuscules, tout en minuscules ou avec une majuscule à l'initiale) est la plus fréquente dans RedditGender. Le tableau contient également leur fréquence brute, ainsi que leur forme longue, la fréquence brute de leurs formes longues et l'*odds ratio* indiquant le rapport entre la fréquence des formes abrégées et des formes longues. Il a été calculé en divisant la fréquence des acronymes par la fréquence des formes longues ; quand il est supérieur à 1, cela signifie que les acronymes sont plus fréquents dans RedditGender que les formes longues correspondantes.

Tableau 11.1 – Acronymes dans RedditGender

Rang	Acronymes	Fréq.	Forme longue	Fréq.	Odds ratio
1	lol	9215	laughing out loud	7	1316.43
2	OP	4366	opening post/poster	3	1455.33
3	SO	2010	significant other	85	23.65
4	Idk	1633	I don't know	4872	0.34
5	IMO	1481	in my opinion	573	2.58
6	Omg	1404	oh my god	578	2.43
7	tbh	1388	to be honest	719	1.93
8	PM	1097	private message	18	60.94
9	lmao	1054	laughing my ass off	9	117.11
10	WTF	997	what the fuck	485	2.06
11	btw	720	by the way	339	2.12
12	af	644	as fuck	673	0.96
13	TLDR	592	too long didn't read	1	592.00
14	irl	487	in real life	415	1.17
15	TIL	369	today I learned	17	21.71
16	POC	358	people of color	142	2.52
17	ASAP	352	as soon as possible	75	4.69
18	IIRC	341	if I remember correctly	93	3.67
19	NSFW	316	not safe for work	2	158
20	FFS	314	for fuck's sake	36	8.72
21	FYI	284	for your information	5	56.80
22	YMMV	268	your mileage may vary	25	10.72
23	FTFY	224	fixed that for you	6	37.33
24	smh	221	shaking my head	7	31.57
25	atm	188	at the moment	592	0.32
26	FWIW	187	for what it's worth	181	1.03
27	MIL	183	mother in law	36	5.08
28	IMHO	161	in my humble opinion	4	40.25
29	SJW	154	social justice warrior	16	9.62
30	rn	137	right now	4156	0.03
31	FWB	136	friends with benefits	24	5.67
32	jk	134	just kidding	67	2.00
33	GTFO	114	get the fuck out	50	2.28

lol est l'acronyme le plus fréquent du corpus, avec 9 185 occurrences. Il est suivi par *OP*, ou *opening post*/*poster*, qui désigne la personne ou le message qui initie un fil de discussion. On trouve ensuite *SO*, ou *significant other*, qui signifie « partenaire » ou « époux·se ». La comparaison des formes courtes avec les formes longues montre que ces dernières sont, dans la majorité des cas, moins fréquentes que les acronymes correspondants. Seules 4 formes longues sont plus fréquentes que leurs acronymes : *I don't know* (*Idk)*, *as fuck* (*af*), *at the moment* (*atm*) et *right now* (*rn*).

Dans le cas des acronymes qui semblent les plus caractéristiques de la CMC, les formes longues sont, sans surprise sans doute, très peu utilisées : 7 fois pour *laughing out loud* (*lol*), 1 fois pour *too long didn't read* (*TLDR*), 9 fois pour *laughing my ass off* (*lmao*) ou 6 fois pour *fixed that for you* (*ftfy*). Plusieurs expressions de l'anglais courant, comme *in my opinion*, *by the way*, *to be honest* ou encore *what the fuck*, sont également plus présentes sous leur forme abrégée que sous leur forme complète. On note le même phénomène en ce qui concerne les acronymes désignant des personnes. *POC* est 2.53 fois plus fréquent que *people*/*person of color*, *MIL* est 5.08 fois plus fréquent que *mother in law*, *SJW* est 9,62 fois plus fréquent que *social justice warrior*, et *FWB* est 5.67 fois plus fréquent que *friends with benefits*.

11.2.2 Réductions (*clippings*)

Types de réductions

5828 occurrences de réductions, dont 19 types différents, ont été identifiées dans le corpus, en suivant la procédure décrite dans la section 5.4.3. Elles sont présentées dans le tableau 11.2, avec les formes standard correspondantes, leurs fréquences respectives et la proportion de formes réduites par rapport aux graphies standard (*odds ratio)*. On voit que certaines abréviations renvoient à la même forme complète : c'est le cas de *bc* et *cuz* (*because*), et de *pls* et *plz* (*please*). Les réductions de *girlfriend* et *boyfriend* et les deux formes de réduction de *because* sont les plus fréquentes dans le corpus. On trouve également dans la liste les noms des réseaux sociaux *Facebook* (*fb*) et *Instagram* (*ig*), ainsi que des adverbes comme *definitely* (*def*), *seriously* (*srs*) et *especially* (*esp*) ainsi que divers noms et adjectifs (*bs* $\rightarrow$ *bullshit*, *dev* $\rightarrow$ *developer*/*pment*, *fave* $\rightarrow$ *favorite*, *masc* $\rightarrow$ *masculine* ou encore *mic* $\rightarrow$ *microphone*).

La colonne « ratio » indique la proportion des formes réduites par rapport à leur forme standard. Deux des réductions étudiées ici sont utilisées plus fréquemment que leur forme complète : il s'agit de *clit*, qui est utilisé 2 fois plus fréquemment que *clitoris*, et de *mic*, qui est utilisé près de 5 fois plus fréquemment que microphone. La réduction *gf* est quant à elle utilisée 2 fois moins fréquemment que la forme standard *girlfriend*. D'autres réductions font une concurrence plus modérée aux graphies standard : c'est notamment le cas de *bf*, *fb*, et *bs*, qui sont utilisées environ 4 fois moins fréquemment que les formes complètes. Les mots plus fréquents dans le corpus, comme *because* (49 942 occurrences), *definitely* (8691 occurrences) ou *especially* (5876 occurrences), font moins souvent l'objet de réductions.

Tableau 11.2 – Réductions dans RedditGender

Rang	Réduction	Fréq.	Forme standard	Fréq.	*Odds ratio*
1	gf	956	girlfriend	1836	0.52
2	bc	737	because	45942	0.02
3	bf	690	boyfriend	2346	0.29
4	cuz	547	because	45942	0.01
5	bs	464	bullshit	1626	0.29
6	def	373	definitely	8691	0.04
7	fb	369	facebook	1406	0.26
8	srs	226	seriously	3350	0.07
9	k	192	okay/ok	8717	0.02
10	pls	191	please	4182	0.05
11	esp	139	especially	5876	0.02
12	dev	137	developer/ment	973	0.14
13	clit	123	clitoris	64	1.92
14	fave	123	favorite	3316	0.04
15	mic	120	microphone	31	3.87
16	masc	116	masculine	1320	0.09
17	plz	113	please	4182	0.03
18	fem	106	feminine	1736	0.06
19	ig	106	instagram	494	0.21
1 à 19	Tous	5828	Total	142 030	0.04

11.2.3 Effets de l'âge, du genre et de l'âge Reddit sur la fréquence des abréviations

Pour réaliser cette analyse, nous avons regroupé réductions et acronymes dans la catégorie « abréviations », considérant que ces deux variables sont proches et participent d'une même tendance à la réduction.

Statistiques descriptives

Seul·es 5 internautes n'ont utilisé aucun des acronymes de notre liste tandis que 161, soit 15.42 %, n'ont utilisé aucune des réductions étudiées ici. Le tableau 11.3 présente la fréquence moyenne et médiane des abréviations par sous-corpus et dans l'ensemble du corpus, ainsi que les mesures de dispersion. Nous avons inclus, en plus du genre et de l'âge, l'âge Reddit, qui est ici une variable d'intérêt.

Ce tableau indique une possible corrélation négative de l'âge avec la production d'abréviations, une médiane de 1.97 pour les plus jeunes et de 1.08 pour les plus âgés. L'âge Reddit semble lui aussi être corrélé négativement avec la fréquence des abréviations. Dans les groupes de genre, les personnes non binaires ont produit le moins d'abréviations (médiane de 1.17), et les femmes cisgenres et transgenres en ont produit le plus (respectivement 1.68 et 1.75 par 1000 tokens).

Tableau 11.3 – Fréquence des abréviations dans le corpus, par 1000 tokens

	Moyenne	ET	Médiane	EI
Femmes cisgenres	2.07	1.52	1.68	1.72
Hommes cisgenres	1.93	1.83	1.38	1.75
Hommes transgenres	1.99	1.34	1.58	1.59
Non-binaires	1.48	1.14	1.17	1.13
Femmes transgenres	2.06	1.52	1.75	1.56
14-20 ans	2.59	2.14	1.97	2.03
21-30 ans	2.16	1.57	1.74	1.89
31 ans et +	1.42	1.18	1.08	1.06
Âge Reddit cat. 1	2.00	1.52	1.70	1.68
Âge Reddit cat. 2	2.33	2.02	1.63	2.05
Âge Reddit cat. 3	2.01	1.54	1.60	1.60
Âge Reddit cat. 4	1.74	1.35	1.35	1.50
Âge Reddit cat. 5	1.38	1.25	1.06	1.15
Tous	1.95	1.59	1.52	1.67

Effets de l'âge, du genre et de l'âge Reddit sur la fréquence des abréviations : modèle de régression

Les données étant, comme pour les autres variables linguistiques, fortement asymétriques, avec une dispersion élevée, nous avons opté pour un modèle binomial négatif. La variable dépendante est la fréquence brute des abréviations (réductions et acronymes) pour chaque personne. Les variables indépendantes sont l'âge, le genre, l'interaction de l'âge et du genre, et l'âge Reddit. Nous avons utilisé la fonction `step()` pour vérifier si toutes les variables contribuent au modèle. Le résultat indique qu'il n'est pas pertinent d'intégrer l'âge Reddit au modèle. La variable ne semble donc pas être corrélée avec la fréquence des acronymes. Les coefficients et intervalles de confiance de 95 % du modèle sont présentés dans le tableau 11.4, qui a pour niveau de référence les femmes cisgenres de 14 à 20 ans.

Le modèle révèle une corrélation négative entre l'âge et la fréquence des abréviations pour les hommes cisgenres, que l'on peut observer dans le graphique 11.1. Cette corrélation est partielle chez les femmes cisgenres et les transgenres ; les femmes de plus de 31 ans emploient moins d'abréviations que les deux groupes les plus jeunes, mais il n'y a pas de différence entre les deux groupes plus jeunes. Il n'y a pas de corrélation chez les personnes non binaires et les hommes transgenres.

Chez les Redditors les plus jeunes, les hommes cisgenres utilisent davantage d'abréviations que tous les autres groupes à l'exception des femmes transgenres. La différence la plus forte est celle entre les hommes cisgenres et les personnes non binaires, qui utilisent environ 2 fois moins d'abréviations. Dans le groupe de 21 à 30 ans, le groupe non-binaire emploie significativement moins d'abréviations que tous les autres groupes. Il n'y a pas d'autre différence significative. Dans le groupe le plus âgé, les femmes cisgenres et les femmes transgenres emploient davantage d'abréviations que

TABLEAU 11.4 – Abréviations, effets de l'interaction de l'âge et du genre

	Variable dépendante :
	Abréviations
Intercept	0.002** (0.002, 0.003)
Hommes cisgenres	1.379* (1.016, 1.856)
Femmes transgenres	1.188 (0.764, 1.896)
Hommes transgenres	0.862 (0.599, 1.245)
Non-binaires	0.648 (0.423, 1.014)
21-30 ans	1.034 (0.789, 1.335)
30 ans et +	0.658** (0.498, 0.857)
21-30 ans :Hommes cisgenres	0.658* (0.471, 0.925)
30 ans et + :Hommes cisgenres	0.609** (0.435, 0.859)
21-30 ans :Femmes transgenres	0.733 (0.439, 1.197)
30 ans et + :Femmes transgenres	0.949 (0.555, 1.597)
21-30 ans :Hommes transgenres	1.006 (0.661, 1.528)
30 ans et + :Hommes transgenres	1.247 (0.742, 2.128)
21-30 ans :Non-binaires	0.983 (0.603, 1.578)
30 ans et + :Non-binaires	1.324 (0.767, 2.273)
Observations	1,044
Log Likelihood	−4,640.832
θ	2.120** (0.095)
Akaike Inf. Crit.	9,311.664
Note :	*p<0.05; **p<0.01

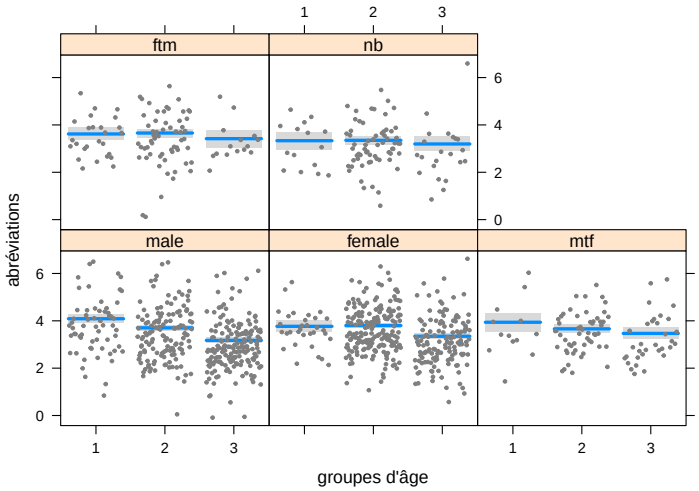

FIGURE 11.1 – Interaction entre genre et âge dans la production des abréviations

les hommes cisgenres. Quand un homme cisgenre produit une abréviation, une femme transgenre en produit 1.34 et une femme cisgenre 1.19.

11.2.4 Effets de l'ethnicité, du genre et de l'âge sur la fréquence des abréviations

Statistiques descriptives

Les boites à moustaches présentées dans la figure 11.2 montrent que le nombre médian d'abréviations est le plus faible chez les hommes blancs, et qu'il est le plus élevé chez les femmes hispaniques. La variation individuelle semble être plus importante dans les groupes hispanique et afro-américain que dans les deux autres groupes, avec un écart interquartile (représenté par la hauteur de la boite) plus élevé.

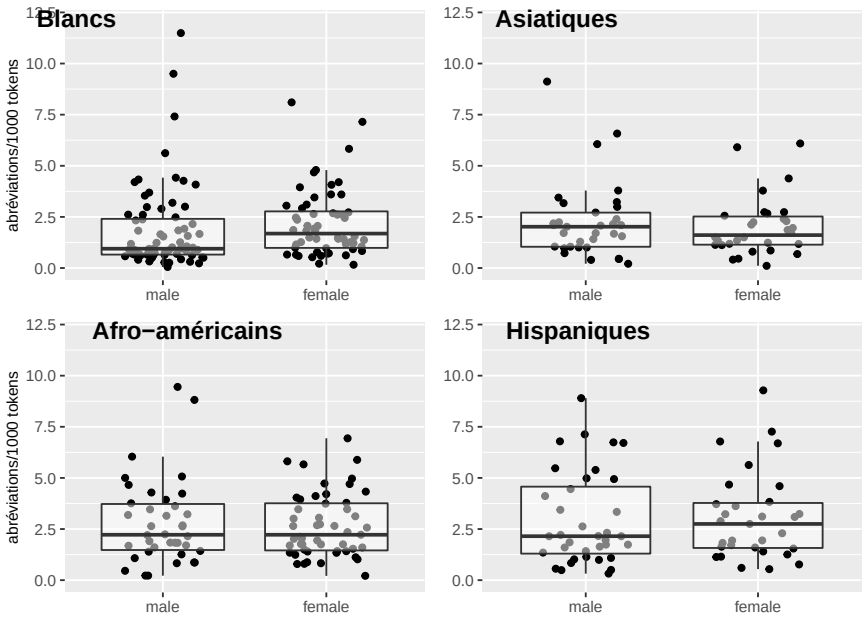

FIGURE 11.2 – Fréquence des abréviations par 1000 tokens : interaction du genre et de l'ethnicité

Effets de l'ethnicité, du genre et de l'âge sur la production d'abréviations : modèle de régression

Nous avons créé un modèle de régression binomial négatif avec la combinaison habituelle de variables. L'interaction du genre et de l'âge ainsi que l'effet principal du genre ont été retirés par le processus de sélection des variables. Nous avons toutefois conservé l'effet du genre. Le modèle, présenté dans le tableau 11.5, a pour niveau de référence les femmes cisgenres.

Le modèle ne renvoie pas de différence significative entre femmes et hommes. Comme l'interaction du genre et de l'ethnicité n'a pas été incluse au modèle, il n'y a de différence significative dans aucun groupe. Le modèle

Tableau 11.5 – Abréviations, effets de l'âge, du genre et de l'ethnicité

	Variable dépendante :
	Abréviations
Intercept	0.003* (0.002, 0.004)
Hommes cisgenres	0.954 (0.819, 1.110)
Afro-américaines	1.343* (1.106, 1.634)
Asiatiques	1.006 (0.808, 1.257)
Hispaniques	1.320 (1.064, 1.642)
21-30 ans	0.753 (0.590, 0.951)
30 ans et +	0.486* (0.378, 0.621)
Observations	347
Log Likelihood	−1,607.066
θ	2.115** (0.161)
Akaike Inf. Crit.	3,228.133
Note :	*p<0.05 ; **p<0.01

indique également une corrélation négative partielle entre fréquence des abréviations et âge ; les Redditors de plus de 31 ans produisent d'abréviations que les 14-20 ans. Les Redditors afro-américain·es et hispaniques emploient significativement plus d'abréviations que les internautes blanc·hes et asiatiques. Quand un·e Redditor blanc·he emploie 1 abréviation, un·e Afro-Américain·e en emploie 1.34 et un·e Hispanique 1.32. Quand une personne du groupe asiatique en utilise 1, une personne afro-américaine en produit 1.34 et une personne hispanique 1.31. Il n'y a pas de différence significative entre le groupe asiatique et le groupe blanc, ni entre le groupe hispanique et le groupe afro-américain.

11.3 Graphies phonétiques

11.3.1 Types de graphies phonétiques

Nous avons recensé 22 types de graphies phonétiques dans le corpus, pour un total de 16 052 tokens. Comme pour les abréviations, ce relevé n'est pas exhaustif ; notre méthode de récolte des données a seulement pris en compte les graphies les plus fréquentes, et il est possible que nous ayons fait des oublis, la frontière entre graphies phonétiques, abréviations et autres graphies non standard étant parfois floue. Les graphies phonétiques relevées dans le corpus sont présentées par ordre décroissant de fréquence dans le tableau 11.6. Le tableau indique également la fréquence de leurs équivalents standard, ainsi que les *odds ratios* (rapport des cotes, *OR* dans le tableau) des graphies phonétiques par rapport à leur forme standard. Les 4 *odds ratios* les plus élevés ont été mis en gras.

On remarque que les trois graphies phonétiques les plus fréquentes sont de graphies synthétiques, qui représentent en seul mot deux ou trois mots (*kinda* pour *kind of*, *dunno* pour *don't know*, *gotta* pour *got to*). Leur ratio est élevé, ce qui signifie que ces formes non standard font concurrence

aux formes standard. C'est tout particulièrement le cas de *gotta*, qui est plus fréquent dans le corpus que son équivalent standard *got to*. Son ratio est de 1.24, ce qui signifie que, pour 1 graphie *got to*, il y a 1.24 *gotta* dans le corpus. La graphie *hella* (un terme d'argot typique de la région de San Francisco, qui tire son origine de l'anglais vernaculaire afro-américain, Bucholtz, 2012), moins présente dans le corpus (297 occurrences), est quasiment aussi fréquente que son équivalent standard *hell of*.

Dans le corpus, les graphies phonétiques sont essentiellement des phénomènes de réduction. Les formes qui génèrent le plus de graphies phonétiques sont composées de plusieurs mots, et la seule graphie phonétique qui est plus longue que son équivalent standard est *nah*. On remarque aussi que les formes standard les plus fréquentes, comme *you*, *the*, *that* ou *this*, entrainent moins de graphies phonétiques que des graphies standard moins fréquentes comme *kind of* ou *don't know*.

Tableau 11.6 – Graphies phonétiques du corpus

Rang	Graphie	Fréq.	Gr. standard	Fréq.	*OR*
1	gonna	4097	going to	14 525	**0.28**
2	kinda	3539	kind of	8394	**0.42**
3	gotta	1566	got to	1260	**1.24**
4	tho	1205	though	17 652	0.07
5	ya	1108	you	282 867	0.00
6	nah	1060	no	40 743	0.03
7	u	846	you	282 867	0.00
8	em	637	them	44627	0.01
9	dunno	633	don't know	9045	0.07
10	hella	297	hell of	262	**0.88**
11	thru	181	through	11 350	0.02
12	lil	170	little	12 461	0.01
13	da	162	the	526 790	0.00
14	dat	146	that	253 243	0.00
15	wut	95	what	61 584	0.00
16	wat	91	what	61 584	0.00
17	doe	60	though	17 652	0.00
18	dis	55	this	101 804	0.00
19	shoulda	46	should have	1718	0.03
20	gon	20	going to	14 525	0.00
21	woulda	20	would have	4280	0.01
22	coulda	18	could have	1812	0.01
Total		16 052		800 450	0.02

11.3.2 Effets du genre et de l'âge sur la production de graphies phonétiques

Statistiques descriptives

Seul·es 38 Redditors, soit 3.70 %, n'ont pas utilisé les graphies phonétiques étudiées ici. Un quart des Redditors a employé plus de 1.12 graphies

phonétiques par 1000 tokens. La fréquence médiane des graphies phonétiques (tableau 11.7) est la plus élevée chez les femmes transgenres. Les femmes cisgenres ont utilisé le moins de graphies phonétiques. L'utilisation de graphies phonétiques semble par ailleurs être corrélée négativement avec l'âge, les Redditors les plus jeunes en employant davantage que les Redditors les plus âgés. La dispersion est importante dans tous les groupes.

Tableau 11.7 – Fréquence des graphies phonétiques par 1000 tokens

	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.96	1.07	0.63	0.97
Femmes cisgenres	0.70	0.77	0.50	0.66
Femmes transgenres	0.93	0.77	0.74	1.00
Hommes transgenres	0.87	1.05	0.52	0.94
Non-binaires	0.83	0.85	0.61	1.00
14 à 20 ans	1.26	1.31	0.88	1.22
21 à 30 ans	0.90	0.88	0.67	0.87
31 ans et +	0.60	0.71	0.41	0.64
Tous	0.84	0.93	0.58	0.96

Effet de l'âge et du genre sur la fréquence des graphies phonétiques : modèle de régression

Les données étant dispersées, nous avons opté pour un modèle de régression binomial négatif. Il a été créé avec comme variables indépendantes le genre, la catégorie d'âge, et leur interaction. Un offset a été ajouté au modèle pour prendre en compte le fait que chaque corpus a une taille différente. Les femmes cisgenres de 14 à 20 ans sont le niveau de référence du modèle. La fonction `step()` indique que toutes les variables contribuent au modèle. Les coefficients et les intervalles de confiance de 95 % du modèle sont présentés dans le tableau 11.8.

Chez les Redditors les plus jeunes, les hommes cisgenres produisent significativement plus de graphies phonétiques que les autres groupes. L'effet est particulièrement fort quand on les compare aux personnes non binaires et aux hommes transgenres : les hommes cisgenres emploient respectivement 2.59 et 2.24 fois plus de graphies phonétiques que les non-binaires et les hommes transgenres. Dans les catégories d'âge 2 (21 à 30 ans) et 3 (31 ans et plus), le genre a un impact moins prononcé sur la fréquence des graphies non standard. Pour la catégorie 2, seules les femmes cisgenres produisent significativement moins de graphies phonétiques que les hommes cisgenres. Chez les Redditors de 31 ans et plus, seuls les hommes transgenres produisent significativement moins de graphies phonétiques que les hommes cisgenres. L'interaction entre âge et genre est présentée dans la figure 11.3. On y remarque que la production de graphies phonétiques est corrélée négativement avec l'âge pour les groupes cisgenres. Cet effet est significatif pour les hommes cisgenres, mais plus limité chez les femmes cisgenres : les femmes de 31 ans et plus produisent moins de

Tableau 11.8 – Graphies phonétiques, effet de l'interaction de l'âge et du genre

	Variable dépendante :
	Graphies phonétiques
Intercept	0.001** (0.001, 0.001)
Hommes cisgenres	1.811** (1.213, 2.665)
Femmes transgenres	1.043 (0.587, 1.936)
Hommes transgenres	0.807 (0.501, 1.310)
Non-binaires	0.698 (0.400, 1.260)
21-30 ans	0.791 (0.554, 1.103)
30 ans et +	0.518** (0.359, 0.732)
Hommes cisgenres :21-30 ans	0.731 (0.473, 1.142)
Femmes transgenres :21-30 ans	1.287 (0.656, 2.442)
Hommes transgenres :21-30 ans	1.345 (0.775, 2.328)
Non-binaires :21-30 ans	1.741 (0.915, 3.229)
Hommes cisgenres :30 ans et +	0.659 (0.424, 1.036)
Femmes transgenres :30 ans et +	1.325 (0.653, 2.617)
Hommes transgenres :30 ans et +	2.724** (1.388, 5.486)
Non-binaires :30 ans et +	1.689 (0.824, 3.430)
Observations	1,044
Log Likelihood	−3,878.750
θ	1.263** (0.058)
Akaike Inf. Crit.	7,787.501
Note :	*p<0.05 ; **p<0.01

graphies phonétiques que les autres, mais il n'y a pas de différence significative entre les deux groupes plus jeunes. En revanche, dans les groupes non binaire et transgenres, l'effet de l'âge n'est pas significatif : les Redditors plus jeunes n'emploient donc pas davantage de graphies phonétiques que les plus âgés.

11.3.3 Effets de l'âge, du genre et de l'ethnicité sur la production de graphies phonétiques

Statistiques descriptives

Les boites à moustaches de la figure 11.4 présentent les fréquences des graphies phonétiques par 1000 tokens pour les femmes et les hommes de chaque groupe ethnique. La médiane est légèrement plus élevée chez les hommes de chaque groupe que chez les femmes ; il y a également plus de variation dans les groupes d'hommes que de femmes, avec des boites plus hautes. Les hommes afro-américains ont la médiane la plus élevée (environ 1.25).

Effets de l'ethnicité, du genre et de l'âge sur la production de graphies phonétiques : modèle de régression

Le premier modèle réalisé incluait l'âge, le genre et l'ethnicité ainsi que toutes les combinaisons d'interactions à deux niveaux possibles. La fonction

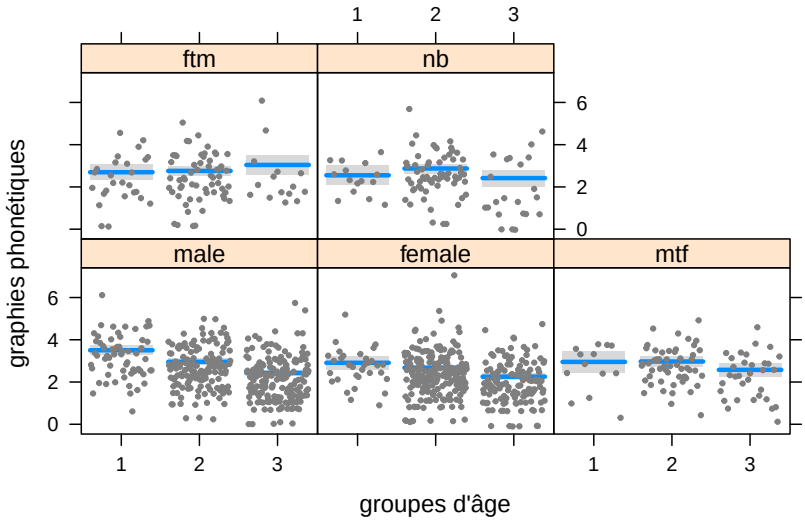

FIGURE 11.3 – Interaction de l'âge et du genre dans la production de graphies phonétiques

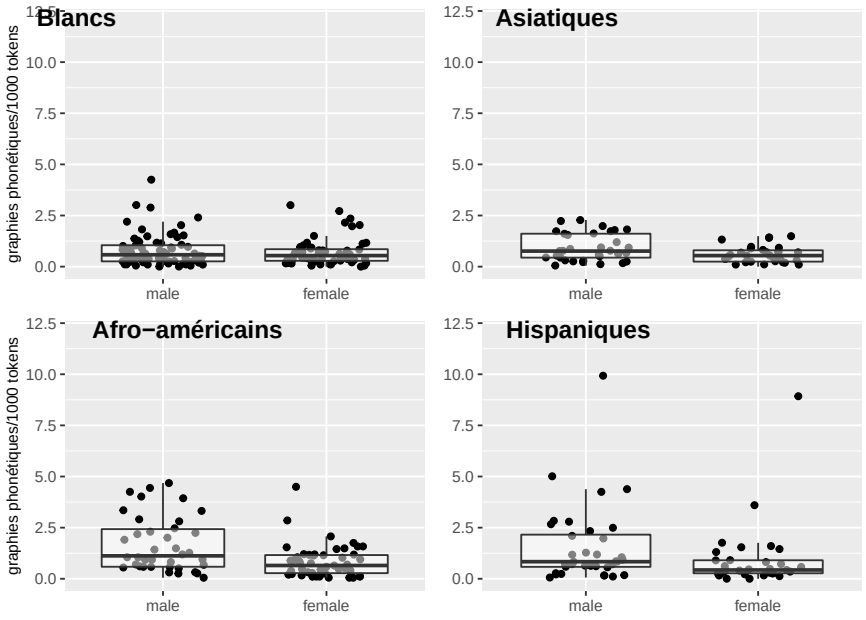

FIGURE 11.4 – Fréquence des graphies phonétiques par 1000 tokens : interaction entre le genre et l'ethnicité

`step()` a été utilisée pour sélectionner les variables indépendantes ; ont été retenus l'âge et l'interaction du genre et de l'ethnicité (tableau 11.9). Le niveau de référence est les femmes blanches. Pour l'âge, le niveau de référence est les 21-30 ans. Le modèle indique qu'il n'y a pas de différence significative dans la fréquence des graphies phonétiques entre hommes blancs et femmes blanches. Dans tous les autres groupes, les hommes produisent davantage de graphies phonétiques que les femmes. L'effet de l'interaction entre ethnicité et genre est le plus marqué dans le groupe afro-américain, où les hommes produisent deux fois plus de graphies phonétiques que les femmes. Ces interactions sont présentées dans la figure 11.5.

Tableau 11.9 – Graphies phonétiques, effets de l'âge, du genre et de l'ethnicité

	Variable dépendante :
	Graphies phonétiques
Intercept	0.001^{**} (0.001, 0.001)
Hommes cisgenres	1.093 (0.801, 1.488)
Afro-américaines	1.133 (0.808, 1.591)
Asiatiques	0.785 (0.530, 1.175)
Hispaniques	1.154 (0.782, 1.724)
14-20 ans	1.513^{**} (1.133, 2.047)
30 ans et +	0.709^{**} (0.572, 0.881)
Hommes cisgenres :Afro-Américains	1.882^{**} (1.170, 3.040)
Hommes cisgenres :Asiatiques	1.486 (0.869, 2.540)
Hommes cisgenres :Hispaniques	1.417 (0.842, 2.383)
Observations	347
Log Likelihood	−1,318.889
θ	1.426^{**} (0.113)
Akaike Inf. Crit.	2,657.777
Note :	*p<0.05 ; **p<0.01

Plusieurs différences significatives entre les groupes ethniques sont également constatées chez les hommes, mais pas chez les femmes. Les hommes hispaniques et afro-américains produisent plus de graphies phonétiques que les hommes blancs. Les hommes asiatiques en produisent moins que les hommes afro-américains. On remarque par ailleurs la même corrélation négative entre âge et production de graphies phonétiques notée précédemment, pour les femmes comme pour les hommes.

11.4 G-droppings

11.4.1 Types de g-droppings

1662 g-droppings, dont 339 types différents, ont été recensés dans le corpus. Le tableau 11.10 présente les 20 g-droppings les plus fréquents. *fuckin* est le g-dropping le plus fréquent du corpus (322 occurrences). *lookin* (94 occurrences) arrive en deuxième position, suivi par *freakin* (80 occurrences) et *friggin* (61 occurrences). Près de la moitié des g-droppings du corpus sont

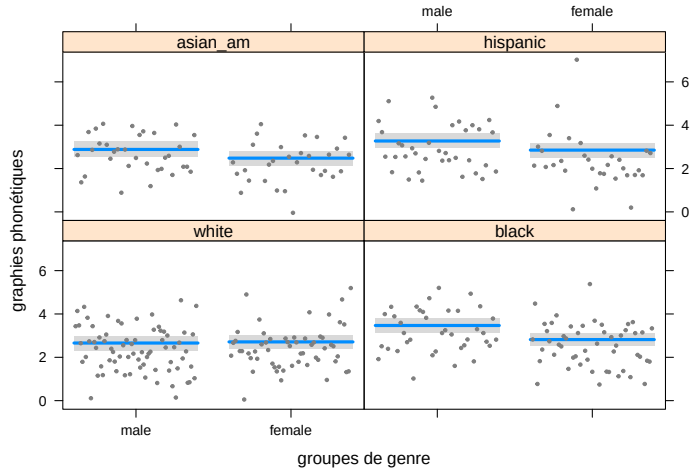

FIGURE 11.5 – Interaction genre et ethnicité dans la production de graphies phonétiques

suivis d'une apostrophe (779 g-droppings, soit 46.87 % de toutes les occurrences).

TABLEAU 11.10 – 20 g-droppings les plus fréquents dans RedditGender

Rang	G-dropping	Fréq.	Rang	G-dropping	Fréq.
1	fuckin	322	11	goin	27
2	lookin	94	12	talkin	25
3	freakin	80	13	comin	21
4	friggin	61	14	feelin	20
5	sayin	56	15	lovin	19
6	doin	54	16	tryin	15
7	frickin	48	17	killin	13
8	nothin	48	18	livin	13
9	gettin	34	19	somethin	13
10	rockin	33	20	makin	11

Prédominance de *fuckin* et de ses synonymes

La liste des g-droppings contient 30 synonymes et variantes orthographiques de *fucking*, soit 8.85 % de tous les types de g-dropping recensés. On compte ainsi 582 occurrences des diverses graphies de *fuckin* et de ses synonymes, soit 35.02 % de l'ensemble des g-droppings du corpus. Ces synonymes et variantes sont présentés dans le tableau 11.11. On note notamment la présence d'euphémismes de *fucking*, comme *freaking*, *frigging* et *fricking* (« Definition of FREAKING », p. d., « frigging », p. d., « fricking », p. d.). La liste comprend également des variantes régionales comme *fook*, utilisée dans le nord de l'Angleterre (« fook », p. d.), et *fark*, originaire d'Australie et de Nouvelle-Zélande (« fark », p. d.), ainsi que des étirements gra-

phiques (*faaaaaarkin*, *fuuuuuckin*). On remarque aussi la présence du suffixe *fuckin* dans le terme *mothafucking* et ses variantes orthographiques *muthafuckin*, *m'fugging* ou encore *mu'fuckin*.

Tableau 11.11 – Liste des occurrences de *fucking* et de ses synonymes, avec g-dropping

G-dropping	Fréq.	G-dropping	Fréq.
fuckin	322	muhfuckin	2
freakin	80	feckin	2
friggin	61	faaaaaarkin	1
frickin	48	faaargin	1
effin	7	fahckin	1
fookin	7	fckin	1
motherfuckin	7	fockin	1
fkin	5	fokkin	1
flippin	5	frikin	1
frikkin	5	fuckfuckityfuckin	1
flippin	5	fuuuuuckin	1
fukin	4	m'fuggin	1
mothafuckin	4	mothafukin	1
muthafuckin	3	mu'fuckin	1
fackin	2	f'in	1
		Total	**582**

Au vu de la fréquence importante de *fuckin* et de ses synonymes dans les g-droppings recensés, il est possible que ces termes engendrent davantage de g-droppings que les autres mots. Pour vérifier si cette hypothèse peut être acceptée, un test du χ^2 a été effectué (→ p. 159). Il est basé sur le tableau de contingence 11.4.1, qui contient les fréquences brutes des formes standard et les formes non standard des 10 g-droppings les plus fréquents de RedditGender. Le résultat est significatif : χ^2 (9 degrés de liberté) = 7683, valeur $p < 0.001$. La taille d'effet a été mesurée avec le test V de Cramer. Au vu du nombre de degrés de liberté (9), la taille d'effet (V = 0.368) peut être considérée comme forte (Brezina, 2018).

Tableau 11.12 – Fréquence des formes standard et non standard des 10 g-droppings les plus fréquents dans RedditGender

	G-dropping	Graphies standard
fucking	322	6851
looking	94	8441
freaking	80	540
frigging	61	9
saying	56	7351
doing	54	12041
fricking	48	17
nothing	48	8094
getting	34	12369
rocking	33	110

Le tableau 11.13 présente les résidus du test du χ^2. Elle indique à la fois la taille de l'effet (avec la présence ou non du signe -) et sa force (mesurée par les valeurs absolues). On y voit qu'effectivement, *fucking* et ses synonymes sont plus susceptibles que les autres mots figurant dans la liste d'entrainer des g-droppings, tout comme le mot *rockin*.

Tableau 11.13 – Résidus du test du χ^2

	G-droppings	Graphies standard
fuckin	21.16	-2.58
lookin	-2.78	0.34
freakin	23.53	-2.87
friggin	59.22	-7.22
sayin	-5.04	0.61
doin	-9.25	1.13
frickin	48.21	-5.88
nothin	-6.53	0.80
gettin	-10.96	1.34
rockin	21.35	-2.60

Importance du contexte

L'observation du contexte des concordances semble révéler que certaines constructions sont plus propices que d'autres au g-dropping. Par exemple, 31 des 94 occurrences de *lookin* sont suivies de *good*, et 13 d'autres quasi-synonymes de *good* comme *great*, *cute*, *sharp* ou *dapper*. En tout, 44 g-droppings de *looking* ont été utilisés dans cette construction qui vise à faire un compliment, soit 46.81 % de toutes les occurrences de *lookin*. Le g-dropping de *saying* semble dépendre encore plus fortement du contexte. 46 occurrences de *sayin* sur 56, soit 78.65 % de toutes les occurrences, se produisent dans l'expression *just sayin'*, par laquelle un locuteur se décharge de la responsabilité de son assertion (Davies, 2015).

11.4.2 Effet de l'âge et du genre sur la production de g-droppings

Statistiques descriptives

526 personnes, soit plus de la moitié des Redditors de RedditGender (50.38 %), n'ont pas utilisé de g-droppings. La fréquence médiane est donc de 0. La moyenne est de 0.01 g-dropping par 1000 tokens (écart type = 0.02). Les moyennes et médianes, par groupe de genre et groupe d'âge, sont présentées dans le tableau 11.14. On voit que seul·es les hommes cisgenres et les Redditors les plus jeunes affichent une médiane supérieure à 0. Il y a deux valeurs extrêmes dans les données, que nous avons retirées pour les analyses qui suivent (la première valeur aberrante correspond à la production d'un homme américain hispanique de 16 ans qui a utilisé 0.33 g-droppings par 1000 tokens, et la seconde à celle d'une femme canadienne

de 30 ans qui a utilisé 0.17 g-droppings par 1000 tokens).

Tableau 11.14 – Fréquence des g-droppings dans le corpus

Sous-corpus	Moyenne	Médiane	1 g-dropping min.
Hommes cisgenres	0.01	0.01	55.26 %
Femmes cisgenres	0.01	0	46.90 %
Hommes transgenres	0.01	0	42 %
Femmes transgenres	0.01	0	46 %
Non-binaires	0.01	0	49 %
14-20 ans	0.01	0.01	52.38 %
21-30 ans	0.01	0	48.94 %
31 ans et +	0.01	0	49.47 %
Tous	0.01	0	49.62 %

Effets du genre et de l'âge sur la fréquence des g-droppings : modèle de régression

Nous avons tenté de créer un modèle *zero-inflated*, mais sans succès. Cela peut être dû à un manque de variation dans les données pour ce groupe, comme le suggère une réponse du créateur du package `pscl` à une question d'un internaute qui a rencontré un problème similaire (Alex, 2016). À la place, nous avons opté pour un modèle de régression binomial négatif, avec la fréquence brute des g-droppings comme variable dépendante, et l'âge, le genre, et l'interaction du genre et de l'âge comme variables indépendantes. La fonction `step()` a retenu uniquement les effets principaux du genre et de l'âge, et non leur interaction. Les hommes cisgenres sont le niveau de référence du modèle.

Le tableau 11.15 présente les coefficients, leurs valeurs p et leurs erreurs standard. Selon le modèle, l'âge n'a pas d'effet significatif sur la production de g-droppings, mais le genre en a un. Tous les coefficients sont significatifs et, une fois exponentialisés, sont inférieurs à 1 (0.63 pour les femmes cisgenres, 0.59 pour les hommes transgenres, 0.64 pour les femmes transgenres, et 0.66 pour les personnes non binaires). Cela signifie que tous les hommes cisgenres, niveau de référence du modèle, ont produit significativement plus de g-droppings que les autres Redditors. Quand les femmes cisgenres sont le niveau de référence du modèle, aucune différence significative n'est notée entre elles d'un côté, et les groupes transgenres et non-binaire de l'autre. Les hommes cisgenres semblent donc se démarquer du reste des Redditors par leur utilisation plus fréquente de g-droppings.

Tableau 11.15 – G-droppings, effets de l'âge et du genre

	Variable dépendante :
	G-droppings
Intercept	0.0001** (0.0001, 0.0002)
Femmes cisgenres	0.628** (0.491, 0.802)
Hommes transgenres	0.589** (0.404, 0.868)
Femmes transgenres	0.644* (0.444, 0.947)
Non-binaires	0.655* (0.452, 0.962)
21-30 ans	0.957 (0.698, 1.300)
30 ans et +	0.755* (0.544, 1.040)
Observations	1,042
Log Likelihood	−1,707.078
θ	0.462** (0.033)
Akaike Inf. Crit.	3,428.155
Note :	*p<0.05 ; **p<0.01

11.4.3 Effet de l'ethnicité et son interaction avec le genre sur la fréquence des g-droppings

Statistiques descriptives

Pour cette analyse, nous avons utilisé le jeu de données réduit décrit dans la section 5.1.7, en supprimant la valeur aberrante qui s'y trouvait. Les données comprennent 346 observations. Dans ce jeu de données réduit, les fréquences moyennes et médianes des g-droppings utilisés par hommes et femmes par 1000 tokens sont les mêmes que celles notées dans l'ensemble du corpus (tableau 11.14). On constate une différence dans les groupes d'âge ; les Redditors de 14 à 20 ans et de 21 à 30 ans ont produit le même nombre médian de g-droppings par 1000 tokens (0.01).

Les fréquences moyennes (0.1) et médianes (0.01) des g-droppings sont les mêmes pour tous les groupes raciaux, à l'exception des Asiatiques (moyenne = 0 ; médiane = 0), comme le montre la figure 11.6. Les pourcentages de Redditors de chaque groupe ayant utilisé au moins un g-dropping sont les suivants : 55.47 % pour les blanc·hes (71 personnes sur 128), 58.43 % pour les Afro-Américain·es (52 personnes sur 89), 54.55 % pour les Hispaniques (36 personnes sur 66), et 34.92 % (22 personnes sur 63) pour les Asiatiques. Ces statistiques descriptives semblent indiquer que les Redditors asiatiques utilisent moins de g-droppings que les autres.

Modèle

Nous avons opté pour un modèle binomial négatif, comprenant comme variables indépendantes l'interaction du genre et de l'ethnicité et l'effet principal de l'âge. Les hommes blancs sont le niveau de référence du modèle (tableau 11.16). L'effet de l'âge n'est pas significatif. L'effet du genre varie selon les groupes ethniques (figure 11.7) et est uniquement significatif dans le groupe hispanique, où les hommes produisent 4.92 g-droppings quand les femmes en produisent 1. Les hommes asiatiques utilisent significativement

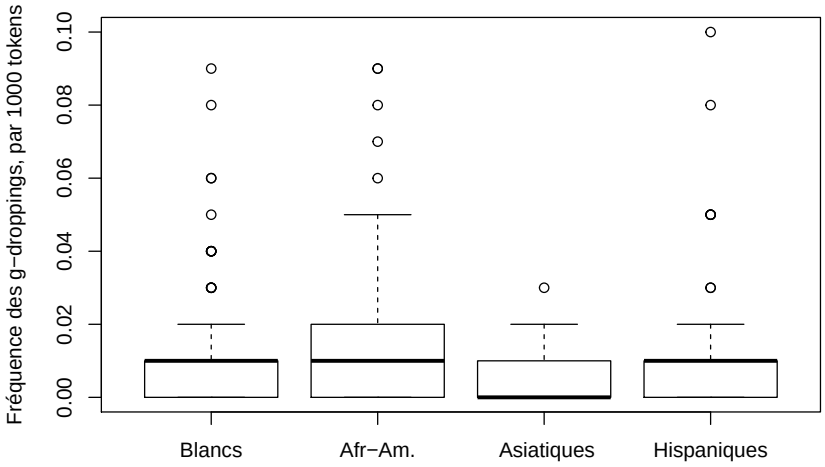

FIGURE 11.6 – Fréquence des g-droppings pour 1000 tokens, par groupe ethnique

moins de g-droppings que tous les autres groupes (5 fois moins que les Afro-Américains et les Hispaniques, 3.21 fois moins que les blancs). Les femmes blanches et afro-américaines emploient davantage de g-droppings que les femmes asiatiques, dans des proportions similaires (environ 3 fois plus). La fréquence des g-droppings est également plus élevée dans le groupe des Afro-américaines que dans le groupe hispanique.

TABLEAU 11.16 – G-droppings, effet de l'âge et de l'interaction entre genre et ethnicité

	Variable dépendante :
	G-droppings
Intercept	0.0001^{**} (0.00004, 0.0001)
Femmes cisgenres	0.968 (0.565, 1.666)
Afro-Américains	1.701 (0.939, 3.150)
Asiatiques	0.322^{**} (0.152, 0.678)
Hispaniques	1.611 (0.870, 3.059)
21-30 ans	1.434 (0.797, 2.534)
30 ans et +	1.308 (0.718, 2.337)
Femmes cisgenres :Afro-américaines	0.610 (0.266, 1.384)
Femmes cisgenres :Asiatiques	1.129 (0.388, 3.284)
Femmes cisgenres :Hispaniques	0.210^{**} (0.078, 0.560)
Observations	346
Log Likelihood	−582.892
θ	0.559^{**} (0.071)
Akaike Inf. Crit.	1,185.785
Note :	*p<0.05 ; **p<0.01

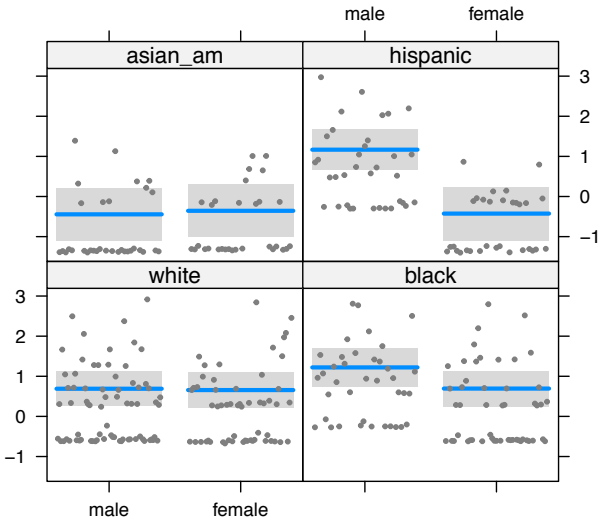

FIGURE 11.7 – Interaction du genre et de l'ethnicité dans la production de g-droppings, valeurs prédites

11.4.4 Analyse du groupe non-binaire

Les g-droppings semblant être des marqueurs masculins (même si l'ethnicité a aussi un impact dans leur utilisation), nous avons essayé de savoir si on retrouvait, au sein du groupe non-binaire, des différences significatives entre les personnes AGAN et AFAN. Sur les 98 Redditors concernés, la moitié (49) ont utilisé au moins un g-dropping. La médiane est de 0 pour les deux groupes, et la moyenne de 0.01. L'âge, le genre assigné à la naissance et leur interaction ont été intégrés au modèle comme variables indépendantes. Le processus de sélection des variables a uniquement retenu le genre assigné à la naissance ; nous avons toutefois conservé l'âge en tant que variable de contrôle. Le modèle, présenté dans le tableau 11.17, ne fait état d'aucune différence significative entre les deux groupes.

TABLEAU 11.17 – G-droppings, effet de l'âge assigné à la naissance

	Variable dépendante :
	G-droppings
Intercept	0.00005** (0.00002, 0.0001)
AGAN	1.417 (0.715, 2.867)
21-30 ans	1.372 (0.489, 3.612)
31 ans et +	1.233 (0.395, 3.742)
Observations	98
Log Likelihood	−154.980
θ	0.497** (0.122)
Akaike Inf. Crit.	317.961
Note :	*p<0.05 ; **p<0.01

11.5 Omissions d'apostrophe

11.5.1 Résultats globaux

Vingt-huit omissions d'apostrophe ont été analysées dans cette étude, après élimination des cas d'ambiguïté. Cela représente un total de 17 332 omissions d'apostrophe. Le tableau 11.18 présente ces omissions d'apostrophes par ordre décroissant de fréquence, ainsi que la fréquence de leurs équivalents avec apostrophe, et le pourcentage occupé par les graphies non standard dans l'ensemble des formes observées (avec et sans apostrophe). Plus ce pourcentage est faible, moins la graphie sans apostrophe est fréquente par rapport à sa forme standard. Ces résultats montrent que *its* est de loin l'omission d'apostrophe la plus fréquente dans RedditGender (7429 occurrences), suivie par *im* (2827 occurrences) et *thats* (1504 occurrences). À elles trois, ces graphies représentent 67.85 % des omissions d'apostrophes étudiées. La fréquence des autres omissions d'apostrophes décline ensuite très rapidement, 22 graphies apparaissant moins de 500 fois et 8 graphies moins de 100 fois dans le corpus.

Tableau 11.18 – Omissions d'apostrophe dans RedditGender

Rang	Omission	Fréquence	Graphie standard	%
1	its	7429	80 034	8.49
2	im	2827	79 953	3.42
3	thats	1504	28 603	5.00
4	cant	795	17 527	4.34
5	ive	627	28 817	2.13
6	id	440	14 956	2.86
7	isnt	422	11 935	3.42
8	wont	409	6116	6.27
9	youre	379	29 425	1.27
10	ill	361	10 813	3.23
11	lets	304	1746	14.83
12	whats	287	3905	6.85
13	wasnt	227	7368	2.99
14	arent	202	6589	2.97
15	wouldnt	187	7497	2.43
16	shes	179	7249	2.41
17	theyre	179	12 835	1.38
18	hes	173	9922	1.71
19	couldnt	143	4083	3.38
20	shouldnt	74	2552	2.82
21	youve	42	4152	1.00
22	youll	40	5348	0.74
23	theyll	31	1843	1.65
24	youd	25	2611	0.95
25	theyve	20	1593	1.24
26	theyd	16	1263	1.25
27	weve	10	1582	0.63
Tous		**17 332**	**390 317**	**4.25**

L'observation des fréquences des graphies avec apostrophe et des pour-

centages révèle que *its* est plus de 2 fois plus fréquent que *im*, dans l'absolu, mais aussi proportionnellement à leurs formes standard, qui ont une fréquence comparable dans le corpus (80034 pour *it's* et 79953 pour *I'm*). La seule autre graphie non standard qui apparait plus fréquemment que *its*, proportionnellement à sa forme avec apostrophe, est *lets* (14.83 % de l'ensemble des formes). Viennent ensuite, *its*, *whats*, *wont* et *thats*.

11.5.2 Effets du genre et de l'âge sur la fréquence des omissions d'apostrophe

Statistiques descriptives

Comme pour toute les variables linguistiques étudiées dans cette thèse, il y a une dispersion importante dans les données, qui indique de grandes différences individuelles. La valeur maximale est de 22 omissions d'apostrophe par 1000 tokens ; la médiane est de 0.15. Deux-cent-trois personnes n'ont pas omis l'apostrophe une seule fois. Le tableau 11.19 indique que la moyenne est de 0.92 omissions par 1000 tokens, avec un écart type de 2.44.

Tableau 11.19 – Fréquence des omissions d'apostrophe, pour 1000 tokens

	Moyenne	ET	Médiane	EI
Hommes cisgenres	1.01	2.21	0.21	0.59
Femmes cisgenres	0.78	2.45	0.11	0.28
Femmes transgenres	1.19	2.52	0.16	1.05
Hommes transgenres	0.92	3.19	0.08	0.29
Non-binaires	0.83	2.25	0.11	0.32
14-20 ans	1.09	2.62	0.21	0.94
21-30 ans	1.00	2.63	0.15	0.47
31 ans et +	0.74	2.06	0.12	0.28
Tous	0.92	2.44	0.15	0.42

Les hommes cisgenres et les femmes transgenres ont produit le plus d'omissions d'apostrophe, avec des médianes respectives de 0.21 et 0.16 omissions d'apostrophe par 1000 tokens. C'est également dans ces groupes que la dispersion est la plus importante, avec des écarts interquartiles de 0.59 et 1.05, (comparé à 0.28 pour les femmes cisgenres par exemple), ce qui indique des différences individuelles plus prononcées.

On remarque par ailleurs une diminution de la fréquence des omissions d'apostrophe avec l'âge. Les plus jeunes affichent une médiane de 0.21 omission par 1000 tokens, et les plus âgé·es une médiane de 0.12 omission. La dispersion diminue également avec l'âge, indiquant une fréquence plus homogène dans les groupes les plus âgés.

Effets de l'âge et du genre sur la fréquence de l'omission d'apostrophe : modèle de régression

Nous avons créé un modèle de régression binomial négatif avec le genre, l'âge et leur interaction comme variables indépendantes, et la fréquence brute des omissions d'apostrophe comme variable dépendante ainsi que, comme d'habitude, un offset qui prend en compte le fait que les sous-corpus ont des tailles différentes (→ p. 165). Nous avons ensuite sélectionné les variables à l'aide de la fonction `step()`. Celle-ci a supprimé l'interaction et l'effet principal du genre ; nous l'avons toutefois conservé comme variable de contrôle. Le modèle est présenté dans le tableau 11.20. Les femmes cisgenres sont le niveau de référence. Le modèle montre que les femmes cisgenres omettent moins fréquemment l'apostrophe que les hommes cisgenres : quand ceux-ci en omettent 1, elles en omettent 0.74. Il n'y a aucune autre différence significative entre les autres groupes de genre. Le modèle indique également un effet limité de l'âge. Il n'y a pas de différence entre le groupe 1 (14-20 ans) et le groupe 2 (21-30 ans). En revanche les Redditors de plus de 31 ans produisent significativement moins d'omissions d'apostrophes que les deux groupes les plus jeunes.

Tableau 11.20 – Omissions d'apostrophe, effet de l'âge et du genre

	Variable dépendante :
	Omissions d'apostrophe
Intercept	0.001^{**} (0.001, 0.001)
Hommes cisgenres	1.348^{*} (1.038, 1.751)
Femmes transgenres	1.413 (0.960, 2.140)
Hommes transgenres	1.057 (0.714, 1.610)
Non-binaires	1.032 (0.702, 1.560)
21-30 ans	0.951 (0.675, 1.315)
31 ans et +	0.677^{*} (0.473, 0.954)
Observations	1,044
Log Likelihood	−3,529.415
θ	0.317^{**} (0.013)
Akaike Inf. Crit.	7,072.831
Note :	*p<0.05 ; **p<0.01

11.5.3 Effets du genre, de l'âge et de l'ethnicité sur la fréquence des omissions d'apostrophe

Statistiques descriptives

Les fréquences moyennes et médianes par groupes ethniques, groupes d'âge et de genre sont présentées dans le tableau 11.21. Dans cet échantillon réduit contenant uniquement les personnes cisgenres, 55 Redditors n'ont pas produit d'omission d'apostrophe.

On remarque, comme c'était le cas dans l'ensemble du corpus, une fréquence plus élevée des omissions d'apostrophe chez les hommes, ainsi qu'une

Tableau 11.21 – Fréquence des omissions d'apostrophe pour 1000 tokens, échantillon réduit

	Moyenne	ET	Médiane	EI
Hommes	1.13	2.51	0.21	0.73
Femmes	0.79	2.25	0.12	0.32
14-20 ans	1.26	2.60	0.26	0.88
21-30 ans	0.98	2.34	0.16	0.58
31 ans et +	0.83	2.41	0.13	0.30
Blancs	0.66	1.89	0.11	0.26
Afro-Américain·es	1.27	2.75	0.21	0.78
Asiatiques	0.90	2.26	0.15	0.34
Hispaniques	1.23	2.81	0.32	0.90
Tous	0.97	2.29	0.16	0.54

possible corrélation négative avec l'âge. Dans les différents groupes ethniques, les Redditors blancs et asiatiques ont utilisé le moins d'omissions d'apostrophe, avec des médianes respectives de 0.11 et 0.15 par 1000 tokens, contre 0.21 et 0.32 pour les groupes afro-américain et asiatique. Encore une fois, c'est dans les groupes où la fréquence des omissions d'apostrophe est la plus élevée que la dispersion est la plus importante, ce qui indique de fortes variations individuelles.

Effet de l'interaction du genre et de l'ethnicité sur la fréquence des omissions d'apostrophe : modèle de régression

Un modèle de régression binomial négatif a été créé avec la combinaison habituelle de variables indépendantes, et la fréquence brute des omissions d'apostrophe comme variable dépendante. L'interaction de l'ethnicité et du genre et l'effet de l'âge ont été retenus. Le niveau de référence du modèle (tableau 11.22) est les femmes cisgenres blanches.

Dans ce modèle, l'âge n'a pas d'effet significatif sur la fréquence des omissions d'apostrophes. L'effet du genre n'est pas le même dans tous les groupes, comme le montre la figure 11.8. Il n'y a ainsi pas de différence significative entre femmes et hommes chez les internautes blanc·hes et hispaniques. En revanche, les femmes afro-américaines et asiatiques produisent significativement moins d'omissions d'apostrophe que les hommes de leurs groupes. Les hommes afro-américains omettent 2 fois plus fréquemment l'apostrophe que les femmes afro-américaines ; les hommes asiatiques les omettent 6.14 fois plus fréquemment que les femmes asiatiques.

La comparaison des groupes d'hommes entre eux montre que les internautes afro-américains omettent 3.02 fois plus fréquemment l'apostrophe que les Redditors blancs. Le groupe asiatique omet également davantage l'apostrophe que le groupe blanc (2.29 fois plus fréquemment). Chez les femmes, les internautes asiatiques produisent significativement moins d'omissions d'apostrophes que tous les autres groupes. C'est quand on les compare au groupe hispanique que la taille d'effet est la plus marquée : les in-

TABLEAU 11.22 – Omissions d'apostrophe, effet de l'âge, du genre et de l'ethnicité

	Variable dépendante :
	Omissions d'apostrophe
Intercept	0.001** (0.001, 0.002)
Hommes cisgenres	0.878 (0.489, 1.557)
Afro-américaines	1.284 (0.652, 2.539)
Asiatiques	0.327** (0.154, 0.729)
Hispaniques	2.058 (0.958, 4.591)
31-30 ans	0.648 (0.359, 1.114)
31 ans et +	0.644 (0.349, 1.137)
Hommes :Afro-Américains	2.353 (0.946, 5.954)
Hommes :Asiatiques	6.998** (2.553, 19.142)
Hommes :Hispaniques	0.855 (0.317, 2.297)
Observations	347
Log Likelihood	−1,204.090
θ	0.384** (0.028)
Akaike Inf. Crit.	2,428.181
Note :	*p<0.05 ; **p<0.01

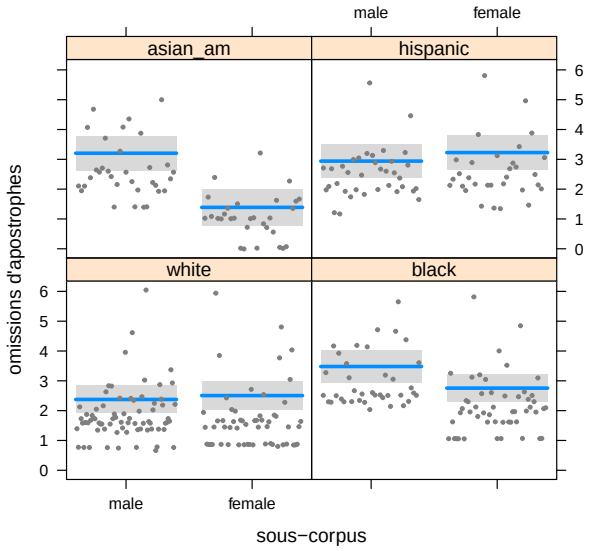

FIGURE 11.8 – Interaction du genre et de l'ethnicité dans la production d'omissions d'apostrophe

ternautes hispaniques omettent 6.28 plus fréquemment l'apostrophe que les asiatiques.

11.5.4 Étude de la variable *it's*/ *its*

L'omission de l'apostrophe de *it's* semblant être un cas à part, du fait de l'ambiguïté avec le possessif *its* (Hokanson & Kemp, 2013), nous avons décidé de l'analyser séparément, en la comparant avec la fréquence de la forme standard *it's*. Comme l'ethnicité semblait avoir un effet significatif sur l'omission d'apostrophe, nous avons choisi d'utiliser l'échantillon réduit analysé dans la section précédente, et non l'ensemble du corpus. Nous avons commencé par calculer l'*odds ratio* de l'utilisation de *its* par rapport à *it's*, en divisant la fréquence de *it's* dans le corpus par celle de *its*. Cette mesure indique la probabilité de l'utilisation de la variante non standard par rapport à la variante standard. L'*odds ratio* est de 0 pour les 111 personnes qui n'ont pas omis l'apostrophe de *it's* une seule fois. Le maximum est de 74. Vingt-et-un Redditors affichent un *odds ratio* supérieur à 1, ce qui signifie qu'ils ont utilisé *its* plus fréquemment que *it's*.

Les moyennes et médianes par groupes et dans l'ensemble du corpus sont présentées dans le tableau 11.23. La médiane, pour l'ensemble de cet échantillon réduit, est de 0.02 ; l'écart interquartile est de 0.04, ce qui signifie que, pour 75 % des Redditors, l'*odds ratio* est égal ou inférieur à 0.06. La moyenne est élevée (0.59), avec un écart type très important (4.57). Tout cela suggère de très grandes différences individuelles dans l'utilisation de *its* et de *it's* : la majorité des Redditors utilise principalement *it's*, et une minorité utilise principalement *its*. Les hommes affichent des valeurs plus importantes que les femmes ; l'*odds ratio* médian (0.03) est 3 fois plus élevé que celui des femmes (0.01). La dispersion est également bien plus importante, avec un écart type de 6.30 pour les hommes contre 0.68 pour les femmes, et un écart interquartile de 0.07 pour les femmes contre 0.04 pour les femmes. Cela indique qu'il y a des différences individuelles plus prononcées chez les hommes.

Tableau 11.23 – Odds ratio de *its* vs. *it's*

	Moyenne	ET	Médiane	EI
Hommes	1.01	6.30	0.03	0.07
Femmes	0.14	0.68	0.01	0.04
14-20 ans	0.10	0.26	0.02	0.07
21-30 ans	0.89	6.10	0.02	0.06
31 ans et +	0.30	1.49	0.02	0.05
Blancs	0.29	2.21	0.01	0.03
Afro-Américain·es	1.43	8.40	0.02	0.18
Asiatiques	0.11	0.29	0.01	0.04
Hispaniques	0.50	1.94	0.03	0.07
Tous	0.59	4.57	0.02	0.06

L'âge ne semble pas avoir d'effet ; la médiane est la même (0.02) dans les

trois groupes. Les Redditors hispaniques ont l'*odds ratio* médian le plus élevé (0.03), et les Redditors blancs et asiatiques l'*odds ratio* médian le plus faible (0.01). La dispersion semble particulièrement importante dans le groupe afro-américain, avec un écart type de 8.40, contre par exemple 0.29 pour le groupe asiatique.

Modèle de régression avec effets mixtes

Nous avons ensuite réalisé un modèle de régression logistique binaire (une méthode traditionnellement utilisée dans la sociolinguistique variationniste, → p. 162) avec effets mixtes. Ce modèle a pour variable dépendante les réalisations de *its* et *it's* dans l'échantillon ; elle comporte donc 2 niveaux. Il y a 27 806 observations, qui correspondent à 25 558 occurrences de *it's* et à 2248 occurrences de *it's*. La forme standard *it's* est le niveau de référence de la variable. Le modèle prédit donc la probabilité des réalisations de *its*, sous forme de *log odds*. Les variables indépendantes sont l'âge, le genre, l'ethnicité, et l'interaction du genre et de l'ethnicité. Comme chacun·e des 347 Redditors qui composent l'échantillon étudié ici contribue à plusieurs observations, nous avons ajouté un effet aléatoire correspondant à l'identifiant de chaque personne (→ p. 162). Le modèle (tableau 11.24) a été créé avec la fonction `glmer()` du package `lme4` (Bates et al., 2015). La fonction `step()` a été utilisée pour sélectionner les variables indépendantes. Toutes les variables ont été conservées par l'algorithme. Les hommes blancs sont le niveau de référence du modèle.

Tableau 11.24 – Production de *its* vs. *it's*, régression logistique binaire

	Variable dépendante :
	it's vs. *its*
Intercept	0.015^{**} (0.006, 0.036)
Femmes	0.481 (0.195-6, 1.181)
Afro-Américains	4.797^{**} (1.853, 12.416)
Asiatiques	2.302 (0.835, 6.353)
Hispaniques	2.837^{*} (1.071, 7.508)
21-30 ans	0.949 (0.413, 2.181)
31 ans et +	1.018 (0.426, 2.433)
Femmes :Afro-américaines	0.570 (0.148, 2.188)
Femmes :Asiatiques	0.226 (0.047, 1.091)
Femmes :Hispaniques	1.212(0.282, 5.197)
Observations	27,806
Log Likelihood	−4,587.529
Akaike Inf. Crit.	9,197.058
Bayesian Inf. Crit.	9,287.621
Note :	*p<0.05 ; **p<0.01

Le modèle indique qu'il n'y a pas de corrélation entre l'âge et l'omission de l'apostrophe de *it's*. Le genre a uniquement un effet significatif dans le groupe afro-américain et le groupe asiatique. La probabilité d'utiliser *its* à la place de *it's* est 3.65 plus importante chez les hommes que chez les

femmes dans le groupe afro-américain ; elle est 9.19 fois plus importante pour les hommes dans le groupe asiatique. La comparaison des groupes d'hommes entre eux révèle que les hommes afro-américains et hispaniques ont davantage tendance à omettre l'apostrophe que les hommes blancs (probabilité 4.79 fois plus importante pour les Afro-Américains, 2.84 fois plus importante pour les hispaniques). Chez les femmes, les afro-américaines sont 2.73 fois susceptibles que les internautes blanches et 5.25 fois plus susceptibles que les internautes asiatiques d'omettre l'apostrophe de *it's*. Les femmes hispaniques omettent également plus souvent l'apostrophe que les femmes asiatiques (probabilité 6.60 fois plus forte) et blanches (probabilité 3.43 plus forte).

11.6 Omission de la majuscule de *I*

11.6.1 Fréquence de *i*

Statistiques descriptives

19 858 occurrences de *i* ont été relevées dans le corpus. Le tableau 11.25 présente les fréquences moyennes et médianes de l'omission de la majuscule de *I*, pour 1000 tokens, ainsi que la proportion de personnes ayant utilisé *i* au moins 1 fois (« Uti. uniques »). Dans l'ensemble du corpus, 373 personnes (35.73 %) n'ont pas utilisé le *i* une seule fois. La médiane de chaque groupe de genre est la même (0.01). On constate en revanche d'apparentes disparités entre les moyennes. Les hommes transgenres ont omis la majuscule plus fréquemment que les autres groupes (0.19 fois par 1000 tokens). Les hommes cisgenres l'ont omise le moins souvent (médiane de 0.07 fois par 1000 tokens). Il y a plus de dispersion dans le groupe des hommes transgenres (écart type = 0.62), ce qui suggère une plus forte dispersion des données, et davantage de valeurs extrêmes.

Le tableau semble également montrer que l'utilisation de *i* diminue avec l'âge. Les adolescents et les jeunes adultes ont utilisé en moyenne 0.17 *i* par 1000 tokens, contre seulement 0.06 *i* pour les 31 ans et plus. La fréquence médiane est deux fois plus élevée pour les plus jeunes que pour les deux autres groupes. Toutefois, il faut noter que le groupe des 14-20 ans présente l'écart type et l'écart interquartile les plus élevés ; c'est donc dans ce groupe que l'on a le maximum de dispersion.

La proportion d'utilisateur·trices uniques de *i* est la plus faible chez les femmes cisgenres (209 utilisatrices uniques, soit 56.18 %). La proportion d'utilisateur·trices uniques dans les autres groupes est similaire (entre 64 % pour les hommes transgenres et les personnes non binaires, et 70.97 % pour les hommes cisgenres). Le nombre d'utilisateur·trices uniques décroit avec l'âge : le *i* a été utilisé par 77.55 % des 14-20 ans, par 64.02 % des 21-30 ans, et par 59.47 % des 31 ans et plus.

TABLEAU 11.25 – Fréquence de *i*, pour 1000 tokens

Sous-corpus	Moyenne	ET	Médiane	EI	Uti. uniques
Hommes cisgenres	0.07	0.23	0.01	0.03	70.97 %
Femmes cisgenres	0.10	0.41	0.01	0.02	56.18 %
Hommes transgenres	0.19	0.62	0.01	0.03	64 %
Femmes transgenres	0.15	0.43	0.01	0.04	70 %
Non-binaires	0.13	0.46	0.01	0.04	64 %
14-20 ans	0.17	0.51	0.02	0.07	77.55 %
21-30 ans	0.12	0.44	0.01	0.03	64.02 %
31 ans et +	0.06	0.23	0.01	0.02	59.47 %
Tous	0.11	0.39	0.01	0.03	64.27 %

11.6.2 Effets de l'âge et du genre sur le choix entre les variantes *i* et *I*

Dans cette sous-section et dans la suivante, nous étudions non pas la fréquence de *i*, mais sa fréquence par rapport à celle de *I*. En d'autres termes, nous étudions le choix des internautes entre les deux variantes, standard et non standard, à leur disposition.

Statistiques descriptives

Nous avons comparé les fréquences relatives de *I* et de *i* avec les boites à moustaches de la figure 11.9. Elles révèlent que la distribution de *I* est bien plus symétrique que la distribution de *i*. Elle ne comporte que 2 valeurs, tandis que la fonction `boxplot.stats` de R indique la présence de 155 valeurs aberrantes possibles dans la distribution de *i*. Cela indique qu'il y a davantage de dispersion dans la distribution *i*, avec des différences individuelles très marquées.

Les *odds ratios* présentés dans le tableau 11.26 indiquent la probabilité d'utiliser un *i* par rapport à un *I*, pour chaque groupe. Plus cette mesure est élevée, plus les Redditors de chaque groupe ont tendance à utiliser *i* à la place de *I*. On remarque que les *odds ratio* sont en général faibles, avec une médiane de 0.002 pour l'ensemble des Redditors. L'écart type et l'écart interquartile sont importants, indiquant la présence d'importantes variations individuelles. Dix-neuf Redditors ont utilisé plus de *i* minuscules que de *I* majuscules. Quatre ont utilisé autant que *i* minuscules que de *I* majuscules. L'immense majorité des Redditors utilise donc peu fréquemment la variante sans majuscule ; la majorité des occurrences de *i* sont dues à un petit nombre d'internautes qui les utilise fréquemment, par rapport à la moyenne et à la médiane. Les médianes des différents groupes de genre et d'âge sont toutes de 0.002 ou de 0.003, à l'exception des 14-20 ans, qui ont une médiane de 0.006, et qui semblent donc utiliser plus fréquemment la variante *i* que les Redditors plus âgés.

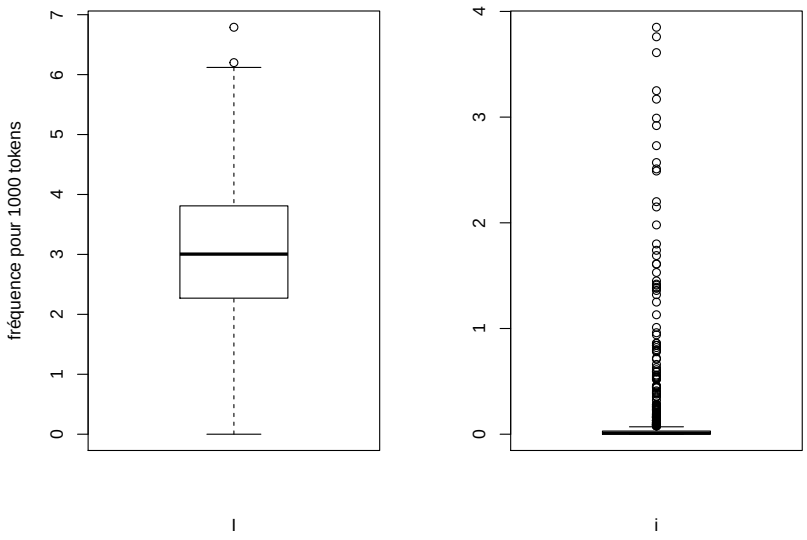

FIGURE 11.9 – Fréquence relative de *I* et de *i* dans RedditGender, pour 1000 tokens

TABLEAU 11.26 – *Odds ratios* de l'utilisation de *i* par rapport à *I*

	Moyenne	ET	Médiane	EI
Hommes cisgenres	0.028	0.084	0.003	0.012
Femmes cisgenres	0.028	0.119	0.002	0.005
Hommes transgenres	0.059	0.193	0.002	0.009
Femmes transgenres	0.048	0.138	0.003	0.014
Non-binaires	0.041	0.145	0.002	0.015
14-20 ans	0.050	0.145	0.006	0.020
21-30 ans	0.039	0.140	0.002	0.008
31 ans et +	0.022	0.080	0.002	0.007
Tous	0.034	0.122	0.002	0.009

Effet de l'âge et du genre sur le choix entre *i* et *I* : modèle de régression logistique binaire avec effets mixts

Pour étudier le choix entre les deux variantes *i* et *I*, nous avons créé, comme pour la variable *it's/its*, un modèle de régression logistique binaire avec un effet aléatoire correspondant à l'identifiant de chaque personne. La variable dépendante comporte deux niveaux, qui correspondent aux réalisations du pronom personnel avec et sans majuscule. Nous avons essayé de créer ce modèle avec l'ensemble des observations, soit 612 676 observations correspondant à 592 818 occurrences de *I* et à 19 858 occurrences de *i*. R n'a pas été capable de réaliser ce modèle, à cause du nombre trop important d'observations. Nous avons donc utilisé la fonction `sample()` pour créer un échantillon aléatoire de 100 000 observations, qui correspondent à 96 739 occurrences de *I* et à 3261 occurrences de *i*. L'interaction du genre et de l'âge a été supprimée suite au processus de sélection des variables avec la fonction `step()`. Le niveau de référence de la variable dépendante est le *I* majuscule. Le modèle prédit donc l'utilisation de *i*.

Ce modèle, présenté dans le tableau 11.27, montre qu'il existe une corrélation entre l'âge et le fait de préférer *i* à *I*. La probabilité d'un Redditor utilise *i* à la place de *I* est 2.51 fois forte chez les moins de 20 ans que chez les 21-30 ans, et 3.59 fois plus forte chez les moins de 20 ans que chez les plus de 31 ans. Il n'y a toutefois pas de différence significative entre les deux groupes les plus âgés. L'effet du genre est lui aussi significatif : les femmes cisgenres omettent moins souvent la majuscule de *I* que les hommes cisgenres et les femmes transgenres. La probabilité qu'une femme cisgenre utilise *i* est environ 2 fois moins forte que pour un homme cisgenre, et environ 3 fois moins que pour une femme transgenre. Aucune autre différence significative n'est révélée par le modèle.

Tableau 11.27 – Effets de l'âge et du genre sur le choix entre *I* et *i*

	Variable dépendante :
	Choix de la variante *i* par rapport à *I*
Intercept	0.005** (0.002, 0.010)
Femmes cisgenres	0.469** (0.267, 0.823)
Hommes transgenres	0.739 (0.319, 1.710)
Femmes transgenres	1.173 (0.526, 2.617)
Non-binaires	1.021 (0.448, 2.325)
21-30 ans	0.398** (0.207, 0.766)
31 ans et +	0.278** (0.139, 0.558)
Observations	100,000
Log Likelihood	−7,178.459
Akaike Inf. Crit.	14,372.920
Bayesian Inf. Crit.	14,449.020
Note :	*p<0.05 ; **p<0.01

Effet de l'âge, du genre et de l'ethnicité sur le choix entre *i* et *I*

Nous avons réalisé le même type de modèle que dans la section précédente, avec un échantillon aléatoire de 100 000 observations tirées du sous-corpus cisgenre (→ p. 136). L'échantillon contient 97186 occurrences de *I* et 2814 occurrences de *i*. Nous avons utilisé la fonction `drop1()` pour sélectionner les variables, et avons enlevé l'interaction de l'âge et du genre. Le modèle est présenté dans le tableau 11.28. Son niveau de référence est les hommes cisgenres.

Tableau 11.28 – Effets de l'âge, du genre et de l'ethnicité sur le choix entre *I* et *i*

	Variable dépendante :
	Choix de la variante *i* par rapport à *I*
Intercept	0.007^{**} (0.003, 0.015)
Femmes cisgenres	0.444^{**} (0.252, 0.7803)
Afro-Américain·es	2.647^{**} (1.274, 5.501)
Asiatiques	1.815 (0.796, 4.141)
Hispaniques	2.077 (0.935, 4.613)
21-30 ans	0.381^{*} (0.164, 0.886)
31 ans et +	0.315^{*} (0.129,0.767)
Observations	100,000
Log Likelihood	−7,054.181
Akaike Inf. Crit.	14,124.360
Bayesian Inf. Crit.	14,200.470
Note :	*p<0.05 ; **p<0.01

Le modèle montre que les femmes sont moins susceptibles que les hommes d'utiliser *i* à la place de *I*. L'effet de l'âge est significatif, mais uniquement quand on compare les plus jeunes (14-20 ans) aux deux groupes les plus âgés. Les Redditors les plus jeunes omettent plus souvent la majuscule que les plus de 21 ans. Enfin, les internautes afro-américain·es ont tendance à omettre la majuscule significativement plus fréquemment que les internautes blancs (2.65 fois plus fréquemment). Il n'y a pas de différence entre les autres groupes.

11.7 Discussion

11.7.1 Abréviations

Acronymes : fréquence et types

Même si notre analyse n'est pas exhaustive, puisque nous ne nous sommes intéressée à 33 types d'acronymes seulement, il semble que les acronymes soient une des caractéristiques principales de la langue de Reddit. Nous avons en effet recensé 31 529 acronymes, contre 21 201 émoticônes et 17 332 omissions d'apostrophe. Seules 5 personnes n'ont pas utilisé ce procédé. *lol* est de loin l'acronyme le plus fréquent du corpus (9215 occurrences), ce qui n'est pas une surprise. Plusieurs autres études de la CMC

font en effet le même constat, comme Baron (2004), Tagliamonte et Denis (2008) et Tagliamonte (2016a), dans leurs corpus de messagerie instantanée, et Hilte (2019), dans son corpus de messages rédigés par des adolescent·es belges néerlandophones.

On trouve dans notre liste des acronymes souvent recensés, comme *Omg* et *WTF*, mais aussi des acronymes qui figurent rarement dans les études de Twitter, de l'IM ou des SMS, comme *SO*, *OP*, *TLDR* et *TIL*, qui sont spécifiques aux forums de discussion. Utilisé dans des contextes où on parle de son expérience personnelle, *SO* permet d'évoquer son ou sa partenaire sans dévoiler son genre ou son orientation sexuelle (« Significant other », 2020), et donc en protégeant son anonymat. *OP* désigne la personne ou le message qui initie un fil de discussion, et n'est donc pas pertinent dans le contexte de la messagerie instantanée ou des SMS. *TIL* est quant à lui spécifique à Reddit, et est dérivé du nom d'un subreddit ; il est généralement utilisé pour commencer une phrase, comme dans « TIL my father is a narcissist ».

Les acronymes recensés sont tous des acronymes non elliptiques (Mattiello, 2013), c'est-à-dire qu'ils ont conservé toutes les initiales des mots sources. Ils sont tous constitués de 2 à 4 caractères ; sur ce point, les acronymes du Netspeak semblent se comporter comme les acronymes issus d'autres domaines, comme ceux étudiés par Cannon (1989), qui a noté que les acronymes de son corpus font entre 1 et 5 lettres. Il semble y avoir une grande diversité dans les acronymes de la CMC en ce qui concerne la nature des syntagmes sources. Mattiello (2013) note que les acronymes sont généralement dérivés de syntagmes nominaux et adjectivaux, ce que l'on retrouve ici avec, par exemple, les acronymes de *opening post*, *significant other*, *not safe for work* et *private message*. Dans notre liste, il y a également des interjections (*Oh my god*, *what the fuck*), des syntagmes verbaux (*I don't know*, *fixed that for you*, *if I remember correctly*), et des syntagmes prépositionnels (*in real life*, *at the moment*). Certains d'entre eux peuvent même être utilisés seuls, comme dans les trois exemples suivants tirés du corpus :

> I don't know. YMMV. My experience might be different from yours.
>
> Emma Watson should have played her. FTFY
>
> You should definitely do this!!! Lol. jk.

En ce qui concerne les variantes typographiques, notre analyse montre que, dans l'ensemble, les graphies en minuscules sont à peu près aussi fréquentes que les graphies en majuscules. Le format typographique privilégié semble varier selon les acronymes. Par exemple, seules 8.90 % des occurrences de *lol* sont en majuscules uniquement, contre 67.49 % en minuscules uniquement. On remarque un phénomène similaire avec *tbh* (70.82 % des occurrences en minuscules, 14.99 % en majuscules). Les acronymes dérivés de syntagmes nominaux, comme *OP*, *PM*, *POC*, *SJW* et *MIL*, sont tous plus fréquents sous forme de majuscules. *TIL* est uniquement présent sous forme de majuscules. Il semble donc que la tendance à la réduction par l'utilisation de minuscules n'affecte pas tous les acronymes de la même façon. Notons, par ailleurs, que les morphologues considèrent généralement que le fait qu'un acronyme s'écrive en minuscules indique qu'il s'est lexicalisé.

Le mot *laser* est ainsi passé du statut d'acronyme à celui de nom (Plag, 2003). Cela ne semble pas être le cas ici, sauf peut-être pour *lol*, qui change parfois de catégorie grammaticale par l'ajout d'un morphème, devenant un nom ou un verbe :

> I go on 4chan just for the lols.
>
> I loled ! Thank god I'm alone in the office rn.

Le fait que, sur les 33 acronymes étudiés, seuls 4 soient moins fréquents que les formes longues correspondantes montre la place importante (et sans doute croissante) occupée par les acronymes dans la langue de Reddit. Dans le cas de certains acronymes, qui n'existent pas ou quasiment pas dans leurs formes complètes dans le corpus et qui sont nées dans la CMC (*OP*, *lol*, *lmao*, *NSFW*), cela n'a rien d'étonnant. Ce ne l'est peut-être pas non plus pour un acronyme comme *FYI*, qui est utilisé dans la langue parlée (on en relève 730 occurrences dans le TV Corpus (« The TV Corpus », p. d.), qui est composé de 75 000 épisodes de télévision). Dans d'autres cas, la prédominance des acronymes sur les formes longues est plus surprenante. Des expressions couramment utilisées en anglais (*to be honest*, *if I remember correctly*, *by the way*, *for what it's worth*) sont ainsi fortement concurrencées par leurs acronymes. Dans le corpus, *tbh*, *IIRC*, *btw* et *FWIW* sont ainsi respectivement utilisés 1.93, 3.67, 2.12 et 1.03 fois plus fréquemment que leurs formes longues. Il semble donc que, pour ces expressions, la forme abrégée soit devenue le standard sur Reddit.

L'outil « How the Internet* Talks » (Olson & King, 2017), qui permet d'obtenir la fréquence relative de mots dans tous les commentaires publiés sur Reddit entre 2008 et juillet 2017, offre une perspective diachronique sur cette question. La figure 11.10 montre la fréquence relative de cinq acronymes présents dans notre liste. On y voit que *FWIW* (en vert clair) a une fréquence stable depuis 2009 environ, et que *IIRC* (en vert foncé) gagne chaque année un peu de terrain. *TBH* (en bleu foncé) et *IDK* (en en bleu clair), par contraste, semblent avoir été très peu utilisés avant 2008, puis avoir connu une forte montée en popularité, allant jusqu'à dépasser *FWIW* et *IIRC*. Entre 2014 et 2017, la fréquence relative de *TBH* a ainsi été multipliée par 4, et celle de *IDK* a doublé. Nous avons également inclus au graphe *rn*, qui est relativement peu employé dans RedditGender (137 occurrences, contre 4156 pour la forme longue *right now*). Le graphique montre que l'acronyme a été davantage utilisé à partir de 2015. Il est donc possible que cette montée se poursuive aujourd'hui, et que, comme *TBH*, *rn* soit en train de progressivement remplacer la forme longue qui lui a donné naissance.

11.7.2 Réductions

Fréquence et types de réductions

Avec seulement 5828 occurrences recensées, dont 19 types différents, les réductions (*clippings*) sont moins fréquentes que les acronymes dans RedditGender. Il semble donc que, sur Reddit, ce procédé soit moins productif

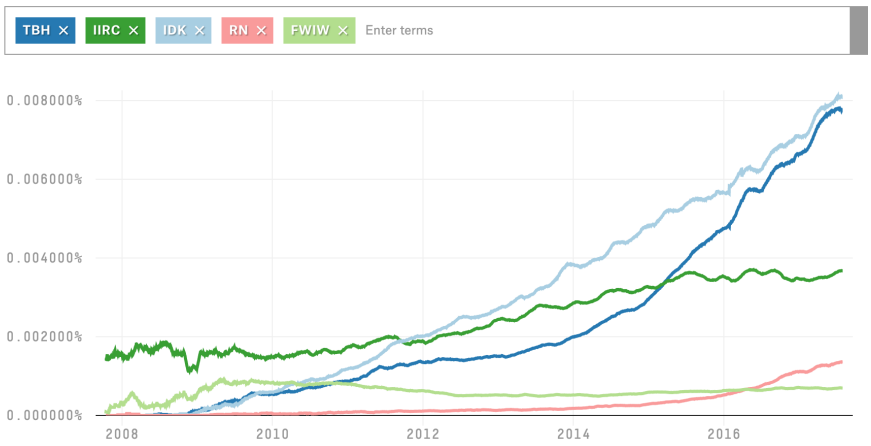

FIGURE 11.10 – Fréquence relative de cinq acronymes sur Reddit, de 2008 à juillet 2017, (Olson & King, 2017)

que l'acronymie. Dans la liste des réductions étudiées, plusieurs thèmes se dégagent, qui reflètent sans doute les préoccupations des Redditors du corpus : la technologie et internet (*ig*, *fb*, *mic*, *dev*), les relations personnelles (*bf*, *gf*), le corps et le genre (*masc*, *fem*, *clit*).

Dans certains cas, les formes abrégées entrent en concurrence avec les formes complètes, mais pas dans des proportions similaires à ce que l'on a observé avec les acronymes. Seuls deux réductions sont plus fréquentes que leurs formes standard, et il s'agit de deux termes peu fréquents : *mic*, avec 120 réductions et 31 formes complètes, et *clit*, avec 123 réductions et 64 formes complètes. Il est possible que ces formes aient été utilisées par une poignée de Redditors, et que les résultats observés ne révèlent pas une tendance plus large. Les *odds ratios* sont généralement faibles (moyenne de 0.04 pour l'ensemble des réductions étudiées ici). Dans certains cas, ils dépassent les 0.20, ce qui pourrait indiquer un possible remplacement progressif des formes complètes par des formes abrégées. Les réductions qui semblent ainsi faire le plus concurrence aux formes standard sont, par ordre croissant, *ig*, *fb*, *bf* et surtout *gf* (odds ratio de 0.52). Notons que tous ces réductions sont composés de deux lettres seulement, et qu'il est possible que la longueur des réductions ait une influence sur leur popularité auprès des internautes.

La fréquence des formes standard semble également avoir une influence sur la probabilité d'une réduction ; l'observation des *odds ratio* des formes abrégées et formes standard montre par ailleurs que moins un mot est fréquent, moins il est susceptible d'être abrégé. L'*odds ratio* de la réduction de *girlfriend* (fréquence de la forme standard : 1836) est ainsi de 0.52, tandis que celui de *bc* par rapport à *because* est de 0.02 (fréquence de *because* : 45 942). Cela peut paraitre logique ; moins abréger les mots très fréquents garantit leur bonne compréhension. Pour les réductions, nous ne pouvons donc pas tirer les mêmes conclusions que pour les acronymes : le phénomène n'a pas la même ampleur sur Reddit, et il ne semble pas que les ré-

ductions fassent une réelle concurrence aux formes non abrégées.

Effets des variables sociales

Nous avions émis l'hypothèque que, à cause du caractère spécifique à Reddit de certains acronymes, l'âge Reddit serait corrélé avec leur fréquence. Notre analyse de l'ensemble des abréviations (acronymes et réductions) montre que ce n'est pas le cas. L'étude de Tagliamonte et Denis (2008), réalisée auprès d'adolescents et de jeunes adultes, suggère que la fréquence de *lol* diminue avec l'âge. Cougnon et François (2010) ont remarqué le même phénomène dans leur corpus de SMS. Notre analyse montre que l'effet de l'âge est limité ; dans l'analyse de l'ensemble du corpus, on le remarque chez les hommes cisgenres, ainsi que de façon partielle chez les femmes cisgenres et transgenres, mais on ne le constate pas chez les personnes non binaires. Cette corrélation est également partielle (les plus de 31 ans emploient moins d'abréviations que les moins de 20 ans, mais pas moins que les 21-30 ans) dans l'analyse de l'échantillon réduit, qui prend en compte l'ethnicité. L'effet significatif de l'interaction de l'âge et du genre peut peut-être expliquer en partie le fait que les études de la CMC qui se sont intéressées aux abréviations ont eu des résultats contrastés, trouvant soit que les femmes en utilisent plus que les hommes (Baron, 2004 ; Cougnon & François, 2010 ; Herring & Zelenkauskaite, 2009), soit qu'il n'y a pas de différence significative entre les groupes (Peersman et al., 2016), soit que les hommes en utilisent davantage (Tagliamonte, 2016a, dans le cas de *lol*).

L'intégration de l'interaction du genre et de l'âge montre que l'effet du genre varie selon les groupes : les hommes cisgenres les plus jeunes emploient davantage d'abréviations que les femmes cisgenres, mais chez le groupe le plus âgé, c'est l'inverse. Les femmes transgenres de 14-20 ans s'alignent sur les hommes cisgenres, et les femmes transgenres les plus âgées sur les femmes cisgenres. Dans le groupe du milieu, il n'y a pas de différence entre femmes et hommes ; seul le groupe non-binaire se démarque par une utilisation limitée des abréviations. Quand l'ethnicité est prise en compte, les différences entre femmes et hommes cisgenres s'effacent. Il se dessine en revanche une démarcation ethnique : les Afro-Américain·es et les Hispaniques sont des utilisateur·trices d'abréviations plus prolifiques que les blanc·hes et les Asiatiques. Ce résultat semble confirmer ce que les rares études quantitatives de l'ethnicité et des graphies non standard suggèrent (Bamman et al., 2014 ; Eisenstein et al., 2011) : les Afro-Américain·es sont la source de nombreux acronymes et d'innovations lexicales. Il est possible que les Hispaniques jouent également un rôle de précurseur·es en matière d'abréviations, ou qu'ils et elles adoptent les usages des Afro-Américain·es, comme cela se produit parfois à l'oral (Fought, 2003 ; Santa Ana & Bayley, 2004).

11.7.3 G-droppings

Des cinq phénomènes étudiés dans ce chapitre, les g-droppings sont les plus rares. On en relève seulement 1662 occurrences (339 types différents),

et plus de moitié des Redditors n'en ont pas utilisé un seul. De plus, les g-droppings semblent être souvent liés à *fucking* et à ses synonymes (plus d'1 g-dropping sur 3), ou à des structures relativement figées (*just sayin'*, ou *lookin' good*/*great*/*dapper*), comme l'avait déjà noté Davies (2015) (→ p. 84). Les g-droppings sont le seul phénomène du Netspeak étudié dans cette thèse qui n'est pas corrélé avec l'âge. Si, en général (chez les hommes cisgenres, tout du moins, comme nous le voyons plus détail plus loin, → p. 318), les graphies non standard de la CMC sont plus fréquemment utilisées par les internautes les plus jeunes, les g-droppings semblent échapper à ce principe. Les jeunes internautes n'emploient donc pas davantage de g-droppings que les plus âgé·es. Dans notre analyse du corpus entier, les g-droppings sont plus fréquemment utilisés par les hommes cisgenres que par les autres groupes. Dans le groupe non-binaire, il n'y a pas de différence entre AFAN et AGAN. Quand l'ethnicité entre en jeu, le contraste entre femmes et hommes est uniquement présent dans le groupe hispanique : une fois de plus, cette variable atténue les différences entre femmes et hommes cisgenres. L'effet de l'ethnicité est assez similaire chez les hommes et les femmes. Nous ne remarquons pas la différence entre les groupes blanc et afro-américain à laquelle nous nous attendions, au vu de l'étude de Eisenstein (2015), qui a montré que la suppression du *g* est moins fréquente dans les comtés principalement habités par des blanc·hes, et plus fréquente dans les comtés où vit une importante population afro-américaine. En revanche nous notons, chez les hommes comme les femmes, une distanciation des Asiatiques par rapport à ce procédé, peut-être à cause de ses associations avec un anglais relâché, avec la classe populaire, et avec les Afro-Américain·es (→ p. 82).

11.7.4 Graphies phonétiques

La comparaison des graphies phonétiques à leurs graphies standard montre que certains mots et groupes de mots sont significativement plus susceptibles que d'autres de générer des graphies phonétiques. Dans certains cas, les graphies phonétiques semblent plus fréquentes que les graphies standard (*gotta*). Dans d'autres, elles ne le sont pas, mais elles leur font fortement concurrence : c'est par exemple le cas de *kinda* ou de *gonna*, qui sont peut-être en train de devenir le « standard » de la CMC. La plupart des graphies étudiées, toutefois, restent relativement rares par rapport à leurs formes standard.

L'analyse inférentielle révèle une corrélation négative et significative entre l'âge et la fréquence des graphies phonétiques chez les groupes cisgenres (même si elle est limitée chez les femmes). L'âge n'a en revanche pas effet chez les groupes transgenres et non binaire. L'effet du genre varie selon les groupes d'âge : dans les deux groupes les plus jeunes, les hommes cisgenres emploient davantage de graphies phonétiques que les femmes cisgenres. Cette démarcation disparait chez les plus âgés, où seuls les hommes transgenres se distinguent pas une utilisation plus faible des graphies phonétiques que les hommes cisgenres. Dans l'analyse du groupe non-binaire,

on ne remarque pas d'effet du genre assigné à la naissance : les AGAN n'utilisent pas davantage de graphies phonétiques que les AFAN.

Quand l'ethnicité est prise en compte, l'effet du genre n'existe pas chez les blanc·hes, mais il est présent dans les autres groupes ethniques, où il semble être un marqueur masculin. Le modèle ne révèle pas d'effet de l'ethnicité chez les femmes ; chez les hommes, il y a une démarcation entre les Afro-Américains, d'un côté, et les blancs et les Asiatiques de l'autre, qui utilisent moins de graphies phonétiques. Plusieurs graphies étudiées sont en effet typiques de l'anglais afro-américain de la CMC : la substitution de *d* à *th* à l'initiale de mots comme *this* (*dis*), *though* (*doe*) ou *that* (*dat*) (Eisenstein et al., 2011 ; Florini, 2014), ou *hella*, un terme d'argot qui tire son origine de l'anglais vernaculaire afro-américain du nord de la Californie Bucholtz (2012). Il est donc possible que les graphies phonétiques indexent plus particulièrement une masculinité afro-américaine, de laquelle les femmes et les hommes blanc·hes et asiatiques se distancient. Il se peut aussi que certaines graphies phonétiques nées dans les pratiques des Afro-Américain·es aient été adoptées par les internautes hispaniques, mais n'aient pas été encore appropriées par les autres groupes ethniques.

11.7.5 Omission d'apostrophe

Types et fréquence des omissions

Nos résultats montrent que certaines contractions sont davantage susceptibles que d'autres de déclencher des omissions d'apostrophe. C'est notamment le cas de *let's* et de *it's*, dans lesquelles l'apostrophe disparait plus fréquemment que pour *I'm* ou *can't*. Cela peut être dû au fait que *let's* et *it's* ont toutes les deux des homographes (*lets*, troisième personne singulier du verbe let au présent, et *its*, pronom possessif de la troisième personne du singulier), ce qui crée des ambiguïtés. Le problème de la confusion entre *it's* et *its*, et de l'omission des apostrophes possessives en général, est déjà bien connu (Connors & Lunsford, 1988 ; Hokanson & Kemp, 2013). Il est également possible que, comme elles ont des homographes, les omissions d'apostrophe de *it's* et *let's* soient moins fréquemment repérées par la frappe prédictive que les autres omissions, qui produisent des graphies n'existant pas dans l'anglais écrit standard, comme *arent*, *isnt*, ou *youre*.

Dans notre corpus, l'apostrophe est omise dans 4.25 % des cas où elle devrait apparaitre selon les normes de l'anglais standard (au moins, puisque nous n'avons pas étudié l'omission de l'apostrophe possessive). À notre connaissance, il n'existe pas d'autre étude récente de ce phénomène, et il est difficile d'en mesurer la magnitude. Nous pouvons simplement constater que l'on est très loin des résultats de Squires (2012), qui a trouvé que l'apostrophe était omise dans 43 % des cas dans son corpus de messagerie instantanée, et qui en conclut que l'omission d'apostrophe est standard dans la CMC. Toutefois, son corpus a été créé en 2004, avant l'apparition des smartphones et la généralisation de la frappe prédictive. Il est fortement possible que celle-ci limite le nombre d'omissions d'apostrophe, dans le cas des messages écrits sur Reddit depuis un smartphone. Notons

que l'usage de l'apostrophe fait polémique dans le monde anglophone (Lukač, 2014), et qu'elle semble de moins en moins utilisée. L'apostrophe en effet est relativement récente en anglais : elle est apparue au 16ème siècle sous l'influence du français, mais n'est devenue standard qu'au 18ème siècle pour le possessif, et au 19ème siècle pour les possessifs pluriels (Little, 1986). *The Cambridge Guide of English Usage* 2004 précise que l'apostrophe n'est plus obligatoire dans plusieurs cas, comme les noms pluriels indiquant l'affiliant (*teachers college*), les nombres et dates (*in his 60s*, *during the 1980s*), les noms de lieux (*Kings cross*), ou encore les noms de marques (*McDonalds*). L'apostrophe a même été interdite dans les noms de rue par le Mid-Devon District Council en 2013, pour éviter les confusions (Lukač, 2014).

Il est fortement possible que les écrits informels des réseaux sociaux, des SMS et de la messagerie instantanée aient contribué à l'essor de graphies sans apostrophe. Toutefois, la frappe prédictive semble ralentir cette disparition annoncée, et, sur Reddit tout du moins, l'omission ne semble pas être devenue le « standard ». Pour avoir une perspective diachronique sur le phénomène, nous avons recherché la fréquence de *thats*, *cant* et *ive* (les trois omissions les plus fréquentes de RedditGender à ne pas avoir d'homographe, contrairement à *its* et *im*) sur le Reddit Ngram (Olson & King, 2017). À titre de comparaison, nous avons inclus l'acronyme *idk*, qui est à peu près aussi fréquent dans notre corpus que *thats*. le graphique généré (figure 11.11) montre que la fréquence de ces trois omissions d'apostrophe augmente jusqu'en 2012. Entre 2012 et 2017, la fréquence reste stable, tandis que celle de *idk* a été multipliée par (environ) 3 ou 4. Reddit a lancé sa première interface mobile en 2010 (Siegler, 2010), mais, à l'époque, l'usage des smartphones ne s'était pas encore généralisé aux États-Unis : 35 % des Américain·es avaient un smartphone en 2011 (Center & Inquiries, 2019). Il est donc possible que la technologie (la frappe prédictive des smartphones) ait stoppé l'essor des omissions d'apostrophe.

Analyse sociolinguistique

Même si l'omission d'apostrophe n'est pas devenue un standard de la CMC informelle, certains internautes y semblent attachés. Comme nous en avions fait l'hypothèse et comme Squires (2007) l'a montré, les hommes cisgenres omettent plus fréquemment l'apostrophe que les femmes cisgenres. Il n'y a pas de contraste dans les groupes transgenres : les femmes transgenres, par exemple, n'omettent pas l'apostrophe moins fréquemment que les hommes cisgenres ou transgenres. L'âge a également un impact sur l'omission d'apostrophe, ce à quoi nous nous attendions également, même si la corrélation n'est que partielle : les plus de 31 ans omettent moins fréquemment l'apostrophe que les deux groupes plus jeunes.

L'interaction entre genre et ethnicité semble particulièrement pertinente dans l'étude des omissions d'apostrophe. Quand on l'inclut au modèle, l'effet de l'âge disparait, et l'effet du genre est plus limité : il n'y a ainsi pas de différence significative entre femmes et hommes dans les groupes blanc et hispanique. En revanche, dans les groupes asiatique et afro-américain, les hommes produisent significativement plus d'omissions que les femmes,

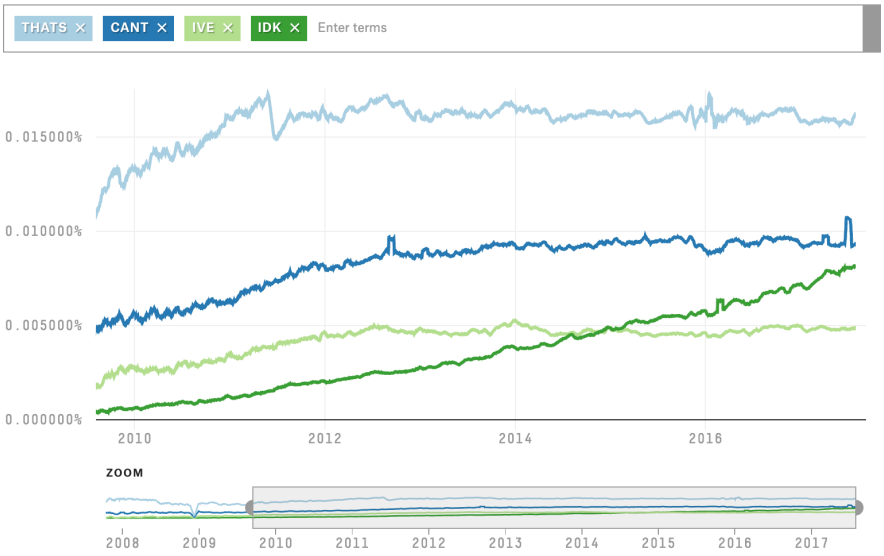

FIGURE 11.11 – Fréquence relative de *thats*, *cant*, *ive* et *idk* sur Reddit, de 2010 à juillet 2017 (Olson & King, 2017)

avec une taille d'effet forte chez le groupe asiatique (les hommes omettent 6 fois plus fréquemment l'apostrophe que les femmes). Les femmes asiatiques sont également le groupe qui utilise le moins d'omissions d'apostrophe quand on le compare aux autres groupes de femmes, avec des tailles d'effet importantes (les Asiatiques omettent 6.28 fois moins d'apostrophes que les femmes afro-américaines). Chez les hommes, c'est le groupe blanc qui se démarque, en omettant significativement moins d'apostrophes que les Asiatiques et les Afro-Américains. Ces résultats montrent que l'omission d'apostrophe n'est pas un marqueur de genre dans tous les groupes. Ils suggèrent aussi que certains groupes semblent se distancier de l'omission d'apostrophe (les femmes asiatiques, notamment), peut-être à cause de son caractère non standard et rebelle.

11.7.6 Omission de la majuscule du *i*

Comme Tagliamonte et Denis (2008) l'ont montré, il y a de fortes différences individuelles dans le choix d'omettre la majuscule du pronom personnel *I*. Plus d'un tiers des Redditors (35.73 %) n'ont pas utilisé ce procédé une seule fois ; un petit groupe d'internautes (24 personnes) utilise aussi ou plus fréquemment *i* que *I*. L'analyse inférentielle, qui n'a pas retenu l'interaction du genre et de l'âge, révèle qu'il n'y a pas de différence entre les deux groupes d'internautes les plus âgé·es. Les moins de 20 ans, en revanche, se démarquent des Redditors plus âgé·es en omettant davantage la majuscule de *I*. L'effet du genre est significatif, avec une démarcation entre, d'un côté, les femmes cisgenres, les hommes transgenres et les personnes non binaires, et, de l'autre, les femmes transgenres et les hommes cisgenres. Ces deux derniers groupes omettent plus fréquemment la majuscule que

les autres internautes, avec une taille d'effet importante : les hommes cisgenres préfèrent *i* à *I* deux fois plus fréquemment que les femmes cisgenres, et les femmes transgenres omettent la majuscule trois fois plus fréquemment que les femmes cisgenres. L'analyse de l'ethnicité (sans interaction avec le genre) montre également la prédilection des jeunes Redditors et des hommes cisgenres pour cette graphie, et met en évidence un contraste ethnique : les internautes blanc·hes omettent moins fréquemment la majuscule que les Afro-Américain·es.

L'omission de la majuscule du *i* pourrait être due au fait que certains internautes utilisent leur ordinateur (où la frappe prédictive n'est pas proposée par défaut) et non leur smartphone (où elle l'est, et corrige donc automatiquement les *i* en *I*). Toutefois, au vu de nos résultats, il semble que l'omission de la majuscule soit intentionnelle : on voit difficilement pourquoi les hommes et les femmes cisgenres, les Afro-Américain·es et les moins de 20 ans utiliseraient davantage leur ordinateur que les autres internautes. Pour une Redditor du corpus, l'omission des majuscules (du *i*, mais aussi en début de mots) est un choix revendiqué. Répondant à la remarque d'un·e autre internaute, elle rétorque (nous traduisons) : « J'ai désactivé les majuscules automatiques sur mon téléphone. Qu'est-ce que ça peut te faire ? ». Notons qu'elle n'utilise pas non plus d'apostrophes, mais que ses messages ne sont pas pour autant courts ni n'ont un style télégraphique. Au contraire : elle emploie également de nombreux procédés d'ajout (notamment des mots en majuscules et des étirements de ponctuation). L'omission des apostrophes et des majuscules fait donc partie intégrante de son « style » d'écriture.

L'omission de la majuscule du *i* semble ainsi être une forme de désinvolture, voire même de rébellion envers les normes orthographiques, et pas uniquement un simple raccourci, résultat d'un besoin de rapidité. Elle indexe l'ethnicité (les Afro-Américain·es l'embrassent, les blanc·hes l'évitent), et l'âge (les jeunes l'adoptent) ; pour le genre, c'est plus compliqué : si l'omission de majuscule indexe la masculinité, pourquoi les femmes transgenres l'utilisent-elles trois plus fréquemment que les femmes cisgenres ? Et si on considère l'omission de majuscule comme une transgression adoptée par les femmes transgenres, pourquoi ne constate-t-on pas le même phénomène chez les hommes transgenres et les personnes non binaires ? Est-il possible que le fait que de nombreuses femmes transgenres ont été socialisées comme des garçons ait un impact sur ce choix stylistique, comme nous en avions fait l'hypothèse pour l'étude des centres d'intérêt, (→ p. 203) ? En tout cas, et comme nous le verrons plus en détail dans notre synthèse (p. → 313), nos résultats montrent qu'il ne faut pas rechercher forcément un alignement des femmes (ou des hommes) transgenres sur les femmes (ou les hommes) cisgenres dans leurs pratiques d'écriture en ligne.

tl ;dr

Contrairement à la plupart des procédés d'ajout, les procédés de réduction sont davantage utilisés par les hommes cisgenres que par les femmes cisgenres, même si, pour les graphies phonétiques et les abréviations, les différences ne sont pas significatives dans tous les groupes d'âge, et si les femmes de plus de 31 ans utilisent davantage d'abréviations que les hommes du même âge.
Les hommes transgenres et les personnes non binaires adoptent une position « neutre », en ne se démarquant ni des hommes cisgenres, ni des femmes cisgenres. Les femmes cisgenres s'alignent sur les hommes cisgenres pour l'omission de la majuscule du *I*, et sur les femmes cisgenres (pour les plus de 31 ans) pour les abréviations.
L'intégration de l'ethnicité aux analyses tempère, comme c'était le cas pour les procédés d'ajout, les différences entre femmes et hommes, à l'exception du *i*. Les différences s'effacent pour les abréviations pour tous les groupes. Elles disparaissent également complètement dans le groupe blanc pour les 3 variables étudiées avec l'interaction. Elles subsistent dans 2 cas sur 3 chez les Asiatiques et les Afro-américain·es.
L'ethnicité n'a pas le même effet chez les hommes et les femmes cisgenres. On remarque une démarcation, chez ces dernières, entre les groupes hispanique et afro-américain et le groupe asiatique, qui utilise peu les procédés de réduction. Cette démarcation est moins prononcée chez les hommes.

Chapitre 12

Synthèse des résultats et discussion

Ce chapitre présente une synthèse des analyses des 11 variables linguistiques étudiées dans les chapitres précédents. Il est structuré en quatre sections. La première compare les usages des femmes et des hommes cisgenres, la seconde est consacrée à ceux des personnes transgenres et non binaire, la troisième examine les effets de l'âge et de son interaction avec le genre, et la dernière porte sur l'interaction du genre et de l'ethnicité.

12.1 Femmes et hommes cisgenres

12.1.1 Synthèse des résultats

Le tableau 12.1 synthétise l'ensemble des résultats pour les 11 variables étudiées, uniquement pour les groupes cisgenres. Il montre les variables les plus utilisées par les femmes et par les hommes, ainsi que celles pour lesquelles aucune différence significative n'a été trouvée. La deuxième ligne du tableau présente ces résultats pour l'ensemble des femmes et des hommes, quel que soit leur âge (parce que l'interaction entre âge et genre n'a pas été retenue dans le processus de sélection des variables, ou parce qu'elle n'était pas significative pour ces deux groupes). Les trois lignes du bas montrent les résultats par groupe d'âge. La taille d'effet est indiquée entre parenthèses. Par exemple, dans la première ligne du tableau, on voit que les femmes utilisent davantage d'émoticônes que les hommes. La taille d'effet est de 1.81, ce qui signifie que quand un homme utilise 1 émoticône, une femme en utilise 1.81. Les procédés d'ajout ont été mis en gras ; les procédés de réduction sont en graisse normale.

Les variables qui sont les moins sensibles à la variation de genre sont deux procédés typographiques : les mots en majuscules et les émojis (que seules les femmes de plus de 31 ans utilisent plus fréquemment que les hommes). La taille d'effet est très élevée dans ce cas (4.41), mais il ne faut pas oublier que les émojis ont été utilisés par une minorité de Redditors, et que leur utilisation est caractérisée par de fortes différences idiosyn-

Tableau 12.1 – Variables les plus fréquemment utilisées par les femme et les hommes cisgenres

Groupes d'âge	Femmes	Hommes	Pas de différence significative
Tous	**Émoticônes** (1.81) **Étirements de ponctuation** (1.57)	G-droppings (1.59) *i* minuscule (2.13) Omissions d'apostrophe (1.34)	***All caps***
14-20 ans	-	Abréviations (1.38) Graphies phonétiques (1.81)	**Étirements** de **lettres** **Interjections** **Émojis** *Tout le Netspeak*
21-30 ans	**Étirements de lettres** (1.36) **Interjections** (1.26)	Graphies phonétiques (1.32)	Abréviations **Émojis** *Tout le Netspeak*
31 ans et +	**Étirements de lettres** (1.62) **Émojis** (4.41) **Interjections** (1.52) Abréviations (1.19) *Tout le Netspeak*	-	Graphies phonétiques

crasiques (comme l'a montré la présence de deux « très gros » utilisateurs dans le corpus). On note également que l'effet de l'âge nuance les différences entre femmes et hommes dans le cas des étirements de lettres, des interjections, des abréviations et des graphies phonétiques. L'interaction du genre avec l'âge montre que c'est chez les plus jeunes qu'il y a le moins de différences significatives (2 différences significatives pour 5 variables), et que c'est chez les plus âgé·es que la variation de genre est la plus forte (4 variables sur 5).

Les trois lignes inférieures du tableau révèlent par ailleurs que les adolescentes et les très jeunes femmes n'utilisent aucune variable plus fréquemment que les adolescents et les très jeunes hommes. Inversement, les hommes de plus de 31 ans n'utilisent aucune variable plus fréquemment que les femmes du même âge. Dans le groupe des 21-30 ans, les différences s'équilibrent, avec 2 variables plus fréquemment utilisées par les femmes (étirements de lettres et interjections), et une plus fréquemment utilisée par les hommes (graphies phonétiques). Nous explorons plus en détail l'interaction du genre et de l'âge dans la section 12.3. Enfin, ce tableau rend clairement visible le constat que nous avons fait dans les deux chapitres précédents : les femmes utilisent les procédés d'ajout plus fréquemment que les hommes, et les hommes emploient les procédés de réduction plus fréquemment que les femmes (à l'exception des abréviations, et uniquement pour le groupe le plus âgé).

12.1.2 Discussion

Différences et similitudes

Avant de tenter d'apporter des interprétations aux différences constatées dans nos analyses, il est important de signaler que nous n'avons pas remarqué uniquement des différences. Sans interaction avec l'âge, il y a une majorité de différences significatives, c'est vrai (pour 5 variables sur 6). Pour les 5 variables où l'interaction avec l'âge a été prise en compte, il y a 2 différences significatives chez les 14-20 ans, 3 chez les 21-30, et 4 chez les 31 ans et plus, soit un total de 9 différences constatées sur 15 comparaisons effectuées. Si on considère l'ensemble des procédés étudiés (sans faire de différence entre les variables, « *Tout le Netspeak* » dans le tableau), on constate uniquement une différence chez les Redditors les plus âgé·es, les hommes de plus de 31 ans utilisant globalement le Netspeak moins fréquemment que les femmes du même âge. Notons également que, chez les hommes comme chez les femmes cisgenres (cela vaut aussi pour les personnes transgenres et non binaires), les procédés du Netspeak sont assez rares. Ils ont toujours une distribution fortement asymétrique (de Zipf), avec quelques utilisateur·trices très prolifiques, et de nombreuses personnes qui l'utilisent très peu, ou pas du tout (dans le cas des émojis ou du *i* minuscule, par exemple). Il n'y a pas de cas où une variable est fréquemment utilisée par de nombreuses femmes, et peu fréquemment utilisée par de nombreux hommes.

La question du standard et de la norme

Tous les procédés que nous avons analysés peuvent être considérés comme « non standard », par rapport à la norme de l'anglais écrit des médias ou de l'école. Les résultats étant contrastés (les femmes ont utilisé plus fréquemment certaines variables, les hommes d'autres, et il n'y a pas ou peu de différences dans d'autres cas), on ne peut pas dire que les femmes utilisent davantage de procédés non standard que les hommes, ou inversement. Les résultats globaux, sur l'ensemble des procédés, le montrent bien : seuls les hommes de plus de 31 ans utilisent moins de « Netspeak » que les femmes de plus de 31 ans.

Dans deux cas, nous avons été capables d'explorer, dans la lignée des études variationnistes sur la langue orale, le choix entre une variante standard et une variante non standard d'une même forme (pour l'omission de l'apostrophe de *it's* et l'omission de la majuscule de *I*). Les modèles réalisés montrent que les hommes sont plus susceptibles que les femmes d'utiliser les variantes non standard *its* et *i*. Pour les internautes, ces formes, qui vont à l'encontre de ce qui est enseigné à l'école, indexent peut-être une identité rebelle et nonchalante, qui est mieux tolérée chez les garçons et les hommes que chez les filles et les femmes (Eckert & McConnell-Ginet, 2003). On pourrait donc comprendre pourquoi celles-ci se distancient de ces usages. Il en va de même pour les g-droppings, plus fréquemment utilisés par les hommes, qui sont associés (dans la langue orale, mais sans doute aussi à l'écrit) avec l'insouciance et la décontraction.

Des cultures séparées ?

Nous avons vu, dans notre étude des centres d'intérêt, que les parcours des femmes et des hommes sur Reddit sont différents. Les femmes de notre corpus fréquentent par exemple deux fois plus les forums consacrés aux thèmes personnels que les hommes, qui se rendent plus volontiers les subreddits dédiés au gaming, à la technologie, à l'actualité et aux sujets « x-rated ». Cette mobilité plus importante pourrait conférer aux hommes une meilleure connaissance de la plateforme, et une meilleure maitrise de ses règles explicites (ce qui est illustré par le fait qu'il y a plus de modérateurs que de modératrices) et implicites (le code langagier de Reddit, qui réprouve par exemple l'utilisation d'étirements graphiques, comme nous l'expliquons p. 259).

Les différences de centres d'intérêt (même si, soulignons-le, il y a aussi de fortes similitudes dans, par exemple la fréquentation des subreddits dédiés à la science et l'éducation, aux thèmes généraux, à l'humour ou à la culture populaire) pourraient également expliquer les différences langagières. Il est fort possible que les points de contact entre femmes et hommes sur Reddit soient relativement peu nombreux, c'est-à-dire que les femmes fréquentent principalement des forums fréquentés par des femmes, et les hommes des forums fréquentés par des hommes. Les subreddits r/AskWomen et r/AskMen sont de bons exemples de cette séparation (même si, bien entendu, certaines femmes commentent sur r/AskMen, et que certains hommes commentent sur r/AskWomen). Si on fait cette hypothèse, il est possible que les femmes adoptent peu le style « sobre », qualifié de « clinical » par un·e Redditor (dandeeo, 2017) préféré par les Redditors hommes, et que ceux derniers imitent peu le style « féminin » embrassé par de nombreuses femmes, et sans doute influencé par leur pratique sur d'autres plateformes et technologies comme Instagram, les SMS, et la messagerie instantanée.

La domination masculine

Par de nombreux aspects, dont sa culture geek et sa composition démographique asymétrique, Reddit est une plateforme dominée par les hommes. L'histoire du site (→ p. 99) montre que sa misogynie peut aller jusqu'au harcèlement, aux atteintes à la vie privée, à la haine, et à l'effacement pur et simple des femmes. Un commentaire de la fondatrice du subreddit r/AskWomen illustre parfaitement ce problème. Elle explique, d'un côté, la difficulté de créer un espace dédié aux femmes sur Reddit, des « power user men » s'accaparant souvent des subreddits créées par et pour elles. Elle souligne aussi que, dans les subreddits généraux (comme r/askreddit), il est difficile d'exister en tant que femme. Elle raconte qu'avant la création de r/AskWomen, de nombreux Redditors pensaient tout simplement qu'il n'y avait pas de femmes sur Reddit ; même si dit-elle, une femme écrivait un post sur les mérites respectifs de Brad Pitt et de Johnny Depp, « they were more likely to assume you were gay than a woman » (Impudence, 2017). Ce constat l'a poussée à lancer r/AskWomen :

> « So this space was created. We existed. A lot of people *still* didn't believe we were women and a lot of people *still didn't believe our opinions mattered*. On anything. At all. [...] The message is gender matters, but only if that gender is male. If it's female, it should be silent and unnoticed. » (Impudence, 2017)

Dans un espace où il n'y a pas de photographies de profil et où les utilisateur·trices sont anonymes, le langage est le seul outil (avec les pseudonymes) que les internautes ont à leur disposition pour exprimer leur identité. Utiliser des émoticônes et des étirements graphiques, procédés à la connotation féminine et jouissant d'une mauvaise réputation sur le site (→ p. 261), ne nait peut-être pas uniquement d'un désir d'expressivité, mais aussi d'un besoin de s'affirmer, de dire « je suis une femme et je suis sur Reddit ». Ce ne serait donc peut-être pas un hasard si les procédés privilégiés par les femmes sont des procédés d'ajout, qui sont très visibles – par opposition aux procédés « minimalistes » comme les omissions d'apostrophes et de majuscule. Ces procédés sont peut-être pour les femmes une façon d'exercer, consciemment ou non, leur « contre-pouvoir linguistique » (Bailly, 2008).

12.2 Personnes transgenres et non binaires

12.2.1 Synthèse des résultats

Pour synthétiser nos résultats sur l'ensemble des groupes de genre, nous avons classé ces groupes en trois catégories dans les tableaux 12.2 et 12.3. La catégorie de gauche contient le(s) groupe(s) avec le(s)quel(s) les différentes variables linguistiques sont associées le plus fortement ; c'est-à-dire, le ou les groupes qui les utilisent plus fréquemment qu'un ou que plusieurs autres groupes. La catégorie de droite présente le(s) groupe(s) qui sont le moins fortement associé(s) avec une variable donnée, c'est-à-dire qui utilisent cette variable significativement moins fréquemment qu'un ou plusieurs groupes.

Dans la catégorie du milieu, nous avons placé le(s) groupe(s) qui, soit :

— utilise(nt) moins fréquemment une variable donnée que le groupe de gauche, et plus fréquemment que le groupe de droite.
— ne présente(nt) pas de différence significative avec le groupe de droite et/ou de gauche. Quand c'est le cas, nous le précisons.

En d'autres termes, le groupe du milieu ne se distingue pas par une utilisation particulièrement élevée ou faible d'un procédé linguistique, par rapport aux autres groupes. Nous avons utilisé une couleur pour chaque groupe, afin de permettre une visualisation plus rapide des tendances qui se dégagent de ces tableaux. Notons, enfin, que nous présentons ici uniquement les cas pour lesquels il y a une différence significative entre femmes et hommes cisgenres. Cela permet aux tableaux d'être plus lisibles, et aide à voir comment les groupes transgenres et non binaire se situent par rapport aux personnes cisgenres. Le premier tableau (12.2) présente les résultats dans les cas où l'interaction entre genre et âge n'a pas été prise en compte.

Le second (12.3) montre les résultats pour chaque groupe d'âge, quand il y a interaction.

Tableau 12.2 – Variables « genrées », sans interaction avec l'âge

Groupe(s) qui les utilisent plus fréquemment	Groupes du « milieu »	Groupe(s) qui les utilisent moins fréquemment
	Étirements de ponctuation	
Femmes cisgenres (+ que tous les autres)	**Hommes cisgenres** **Hommes transgenres** **Non-binaires** (- que les femmes cis)	**Femmes transgenres** (- que tous les autres)
	Émoticônes	
Femmes transgenres (+ que tous les autres)	**Femmes cisgenres** **Hommes transgenres** **Non-binaires** (- que les femmes trans)	**Hommes cisgenres** (- que les tous les autres)
	Omissions d'apostrophe	
Hommes cisgenres	**Femmes transgenres** **Hommes transgenres** **Non-binaires** (pas de diff. significative avec les hommes et femmes cis)	**Femmes cisgenres** (- que les hommes cis)
	i minuscule	
Hommes cisgenres **Femmes transgenres**	**Hommes transgenres** **Non-binaires** (pas diff. significative avec les groupes de gauche et droite)	**Femmes cisgenres** (- que les hommes cis et femmes trans)
	G-droppings	
Hommes cisgenres (+ que tous les autres)	**Femmes cisgenres** **Femmes transgenres** **Hommes transgenres** **Non-binaires**	-

Quand il n'y a pas d'interaction (tableau 12.2), deux tendances se profilent. Tout d'abord, hommes transgenres et personnes non binaires sont toujours dans le même groupe, et ce groupe est toujours celui du milieu. Notons que cela ne signifie pas forcément que ces Redditors utilisent une variable plus ou moins fréquemment que le groupe de gauche ou de droite ; ils produisent ainsi autant d'émoticônes que les femmes cisgenres, et autant d'omissions d'apostrophes et *i* minuscules que les hommes cisgenres. Le comportement des femmes transgenres est différent pour 3 variables sur 5, pour lesquelles elles sont soit le groupe (ou font partie du groupe) qui utilise le plus une variable (émoticônes, par rapport à tous les autres Redditors ; *i* minuscule et omissions d'apostrophes, par rapport aux femmes cisgenres),

soit celles qui utilisent le moins une variable (étirements de ponctuation).

Quand il y a interaction avec l'âge (tableau 12.3), le constat est le même pour les hommes transgenres et les non-binaires, mais il est plus nuancé pour les femmes transgenres. Il n'y a jamais de différence significative entre hommes transgenres et personnes non binaires (dans le cas des interjections, ils et elles ne sont pas dans le même groupe parce que les hommes transgenres utilisent davantage ce procédé que les hommes cisgenres, ce qui n'est pas le cas des non-binaires, mais il n'y a pas de différence dans leur production d'interjections). Dans 6 cas sur 9, hommes transgenres et personnes non binaires sont dans le groupe « du milieu ».

Les femmes transgenres de moins de 30 ans se classent elles aussi dans le groupe le plus indistinct dans 5 cas sur 6. Dans le groupe des 31 ans et plus, cela change : dans 3 cas sur 4, elles se démarquent de 1 ou plusieurs groupes. Elles s'alignent sur les femmes cisgenres à deux reprises (pour abréviations et interjections), et sur les hommes cisgenres pour les étirements de lettres. On note également que les femmes transgenres, quand elles se distinguent des personnes non binaires et des hommes transgenres, se comportent de façon imprévisible. Parfois, elles utilisent plus fréquemment une variable « féminine » que les hommes cisgenres (émoticônes) ; parfois, elles utilisent moins fréquemment une variable « féminine » que tous les autres groupes (étirements de ponctuation et de lettres) ; parfois, elles utilisent une variable « masculine » plus fréquemment que les femmes cisgenres (*i* minuscule) ; dans d'autres cas enfin, elles s'alignent sur les femmes cisgenres (pour les interjections et abréviations).

12.2.2 Discussion

Personnes non binaires et hommes transgenres : un parallèle presque parfait

La plupart de nos analyses mettent en évidence un parallèle frappant entre les personnes non binaires et les hommes transgenres. Dans le chapitre 8, nous avons vu que ces deux groupes ont des centres d'intérêt similaires sur Reddit, qu'ils écrivent des messages plus longs que les autres groupes de genre, et que leur mobilité sur le site est limitée. Les résultats de nos analyses linguistiques confirment ce parallèle. Ils positionnent les non-binaires et les hommes transgenres dans un groupe « médian », qui n'est caractérisé que de façon exceptionnelle par un usage faible ou élevé d'une variable par rapport à un ou à plusieurs autres groupes.

Personnes non binaires Dans le cas des personnes non binaires, ce résultat peut paraitre peu étonnant. Ce groupe défie une vision normative du genre. Comme il est hétérogène (il réunit en fait plusieurs identités de genre, comme *genderqueer*, agenre ou *genderfluid*), il est difficile de dire exactement comment cela se traduit sur le plan linguistique. Il est possible qu'une personne agenre n'utilise pas les ressources linguistiques de la CMC de la même façon qu'une personne *genderfluid* pour construire son identité de genre sur Reddit. Nos données ne fournissent qu'une vue d'ensemble de

Tableau 12.3 – Variables « genrées », avec interaction avec l'âge

Groupe(s) qui les utilisent plus fréquemment	Groupes du « milieu »	Groupe(s) qui les utilisent moins fréquemment
14-20 ans		
Abréviations		
Hommes cisgenres	**Femmes transgenres** (pas de diff. significative avec gauche et droite)	**Femmes cisgenres** **Hommes transgenres** **Non-binaires**
Graphies phonétiques		
Hommes cisgenres (+ que tous les autres)	**Femmes cisgenres** **Femmes transgenres** **Hommes transgenres** **Non-binaires**	-
21-30 ans		
Graphies phonétiques		
Hommes cisgenres	**Femmes transgenres** **Hommes transgenres** **Non-binaires** (pas de diff. significative avec gauche et droite)	**Femmes cisgenres**
Étirements de lettres		
Femmes cisgenres	**Femmes transgenres** (pas de diff. significative)	**Hommes cisgenres** **Hommes transgenres** **Non-binaires** (- que femmes cis)
Interjections		
Femmes cisgenres	**Femmes transgenres** **Hommes transgenres** **Non-binaires** (pas de diff. significative)	**Hommes cisgenres**
31 ans et +		
Émojis		
-	**Femmes cisgenres** **Femmes transgenres** **Hommes transgenres** **Non-binaires**	**Hommes cisgenres**
Abréviations		
Femmes cisgenres **Femmes transgenres**	**Hommes transgenres** **Non-binaires** (pas de diff. significative avec gauche et droite)	**Hommes cisgenres**
Étirements de lettres		
Femmes cisgenres	**Hommes transgenres** **Non-binaires** (pas de diff. significative)	**Femmes transgenres** **Hommes cisgenres**
Interjections		
Femmes cisgenres **Femmes transgenres** **Hommes transgenres**	**Non-binaires** (pas de diff. significative avec gauche ou droite)	**Hommes cisgenres**

la non-binarité ; de plus, nous manquons encore d'études sociolinguistiques de la non-binarité, qualitatives comme quantitatives, auxquelles nous pourrions comparer nos résultats. Toutefois, les analyses que nous avons réalisées sur le groupe non binaire uniquement, qui visaient à savoir si le genre assigné à la naissance avait un impact sur la fréquence d'utilisation de certaines variables « genrées » (les émoticônes → p. 223, les étirements de lettres → p. 236 et de ponctuation → p. 241 et les g-droppings → p. 284), fournissent une information supplémentaire. Le genre assigné à la naissance ne semble avoir qu'une influence extrêmement limitée (uniquement pour les étirements graphiques, et seulement pour les internautes de 21 à 30 ans). Dans ce cas, les personnes AFAN se rapprochaient, par leur usage, des femmes cisgenres.

Cette position médiane des personnes binaires que nous constatons peut être rapprochée du travail sociophonétique qualitatif de Gratton (2016) sur la variable (ING) chez des personnes non binaires au Canada. Cette étude montre que celles-ci utilisent diverses stratégies pour indexer leur identité de genre. En fonction du contexte, elles optent pour la variante standard « féminine » ou pour la variante non standard « masculine » pour se distancier de leur genre assigné à la naissance. Ce choix est lié au concept de « safe-space », « where marginalized groups can feel secure expressing themselves without being subject to mainstream norms and stereotypes » (Gratton, 2016, p. 55). Dans un « safe space », le besoin de distanciation par rapport au genre assigné à la naissance est moins fort que dans l'espace public. Nous pourrions donc faire l'hypothèse que, pour les personnes non binaires, Reddit est un « safe space ». Ces internautes fréquentent essentiellement des subreddits consacrés aux thèmes transgenres, où ils et elles échangent avec d'autres personnes transgenres et non binaires. De plus, ils et elles le font de manière virtuelle, sans que leur corps ou leur identité « réelle » ne soit révélés. Présenter une identité non binaire (par la distanciation avec les variables connotées « féminines » et « masculines ») n'est peut-être pas très important sur ces espaces virtuels.

Hommes et femmes transgenres Comme les personnes non binaires, les hommes transgenres évitent les extrêmes. Il se peut que, encore une fois, cela soit lié au concept de « safe space » : les hommes transgenres se déplacent sur Reddit de façon limitée, et, sur les subreddits qu'ils fréquentent (où ils échangent avec des personnes qui partagent la même expérience qu'eux), il est possible que leur présentation de genre ne soit pas une priorité. Ils ne chercheraient donc pas spécialement à éviter les variables associées à la féminité, ou à « sur-utiliser » celles qui indexent la masculinité.

Nos résultats font écho à l'étude de Hazenberg (2015), qui a comparé l'utilisation des adverbes d'intensité *pretty* et *so* et la production phonétique du [s] chez 6 groupes de locuteurs : des hommes et des femmes cisgenres hétérosexuel·les, des hommes et des femmes cisgenres gays, et des hommes et des femmes transgenres. Il a remarqué que les hommes et les femmes cisgenres se servent de ces ressources linguistiques pour indexer

leur identité de genre. Les femmes et les hommes gays, par exemple, utilisent plus beaucoup plus fréquemment *so*, marqué « féminin » que *pretty*, associé à la masculinité ; les hommes cisgenres font l'inverse. Les hommes et les femmes transgenres sont dans une position intermédiaire, pour les deux variables. Cette position médiane est occupée, dans notre étude, par les personnes non binaires, les hommes transgenres, et souvent, mais pas toujours, par les femmes transgenres. Hazenberg avance plusieurs explications possibles. Il émet l'hypothèse qu'être « au milieu » est une façon pour les personnes transgenres de se distancier des personnes cisgenres ; dans notre corpus, on voit que cette distanciation ne s'effectue pas autant que chez Hazenberg. Parfois, les hommes transgenres se distancient des hommes cisgenres mais se comportent comme les femmes cisgenres (émoticônes, g-droppings) ; parfois c'est l'inverse (étirements de ponctuation, étirements de lettres pour les 21-30 ans). Dans une majorité des cas enfin, ils ne se distancient d'aucun groupe cisgenre (*i* minuscule, omissions d'apostrophe, et abréviations, étirements de lettres et interjections pour les 31 ans et plus, notamment). C'est dans ces cas qu'ils sont véritablement au milieu : il y a une différence significative entre femmes et hommes cisgenres, mais, pour les hommes transgenres, les différences de médianes et de moyennes avec les autres groupes, constatées lors de la phase exploratoire des analyses, n'ont pas été confirmées par les modèles de régression.

La seconde hypothèse de Hazenberg s'applique sans doute mieux à nos résultats. Il écrit qu'il est possible que les personnes transgenres fassent leurs choix linguistiques en fonction de ce qui est considéré comme « acceptable » de la part d'un homme ou d'une femme, c'est-à-dire, ce que l'on peut faire sans que son identité de genre ne soit remise en question. C'est ce que semblent faire les personnes transgenres, en grande partie, « choosing a path that is neither markedly feminine nor markedly masculine, but nevertheless falls within the acceptable ranges of both » (Hazenberg, 2015, p. 289). En ne se distanciant pas des groupes cisgenres, les personnes transgenres profitent quelque part d'une plus grande liberté. Il semble qu'employer à la fois fréquemment émoticônes et étirements de lettres (marqués « féminins ») et omissions d'apostrophe et *i* minuscules (marqués « masculins ») ne soit pas acceptable pour la majorité des hommes et des femmes cisgenres, peut-être parce que cela créerait une sorte de « dissonance » dans la présentation de genre. Pour les hommes transgenres, les personnes non binaires et, dans une moindre mesure, pour les femmes transgenres, c'est acceptable.

Cette interprétation peut être rapprochée de ce qui a été écrit sur les hijras, ces personnes AGAN qui adoptent une autre identité de genre, plus proche de la féminité (Hall & O'Donovan, 1996), et sur les travestis brésiliens, (Borba & Ostermann, 2007), qui n'utilisent pas forcément les ressources de la langue (le genre grammatical) pour indexer en permanence une identité féminine. Zimman (2017), qui a étudié la voix des hommes transgenres, parle de « stylistic bricolage » pour expliquer le fait que ceux-ci réincorporent parfois, après leur transition, des éléments « féminins » dans leur façon de parler ou de s'habiller qu'ils n'utilisaient pas avant leur tran-

sition. Le fait que leur voix soit perçue comme étant « masculine », grâce à l'effet de la testostérone leur suffisait ; pour beaucoup (et surtout pour les hommes transgenres gays), avoir une voix reflétant des stéréotypes hétéronormatifs n'était pas important.

Les femmes transgenres : un autre type de bricolage Le bricolage stylistique est également à l'œuvre chez les femmes transgenres, qui se comportent souvent, sur le plan linguistique, comme les hommes transgenres et les personnes non binaires. Toutefois, leur usage des variables du Netspeak est parfois surprenant. Tout d'abord, contrairement aux hommes transgenres, elles vont plus volontiers vers les extrêmes en se distanciant d'un groupe cisgenre. De plus, celui-ci n'est pas toujours le même. Pourquoi les femmes transgenres surutilisent-elles les émoticônes, marqueurs de féminínité, alors qu'elles sous-utilisent les étirements de ponctuation, autre marqueur féminin ? Pourquoi produisent-elles autant de *i* minuscules que les hommes cisgenres, alors que ce procédé est boudé par les femmes cisgenres ? Les femmes transgenres de plus de 31 ans semblent particulièrement susceptibles de « bricoler » avec les variables « genrées » ; elles produisent aussi peu d'étirements de lettres que les hommes cisgenres, mais autant d'interjections que les femmes cisgenres. Les étirements de lettres et de ponctuation auraient-ils une connotation que les femmes transgenres cherchent à éviter ?

Nos analyses de la Redditentité, des centres d'intérêt et de la mobilité éclairent peut-être ces résultats surprenants. Nous avons vu que les femmes transgenres adoptent beaucoup plus volontiers que tous les autres groupes des pseudonymes genrés (féminins, dans leur cas). De la même manière, elles font un usage excessif, par rapport aux autres groupes de genre, dont les femmes cisgenres, des émoticônes, qui indexent la féminité. D'un autre côté, les femmes transgenres se déplacent plus sur Reddit que les hommes cisgenres, qui eux-mêmes sont plus mobiles que les trois autres groupes. Les centres d'intérêt des femmes transgenres sont par ailleurs plus diversifiés que ceux des autres groupes, à l'exception des hommes cisgenres. Si on considère que les hommes transgenres et les personnes non binaires choisissent la sécurité sur Reddit, par les forums qu'ils fréquentent et leurs choix linguistiques, ce n'est pas le cas des femmes transgenres. C'est particulièrement intrigant parce que les femmes transgenres sont celles qui courent le plus de risques dans la vie « réelle » (à cause du harcèlement et de la discrimination auxquels elles sont confrontées) et sur Reddit (site qui peut être misogyne et haineux). Nos analyses montrent en tout cas que l'asymétrie entre l'expérience des femmes transgenres et celle des hommes transgenres dans la vie « réelle », soulignée par Schilt (2010) et Rankin et Beemyn (2012), et évoquée plus en détail dans notre chapitre sur la Redidentité, (→ p. 203), se reflète en ligne.

Autres explications Terminons par évoquer, enfin, d'autres hypothèses qui pourraient expliquer nos résultats sur les personnes transgenres. Il est possible que les personnes transgenres de notre corpus soient à différentes

étapes de leur processus de transition (même si elles ont toutes fait leur transition dans le monde virtuel). Dans ce cas, elles n'ont peut-être pas encore toutes acquis un répertoire linguistique genré :

> « We may be examining a snapshot of a linguistic system in flux, as individual speakers move from sounding 'masculine' to sounding 'feminine', or vice versa » (Hazenberg, 2015, p. 288).

Soulignons également que les pratiques langagières des personnes transgenres de notre corpus peuvent refléter des différences générationnelles entre les personnes transgenres qui ont fait leur transition très jeunes, comme c'est de plus en plus le cas en Amérique du Nord, et les personnes plus âgées, qui bien souvent ont mis plus de temps à s'identifier ou à vivre en tant que transgenres (Rankin & Beemyn, 2012).

Il se peut aussi que l'orientation sexuelle joue un rôle ; la moitié des hommes transgenres (51 %) et des femmes transgenres (48 %) de notre échantillon ne sont pas hétérosexuel·les (ils ou elles sont gays, bisexuel·les, asexuel·les ou pansexuel·les). La proportion de non-hétérosexuel·les est plus faible dans les groupes cisgenres (23.65 % pour les femmes et 21.50 % et pour les hommes dont nous connaissons l'orientation sexuelle). Enfin, notons qu'il ne nous a pas été possible d'étudier l'ethnicité des personnes transgenres (parce qu'il a été difficile de recueillir des informations à ce sujet). Il est toutefois fort probable que la majorité des personnes transgenres de notre échantillon soient blanches, Reddit étant un site geek et blanc. L'identité transgenre n'est évidemment pas monolithique, et est également façonnée par l'ethnicité (entre autres) ; nos résultats ne reflètent donc sans doute qu'une facette de l'identité transgenre en ligne.

12.3 Effet de l'âge et de son interaction avec le genre

12.3.1 Synthèse des résultats

Le tableau 12.4 indique l'absence ou la présence d'une corrélation entre âge et fréquence des variables du Netspeak, par groupe de genre (de gauche à droite, hommes cisgenres, femmes cisgenres, femmes transgenres, hommes transgenres et non binaire). Un signe moins (-) indique la présence d'une corrélation négative (la fréquence des variables diminue quand l'âge augmente). Un point (.) indique la présence d'une corrélation négative partielle (par exemple, quand les 14-20 ans utilisent davantage un procédé plus fréquemment que les plus de 31 ans, mais pas plus fréquemment que les 21-30 ans). Un x indique l'absence totale de corrélation. La partie supérieure du tableau présente les variables pour lesquelles l'interaction du genre et de l'âge a été intégrée au modèle, et la partie inférieure celles qui ont été analysées sans interaction. La dernière présente les résultats pour l'ensemble des variables, avec interaction.

Ce tableau nous permet de dresser plusieurs conclusions :

Tableau 12.4 – Présence de corrélations entre âge et fréquence des variables du Netspeak, par groupe de genre

	H cis	F cis	MTF	FTM	NB
Avec interaction du genre et de l'âge					
Émojis	-	x	x	x	x
Mots en majuscules	-	.	x	x	x
Étirements de lettres	-	x	.	x	x
Interjections	-	.	.	x	x
Abréviations	-	.	.	x	x
Graphies phonétiques	-	-	x	x	x
Sans interaction du genre et de l'âge					
Émoticônes	-	-	-	-	-
Étirements de ponctuation	.	.	.	.	.
G-droppings	x	x	x	x	x
Omission d'apostrophe	.	.	.	.	.
i minuscule	.	.	.	.	.
Tout le Netspeak	-	-	x	x	x

— Les procédés du Netspeak ne sont jamais corrélés positivement avec l'âge, même partiellement. Il n'y a donc pas de cas où la fréquence d'une variable augmente avec l'âge.
— Seuls les g-droppings ne sont jamais corrélés avec l'âge.
— Seules les émoticônes sont corrélées négativement avec l'âge pour tous les groupes.
— Pour l'ensemble des variables (« *Tout le Netspeak* »), on note une corrélation négative pour les groupes cisgenres, et une absence de corrélation pour les groupes transgenres.

Pour les 6 variables où nous avons étudié l'interaction entre genre et âge (partie supérieure du tableau), nous notons que :

— Toutes les variables sont corrélées négativement avec l'âge pour les hommes cisgenres.
— Il n'y a jamais de corrélation entre âge et fréquence des procédés du Netspeak pour les hommes transgenres et les personnes non binaires.
— Il n'y a aucune corrélation entre âge et fréquence dans 3 cas sur 6 pour les femmes transgenres, et une corrélation partielle dans 3 cas.
— Il y a une corrélation négative pour les femmes cisgenres, 3 corrélations négatives partielles, et 2 cas où il n'y a pas de corrélation.

12.3.2 Discussion

Il a été montré à de nombreuses reprises par les sociolinguistes que l'adolescence est une période de la vie fortement marquée par l'innovation linguistique ; les adolescent·es utilisent des formes non standard pour se différencier de leurs parents et enseignants, et pour se rapprocher des autres adolescent·es (Tagliamonte, 2016b). Même si notre corpus ne cap-

ture que de manière incomplète les usages des adolescent·es (à cause de la difficulté de trouver des Redditors de moins de 18 ans, → p. 117), nous nous attendions à constater des différences significatives entre les différents groupes : les adolescent·es et très jeunes adultes (14 à 20 ans), les jeunes adultes (21 à 30 ans), et les adultes plus âgés (31 ans et plus).

Quand on regarde la partie inférieure du tableau 12.4, on remarque que c'est le cas pour les émoticônes, dont la fréquence est corrélée négativement avec l'âge pour tous les groupes de genre. C'est aussi le cas, mais dans une moindre mesure (c'est-à-dire qu'il n'y a pas de différence significative entre tous les groupes d'âge) pour les étirements de ponctuation et les omissions d'apostrophes et de majuscules. Mais, quand on s'intéresse à la partie supérieure du tableau, c'est-à-dire lorsqu'il y a interaction entre genre et âge, les résultats sont contrastés. D'un côté, la situation est remarquablement homogène pour les hommes cisgenres : plus les hommes sont âgés, moins ils produisent de Netspeak. Les adolescents et les très jeunes hommes produisent plus d'émojis, de mots en majuscules, d'abréviations, etc. que les jeunes hommes, qui eux même en produisent plus que les hommes de plus de 31 ans. On remarque également une homogénéité parfaite, mais un effet différent, chez les hommes transgenres et les personnes non binaires. Ces deux groupes ont le même comportement vis-à-vis de chaque variable : l'âge, pour eux, n'est pas un facteur de variation. Les femmes transgenres et les femmes cisgenres se retrouvent « au milieu » de ces deux extrêmes : les premières sont plus proches, dans leur comportement, des personnes non binaires et des hommes transgenres (avec trois corrélations partielles), et les secondes sont plus proches des hommes cisgenres (avec trois corrélations « partielles » et une corrélation « totale »).

Nos résultats montrent que, comme l'a écrit Eckert, « gender is quite explicitly constructed partially in its interaction with age » (Eckert, 1998, p. 156). Les systèmes d'âge et les grandes étapes de la vie, dans les différentes sociétés, n'affectent en effet pas tous les individus de la même façon. Dans les sociétés occidentales, les hommes seraient contraints par un cycle de vie plus rigide. Les femmes bénéficieraient de davantage de fluidité (Eckert, 1998). Cette construction du genre en interaction avec l'âge (et avec d'autres variables, bien entendu) semble à l'œuvre dans notre corpus pour plus de la moitié des variables étudiées. Les hommes les plus âgés utilisent peu de graphies non standard, que les adolescents et les jeunes adultes emploient plus fréquemment. Les hommes les plus âgés s'alignent ainsi avec les usages normatifs de l'anglais écrit standard. Pour les femmes, en revanche, ce n'est pas forcément le cas.

C'est toutefois pour les groupes transgenres et non binaire que les résultats sont les plus étonnants, avec une absence d'effet de l'âge pour 6 variables pour les hommes transgenres et les non-binaires, et une absence d'effet de l'âge pour 4 variables (émojis, mots en majuscules, étirements de lettres et graphies phonétiques) pour les femmes transgenres. Si on regarde la dernière ligne du tableau 12.4, qui montre l'effet de l'âge sur l'ensemble des phénomènes étudiés, la démarcation entre groupes cisgenres (sur qui l'âge a un effet significatif) et les groupes transgenres et non bi-

naire (pour qui l'âge n'a pas d'effet significatif sur la production de graphies du Netspeak) est encore plus nette. On pourrait émettre l'hypothèse que le caractère transgressif de la transidentité (transgression des normes du genre) annule ou réduit fortement la corrélation négative attendue entre fréquence des transgressions linguistiques et âge (que l'on constate par contraste parfaitement chez les hommes) : l'usage des procédés non standard par les personnes transgenres ne semble pas affecté par leur âge. Il est également possible que les jeunes internautes transgenres et non binaires préfèrent ne pas utiliser fréquemment des procédés qui indexent la féminité ou la masculinité, contrairement aux jeunes femmes et hommes cisgenres.

12.4 L'ethnicité entre dans l'équation

12.4.1 Synthèse des résultats

Comparaison entre femmes et hommes de chaque groupe ethnique

Le tableau 12.5 présente, pour chaque groupe ethnique, la présence de différence significative entre femmes et hommes, avec interaction du genre et de l'ethnicité, et sans interaction. Quand les femmes utilisent davantage une variable, nous avons indiqué « F ». Un « H » représente les hommes. La taille d'effet est indiquée entre parenthèses. L'absence de différence significative est indiquée par un signe moins.

Tableau 12.5 – Différences significatives entre femmes et hommes, par groupe ethnique

Variables	Blancs	Afr.Am.	Asiatiques	Hispaniques
	Avec interaction			
Émoticônes	-	**F** (2.94)	-	**F** (2.69)
Émojis	-	**F** (3.76)	**F** (10.37)	-
Étirements de ponctuation	**F** (1.81)	**F** (2.31)	**F** (2.44)	-
G-droppings	-	-	-	**H** (4.92)
Interjections	**F** (1.52)	-	**F** (1.40)	-
Graphies phonétiques	-	**H** (2.06)	**H** (1.62)	**H** (1.55)
Omissions d'apostrophe	-	**H** (2.06)	**H** (6.14)	-
Mots en majuscules	-	-	-	-
	Sans interaction			
i	**H** (2.25)			
Étirements de lettres	**F** 21-30 ans et 31 ans et + (1.41 ; 1.71)			
Abréviations	-			

Quand l'ethnicité est prise en compte, les différences entre les femmes et les hommes sont plus nuancées encore que dans le tableau présenté dans la section 12.1. Dans ce tableau, qui synthétisait des analyses réalisées sur l'ensemble du corpus, il y avait une différence significative entre femmes et hommes (quel que soit leur âge) pour 5 variables : émoticônes et étirements

de ponctuation (plus fréquemment utilisés par les femmes) et g-droppings, *i* minuscule et omissions d'apostrophes (plus fréquemment utilisés par les hommes). Si on écarte le *i* minuscule, pour lequel l'interaction du genre et l'ethnicité n'a pas été intégrée et qui reste plus fréquemment employé par les hommes, il n'y a aucun cas où toutes les femmes utilisent ces 4 variables plus que tous les hommes, ou vice versa : il y a une différence significative dans 2 groupes pour les émoticônes (afro-américain et hispanique), dans 2 groupes pour les émojis (afro-américain et asiatique), et dans 2 groupes pour les omissions d'apostrophes (afro-américains et asiatiques). Pour le cas des g-droppings, qui étaient utilisés plus fréquemment par les hommes dans l'analyse du corpus entier, le constat est particulièrement frappant : dans l'échantillon réduit, ce sont uniquement les hommes hispaniques qui emploient cette variable davantage que les femmes (hispaniques), avec une taille d'effet forte (4.92).

Pour les 8 variables où il y a interaction entre genre et ethnicité, nous remarquons une différence significative entre femmes et hommes dans près de la moitié des cas (15 cas sur 32). Les femmes sont les utilisatrices les plus prolifiques dans 9 cas ; elles sont devancées par les hommes dans 6 cas. C'est dans les groupes afro-américain et asiatique qu'il y a le plus de variation de genre, avec respectivement 6 et 5 différences significatives. C'est dans le groupe blanc qu'il y a le moins de différences, avec seulement 2 résultats significatifs. De plus, les tailles d'effet sont plus faibles dans le groupe blanc que dans les autres groupes (inférieures à 2, alors que les tailles d'effet dépassent 2 dans 10 cas sur 13 dans les autres groupes). On remarque par ailleurs que les hommes blancs n'utilisent jamais (quand il y a interaction) une variable plus fréquemment que les femmes blanches, alors que les hommes des autres groupes utilisent tous 2 variables plus fréquemment que les femmes.

Comparaisons entre groupes ethniques : femmes

Les tableaux ci-dessous ont été réalisés sur le même principe que les tableaux présentant les différences entre les cinq groupes de genre (→ p. 311). Ils classent les groupes ethniques en trois catégories, en fonction de la présence (ou de l'absence) de différence significative. Pour les trois variables où l'interaction entre genre et ethnicité n'a pas été prise en compte (tableau 12.6), les blanc·hes et les Asiatiques ne se démarquent jamais par une utilisation plus fréquente d'une variable, par rapport aux groupes hispanique et afro-américain. Les groupe afro-américain utilise 2 variables sur 3 fréquemment que les Asiatiques, et plus fréquemment que les blanc·hes dans deux cas (abréviations et *i* minuscule). Une première démarcation ethnique se dessine.

Le tableau 12.7, consacré aux femmes (pour les 5 variables où il y avait une différence entre les groupes ethniques), montre une démarcation forte entre femmes hispaniques et afro-américaines, qui ont souvent un comportement similaire, d'une part, et femmes asiatiques d'autre part. Les premières font généralement partie du groupe qui utilise une variable le plus fréquemment. Les secondes utilisent moins fréquemment que les autres

Tableau 12.6 – Effet d'ethnicité, sans interaction, femmes et hommes

Groupe(s) qui les utilisent plus fréquemment	Groupes du « milieu »	Groupe(s) qui les utilisent moins fréquemment
	Étirements de lettres	
-	**Blancs** **Afro-Américain·es** **Hispaniques**	**Asiatiques**
	Abréviations	
Afro-Américain·es **Hispaniques**	-	**Blancs** **Asiatiques**
	i minuscule	
Afro-Américain·es	**Hispaniques** **Asiatiques** (pas de diff. sign.)	**Blancs**

Tableau 12.7 – Effet de l'ethnicité, femmes cisgenres

Groupe(s) qui les utilisent plus fréquemment	Groupes du « milieu »	Groupe(s) qui les utilisent moins fréquemment
	Émoticônes	
Hispaniques **Afro-américaines**	**Blanches** (pas de diff. sign. avec gauche et droite)	**Asiatiques**
	Émojis	
Hispaniques **Afro-américaines**	**Asiatiques** (pas de diff. sign.)	**Blanches**
	Étirements de ponctuation	
Hispaniques **Afro-américaines**	**Blanches** (pas de diff. sign.)	**Asiatiques**
	Omissions d'apostrophes	
-	**Blanches** **Hispaniques** **Afro-américaines**	**Asiatiques**
	G-droppings	
Blanches **Afro-américaines**	**Hispaniques** (pas de diff. avec blanches et asiatiques, mais diff avec afro-américaines)	**Asiatiques**

(mais attention, pas forcément moins fréquemment que *toutes* les autres) 4 variables sur 5. Les femmes asiatiques utilisent ainsi moins fréquemment que les femmes afro-américaines, et de façon significative, ces 4 variables. Dans 3 cas, elles les utilisent moins fréquemment que les femmes hispaniques. Les femmes blanches sont généralement dans une position intermédiaire. Dans deux cas (émoticônes et émojis), elles ne présentent de différence significative avec aucun des groupes ; dans un cas (omissions d'apostrophe), elles se démarquent des femmes asiatiques. Dans un cas (émojis), elles se démarquent des femmes hispaniques et afro-américaines, et dans un cas (g-droppings), elles se distinguent des femmes asiatiques.

Comparaisons entre groupes ethniques : hommes

Avec interaction, il y a davantage de différences significatives entre les groupes d'hommes qu'entre les groupes de femmes : on en compte 8, contre 5 chez les femmes, ce qui signifie que l'ethnicité a un effet sur toutes les variables étudiées pour les hommes. Des tendances similaires à celles notées chez les femmes apparaissent. Tout d'abord, les hommes afro-américains et hispaniques sont souvent regroupés ; dans 6 cas sur 8, il n'y a pas de différence significative entre ces deux groupes. Ensuite, Afro-Américains et Hispaniques se démarquent généralement des Asiatiques. Les hommes asiatiques utilisent moins fréquemment que les hommes afro-américains, et de façon significative, 6 variables sur 8. Ils utilisent 5 variables moins fréquemment que les Hispaniques. Ils ne sont jamais ceux qui utilisent une variable plus qu'un ou plusieurs autres groupes. Les hommes blancs sont les utilisateurs les plus prolifiques d'une seule variable : les émoticônes. Pour 3 variables, ils font partie de ceux qui se distinguent par une utilisation limitée, par rapport aux autres groupes. Tout comme dans le cas des femmes blanches, il ne semble pas y avoir de tendance claire chez les hommes blancs, contrairement aux autres groupes.

On remarque par ailleurs des comportements différents entre femmes et hommes d'un même groupe, quand on les compare aux autres groupes ethniques. C'est le cas, notamment, pour les émoticônes. Chez les femmes, les groupes hispanique et afro-américain se distinguaient du groupe asiatique par leur utilisation plus fréquente de cette variable. Il n'y avait pas de différence entre le groupe blanc et les groupes hispanique et afro-américain. Chez les hommes, c'est le groupe afro-américain qui se différencie du groupe blanc par une production moins importante d'émoticônes. Dans d'autres cas, on remarque des tendances similaires d'un tableau à l'autre. Pour les étirements de ponctuation, il n'y a pas de différence significative entre hommes hispaniques, afro-américains et blancs. Le groupe asiatique utilise cette variable moins fréquemment que tous les autres groupes. Chez les femmes, c'est la même chose, à une différence près : il n'y a pas de différence entre blanches et Asiatiques.

Tableau 12.8 – Effet de l'ethnicité, hommes cisgenres

Groupe(s) qui les utilisent plus fréquemment	Groupes du « milieu »	Groupe(s) qui les utilisent moins fréquemment
	Émoticônes	
Blancs	**Hispaniques** **Asiatiques** (pas de diff. sign.)	**Afro-américains**
	Émojis	
-	**Blancs** **Hispaniques** **Afro-américains**	**Asiatiques**
	All caps	
Hispaniques	**Blancs** **Asiatiques** **Afro-Américains**	-
	Étirements de ponctuation	
-	**Blancs** **Afro-Américains** **Hispaniques**	**Asiatiques**
	Omissions d'apostrophes	
Afro-Américains	**Hispaniques** (pas de diff) **Asiatiques** (+ que les blancs, - que Afro-Am.)	**Blancs**
	Graphies phonétiques	
Hispaniques **Afro-Américains**	**Asiatiques** (- que Afro-am., mais pas de diff. avec les autres groupes)	**Blancs**
	G-droppings	
-	**Blancs** **Afro-Américains** **Hispaniques**	**Asiatiques**
	Interjections	
Hispaniques	**Blancs** **Afro-Américains** (pas de diff. sign.)	**Asiatiques**

12.4.2 Discussion

De la pertinence de l'étude de l'interaction du genre et de l'ethnicité

Nos résultats montrent que l'interaction du genre et de l'ethnicité est particulièrement pertinente pour l'analyse des graphies non standard et procédés typographiques de la CMC. Sa contribution est double. Tout d'abord, elle nuance la variation de genre, parce que, à chaque fois que nous l'avons intégrée, nous n'avons pas remarqué de différence significative entre femmes et hommes de tous les groupes. Le cas des g-droppings est sans doute le plus parlant : la différence significative entre femmes et hommes constatée dans nos premières analyses est uniquement confirmée dans le groupe hispanique (c'est-à-dire, chez une minorité d'internautes).

L'interaction entre genre et ethnicité montre qu'il n'y a parfois pas de différence significative entre les femmes et hommes de tous les groupes, mais cette interaction n'a jamais un effet inverse. Il n'y a pas de cas où une même variable est préférée par les femmes dans un groupe, et par les hommes dans un autre. Il semble donc que certaines variables de la CMC permettent d'indexer un style d'écriture « féminin » ou « masculin », en interaction avec l'ethnicité (mais aussi l'âge, l'orientation sexuelle, et sans doute d'autres variables).

Ensuite, l'interaction révèle que les différences entre femmes et hommes sont plus ou moins nombreuses ou fortes selon les groupes ethniques. C'est dans le groupe blanc, qui domine Reddit (en termes de composition démographique) que les différences sont les plus rares et les moins marquées. Elles concernent uniquement 4 variables sur 11 (en prenant l'ensemble des variables en compte). Les Afro-Américain·es et les Asiatiques (7 variables sur 11) utilisent davantage ces variables pour indexer des identités de genre, avec des différences non seulement significatives, mais importantes en termes de taille d'effet : les femmes asiatiques produisent 10 fois plus d'émojis que les hommes asiatiques, qui utilisent 6 fois plus d'omissions d'apostrophe qu'elles. Les femmes afro-américaines emploient près de 4 fois plus d'émoticônes que les hommes afro-américains, qui utilisent 2 fois plus de graphies phonétiques qu'elles.

Il ressort donc de nos résultats qu'une étude variationniste quantitative qui considérerait ces variables isolément, dans le contexte américain, ferait en partie fausse route. On ne peut pas prédire l'usage des femmes afro-américaines en additionnant l'effet du genre (la différence éventuelle entre les femmes et hommes) et l'effet de l'ethnicité (la différence éventuelle entre hommes blancs et hommes afro-américains). Intégrer les interactions dans les modèles de régression permet donc de mettre en lumière des résultats étonnants, que l'on ne pouvait pas anticiper ici à cause de la rareté des études quantitatives du genre et de l'ethnicité en ligne. C'est le cas des émoticônes : les hommes blancs en produisent plus fréquemment que tous les autres groupes, mais pas plus fréquemment que les femmes blanches. Chez les hommes, les Afro-Américains les utilisent moins fréquemment que les autres hommes (et aussi que les femmes afro-américaines). Chez

les femmes, ce sont les Asiatiques qui se démarquent par une production d'émoticônes plus faible que les Hispaniques et les Afro-Américaines.

L'anglais afro-américain, source d'innovations du Netspeak

Même si l'ethnicité est rarement prise en compte par les études quantitatives de la CMC (et que, quand elle l'est, c'est souvent indirectement, par les données de géolocalisation, → p. 54), plusieurs travaux (Bamman et al., 2014 ; Eisenstein et al., 2010 ; Eisenstein et al., 2011) suggèrent que bon nombre des innovations lexicales et orthographiques de la CMC viennent de l'anglais afro-américain. Nos résultats montrent qu'effectivement, les internautes afro-américain·es sont en tête de file dans l'utilisation de nombreuses variables. La prise en compte de l'interaction avec le genre apporte une précision : sur Reddit en tout cas, il semble que ce soit surtout les femmes afro-américaines qui soient les principales innovatrices. Quand il y a une différence entre elles et au moins un groupe d'autres femmes (c'est-à-dire, dans 8 cas sur 11), elles sont toujours les utilisatrices les plus prolifiques des procédés du Netspeak, avec les femmes hispaniques. Les femmes afro-américaines et hispaniques ne sont ainsi jamais celles qui utilisent une variable le moins fréquemment – alors que c'est le cas pour les hommes afro-américains, qui utilisent peu les émoticônes, et emploient les mots en majuscules moins fréquemment que les Hispaniques.

Le fait que l'anglais afro-américain soit une source d'innovation dans la langue d'internet est reconnu par de nombreux internautes, journalistes et blogueur·ses (ComingofFaith, 2016 ; Natalie, 2018 ; Rose, 2020 ; Tenbarge, 2020), mais aussi par des travaux universitaires (Florini, 2014 ; Jones, 2015). Internet et les réseaux sociaux ont en effet changé la donne pour l'anglais afro-américain. Avant, l'anglais afro-américain, un dialecte souvent dénigré par les institutions et l'école (→ p. 44), ne s'écrivait que très peu, et les Afro-Américain·es avaient peu d'opportunités de « lire » de l'anglais afro-américain. C. A. Thompson et al. (2004) ont ainsi montré que les productions écrites d'enfants afro-américain·es reflétaient peu leurs pratiques orales. Internet a permis le passage à l'écrit de l'anglais afro-américain, ainsi qu'une certaine standardisation de son orthographe ; la graphie *doe* représente ainsi la prononciation afro-américaine de *though* (Florini, 2014). Il est même possible, en utilisant Twitter, de cartographier les différents dialectes afro-américains (Jones, 2015).

Notons ici que l'innovation linguistique afro-américaine semble provenir en grande partie de « Black Twitter », que le mouvement Black Lives Matter a mis sur le devant de la scène, mais qui existe depuis au moins une décennie (Brock, 2012). Le terme « Black Twitter » désigne la forte présence afro-américaine sur le site de microblogage ; en 2013, 40 % des internautes afro-américain·es de 18 à 29 ans utilisaient Twitter contre 28 % des internautes blancs (Smith, 2014). Par les interactions qu'il permet et la taille courte des messages, Twitter a été décrit comme un espace propice à la pratique du *signifyin'*, une tradition orale afro-américaine qui repose sur l'humour et la virtuosité verbale, et qui utilise généralement l'anglais vernaculaire afro-américain. Sur Twitter, le *signifyin'* fait appel à de nom-

breux procédés non standard, comme les graphies phonétiques et l'argot, mais aussi sur des éléments paralinguistiques, comme les émoticônes, qui sont pour les internautes afro-américain·es une façon d'indexer leur identité ethnique (Florini, 2014). Certaines pratiques spécifiques aux femmes afro-américaines ont été « traduites » en Netspeak : c'est notamment le cas du « black girl clap », un procédé qui consiste à ponctuer chaque mot par un claquement de main, pour mettre de l'emphase. Sur internet, le geste est représenté par l'émoji 👏, de la façon suivante : « yes 👏 just 👏 like 👏 this 👏 » (Brown, 2016).

Ces pratiques linguistiques s'exportent ensuite en dehors de la communauté afro-américaine : elles sont remarquées et empruntées par la masse des internautes, qu'ils soient américain·es ou non (Ilbury, 2020 ; Jones, 2015). L'appropriation de l'anglais afro-américain par les Américain·es blancs n'est pas nouvelle (Cutler, 1999 ; Fix, 2010), mais elle n'a sans doute jamais été aussi visible. Grâce aux réseaux sociaux, on peut retracer l'itinéraire d'un mot ou d'une expression, et voir à quelle période il a été approprié par la masse des internautes. McCulloch (2019) estime ainsi que *bae* (une abréviation de *babe*, « Urban Dictionary », 2017), était principalement utilisé par les Afro-Américain·es jusqu'en 2014, date à laquelle où le mot apparait dans des médias en ligne non afro-américains. Soulignons également le caractère problématique de certains de ces emprunts, qui reviennent parfois à faire de la « digital blackface » (Parham, 2020), c'est-à-dire à s'approprier des pratiques ou des comportements afro-américains de façon stéréotypée et souvent (pas forcément intentionnellement) raciste.

La position des Hispaniques

Le rapprochement entre Hispaniques et Afro-américain·s (femmes et hommes, dans de nombreux cas) indique des pratiques linguistiques partagées, et sans doute une influence de l'anglais afro-américain sur les Hispaniques. Les études de la langue orale ont en effet montré que les populations latino adoptent souvent des caractéristiques de l'anglais afro-américain (Carter, 2013 ; Fought, 2003 ; Santa Ana & Bayley, 2004 ; Wolfram, 1974) ; une étude de la CMC suggère aussi de nombreux usages communs entre Hispaniques et Afro-Américain·es (Eisenstein et al., 2011), notamment dans l'utilisation d'acronymes comme *lmaoo*, *smfh* et *lml*. Il est toutefois possible que les internautes hispaniques influencent parfois les usages afro-américains ; c'est en tout cas ce que semblent suggérer les résultats d'Eisenstein et al. (2010), qui a retracé l'itinéraire géographiques de graphies non standard aux États-Unis. Il montre notamment comment l'acronyme *af* (*as fuck*), qui indique l'emphase, semble être apparu autour de 2009 à Los Angeles et Miami (deux villes à la forte population hispanique) pour se répandre en 2011 dans des régions à la forte présence afro-américaine. McCulloch (2019) ajoute que l'acronyme s'est ensuite étendu à la masse des internautes en 2014, date à laquelle il commence à apparaitre dans les titres de médias populaires en ligne.

La distanciation des Asiatiques

Dans 7 cas sur 11, pour les hommes comme pour les femmes, les Asiatiques sont plus conservateur·trices que les Afro-Américain·es. Il est donc possible que la faible adoption de certaines variables du Netspeak, entre dans le cadre d'une stratégie de distanciation par l'hypercorrection. Les études sociolinguistiques réalisées sur le « Asian English » (l'anglais des Asiatiques aux États-Unis) nous offrent des clés d'interprétation.

Il a en effet été montré par, entre autres, Bucholtz (2004) et Reyes (2005) que les Asiatiques américain·es puisent dans l'argot et l'anglais afro-américains pour construire des identités « cool » et urbaines, ou, au contraire, les évitent pour créer des identités plus en phase avec l'image stéréotypée de « minorité modèle » des Asiatiques. Si le Netspeak est une sorte « d'argot » du web, et si certains éléments, comme les acronymes, sont directement en lien avec les usages des Afro-Américain·es en ligne, le non-recours à ces éléments indexerait, pour les femmes asiatiques, une identité racialisée en opposition aux autres Américain·es non blanc·hes. Le même phénomène a été mis en évidence chez des lycéens blancs californiens, qui se décrivent comme des « nerds » (Bucholtz, 2001). Ils utilisent un anglais « superstandard », évitant l'argot et les formes de l'anglais afro-américain. Cela leur permet de construire leur identité d'« intellos » blancs par opposition aux Afro-Américain·es et autres lycéens blancs qui adoptent les innovations de la culture urbaine et « cool » afro-américaine. Dans notre corpus, il semble qu'une partie des femmes et des hommes asiatiques sont conscient·es des dimensions idéologiques attachées au Netspeak, qu'il s'agisse de son association aux Afro-Américain·es ou à un anglais relâché, et s'en détachent par leur usage superstandard de la CMC.

tl ;dr

En rassemblant les résultats des analyses présentées dans les deux chapitres précédents, de grandes tendances se dessinent. Les hommes cisgenres utilisent davantage les procédés de réduction que les femmes cisgenres, qui emploient davantage les procédés d'ajout qu'eux. Hommes transgenres et personnes non binaires semblent adopter une stratégie similaire, restant dans une position « médiane » qui ne se rapproche ou ne se distancie que rarement des femmes et des hommes cisgenres. Les femmes transgenres sont souvent dans la même position, mais elles s'en détachent parfois de façon inattendue, par une utilisation fréquente des variables marquées « féminines » ou « masculines ».

L'exploration de l'interaction entre genre et âge révèle qu'elle n'est pas la même dans tous les groupes. L'effet de l'âge se manifeste le plus clairement chez les hommes cisgenres, les plus jeunes utilisant significativement davantage de Netspeak que les plus âgés. Cet effet est présent, dans une moindre mesure, chez les femmes cisgenres et transgenres. Il est en revanche complètement absent chez les hommes transgenres et les personnes non binaires.

L'interaction du genre et de l'ethnicité permet de nuancer les différences constatées, dans l'étude du corpus entier, entre femmes et hommes cisgenres. Les différences s'effacent souvent complètement dans le groupe blanc. On remarque par ailleurs, dans plusieurs cas, une distanciation entre les internautes afro-américain·es et hispaniques d'un côté, et les internautes asiatiques de l'autre, ce qui suggère que l'utilisation ou la non-utilisation de Netspeak peut être une façon d'indexer non seulement le genre, mais aussi l'ethnicité.

Cinquième partie

Conclusion

Cette thèse s'est donné comme ambition de réaliser une étude intersectionnelle du genre et de l'utilisation de 11 procédés langagiers qui s'écartent de l'anglais écrit standard, et qui sont typiques de la langue d'internet. Nous avons choisi, pour créer notre corpus, le site communautaire américain Reddit, qui présentait l'avantage d'être librement accessible et de privilégier un style d'écriture informel. Le corpus, construit avec l'aide d'un ingénieur de recherche de l'ATILF, comprend 460 707 commentaires produits par 1044 personnes, majoritairement américaines, soit près de 20 millions de tokens.

RedditGender est un des rares grands corpus de CMC à contenir des annotations sociodémographiques aussi riches et « qualitatives », par opposition aux annotations obtenues, par exemple, en inférant l'identité de genre à partir des noms d'utilisateur·trices des internautes, ou l'ethnicité à partir d'informations de géolocalisation. Le choix de Reddit s'est avéré essentiel pour réussir ce pari. Grâce à sa liberté de ton, due au système du pseudonymat sous lequel le site fonctionne, et à son organisation en communautés thématiques nous avons pu cibler les internautes et tirer les données sociodémographiques directement dans le contenu de leurs commentaires. Conçu dans une perspective intersectionnelle, RedditGender nous a permis d'étudier les variables en interaction, même si nous n'avons pas réussi à recueillir des informations « complètes » sur tou·tes les Redditors. Il s'agit également, à notre connaissance, d'un des seuls grands corpus de CMC (voire peut-être le seul) à faire une part aussi importante (près d'un tiers du contenu et des contributeur·trices) aux internautes transgenres et non binaires.

Notre analyse des procédés langagiers a été précédée d'une exploration de la façon dont les internautes créent leur identité et investissent l'espace virtuel sur Reddit. Nous nous sommes principalement intéressée à l'effet du genre (pseudonymes, centres d'intérêt, profils supprimés), parfois en combinaison ou en interaction avec l'âge, l'ethnicité et des variables de la Reddidentité, comme l'âge Reddit ou le fait d'être modérateur·trice ou non (pour la mobilité, la longueur des commentaires, et le karma). Ces analyses confirment le statut de Reddit comme site principalement occupé et contrôlé (par le système de la modération) par les hommes (Massanari, 2017). Les hommes cisgenres y sont installés depuis plus longtemps que les femmes cisgenres, s'y déplacent davantage, fréquentent une plus large variété de forums, et occupent plus fréquemment la fonction de modérateurs. Dans RedditGender, les femmes cisgenres écrivent des messages plus longs que les hommes cisgenres, et écrivent davantage de commentaires sur des forums consacrés aux thèmes personnels ; les hommes parlent davantage de politique, d'actualité ou de jeux vidéo qu'elles. Les centres d'intérêt traditionnellement considérés comme « féminins » ou « masculins » se reflètent donc dans les pratiques des internautes de RedditGender. Le fonctionnement du site, qui privilégie les intérêts des hommes geeks, met les femmes à la marge de Reddit : comme il peut être dangereux pour une femme, dans une ville, de s'aventurer dans des ruelles sombres la nuit, il peut être risqué pour les femmes Redditors de faire des incursions dans le territoire des

hommes.

Évidemment, sur internet, les choses changent très vite. En 2020, Reddit, comme une large partie de la société américaine, s'est remis en question suite à la mort de George Floyd et aux émeutes qu'elle a déclenchées. La « domination » des hommes blancs sur le site a été mise en lumière par de nombreux·ses Redditors, qui réclament du changement. Les médias se sont fait l'écho du racisme et du harcèlement auxquels sont confronté·es les modératrices et les modérateur·trices afro-américain·es en particulier (Hussain, 2020). Les femmes Redditors s'organisent pour faire bannir les subreddits les plus toxiques avec, par exemple, la création du forum r/BanFemaleHateSubs. Il reste à voir si les actions des administrateurs de Reddit réussiront à protéger plus efficacement les femmes, les Afro-Américain·es et tou·tes les internautes qui sont la cible d'attaques. Une étude récente (Ribeiro et al., 2020) suggère toutefois que, sur Reddit, les communautés de la « Manosphere » se radicalisent, devenant de plus en plus violentes et misogynes.

Et pourtant, malgré la haine qui s'affiche dans de nombreux subreddits, Reddit est aussi un lieu où les personnes transgenres et non binaires ont créé des communautés dynamiques, qui viennent en aide à leurs membres, et qui sont relativement préservées : c'est un des nombreux paradoxes de Reddit. Ce sont ces communautés que fréquentent principalement les hommes transgenres et les personnes non binaires, qui s'aventurent peu en dehors des forums transgenres. Ces internautes écrivent des messages longs, dans un nombre réduit de forums, en lien avec des thèmes personnels. Les femmes transgenres, quant à elles, nous ont surprise : nous émettions l'hypothèse d'un possible alignement sur les femmes cisgenres ou sur les autres personnes transgenres, mais nos résultats ont révélé une autre réalité. Les femmes transgenres se déplacent beaucoup sur le site, davantage que les hommes cisgenres, et elles fréquentent des forums variés. De plus, ce sont elles qui choisissent le plus volontiers des pseudonymes genrés, devant tou·tes les autres Redditors. Reddit, malgré sa misogynie, semble donc pour elles un espace de liberté ; plus marginalisées que les hommes transgenres dans la vie « hors ligne », elles peuvent vivre dans le cyberespace en tant que femmes.

L'analyse des 11 variables linguistiques que nous avons choisies et classées en deux catégories esquisse un panorama des graphies et procédés non standard utilisés sur Reddit. Nous retenons, notamment, la place importante occupée par les abréviations, qui font concurrence aux formes standard longues, et la préférence des Redditors pour les émoticônes plutôt que pour les émojis. Même si ces 11 variables n'offrent qu'un aperçu partiel des pratiques langagières des internautes, leur analyse a révélé des tendances qui, dans certains cas, ont confirmé les résultats d'autres études, et qui, dans d'autres, leur apportent des précisions ou des nuances. Notre approche intersectionnelle offre ainsi un éclairage nouveau sur certains phénomènes qui, comme les émoticônes ou les émojis, ont été abondamment étudiés. Nous n'avons en effet jamais analysé le genre isolément ; grâce à la méthode de la régression multiple, nous avons étudié son effet, mais aussi

celui d'autres variables, et, dans de nombreux cas, celui de leur interaction.

Nous avons trouvé des différences significatives entre femmes et hommes, mais elles ont souvent été tempérées par l'interaction avec l'âge. Cette variable a l'effet le plus significatif et prononcé chez les hommes cisgenres, qui pourraient soit délaisser les procédés du Netspeak en vieillissant (comme notre étude n'est pas diachronique, nous ne le savons pas), soit, pour les plus âgés, ne jamais les avoir intégrés. Ce phénomène explique bien des différences significatives avec les femmes cisgenres, car l'âge n'a pas toujours cet effet chez elles. Sur Reddit, tout du moins, les hommes cisgenres semblent plus sujets à la pression de l'orthographe standard. Ils préfèrent également généralement les procédés de réduction, plus « sobres », qui s'inscrivent peut-être mieux dans le style langagier geek que les procédés d'ajout. Émoticônes, étirements graphiques, et, dans une moindre mesure, émojis et interjections, sont par contraste privilégiés par les femmes cisgenres.

En intégrant l'ethnicité des Redditors à notre analyse, nous avons pu nuancer ces résultats : le genre n'a pas le même effet dans tous les groupes ethniques, et les différences significatives entre femmes et hommes sont parfois limitées à 1 ou 2 groupes sur 4. Il semble que les femmes afro-américaines et hispaniques jouent un rôle de premier plan dans la diffusion (et peut-être dans l'invention) des formes du Netspeak. L'autre résultat marquant de nos analyses est la distanciation opérée par les internautes asiatiques qui utilisent, comparé·es aux groupes hispanique et afro-américain, peu fréquemment certaines variables non standard que nous avons étudiées. Est-il possible que les femmes asiatiques, en particulier, se distancient de certaines formes du Netspeak parce que celles-ci pourraient entrer dans la construction d'un type d'identité féminine dont elles souhaitent se dissocier, à cause de l'hypersexualisation dont elles font l'objet dans la culture patriarcale blanche américaine (Seethaler, 2013) ? Aux États-Unis, les Asiatiques sont considéré·es comme d'« éternels étranger·es », qui ne maitrisent pas l'anglais même si c'est leur langue maternelle (Tuan, 1998) ; le rejet de certaines formes de Netspeak et l'hypercorrection des Asiatiques pourraient-ils être liés à un désir d'affirmer leur américanité par la langue ?

Notre exploration de la non-conformité de genre a montré que les personnes transgenres s'alignent rarement sur les personnes cisgenres. La vue d'ensemble offerte par le chapitre 12 révèle que les hommes transgenres et les personnes non binaires utilisent les variables que nous avons examinées souvent de façon similaire. Ils occupent une position médiane dans les 5 groupes étudiés ; ils ne se distancient en général ni des hommes cisgenres, ni des femmes cisgenres, peut-être parce qu'ils veulent éviter un usage « genré » de la langue, ou parce qu'ils ne ressentent pas le besoin de marquer, par exemple, une identité masculine en omettant les apostrophes et les majuscules. Les femmes transgenres se démarquent parfois de façon frappante des groupes transgenres et des groupes cisgenres. Elles puisent à la fois dans les ressources langagières marquées comme « féminines » et « masculines », s'alignant sur les femmes cisgenres, ou se distinguant

d'elles. Elles « bricolent avec le style », pour reprendre l'expression de Zimman (2017), et construisent une identité de genre qui n'est pas normative. Chez les personnes transgenres, étonnamment, il y a rarement de corrélation entre fréquence des formes non standard et âge, soit parce que l'effet « standardisant » de l'âge n'a pas prise sur elles, ou parce qu'elles dévient relativement peu, lorsqu'elles sont adolescentes ou jeunes adultes, des usages standard.

On pourrait également se demander, au-delà du genre, de l'âge et de l'ethnicité, si l'adoption des formes de la CMC par les internautes signale leur adhésion aux valeurs de la communauté. Il faudrait toutefois sans doute parler de « communautés » plutôt que de « communauté ». Comme nous l'avons vu, les Redditors ont des territoires très différents sur Reddit, et le fonctionnement décentralisé du site permet à des communautés d'exister de façon parallèle, sans qu'elles ne se rencontrent jamais. Dans la galaxie de subreddits existants (1 million quand nous avons créé le corpus, plus de 2 millions en octobre 2020) coexistent des idéologies très différentes, ce qui peut avoir un lien avec les pratiques linguistiques qui s'y déploient. Il est probable que le style masculin soit plus en adhésion avec le style « geek » de Reddit, qualifié de « clinique » par un Redditor (→ p. 310). Le style adopté par de nombreuses femmes de RedditGender, et notamment par des femmes hispaniques et afro-américaines, caractérisé par l'utilisation d'émoticônes et d'étirements graphiques, refléterait d'autres pratiques et valeurs, peut-être importées d'autres communautés du web, comme Instagram ou Twitter. Ces valeurs et pratiques langagières sont aujourd'hui devenues, également, celles d'une partie de Reddit.

Notre thèse, évidemment, a des limites. Elle ne fournit qu'un aperçu incomplet des pratiques langagières en ligne. Elle ne prend pas en compte, à cause de l'immense complexité de la méthode, des interactions à trois termes (entre âge, ethnicité et identité de genre, par exemple). Elle n'a pas réussi à explorer l'interaction de l'ethnicité et de la non-conformité de genre, à cause de la difficulté posée par le recueil des données. Enfin, elle n'apporte pas d'éclairage qualitatif sur les usages des internautes. Elle ouvre toutefois de nombreuses perspectives. Par son annotation riche et sa grande taille, RedditGender peut se prêter à d'autres études quantitatives du genre, de l'âge et de l'ethnicité, mais aussi de l'orientation sexuelle et des catégories socioprofessionnelles des internautes. Même si les internautes américain·es sont majoritaires, il y a suffisamment d'individus canadien·nes et britanniques pour comparer des variétés d'anglais. On pourrait également adopter une perspective diachronique pour voir si (et comment), dans les mois qui suivent l'arrivée d'un·e Redditor sur le site, sa façon d'utiliser les formes du Netspeak change sous l'influence du « dialecte » utilisé sur Reddit.

Dans une perspective qualitative, on pourrait se servir des résultats de cette thèse pour choisir des internautes qui transgressent les normes du genre ou de l'orthographe, par exemple des hommes cisgenres qui utilisent beaucoup de procédés d'ajout, des femmes cisgenres qui emploient très fréquemment les procédés de réduction, ou des internautes qui utilisent énor-

mément ou très peu les procédés du Netspeak. On pourrait ensuite examiner comment ces internautes utilisent (ou non) le Netspeak pour indexer des stances (c'est-à-dire la façon dont ils se positionnent dans les interactions par leurs choix linguistiques) et construire des styles (c'est-à-dire la façon dont ils créent leur identité par le langage), comme Bucholtz l'a par exemple fait dans son étude de l'argot des immigrants mexicains aux États-Unis (2009), ou de la façon dont deux jeunes filles d'origine laotienne s'approprient ou non les formes de l'anglais afro-américain (2004). L'approche qualitative pourrait également faire émerger d'autres facettes des identités des Redditors, que les méthodes quantitatives que nous avons utilisées ne permettent pas d'identifier : les personnes qui écrivent sur Reddit pour partager leur expertise, pour raconter leur vie, pour faire des rencontres (virtuelles ou non), pour obtenir des conseils, pour « polluer » les fils de discussion par des interventions peu pertinentes, etc.

Enfin, notre choix de Reddit a montré que ce site, beaucoup moins étudié que Twitter, est d'une grande richesse pour l'étude sociolinguistique. Il est possible, en explorant ses communautés thématiques, d'étudier les pratiques, par exemple, des Américain·es asiatiques, délaissés par la recherche sociolinguistique (Wolfram & Schilling, 2016), et des Afro-Américains·es, mais aussi des personnes transgenres et non hétérosexuelles. La popularité grandissante de r/France, qui compte aujourd'hui plus de 400 000 membres (subredditstats, p. d.), et l'expansion de la sphère francophone sur Reddit offrent également des perspectives intéressantes pour la recherche sociolinguistique sur le français de la CMC. Notre méthodologie pourrait ainsi être utilisée pour étudier, de façon diachronique ou synchronique, les usages de l'écriture inclusive par les internautes, sur Reddit ou d'autres sites, ou de mesurer l'influence des emprunts anglais sur le français à l'heure d'internet.

Bibliographie

(I'm) just saying. (p. d.). Cambridge English Dictionary. Récupérée 30 mars 2020, à partir de https://dictionary.cambridge.org/us/dictionary/english/i-m-just-saying

About Twitter's APIs. (p. d.). Help Center. Récupérée 3 août 2020, à partir de https://help.twitter.com/en/rules-and-policies/twitter-api

Account status. (p. d.). Reddit Help. Récupérée 30 avril 2020, à partir de https://www.reddithelp.com/en/categories/account-status/my-account-was-suspended-violating-reddits-content-policy

Acronym definition & 3000+ acronyms list from a-z. (2019, septembre 6). 7 E S L. Récupérée 20 avril 2020, à partir de https://7esl.com/acronyms/

Adams, V. (1973). *An Introduction to Modern English word-formation*. London, Longman.

Admiral Lord Fisher to Churchill : OMG. (2012, août 7). The International Churchill Society. Récupérée 16 janvier 2020, à partir de https://winstonchurchill.org/publications/churchill-bulletin/bulletin-050-aug-2012/admiral-lord-fisher-to-churchill-omg/

Aebischer, V. & Forel, C. (Éd.). (1983). *Parlers masculins, parlers féminins ?* Paris, Delachaux-Niestlé.

Aijmer, K. (2009). Interjections in the COLT corpus. In *From Will to Well : Studies in Linguistics, Offered to Anne-Marie Simon-Vandenbergen* (p. 11-20). Academia Press.

Alex. (2016, juin 7). *R hurdle : In sqrt(diag(object$vcov)) : NaNs produced*. Stack Overflow. Récupérée 4 septembre 2020, à partir de https://stackoverflow.com/questions/37670236/r-hurdle-in-sqrtdiagobjectvcov-nans-produced

Alexa. (p. d.). *Top sites in the United States*. Alexa.com. Récupérée 13 mai 2017, à partir de https://www.alexa.com/topsites/countries/US

Alexis Ohanian - Wikipedia. (p. d.). Récupérée 18 mars 2020, à partir de https://en.wikipedia.org/wiki/Alexis_Ohanian

All caps. (p. d.). Cambridge Dictionary. Récupérée 21 juin 2020, à partir de https://dictionary.cambridge.org/dictionary/english/all-caps

Ameka, F. (1992). Interjections : the universal yet neglected part of speech. *Journal of Pragmatics*, *18*, 101-118.

Amg137. (2019). *A short-ish history of new features on reddit : Announcements*. Récupérée 20 janvier 2020, à partir de https://www.reddit.com/r/announcements/comments/84nyj6/a_shortish_history_of_new_features_on_reddit/

Androutsopoulos, J. (2014). Computer-mediated communication and linguistic landscapes (J. Holmes & K. Hazen, Éd.). In J. Holmes & K. Hazen (Éd.), *Research methods in sociolinguistics : A practical guide*. Wiley-Blackwell Oxford.

Anglosphere. (p. d.). Merriam-Webster. Récupérée 29 avril 2020, à partir de https://www.merriam-webster.com/dictionary/Anglosphere

Anthony, L. (p. d.). *AntConc*. Tokyo, Japan. https://www.laurenceanthony.net/software

Arcep. (2020, mai 19). *Marché des communications électroniques en France - Année 2017 - Résultats définitifs*. Arcep. Récupérée 5 septembre 2020, à partir de https://www.arcep.fr/cartes-et-donnees/nos-publications-chiffrees/observatoire-des-marches-des-communications-electroniques-en-france/obs-marches-an2017-def.html

Argamon, S., Koppel, M., Fine, J. & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, *23*(3). https://doi.org/10.1515/text.2003.014

Aries, E. J. & Johnson, F. L. (1983). Close friendship in adulthood : Conversational content between same-sex friends. *Sex Roles*, *9*(12), 1183-1196. https://doi.org/10.1007/BF00303101

Ashcraft, C., McLain, B. & Eger, E. (2016). *Women in tech. The facts*. National Center for Women Information Technology.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics*. Cambridge, Cambridge University Press.

Bae. (2017, février 10). Urban Dictionary. Récupérée 10 septembre 2020, à partir de https://www.urbandictionary.com/define.php?term=Bae

Baider, F. H. (2004). *Hommes galants, femmes faciles : Étude socio-sémantique et diachronique*. L'Harmattan.

Bailly, S. (2008). *Les hommes, les femmes et la communication. Mais que vient faire le sexe dans la langue ?* L'Harmattan.

Baker, P. (2014). *Using corpora to analyze gender*. London, Bloomsbury.

Baker, P., Hardie, A. & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh, Edinburgh University Press.

Bamman, D., Eisenstein, J. & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135-160. https://doi.org/10.1111/josl.12080

banned - ListOfSubreddits. (p. d.). Récupérée 25 juillet 2018, à partir de https://www.reddit.com/r/ListOfSubreddits/wiki/banned

Barack Obama surprises internet with Ask Me Anything session on reddit. (2012, août 29). The Guardian. Récupérée 18 mars 2020, à partir de https://www.theguardian.com/technology/us-news-blog/2012/aug/29/barack-obama-ask-me-anything-reddit

Baron, N. S. (2004). See you online : Gender issues in college student use of Instant Messaging. *Journal of Language and Social Psychology*, *23*(4), 397-423. https://doi.org/10.1177/0261927X04269585

Barrett, R. (1999). Indexing polyphonous identity in the speech of African American drag queens. In *Reinventing identities : The gendered self in discourse* (p. 313-331). Oxford University Press.

Barthel, M., Stocking, G., Holcomb, J. & Mitchell, A. (2016a, février 25). *Discussion in Reddit circles more likely to focus on one candidate and one party*. Pew Research Center's Journalism Project. Récupérée 28 octobre 2017, à partir de http://www.journalism.org/2016/02/25/discussion-in-reddit-circles-more-likely-to-focus-on-one-candidate-and-one-party/

Barthel, M., Stocking, G., Holcomb, J. & Mitchell, A. (2016b, février 25). *Reddit news users more likely to be male, young and digital in their news preferences*. Pew Research Center's Journalism Project. Récupérée 28 octobre 2017, à partir de http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software67*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Baudelot, C. & Establet, R. (2006). *Allez les filles ! Une révolution silencieuse*. Paris, Seuil.

Bauer, G. R. (2014). Incorporating intersectionality theory into population health research methodology : Challenges and the potential to advance health equity. *Social Science & Medicine*, *110*, 10-17. https://doi.org/10.1016/j.socscimed.2014.03.022

Baugh, J. (1983). *Black street speech : Its history, structure, and survival*. University of Texas Press.

Baumgartner, J. (p. d.). *The day I posted Reddit to Reddit*. pushshift.io. Récupérée 9 décembre 2019, à partir de https://pushshift.io/author/stuck_in_the_matrix/

Bauvois, C. (2002). *Ni d'Eve ni d'Adam : étude sociolinguistique de douze variables du français*. Editions L'Harmattan.

Bechar-Israeli, H. (1995). From Bonehead to cLoNehEad : Nicknames, play, and identity on Internet Relay Chat. *Journal of Computer-Mediated Communication*, *1*(2). https://doi.org/10.1111/j.1083-6101.1995.tb00325.x

Beeching, K. (2002). *Gender, politeness and pragmatic particles in French*. John Benjamins Publishing.

Bergstrom, K. (2011). "Don't feed the troll" : Shutting down debate about community expectations on reddit.com. *First Monday*, *16*(8). Récupérée 6 novembre 2017, à partir de http://firstmonday.org/ojs/index.php/fm/article/view/3498

Bershtling, O. (2014). Speech creates a kind of commitment (L. Zimman, J. Davis & J. Raclaw, Éd.). In L. Zimman, J. Davis & J. Raclaw (Éd.), *Queer excursions : Retheorizing binaries in language, gender, and sexuality*. New York, Oxford University Press.

Biber, D. (1995). *Dimensions of register variation : A cross-linguistic comparison*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Bieswanger, M. (2007). A contrastive analysis of different shortening strategies in English and German text messages.

Bischoping, K. (1993). Gender differences in conversation topics, 1922-1990. *Sex Roles*, *28*(1-2), 1-18. https://doi.org/10.1007/BF00289744

Black Canadians. (2020, juin 6). In *Wikipedia*. Récupérée 8 juillet 2020, à partir de https://en.wikipedia.org/w/index.php?title=Black_Canadians&oldid=966251332
Page Version ID : 966251332

Blackless, M., Charuvastra, A., Derryck, A., Fausto-Sterling, A., Lauzanne, K. & Lee, E. (2000). How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology*, *12*, 151-166.

Bock, T. (2017, mai 19). *How correspondence analysis works (a simple explanation)*. Displayr. Récupérée 10 mars 2020, à partir de https://www.displayr.com/how-correspondence-analysis-works/

Borba, R. & Ostermann, A. C. (2007). Do bodies matter? Travestis' embodiment of (trans) gender identity through the manipulation of the Brazilian Portuguese grammatical gender system. *Gender and Language*, *1*(1), 131-147.

Bornstein, K. (1994). *Gender outlaw : On men, women and the rest of us* (Routledge).

Bosson, J. K., Vandello, J. A. & Buckner, C. E. (2019). *The psychology of sex and gender*. Thousand Oaks, SAGE Publications.

Boutet, J. (2017). La pensée critique dans la sociolinguistique en France. *Langage et societe*, *2-3*(160-161), 23-42. Récupérée 16 août 2020, à partir de https://www.cairn.info/journal-langage-et-societe-2017-2-page-23.htm

Boutin, M. B. P. E. (2012). La cyberlangue dans les forums de discussion : étude exploratoire dans le domaine de la télé réalité. Colloque de la SFSIC - Rennes. Récupérée 21 mars 2017, à partir de https://halshs.archives-ouvertes.fr/sic_00827718/document

Bowleg, L. (2008). When Black + Lesbian + Woman =/= Black Lesbian Woman : The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles*, *59*(5-6), 312-325. https://doi.org/10.1007/s11199-008-9400-z

Bowleg, L. & Bauer, G. (2016). Invited reflection : Quantifying intersectionality. *Psychology of Women Quarterly*, *40*(3), 337-341. https://doi.org/10.1177/0361684316654282

Branca-Rosoff, S. (2018). Modes langagières : le style des radios jeunes, In *Modes langagières dans l'histoire : Processus mimétiques et changements linguistiques*, Montpellier.

Braun, S. (2019). Les femmes, les hommes. Et les autres… Lexique. *Cahiers jungiens de psychanalyse*, *149*(1), 50-68. Récupérée 22 août 2020, à partir de https://www.cairn.info/revue-cahiers-jungiens-de-psychanalyse-2019-1-page-50.htm

Breheny, P. & Burchett, W. (2017). Visualization of regression models using visreg. *The R Journal*, *9*(2), 56-71.

Brezina, V. (2018). *Statistics in corpus linguistics : A practical guide*. Cambridge University Press.

Brock, A. (2012). From the blackhand side : Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media*, *56*(4), 529-549.

Brown, K. (2016, juin 4). *Your Twitter trend analysis is not deep, and it's probably wrong*. Jezebel. Récupérée 10 septembre 2020, à partir de https://jezebel.com/your-twitter-trend-analysis-is-not-deep-and-it-s-proba-1769411909

Bucholtz, Mary. (2009). From stance to style : gender, interaction and indexicality in Mexican immigrant youth slang, In *Stance : sociolinguistic perspectives,* Oxford, Oxford University Press.

Bucholtz, M. (1995). From Mulatta to Mestiza : language and the reshaping of ethnic identity. In *Gender articulated : Language and the socially constructed self* (p. 351-374). Psychology Press.

Bucholtz, M. (1999a). Bad examples : Transgression and progress in language and gender studies (M. Bucholtz, A. C. Liang & L. A. Sutton, Éd.). In M. Bucholtz, A. C. Liang & L. A. Sutton (Éd.), *Reinventing identities : The gendered self in discourse*. Oxford University Press.

Bucholtz, M. (1999b). You da man : Narrating the racial other in the production of white masculinity. *Journal of sociolinguistics*, *3*(4), 443-460.

Bucholtz, M. (2001). The whiteness of nerds : superstandard English and racial markedness. *Journal of linguistic anthropology*, *11*(1), 84-100.

Bucholtz, M. (2002). Geek feminism (S. Benor, M. Rose, D. Sharma, J. Sweetland & Q. Zhang, Éd.). In S. Benor, M. Rose, D. Sharma, J. Sweetland & Q. Zhang (Éd.), *Gendered practices in language*. Stanford, Center for the Study of Language and Information.

Bucholtz, M. (2004). Styles and stereotypes : the linguistic negotiation of identity among Laotian American youth. *Pragmatics*, *14*(2-3), 127-147. https://doi.org/10.1075/prag.14.2-3.02buc

Bucholtz, M. (2010). *White kids : Language, race, and styles of youth identity*. Cambridge University Press.

Bucholtz, M. (2012). Word up : Social meanings of slang in California youth culture (L. Monaghan, J. E. Goodman & J. M. Robinson, Éd.). In L. Monaghan, J. E. Goodman & J. M. Robinson (Éd.), *A cultural approach to interpersonal communication. Essential readings*. Wiley Blackwell.

Bucholtz, M., Liang, A. C. & Sutton, L. A. (1999). *Reinventing identities : The gendered self in discourse*. Oxford University Press.

Burch, B. & Egbert, J. (2020). Zero-inflated beta distribution applied to word frequency and lexical dispersion in corpus linguistics. *Journal of Applied Statistics*, *47*(2), 337-353. https://doi.org/10.1080/02664763.2019.1636941

Bureau, U. S. C. (p. d.). *Race*. United States Census Bureau. Récupérée 14 août 2020, à partir de https://data.census.gov/cedsci/table?q=asian%20race&tid=ACSDT1Y2018.B02001&hidePreview=false

Burger, J. D., Henderson, J., Kim, G. & Zarrella, G. (2011). Discriminating gender on Twitter, In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh.

Buridant, C. (2006). L'interjection : jeux et enjeux. *Langages*, *161*(1), 3-9. Récupérée 6 juin 2020, à partir de https://www.cairn.info/revue-langages-2006-1-page-3.htm

Burlage, B. (2019, mai 31). *Confessions of a Reddit 'Karma Whore'*. Vice. Récupérée 15 juin 2020, à partir de https://www.vice.com/en_us/article/3k359n/confessions-of-a-reddit-karma-whore

Bury, R. (2011). She's geeky : The performance of identity among women working in IT. *International Journal of Gender, Science and Technology*, *3*(1), 33-53.

Butler, J. (1991). Gender is burning : Questions of appropriation and subversion. *Cultural Politics*, *11*, 381-395.

Butler, J. (2006). *Gender trouble* (3rd). Routledge.

Cameron, D. (2005). Language, gender, and sexuality : Current issues and new directions. *Applied Linguistics*, *26*(4), 482-502. https://doi.org/10.1093/applin/ami027

Cameron, D. (2011). Sociophonetics and sexuality : Discussion. *American Speech*, *86*(1), 98-103. https://doi.org/10.1215/00031283-1277537

Cameron, D. (2014). Straight talking : The sociolinguistics of heterosexuality. *Langage et société*, *148*(2), 75. https://doi.org/10.3917/ls.148.0075

Cameron, D. & Kulick, D. (2003). *Language and sexuality*. Cambridge University Press.

Campbell, J. E. (2014). *Getting it on online : Cyberspace, gay male sexuality, and embodied identity.* New York, Routledge.

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech*, *82*(1), 32-64. https://doi.org/10.1215/00031283-2007-002

Canada.ca. (2017, février 2). *Choisissez ou mettez à jour l'identifiant de genre sur votre passeport ou document de voyage*. Canada.ca - Gouvernement du Canada. Récupérée 25 août 2020, à partir de https://www.canada.ca/fr/immigration-refugies-citoyennete/services/passeports-canadiens/changer-sexe.html
Last Modified : 2019-07-11

Cannon, G. (1989). Abbreviations and acronyms in English word-formation. *American Speech*, *64*(2), jstor 455038, 99-127. https://doi.org/10.2307/455038

Carlson, J. S., Cook, S. W. & Stromberg, E. L. (1936). Sex differences in conversation. *Journal of Applied Psychology*, *20*(6), 727.

Carter, P. M. (2013). Shared spaces, shared structures : Latino social formation and African American English in the U.S. south. *Journal of Sociolinguistics*, *17*(1), 66-92. https://doi.org/10.1111/josl.12015

Cedergren, H. J. & Sankoff, D. (1974). Variable rules : Performance as a statistical reflection of competence. *Language*, 333-355.

Center, P. R. & Inquiries. (2019, juin 12). *Demographics of mobile device ownership and adoption in the United States*. Pew Research Center : Internet, Science & Tech. Récupérée 12 septembre 2020, à partir de https://www.pewresearch.org/internet/fact-sheet/mobile/

Cheshire, J. (1982). Variation in an English dialect : A sociolinguistic study. *Cambridge Studies in Linguistics London*, *37*, x-v.

Chess, S. & Shaw, A. (2015). A Conspiracy of fishes, or, how we learned to stop worrying about #GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting & Electronic Media*, *59*(1), 208-220. https://doi.org/10.1080/08838151.2014.999917

Chetcuti, N. (2012). *La face cachée du genre : Langage et pouvoir des normes*. Presses Sorbonne Nouvelle.

Christ, O., Schulze, B., Hoffmann, A. & König, E. (1999). *The IMS Corpus Workbench, Corpus Query Processor (CQP)*. Universität Stuttgart, Stuttgart, Institut für Maschinelle Sprachverarbeitung.

Chun, E. W. (2001). The construction of White, Black, and Korean American identities through African American Vernacular English. *Journal of Linguistic Anthropology*, *11*(1), jstor 43103953, 52-64.

Circle jerk. (p. d.). Dictionary.com. Récupérée 14 mai 2020, à partir de https://www.dictionary.com/e/slang/circle-jerk/

Coaston, J. (2019, mai 20). *The intersectionality wars*. Vox. Récupérée 6 février 2020, à partir de https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination

Coats, S. (2017a). Gender and grammatical frequencies in social media English from the Nordic countries (D. Fišer & M. Beißwenger, Éd.). In D. Fišer & M. Beißwenger (Éd.), *Investigating Computer-Mediated Communication : Corpus-based Approaches to Language in the Digital World*. Ljubljana University Press.

Coats, S. (2017b). Gender and lexical type frequencies in Finland Twitter English. *Studies in Variation, Contacts and Change in English*, *19*. Récupérée 1 janvier 2019, à partir de http://www.helsinki.fi/varieng/series/volumes/19/coats/

Cole, J. R., Ghafurian, M. & Reitter, D. (2017, juillet 5). Is word adoption a grassroots process? An analysis of Reddit communities, In *Social, Cultural, and Behavioral Modeling*. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, Cham. https://doi.org/10.1007/978-3-319-60240-0_28

Colley, A. & Todd, Z. (2002). Gender-Linked Differences in the Style and Content of E-Mails to Friends. *Journal of Language and Social Psychology*, *21*(4), 380-392. https://doi.org/10.1177/026192702237955

Collins, P. H. & Chepp, V. (2013). Intersectionality (G. Waylen, K. Celis, J. Kantola & S. L. Weldon, Éd.). In G. Waylen, K. Celis, J. Kantola &

S. L. Weldon (Éd.), *The Oxford Handbook of Gender and Politics*. Oxford University Press.

The Combahee River Collective Statement. (p. d.). Récupérée 6 février 2020, à partir de http://circuitous.org/scraps/combahee.html

ComingofFaith. (2016, mai 1). *The internet's love of Black slang makes some of us uncomfortable*. HuffPost. Récupérée 11 septembre 2020, à partir de https://www.huffpost.com/entry/the-internets-love-of-bla_b_8903778

The Comprehensive R Archive Network. (p. d.). The Comprehensive R Archive Network. Récupérée 2 janvier 2020, à partir de https://cran.r-project.org/

Connors, R. J. & Lunsford, A. A. (1988). Frequency of formal errors in current college writing. *College Composition and Communication*, *39*(4), 395-409. https://doi.org/10.2307/357695

Cop, M. & Hatfield, H. (2017). An athletes [sic] performance : Can a possessive apostrophe predict success ? : Misplace apostrophes, miss out on med school ? *English Today*, *33*(3), 39-45. https://doi.org/10.1017/S026607841600064X

Cornetto, K. M. & Nowak, K. L. (2006). Utilizing usernames for sex categorization in computer-mediated communication : Examining perceptions and accuracy. *CyberPsychology & Behavior*, *9*(4), 377-387. https://doi.org/10.1089/cpb.2006.9.377

Cougnon, L.-A. & François, T. (2010). Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS. JADT : 10th International Conference on Statistical Analysis of Textual Data.

Coupland, N. (2001). Age in social and sociolinguistic theory (N. Coupland, S. Sarangi & C. N. Candlin, Éd.). In N. Coupland, S. Sarangi & C. N. Candlin (Éd.), *Sociolinguistics and social theory*. Pearson Education Limited.

Covarrubias, A. (2011). Quantitative intersectionality : A critical race analysis of the chicana/o educational Pipeline. *Journal of Latinos and Education*, *10*(2), 86-105. https://doi.org/10.1080/15348431.2011.556519

Crawl Wiki for Acronyms. (p. d.). https://raw.githubusercontent.com/krishnakt031990/Crawl-Wiki-For-Acronyms/master/AcronymsFile.csv

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex : A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University Of Chicago Legal Forum*, (1), 140-167.

Crystal, D. (2006). *Language and the Internet* (2nd edition). Cambridge, Cambridge University Press. Récupérée 19 mars 2017, à partir de http://public.eblib.com/choice/publicfullrecord.aspx?p=274901

Crystal, D. (2008). *Txtng : the Gr8 Db8*. New York, Oxford University Press.

Cunha, E., Magno, G., Almeida, V., Gonçalves, M. A. & Benevenuto, F. (2012). A gender based study of tagging behavior in twitter, In *Pro-*

ceedings of the 23rd ACM conference on Hypertext and social media - HT '12. the 23rd ACM conference, Milwaukee, Wisconsin, USA, ACM Press. https://doi.org/10.1145/2309996.2310055

Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C. & Almeida, V. (2014). He votes or she votes ? Female and male discursive strategies in Twitter political hashtags (T. Preis, Éd.). *PLoS ONE*, *9*(1), e87041. https://doi.org/10.1371/journal.pone.0087041

Cutler, C. A. (1999). Yorkville Crossing : White teens, hip hop and African American English. *Journal of Sociolinguistics*, *3*(4), 428-442. https://doi.org/10.1111/1467-9481.00089

Dagorn, J. (2011). Les trois vagues féministes – une construction sociale ancrée dans une histoire. *Diversité : ville école intégration*. Récupérée 13 septembre 2020, à partir de https://hal.archives-ouvertes.fr/hal-02053657

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A. & Swinton, J. (2019). *xtable : Export tables to LaTeX or HTML*. manual. https://CRAN.R-project.org/package=xtable

daisygirl_79. (2013). *I'm new to reddit. Tell me the acronyms !* Reddit. Récupérée 14 mai 2020, à partir de https://www.reddit.com/r/help/comments/1a0mq2/im_new_to_reddit_tell_me_the_acronyms/

dandeeo. (2017, mars 31). *Why is there an unspoken agreement to never use emojis on Reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/AskReddit/comments/62l8q2/why_is_there_an_unspoken_agreement_to_never_use/

Danet, B. (2001). *Cyberpl@y : Communicating online*. Oxford, Berg.

Danet, B. & Herring, S. C. (Éd.). (2007). *The multilingual Internet : Language, culture, and communication online*. Oxford ; New York, Oxford University Press
OCLC : ocm67405669.

datterHFX. (2017, mai 25). *What percentage of your reddit usage is desktop vs mobile ?* reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/AskReddit/comments/6db6pz/what_percentage_of_your_reddit_usage_is_desktop/

Davies, C. E. (2015). Twitter as political discourse (J. Wilson & D. Boxer, Éd.). In J. Wilson & D. Boxer (Éd.), *Discourse, politics and women as global leaders*. John Benjamins Publishing Company.

Davis, A. Y. (1983). *Women, race, & class*. Vintage.

Davis, M. & Edberg, P. (2020, septembre 18). *Unicode Emoji*. Récupérée 22 septembre 2020, à partir de https://www.unicode.org/reports/tr51/#Introduction

De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W. & Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness, In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. the 2017 ACM Conference, Portland, Oregon, USA, ACM Press. https://doi.org/10.1145/2998181.2998220

De Jonge, S. & Kemp, N. (2012). Text-message abbreviations and language skills in high school and university students. *Journal of Research in Reading*, *35*(1), 49-68. https://doi.org/10.1111/j.1467-9817.2010.01466.x

defaults - ListOfSubreddits. (p. d.). Reddit. Récupérée 7 juillet 2020, à partir de https://www.reddit.com/r/ListOfSubreddits/wiki/defaults

Definition of BOO. (p. d.). Merriam-Webster. Récupérée 20 avril 2020, à partir de https://www.merriam-webster.com/dictionary/boo

Definition of emoticon. (p. d.). Merriam-Webster. Récupérée 9 avril 2020, à partir de https://www.merriam-webster.com/dictionary/emoticon

Definition of FREAKING. (p. d.). Merriem Webster. Récupérée 28 mars 2020, à partir de https://www.merriam-webster.com/dictionary/freaking

Del Valle, M. E., Gruzd, A., Kumar, P. & Gilbert, S. (2020). Learning in the wild : Understanding networked ties in Reddit (N. B. Dohn, P. Jandrić, T. Ryberg & M. de Laat, Éd.). In N. B. Dohn, P. Jandrić, T. Ryberg & M. de Laat (Éd.), *Mobility, Data and Learner Agency in Networked Learning*. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-36911-8_4

[deleted]. (2010, août 17). *DAE hate it when people add extra letters to the end of a word to be cute like thisss ?* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/DoesAnybodyElse/comments/d23du/dae_hate_it_when_people_add_extra_letters_to_the/

Del-Teso-Craviotto, M. (2008). Gender and sexual identity authentication in language use : the case of chat rooms. *Discourse Studies*, *10*(2), 251-270. https://doi.org/10.1177/1461445607087011

Demographics of Social Media Users and Adoption in the United States. (2019, juin 12). Pew Research Center. Récupérée 14 mai 2020, à partir de https://www.pewresearch.org/internet/fact-sheet/social-media/

Derks, D., Bos, A. E. R. & von Grumbkow, J. (2008). Emoticons in computer-mediated communication : Social motives and social context. *CyberPsychology & Behavior*, *11*(1), 99-101. https://doi.org/10.1089/cpb.2007.9926

Desagulier, G. (2017). *Corpus linguistics and statistics with R*. Springer International Publishing.

Digest of Education Statistics. (2016). National Center for Education Statistics. Récupérée 8 juillet 2020, à partir de https://nces.ed.gov/programs/digest/d16/tables/dt16_325.35.asp

Dill, B. T. (1983). Race, class, and gender : Prospects for an all-inclusive sisterhood. *Feminist Studies*, *9*(1), 131-150.

Dorlin, E. (Éd.). (2009). *Sexe, race, classe : pour une épistémologie de la domination*. Presses universitaires de France.

Dorlin, E. (2014). *La matrice de la race : généalogie sexuelle et coloniale de la nation française*. La Découverte.

Dosono, B. & Semaan, B. (2019). Moderation practices as emotional labor in sustaining online communities : The case of AAPI identity

work on Reddit, In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. the 2019 CHI Conference, Glasgow, Scotland Uk, ACM Press. https://doi.org/10.1145/3290605.3300372

Dresner, E. & Herring, S. C. (2010). Functions of the nonverbal in CMC : Emoticons and illocutionary force. *Communication Theory*, *20*(3), 249-268. https://doi.org/10.1111/j.1468-2885.2010.01362.x

Drouin, M. & Driver, B. (2014). Texting, textese and literacy abilities : A naturalistic study. *Journal of Research in Reading*, *37*(3), 250-267. https://doi.org/10.1111/j.1467-9817.2012.01532.x

Drummond, R. & Schleef, E. (2016). Identity in variationist sociolinguistics (S. Preece, Éd.). In S. Preece (Éd.), *The Routledge handbook of language and identity*. Routledge.

Drunken_Economist. (p. d.). *Moderators : New subreddit settings for mobile*. reddit. Récupérée 14 mai 2020, à partir de https://www.reddit.com/r/modnews/comments/41054l/moderators_new_subreddit_settings_for_mobile/

dsamanthas. (2018, décembre 28). *Why don't people use emojis on Reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/TooAfraidToAsk/comments/aad2ao/why_dont_people_use_emojis_on_reddit/

Dunbar, R. I. M., Marriott, A. & Duncan, N. D. C. (1997). Human conversational behavior. *Human Nature*, *8*(3), 231-246. https://doi.org/10.1007/BF02912493

Eckert, P. (1989a). *Jocks and burnouts : Social categories and identity in the high school*. Teachers college press.

Eckert, P. (1989b). The whole woman : Sex and gender differences in variation. *Language Variation and Change*, (1), 245-267.

Eckert, P. (1998). Age as a sociolinguistic variable (F. Coulmas, Éd.). In F. Coulmas (Éd.), *The handbook of sociolinguistics*. Wiley Blackwell.

Eckert, P. (2000). *Language variation as social practice : The linguistic construction of identity in Belten High*. Wiley-Blackwell.

Eckert, P. (2003). Language and gender in adolescence (J. Holmes & M. Meyerhoff, Éd.). In J. Holmes & M. Meyerhoff (Éd.), *The Handbook of language and gender*. Blackwell Publishing.

Eckert, P. (2012). Three waves of variation study : The emergence of meaning in the study of variation. *Annual Review of Anthropology*, *41*, 87-100.

Eckert, P. (2014). The problem with binaries : Coding for gender and sexuality. *Language and Linguistics Compass*, *8*(11), 529-535. https://doi.org/10.1111/lnc3.12113

Eckert, P. & McConnell-Ginet, S. (2003). *Language and gender*. Cambridge, Cambridge University Press.

Eckert, P. & McConnell-Ginet, S. (2012). Think practically and look locally : Language and gender as community-based practice. *Annu. Rev. Anthropol.*, (21), 461-490.

Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, *19*(2), 161-188. https://doi.org/10.1111/josl.12119

Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. (2010). A latent variable model for geographic lexical variation, In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, MIT, Massachusetts.

Eisenstein, J., Smith, N. A. & Xing, E. P. (2011). Discovering sociolinguistic associations with structured sparsity, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. ACL-HLT 2011, Portland, Oregon, USA, Association for Computational Linguistics. Récupérée 12 août 2020, à partir de https://www.aclweb.org/anthology/P11-1137

emojitracker : realtime emoji use on twitter. (p. d.). emojitracker.com. Récupérée 19 juin 2020, à partir de https://emojitracker.com/

Encrevé, P. (1976). Labov, linguistique, sociolinguistique (W. Labov, Éd.). In W. Labov (Éd.), *Sociolinguistique*. Paris, Minuit.

Encrevé, P. (1988). *La liaison avec et sans enchainement : phonologie tridimensionnelle et usages du français*. Paris, Editions du Seuil.

Epstein, C. F. (1986). Symbolic segregation : Similarities and differences in the language and non-verbal communication of women and men. *Sociological Forum*, *1*(1), jstor 684552, 27-49.

Evans, C. L. (2018). *Broad Band. The untold story of the women who made the internet*. New York, Penguin.

Evans, H. (2016). Do women only talk about "female issues"? Gender and issue discussion on Twitter. *Online Information Review*, *40*(5), 660-672. https://doi.org/10.1108/OIR-10-2015-0338

FAQ - Emoji & pictographs. (p. d.). Récupérée 1 novembre 2017, à partir de https://unicode.org/faq/emoji_dingbats.html

faq - reddit.com. (p. d.). Récupérée 18 janvier 2020, à partir de https://www.reddit.com/wiki/faq

Faraway, J. J. (2016). *Linear models with R*. Chapman and Hall/CRC.

fark. (p. d.). Wiktionary. Récupérée 29 mars 2020, à partir de https://en.wiktionary.org/wiki/fark#English

Farrell, T., Fernandez, M., Novotny, J. & Alani, H. (2019). Exploring misogyny across the Manosphere in Reddit, In *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*. the 10th ACM Conference, Boston, Massachusetts, USA, ACM Press. https://doi.org/10.1145/3292522.3326045

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

Fausto-Sterling, A. (2012). *Sex/gender : Biology in a social world*. Routledge.

Ferguson, R. A. (2012). Reading intersectionality. *Trans-Scripts*, *2*, 91-99. Récupérée 7 juin 2017, à partir de http://sites.uci.edu/transscripts/files/2014/10/2012_02_08.pdf

Ferson, P. (2016, août 9). *MTE explains : Full-width text on the Internet and its origins*. Make Tech Easier. Récupérée 4 septembre 2020, à partir de https://www.maketecheasier.com/full-width-text-and-origins/

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications, Inc.

Finlay, S. C. (2014). Age and gender in Reddit commenting and success. *Journal of Information Science Theory and Practice*, *2*(3), 18-28. https://doi.org/10.1633/JISTaP.2014.2.3.2

Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, *14*(1), 47-56. https://doi.org/10.1080/00437956.1958.11659655

Fisher, O. (2019). On trans issues, Iceland has just put Britain to shame | Owl Fisher [newspaper]. *The Guardian : Opinion*. Récupérée 25 août 2020, à partir de https://www.theguardian.com/commentisfree/2019/jun/21/trans-issues-britain-iceland-law-intersex-rights

Fishman, P. M. (1978). Interaction : The work women do. *Social Problems*, *25*(4), 397-406. https://doi.org/10.2307/800492

Fix, S. (2010). Representations of blackness by white women : Linguistic practice in the community versus the media. *University of Pennsylvania Working Papers in Linguistics*, *16*(2), 56-65.

Flesch, M. (2016). *Acronyms and emoticons on a popular web forum : Does gender make a difference ? A corpus-based study of Reddit*. Université de Lorraine – ERUDI.

Flood, R. (2018, février 19). *What does SMH mean ? All the Reddit slang you pretend to understand actually explained*. PinkNews - Gay news, reviews and comment from the world's most read lesbian, gay, bisexual, and trans news service. Récupérée 15 mai 2020, à partir de https://www.pinknews.co.uk/2018/02/19/smh-what-does-smh-mean/

Flores, A. R., Herman, J. L., Gates, G. J. & Brown, T. N. T. (2016, juin). *How many adults identify as transgender in the United States ?* The Williams Institute. http://williamsinstitute.law.ucla.edu/wp-content/uploads/How-Many-Adults-Identify-as-Transgender-in-the-United-States.pdf

Florini, S. (2014). Tweets, tweeps, and signifyin' : communication and cultural performance on "Black Twitter". *Television & New Media*, *15*(3), 223-237. https://doi.org/10.1177/1527476413480247

fook. (p. d.). Wiktionary. Récupérée 29 mars 2020, à partir de https://en.wiktionary.org/wiki/fook

Foster, A. (2018, mai 10). *Cara Delevingne : Realising I am gender fluid was a milestone*. Evening Standard. Récupérée 25 août 2020, à partir de https://www.standard.co.uk/showbiz/celebrity-news/cara-delevingne-realising-i-am-gender-fluid-was-a-breakthrough-moment-for-me-a3835986.html

Fought, C. (2003). *Chicano English in context*. Palgrave Macmillan.

Fought, C. (2006). *Language and ethnicity*. Cambridge, Cambridge University Press. Récupérée 19 mars 2017, à partir de http://dx.doi.org/

10.1017/CBO9780511791215
OCLC : 252528937

Fournier, P. (2010). Âge (S. Paugam, Éd.). In S. Paugam (Éd.), *Les 100 mots de la sociologie*. Paris, Presses universitaires de France. Récupérée 17 octobre 2017, à partir de https://sociologie.revues.org/522

Fradin, B. (2003). *Nouvelles approches en morphologie*. Paris, Presse Universitaires de France.

Francis, W. N. & Kucera, H. (1964). Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island*, *1*.

Francis, W. N., Kučera, H. & Mackie, A. W. (1982). *Frequency analysis of English usage : Lexicon and grammar*. Houghton Mifflin.

Frey, J.-C., König, A. & Stemle, E. W. (2019). How FAIR are CMC corpora ?, In *Proceedings of the 4th conference on CMC and social media corpora for the humanities*. CMC-Corpora2019, Université de Cergy-Pontoise. https://cris.unibo.it/retrieve/handle/11585/723943/548469/2019_How%20FAIR%20are%20CMC%20corpora.pdf

fricking. (p. d.). Online Etymology Dictionary. Récupérée 29 mars 2020, à partir de https://www.etymonline.com/search?q=fricking

Frigging. (p. d.). Online Etymology Dictionary. Récupérée 28 mars 2020, à partir de https://www.etymonline.com/word/frigging

Full Emoji List, v13.0. (p. d.). Unicode.org. Récupérée 4 juillet 2020, à partir de https://unicode.org/emoji/charts/full-emoji-list.html

Fullwood, C., Orchard, L. J. & Floyd, S. A. (2013). Emoticon convergence in Internet chat rooms. *Social Semiotics*, *23*(5), 648-662. https://doi.org/10.1080/10350330.2012.739000

Gadet, F. (2003). Is there a French theory of variation ? *International Journal of the Sociology of Language*, *2003*(160). https://doi.org/10.1515/ijsl.2003.017

Galwey, N. W. (2014). *Introduction to mixed modelling*. John Wiley & Sons.

Garrison, A., Remley, D., Thomas, P. & Wierszewski, E. (2011). Conventional faces : Emoticons in instant messaging discourse. *Computers and Composition*, *28*(2), 112-125. https://doi.org/10.1016/j.compcom.2011.04.001

Gaudio, R. P. (1994). Sounding gay : Pitch properties in the speech of gay and straight men. *American Speech*, *69*(1), jstor 455948, 30-57. https://doi.org/10.2307/455948

Gauthier, M. (2017). *Age, gender, fuck, and Twitter : A sociolinguistic analysis of swear words in a corpus of British tweets*. Université Lumière Lyon 2.

Geek. (p. d.). Online Etymology Dictionary. Récupérée 8 juillet 2020, à partir de https://www.etymonline.com/word/geek

ginganinja2507. (2010, août 17). *DAE hate it when people add extra letters to the end of a word to be cute like thisss ?* reddit. Récupérée 1 août 2020, à partir de https://www.reddit.com/r/DoesAnybodyElse/comments/d23du/dae_hate_it_when_people_add_extra_letters_to_the/

glossary - TheoryOfReddit. (p. d.). Reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/TheoryOfReddit/wiki/glossary

Glynn, D. (2014). Correspondence analysis : Exploring data and identifying patterns (D. Glynn & J. A. Robinson, Éd.). In D. Glynn & J. A. Robinson (Éd.), *Human Cognitive Processing*. Amsterdam, John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.17gly

Gonzalez-Barrera, A. & Lopez, M. H. (2015, juin 15). *Is being Hispanic a matter of race, ethnicity or both ?* Pew Research Center. Récupérée 28 avril 2018, à partir de http://www.pewresearch.org/fact-tank/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/

Gratton, C. (2016). Resisting the gender binary : The use of (ING) in the construction of non-binary transgender identities. *University of Pennsylvania Working Papers in Linguistics*, *22*(2), 7. Récupérée 23 mars 2017, à partir de http://repository.upenn.edu/pwpl/vol22/iss2/7/

Gray, J. (1992). *Men are from Mars, women are from Venus*. HarperCollins.

Greco, L. (2014). Les recherches linguistiques sur le genre : un état de l'art. *Langage et société*, *148*(2), 11. https://doi.org/10.3917/ls.148.0011

Greco, L. (2018). *Dans les coulisses du genre : la fabrique de soi chez les Drag Kings*.

Green, L. J. (2002). *African American English : A linguistic introduction*. Cambridge University Press.

Gries, S. T. (2013). *Statistics for linguistics with R*. Berlin, De Gruyter Mouton.

Guiller, J. & Durndell, A. (2007). Students' linguistic behaviour in online discussion groups : Does gender matter ? *Computers in Human Behavior*, *23*(5), 2240-2255. https://doi.org/10.1016/j.chb.2006.03.004

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.

Guy, G. R. (2018). Saks vs. Macys : (r-1) marches on in New York City department stores. *University of Pennsylvania Working Papers in Linguistics*, *24*(2), 49-55.

ha-ha | Origin and meaning of ha-ha by Online Etymology Dictionary. (p. d.). Récupérée 28 février 2020, à partir de https://www.etymonline.com/word/ha-ha

Haas, A. & Sherman, M. A. (1982). Reported topics of conversation among same-sex adults. *Communication Quarterly*, *30*(4), 332-342. https://doi.org/10.1080/01463378209369469

The Hackers Acronym Chart. (2000, mars 31). Récupérée 20 avril 2020, à partir de http://www.textfiles.com/magazines/PHANTASY/iirg-acronyms-v12.txt

Hall, K. (1996). Cyberfeminism (S. Herring, Éd.). In S. Herring (Éd.), *Computer-mediated communication : linguistic, social and cross-cultural perspectives*. Amsterdam, Benjamins.

Hall, K. & O'Donovan, V. (1996). Shifting gender positions among Hindi-speaking hijras (V. Bergvall, J. Bing & A. F. Freed, Éd.). In V. Bergvall, J. Bing & A. F. Freed (Éd.), *Rethinking language and gender research : Theory and practice*. London, Longman.

Hall-Lew, L. (2009). *Ethnicity and phonetic variation in a San Francisco neighborhood*. Stanford University.

Haney-Lopez, I. F. (1994). The social construction of race : Some observations on illusion, fabrication, and choice. *Harvard Civil Rights-Civil Liberties Law Review*, *29*, 1-63.

Hanna, D. B. (1997). Do I sound "Asian" to you ? : Linguistic markers of Asian American identity. *University of Pennsylvania working papers in linguistics*, *4*(2), 141-53.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616. https://doi.org/10.7717/peerj.616

Hazenberg, E. (2015). Walking the straight and narrow : linguistic choice and gendered presentation. *Gender and Language*, *10*(2), 270-294. https://doi.org/10.1558/genl.v10i2.19812

Heaney, K. (2013, février 18). *How men and women tweet*. BuzzFeed News. Récupérée 1 août 2020, à partir de https://www.buzzfeednews.com/article/katieheaney/how-men-and-women-tweet

The Heartbreaking Backstory To The Founding Of Reddit. (p. d.). Récupérée 18 mars 2020, à partir de https://www.businessinsider.fr/us/alexis-ohanians-mother-diagnosed-with-brain-cancer-within-month-of-reddits-founding-2012-5

Heiden, S., Magué, J.-P. & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement, In *10th International Conference on the Statistical Analysis of Textual Data*. JADT 2010, Rome.

Henley, N. & Kramarae, C. (1991). Gender, power and miscommunication (N. Coupland, H. Giles & J. Wiemann, Éd.). In N. Coupland, H. Giles & J. Wiemann (Éd.), *Miscommunication and problematic talk, Sage, (Calif.)* Newbury Park, Sage.

Herring, S. C. (1994). Politeness in computer culture : Why women thank and men flame, In *Cultural Performances : Proceedings of the Third Berkeley Women and Language Conference*, Berkeley Women and Language Group.

Herring, S. C. (1996). Bringing familiar baggage to the new frontier : Gender differences in computer-mediated communication (V. Vitanza, Éd.). In V. Vitanza (Éd.), *CyberReader*. Boston, Allyn & Bacon.

Herring, S. C. (2003). Gender and power in online communication (J. Holmes & M. Meyerhoff, Éd.). In J. Holmes & M. Meyerhoff (Éd.), *The Handbook of Language and Gender*. Blackwell.

Herring, S. C. (2012). Grammar and electronic communication (C. A. Chapelle, Éd.). In C. A. Chapelle (Éd.), *The Encyclopedia of Applied Linguistics*. Oxford, UK, Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0466

Herring, S. C. (2015). New frontiers in interactive multimodal communication. *The Routledge handbook of language and digital communication*, 398-402.

Herring, S. C. & Zelenkauskaite, A. (2008). Gendered typography : Abbreviation and insertion in Italian iTV SMS. *IULC Working Papers*, *8*(3). Récupérée 21 mars 2017, à partir de https://www.indiana.edu/~iulcwp/wp/article/view/08-22A

Herring, S. C. & Zelenkauskaite, A. (2009). Symbolic capital in a virtual heterosexual market : Abbreviation and insertion in Italian iTV SMS. *Written Communication*, *26*(1), 5-31. https://doi.org/10.1177/0741088308327911

HideHideHidden. (2017, mai 17). *Try the new profiles page yourselves and tell us what you think*. reddit. Récupérée 7 janvier 2020, à partir de https://www.reddit.com/r/beta/comments/6bqemt/try_the_new_profiles_page_yourselves_and_tell_us/

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Hilte, L. (2019). *The social in social media writing : The impact of age, gender and social class indicators on adolescents' informal online writing practices*. Universiteit Antwerpen. Anvers.

Hlavac, M. (2018). *stargazer : Well-formatted regression and summary statistics tables. R package version 5.2.1*. https://CRAN.R-project.org/package=stargazer

Hoffa, F. (2017, août 2). *The most famous reddit accounts*. Medium. Récupérée 7 février 2020, à partir de https://medium.com/@hoffa/the-most-famous-reddit-accounts-c9958b5bc376

Hokanson, L. & Kemp, N. (2013). Adults' spelling and understanding of possession and plurality : an intervention study. *Reading and Writing*, *26*(2), 241-261. https://doi.org/10.1007/s11145-012-9366-7

Holmberg, K. & Hellsten, I. (2015). Gender differences in the climate change communication on Twitter. *Internet Research*, *25*(5), 811-828. https://doi.org/10.1108/IntR-07-2014-0179

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, *45*(1), 92-99. https://doi.org/10.1016/j.jrp.2010.11.015

Houdebine, A.-M. (1979). La différence sexuelle et la langue. *Langage & société*, *7*(1), 3-30.

Houdebine-Gravaud, A.-M. (1978). *La variété et la dynamique d'un français régional : étude phonologique, analyse des facteurs de variations à partir d'une enquête à grande échelle dans le département de la Vienne (Poitou)*. Université Paris Descartes.

How do I get flair (the text / image next to my username) ? (p. d.). Reddit Help. Récupérée 29 décembre 2019, à partir de https://www.reddithelp.com/en/categories/using-reddit/your-reddit-account/how-do-i-get-flair-textimage-next-my-username

How do I get karma ? | Reddit Help. (p. d.). Récupérée 18 mars 2020, à partir de https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/how-do-i-get-karma

Huffaker, D. A. (2010). Dimensions of leadership and social influence in online communities. *Human Communication Research*, *36*(4), 593-617. https://doi.org/10.1111/j.1468-2958.2010.01390.x

Huffaker, D. A. & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, *10*(2), 00-00. Récupérée 12 avril 2019, à partir de https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2005.tb00238.x

Hussain, S. (2020, juillet 8). *Reddit moderators spent years asking for help fighting hate. The company may finally be listening*. Los Angeles Times. Récupérée 9 octobre 2020, à partir de https://www.latimes.com/business/technology/story/2020-07-08/reddit-hate-harassment-moderators-gain-ground

Husson, F., Lê, S. & Pagès, J. (2011). *Exploratory multivariate analysis by example using R*. Chapman and Hall/CRC.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581-592. https://doi.org/10.1037/0003-066X.60.6.581

IANAL. (2019, septembre 21). In *Wikipedia*. Récupérée 14 mai 2020, à partir de https://en.wikipedia.org/w/index.php?title=IANAL&oldid=917035032
Page Version ID : 917035032

Ilbury, C. (2020). Sassy Queens : Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, *24*(2), 245-264. https://doi.org/10.1111/josl.12366

Impudence. (2017, avril 15). *Why do you think that "there are no girls on the internet" is a thing ?* reddit. Récupérée 10 août 2020, à partir de https://www.reddit.com/r/AskWomen/comments/65idnb/why_do_you_think_that_there_are_no_girls_on_the/

[inconnu]. (2017). *Why do some girls repeat multiple same letters in a word when typing ? - Quora*. Quora. Récupérée 5 août 2020, à partir de https://www.quora.com/Why-do-some-girls-repeat-multiple-same-letters-in-a-word-when-typing

[inconnu]. (2020). *Why does it mean when a girl message or texts you with extra letters like hiii, thanksss and bossss ? - quora*. Quora. Récupérée 5 août 2020, à partir de https://www.quora.com/Why-does-it-mean-when-a-girl-message-or-texts-you-with-extra-letters-like-Hiii-thanksss-and-bossss

Initiative, T. E. (p. d.). *OxGarage*. Récupérée 3 octobre 2020, à partir de https://oxgarage2.tei-c.org/#

Isaac, M. & Streitfeld, D. (2015). It's Silicon Valley 2, Ellen Pao 0 : Fighter of sexism is out at Reddit [newspaper]. *The New York Times*. Récupérée 18 mars 2020, à partir de https://www.nytimes.com/2015/07/11/technology/ellen-pao-reddit-chief-executive-resignation.html?smid=tw-nytimes&_r=0

Jackman, S. (2020). *pscl : Classes and Methods for R Developed in the Political Science Computational Laboratory*. Sydney.

Jaunait, A. & Chauvin, S. (2012). Représenter l'intersection : Les théories de l'intersectionnalité à l'épreuve des sciences sociales. *Revue française de science politique*, *62*(1), 5. https://doi.org/10.3917/rfsp.621.0005

Jespersen, O. (1922). *Language, its nature, development, and origin*. Allen & Unwin.

Johnson, R. I. (1917). The persistency of error in English composition. *The School Review*, *25*(8), 555-580.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, *90*(4), 403-440. https://doi.org/10.1215/00031283-3442117

JustALivingHuman. (2019, juillet 7). *What does it REALLY mean if she adds multiple letters at the end of the words like 'heyyy'?* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/dating_advice/comments/ca38e3/what_does_it_really_mean_if_she_adds_multiple/

Kalman, Y. M. & Gergle, D. (2014). Letter repetitions in computer-mediated communication : A unique link between spoken and online language. *Computers in Human Behavior*, *34*, 187-193. https://doi.org/10.1016/j.chb.2014.01.047

Kavanagh, B. (2016). Emoticons as a medium for channeling politeness within American and Japanese online blogging communities. *Language & Communication*, *48*, 53-65. https://doi.org/10.1016/j.langcom.2016.03.003

Kaye, L. K., Wall, H. J. & Malone, S. A. (2016). Turn that frown upside-down : A contextual account of emoticon usage on different virtual platforms. *Computers in Human Behavior*, *60*, 463-467. https://doi.org/10.1016/j.chb.2016.02.088

Kemp, N. (2010). Texting versus txtng : reading and writing text messages, and links with other linguistic skills. *Writing Systems Research*, *2*(1), 53-71. https://doi.org/10.1093/wsr/wsq002

Kendall, L. (2011). White and nerdy : Computers, race, and the nerd stereotype. *The Journal of Popular Culture*, *44*(3), 505-524. https://doi.org/10.1111/j.1540-5931.2011.00846.x

Kennedy, H. W. (2006). Illegitimate, monstrous and out there : Female'Quake'players and inappropriate pleasures (J. Hollows & R. Mosley, Éd.). In J. Hollows & R. Mosley (Éd.), *Feminism in popular culture*. Oxford, UK, Berg.

Kergoat, D. (1978). Ouvriers = ouvrières ? Propositions pour une articulation théorique de deux variables : sexe et classe sociale. *Critiques de l'économie politique*, (5), 65-97.

Kergoat, D. (2000). Division sexuelle du travail et rapports sociaux de sexe (H. Hirata, F. Laborie, H. Le Doaré & D. Senotier, Éd.). In H. Hirata, F. Laborie, H. Le Doaré & D. Senotier (Éd.), *Dictionnaire critique du féminisme*. Paris, PUF.

Kesseler, A. & Bergs, A. (2003). Literacy and the new media : vita brevis, lingua brevis (J. Aitchison & D. Lewis, Éd.). In J. Aitchison & D. Lewis (Éd.), *New Media Language*. London, Routledge.

Kidd, D. (2018). *Social media freaks : Digital identity in the network society*. New York, Routledge.

Kiesler, S., Siegel, J. & McGuire, T. W. (1984). Social psychological aspects of Computer-Mediated Communication. *American Psychologist*, *39*, 1123-1234.

Kiesling, S. F. (1998). Men's identities and sociolinguistic variation : The case of fraternity men. *Journal of Sociolinguistics*, *2*(1), 69-99. https://doi.org/10.1111/1467-9481.00031

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263-276.

Kilgo, D. K., Yoo, J. J., Sinta, V., Geise, S., Suran, M. & Johnson, T. J. (2016). Led it on Reddit : An exploratory study examining opinion leadership on Reddit. *First Monday*. https://doi.org/10.5210/fm.v21i9.6429

kinsi55. (2014). *What are the valid username-characters ?* r/help. Récupérée 4 avril 2020, à partir de https://www.reddit.com/r/help/comments/1ttv80/what_are_the_valid_usernamecharacters/

Kipers, P. S. (1987). Gender and topic. *Language in Society*, *16*(4), jstor 4167882, 543-557.

Kircher, M. M. (2017, mai 17). *WhAt Is Up WiTh ThAt WeIrD, NeW SpOnGeBoB MeMe ?* Intelligencer. Récupérée 4 septembre 2020, à partir de https://nymag.com/intelligencer/2017/05/what-is-the-mocking-spongebob-capitalized-letters-chicken-meme.html

Kitzinger, C. (2005). Speaking as a heterosexual : (How) does sexuality matter for talk-in-Interaction ? *Research on language and social interaction*, *38*(3), 221-265.

Koerber, B. (2014, mars 10). *The TL ;DR guide to Reddit lingo*. Mashable. Récupérée 15 mai 2020, à partir de https://mashable.com/2014/03/10/reddit-lingo-guide/

Kolko, B. E., Nakamura, L. & Rodman, G. B. (Éd.). (2000). *Race in cyberspace*. New York, Routledge.

Konzack. (2014). The origins of geek culture : Perspectives on a parallel intellectual milieu (WyrdCon), In *Wyrd Con Companion Book* (WyrdCon). Costa Mesa.

Krapp, G. P. (1925). *The English language in America*. The Century Co., for the Modern Language Association of America.

Kreidler, C. W. (2000). Clipping and acronymy (G. E. Booij, C. Lehmann, J. Mugdan, W. Kesselheim & S. Skopeteas, Éd.). In G. E. Booij, C. Lehmann, J. Mugdan, W. Kesselheim & S. Skopeteas (Éd.), *Morphologie–Morphology : An international handbook of inflection and word-formation*. Berlin/New York, Walter de Gruyter.

Kunneman, F., Liebrecht, C., van Mulken, M. & van den Bosch, A. (2015). Signaling sarcasm : From hyperbole to hashtag. *Information Pro-*

cessing & Management, *51*(4), 500-509. https://doi.org/10.1016/j.ipm.2014.07.006

Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC, Center for applied linguistics.

Labov, W. (1972). *Language in the inner city : Studies in the Black English vernacular*. University of Pennsylvania Press.

Labov, W. (1989). The child as linguistic historian. *Language Variation and Change*, *1*(1), 85-97. https://doi.org/10.1017/S0954394500000120

Labov, W. (2001). *Principles of linguistic change, volume 2 : Social factors*. Malden, Blackwell.

Labov, W. (2006). *The social stratification of English in New York City* (2e éd.). Cambridge University Press.

Lakoff, R. T. (1973). Language and woman's place. *Language in Society*, *2*(1), 45-79. https://doi.org/doi:10.1017/S0047404500000051

Lakoff, R. T. (2004). *Language and woman's place : Text and commentaries* (T. 3). Oxford University Press, USA.

Laks, B. (1977). Contribution empirique à l'analyse socio-différentielle de la chute des /r/ dans les groupes consonantiques finals. *Langue française*, *34*(1), 109-125. https://doi.org/10.3406/lfr.1977.4819

Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L. & Riedl, J. (2011). WP : clubhouse ? : an exploration of Wikipedia's gender imbalance, In *Proceedings of the 7th international symposium on Wikis and open collaboration*, ACM. Récupérée 21 mars 2017, à partir de http://dl.acm.org/citation.cfm?id=2038560

Landis, C. (1927). National differences in conversations. *The Journal of Abnormal and Social Psychology*, *21*(4), 354.

Landis, M. & Burtt, H. E. (1924). A study of conversations. *Journal of Comparative Psychology*, *4*(1), 81.

larbarr. (2014, mars 14). *Girls of Reddit, what's the deal with repeating multiple letters in texts to guysssssss ?* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/AskReddit/comments/20f2qi/girls_of_reddit_whats_the_deal_with_repeating/

Larmarange, J. (p. d.). *Régression logistique binaire, multinomiale et ordinale*. analyse-R. Récupérée 9 juillet 2020, à partir de https://larmarange.github.io/analyse-R/regression-logistique.html

Lave, J. & Wenger, E. (1991). *Situated learning : Legitimate peripheral participation*. Cambridge University Press.

Le règlement général sur la protection des données - RGPD | CNIL. (p. d.). Récupérée 5 juillet 2020, à partir de https://www.cnil.fr/fr/reglement-europeen-protection-donnees

Lê, S., Josse, J. & Husson, F. (2008). FactoMineR : An R package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1-18.

Leap, W. (1996). *Word's out : Gay men's english*. University of Minnesota Press.

Leap, W. L. (1993). *American Indian English*. University of Utah Press.

Learn how to speak Reddit. (2012, avril 11). The Daily Dot. Récupérée 15 mai 2020, à partir de https://www.dailydot.com/unclick/reddit-glossary-ama-iama-redditors-guide/

Leavitt, A. (2015). "This is a throwaway account" : temporary technical identities and perceptions of anonymity in a massive online community, ACM Press. https://doi.org/10.1145/2675133.2675175

Leech, G. N. (p. d.). 100 million words of English : the British National Corpus (BNC). *Language Research*, *1*, 1-13.

Lester, J. A. (1922). A study of high school spelling material. *Journal of Educational Psychology*, *13*(2), 65.

Levon, Erez. (2012). The voice of others : identity, alterity and gender normativity among gay men in Israel. *Language in Society*, *41*(2), 187-211.

Levon, E. (2014). The politics of prosody : language, sexuality and national belonging in Israel. *Queer Excursions : retheorizing binaries in language, gender and sexuality*, 101-28. Récupérée 1 août 2017, à partir de http://linguistics.sllf.qmul.ac.uk/media/sllf-migration/department-of-linguistics/16-QMOPAL-Levon-Prosody.pdf

Levon, E. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, *9*(7), 295-308. https://doi.org/10.1111/lnc3.12147

Levon, E. & Mendes, R. B. (2015). Introduction (E. Levon & R. B. Mendes, Éd.). In E. Levon & R. B. Mendes (Éd.), *Language, Sexuality, and Power*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190210366.003.0001

Levshina, N. (2015). *How to do linguistics with R : Data exploration and statistical analysis*. John Benjamins Publishing Company.

Levy, I. (2019, décembre 16). *Eugenics and the ethics of statistical analysis*. Georgetown Public Policy Review. Récupérée 20 juillet 2020, à partir de http://gppreview.com/2019/12/16/eugenics-ethics-statistical-analysis/

Lin, Y.-L. ((2016). Non-standard capitalisation and vocal spelling in intercultural computer-mediated communication. *Corpora*, *11*(1), 63-82. https://doi.org/10.3366/cor.2016.0085

Linneman, T. J. (2013). Gender in Jeopardy ! : Intonation variation on a television game show. *Gender & Society*, *27*(1), 82-105. https://doi.org/10.1177/0891243212464905

List of emoticons. (2020, avril 8). In *Wikipedia*. Récupérée 20 avril 2020, à partir de https://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=949712309
Page Version ID : 949712309

Little, G. D. (1986). The ambivalent apostrophe. *English Today*, *2*(4), 15-17. https://doi.org/10.1017/S0266078400002388

lme4 convergence warnings : troubleshooting. (p. d.). rpubs.com. Récupérée 22 avril 2020, à partir de https://rstudio-pubs-static.s3.amazonaws.com/33653_57fc7b8e5d484c909b615d8633c01d51.html

Logan, J. R. (2007). Who are the other African Americans ? Contemporary African and Caribbean immigrants in the United States (Y. Shaw-Taylor & S. A. Tuch, Éd.). In Y. Shaw-Taylor & S. A. Tuch (Éd.), *The other African Americans : Contemporary African and Caribbean immigrants in the United States*. Rowman & Littlefield.

López, N., Erwin, C., Binder, M. & Chavez, M. J. (2018). Making the invisible visible : advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*(2), 180-207. https://doi.org/10.1080/13613324.2017.1375185

Lukač, M. (2014). Apostrophe(')s, who needs them ? : A further invitation to contribute to questions studied by the 'Bridging the Unbridgeable' Project at the Leiden Centre for Linguistics. *English Today*, *30*(3), 3-4. https://doi.org/10.1017/S0266078414000200

Macaulay, R. K. (1977). *Language, social class, and education : A Glasgow study*. Edinburgh University Press.

Maltz, D. N. & Borker, R. A. (1982). A cultural approach to male-female miscommunication, In *A cultural approach to interpersonal communication : Essential readings*. Wiley Malden, MA.

Manuel de TXM. (2018). ENS de Lyon & Université de Franche-Comté. http : / / textometrie . ens - lyon . fr / files / documentation / Manuel % 20de%20TXM%200.7%20FR.pdf

Marcellin, R., R. Bauer, G. & I. Scheim, A. (2013). Intersecting impacts of transphobia and racism on HIV risk among trans persons of colour in Ontario, Canada (D. Carol Mutch and Dr Jay Marlowe, Éd.). *Ethnicity and Inequalities in Health and Social Care*, *6*(4), 97-107. https://doi.org/10.1108/EIHSC-09-2013-0017

Marciano, A. (2014). Living the VirtuReal : Negotiating transgender identity in cyberspace. *Journal of Computer-Mediated Communication*, *19*(4), 824-838. https://doi.org/10.1111/jcc4.12081

Marger, M. N. (2014). *Race and ethnic relations : American and global perspectives* (10ème). Cengage Learning.

Margolis, J. & Fisher, A. (2002). *Unlocking the clubhouse : women in computing*. Cambridge, Mass, MIT Press.

Markdown. (2019, décembre 24). In *Wikipedia*. Récupérée 20 janvier 2020, à partir de https://en.wikipedia.org/w/index.php?title=Markdown&oldid=932190046
Page Version ID : 932190046

markdown - reddit.com. (p. d.). Récupérée 20 janvier 2020, à partir de https://www.reddit.com/wiki/markdown

Marwick, A. E. (2017). Scandal or sex crime ? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, *19*(3), 177-191. https://doi.org/10.1007/s10676-017-9431-7

Massanari, A. (2017). #Gamergate and The Fappening : How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, *19*(3), 329-346. Récupérée 6 juillet 2018, à partir de http://journals.sagepub.com/doi/10.1177/1461444815608807

Mather, P.-A. (2012). The social stratification of /r/ in New York city : Labov's department store study revisited. *Journal of English Linguistics*, *40*(4), 338-356. https://doi.org/10.1177/0075424211431265

Mattiello, E. (2013). *Extra-grammatical morphology in English : Abbreviations, blends, reduplicatives, and related phenomena*. Berlin/Boston, De Gruyter Mouton.

MC808. (2013, octobre 31). *r - How do I add confidence intervals to odds ratios in stargazer table ?* Stack Overflow. Récupérée 9 juillet 2020, à partir de https://stackoverflow.com/questions/19576356/how-do-i-add-confidence-intervals-to-odds-ratios-in-stargazer-table

McCall, L. (2005). The complexity of intersectionality. *Signs : Journal of Women in Culture and Society*, *30*(3), 1771-1800.

McCarthy, M. M. & Konkle, A. T. (2005). When is a sex difference not a sex difference ? *Frontiers in neuroendocrinology*, *26*(2), 85-102.

McCulloch, G. (2019). *Because internet : Understanding the new rules of language*. New York, Riverhead Books.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics : method, theory and practice*. Cambridge ; New York, Cambridge University Press OCLC : ocn732967848.

McGlashan, H. & Fitzpatrick, K. (2018). I use any pronouns, and I'm questioning everything else : transgender youth and the issue of gender pronouns. *Sex Education*, *18*(3), 239-252. https://doi.org/10.1080/14681811.2017.1419949

meemersbarnhart. (2013). *A New User's Guide to Reddit : redditguides*. Reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/redditguides/comments/u6e0r/a_new_users_guide_to_reddit/

Meil, P. (1984). *A systematic observational study of sex differences in conversation topic*. Ann Arbor, Manuscrit non publié. University of Michigan.

Mendoza-Denton, N. (2012). Pregnant pauses : Silence and authority in the Anita Hill–Clarence Thomas hearings (K. Hall & M. Bucholtz, Éd.). In K. Hall & M. Bucholtz (Éd.), *Gender articulated*. Routledge.

Mendoza-Denton, N. (2014). *Homegirls : Language and cultural practice among Latina youth gangs*. John Wiley & Sons.

Menking, A. & Erickson, I. (2015). The heart work of Wikipedia : Gendered, emotional labor in the world's largest online encyclopedia, In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*.

Meyer, D., Zeileis, A. & Hornik, K. (2020). *vcd : Visualizing categorical data* (Version R package version 1.4-7.).

Meyerhoff, M. & Ehrlich, S. (2019). Language, gender, and sexuality. *Annual Review of Linguistics*, *5*, 455-475.

Meyerhoff, M. & Stanford, J. (2015). "Tings change, all tings change" : the changing face of sociolinguistics with a global perspective (D. Smakman & P. Heinrich, Éd.). In D. Smakman & P. Heinrich (Éd.), *Globalising sociolinguistics : Challenging and expanding theory*. London, Routledge.

Milroy, L. (1980). *Language and social networks*. Oxford, Blackwell.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users, In *ICWSM*.

Mitchell, C. & Reid-Walsh, J. (2005). Theorizing tween culture within girlhood studies. *Counterpoints*, *245*jstor 42978689, 1-21.

Mock, C. (1979). The social maturation of pronunciation : a family case study. *The Rural Learner*, *1*, 23-27.

Money, J. & Ehrhardt, A. A. (1972). *Man and woman, boy and girl : Differentiation and dimorphism of gender identity from conception to maturity*. John Hopkins University Press.

Moore, H. T. (1922). Further data concerning sex differences. *The Journal of Abnormal Psychology and Social Psychology*, *17*(2), 210-214. https://doi.org/10.1037/h0064645

Moreau, M.-L. (Éd.). (1997). *Sociolinguistique : les concepts de base*. Mardaga.

Morgan, M. (1996). Conversational signifying : Grammar and indirectness among African American women (E. Ochs, E. Schegloff & S. Thompson, Éd.). *Interaction and Grammar*, 405-433.

Morgan, M. (2004). "I'm every woman" black women's (dis) placement in women's language study. In M. Bucholtz (Éd.), *Language and woman's place : Text and commentaries* (p. 252-259). Oxford University Press.

Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. Chapman & Hall/CRC.

Natalie. (2018, mai 7). *Keeping it one hundred : Urban Dictionary, mining Black internet culture and why "fleek" still...* Medium. Récupérée 11 septembre 2020, à partir de https://medium.com/@natalieisonline/keeping-it-one-hundred-b46158511e37

Nelson, A. (2005). Children's toy collections in Sweden : A less gender-typed country ? *Sex Roles*, *52*(1-2), 93-102. https://doi.org/10.1007/s11199-005-1196-5

Nevalainen, T. (1996). Gender difference (T. Nevalainen & H. Raumolin-Brunberg, Éd.). In T. Nevalainen & H. Raumolin-Brunberg (Éd.), *Sociolinguistics and language history : Studies based on the Corpus of Early English Correspondence*. Amsterdam, Rodopi.

Newport, F. (2018, mai 22). *In U.S., Estimate of LGBT Population Rises to 4.5%*. Gallup.com. Récupérée 5 juillet 2020, à partir de https://news.gallup.com/poll/234863/estimate-lgbt-population-rises.aspx

Newton, C. (2020, juin 29). *Reddit bans r/The_Donald and r/ChapoTrapHouse as part of a major expansion of its rules*. The Verge. Récupérée 7 juillet 2020, à partir de https://www.theverge.com/2020/6/29/21304947/reddit-ban-subreddits-the-donald-chapo-trap-house-new-content-policy-rules

Nguyen, D., Doğruöz, A. S., Rosé, C. P. & de Jong, F. (2016). Computational sociolinguistics : A survey. *Computational Linguistics*, *42*(3), 537-593. https://doi.org/10.1162/COLI_a_00258

Nguyen, D., Gravel, R., Trieschnigg, D. & Meder, T. (2013). How old do you think I am ? : A study of language and age in Twitter, In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, Boston, MA.

Nguyen, D., Trieschnigg, D., Dogruoz, A. S., Gravel, R., Theune, M., Meder, T. & Jong, F. D. (2014). Why gender and age prediction from tweets is hard : Lessons from a crowdsourcing experiment, In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, Dublin.

Nichols, P. C. (1978). Black Women in the Rural South : Conservative and Innovative. *International Journal of the Sociology of Language*, *1978*(17). https://doi.org/10.1515/ijsl.1978.17.45

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of Pragmatics*, *41*(5), 866-891. https://doi.org/10.1016/j.pragma.2008.08.005

Norwood, C. (2019, juin). *How governments are transitioning their gender policies to nonbinary*. Governing.com. Récupérée 25 août 2020, à partir de https://www.governing.com/topics/health-human-services/gov-nonbinary-lgbtq-legislation-regulations.html

Novaraa. (2017, mars 31). *Why is there an unspoken agreement to never use emojis on Reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/AskReddit/comments/62l8q2/why_is_there_an_unspoken_agreement_to_never_use/

Oakley, A. (2016). Disturbing hegemonic discourse : Nonbinary gender and xexual orientation labeling on Tumblr. *Social Media + Society*, *2*(3), 2056305116664217. https://doi.org/10.1177/2056305116664217

O'Barr, W. M. & Atkins, B. K. (1998). "Women's language or "powerless language"? (J. Coates, Éd.). In J. Coates (Éd.), *Language and gender : A reader*. Oxford, Blackwell.

O'Connor, M. (2013, février 21). *The 5 reasons girls type like thissss*. The Cut. Récupérée 1 août 2020, à partir de https://www.thecut.com/2013/02/5-reasons-girls-type-like-thissss.html

Office, U. C. B. P. I. (2011, septembre 29). *2010 Census shows Black population has highest concentration in the South*. Récupérée 8 juillet 2020, à partir de https://www.census.gov/newsroom/releases/archives/2010_census/cb11-cn185.html

Ogletree, S. M., Fancher, J. & Gill, S. (2014). Gender and texting : Masculinity, femininity, and gender role ideology. *Computers in Human Behavior*, *37*, 49-55. https://doi.org/10.1016/j.chb.2014.04.021

Oleszkiewicz, A., Karwowski, M., Pisanski, K., Sorokowski, P., Sobrado, B. & Sorokowska, A. (2017). Who uses emoticons? Data from 86702 Facebook users. *Personality and Individual Differences*, *119*, 289-295. https://doi.org/10.1016/j.paid.2017.07.034

Olson, R. & King, R. (2017, septembre 22). *How The Internet* Talks*. FiveThirtyEight. Récupérée 28 avril 2018, à partir de https://projects.fivethirtyeight.com/reddit-ngram/

omnikan. (2017, août 24). *What are some reddit slang everyone should know ?* Reddit. Récupérée 14 mai 2020, à partir de https://www.reddit.com/r/AskReddit/comments/6vn8hc/what_are_some_reddit_slang_everyone_should_know/

Original Bboard Thread in which :-) was proposed. (p. d.). Récupérée 30 septembre 2020, à partir de https://www.cs.cmu.edu/~sef/Orig-Smiley.htm

Overbeck, A. (2015). La communication dans les médias électroniques. In *Manuel de linguistique française* (p. 275-292).

Overdorf, R. & Greenstadt, R. (2016). Blogs, Twitter feeds, and Reddit comments : Cross-domain authorship attribution. *Proceedings on Privacy Enhancing Technologies*, *2016*(3), 155-171. https://doi.org/10.1515/popets-2016-0021

Palander-Collin, M. (1999). Male and female styles in 17th century correspondence : I THINK. *Language Variation and Change*, *11*(2), 123-141. https://doi.org/10.1017/S0954394599112018

Parham, J. (2020). TikTok and the evolution of digital blackface [magazine]. *Wired*. Récupérée 11 septembre 2020, à partir de https://www.wired.com/story/tiktok-evolution-digital-blackface

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., Stillwell, D., Ungar, L. H. & Seligman, M. E. P. (2016). Women are warmer but no less assertive than men : Gender and language on Facebook (C. M. Danforth, Éd.). *PLOS ONE*, *11*(5). https://doi.org/10.1371/journal.pone.0155885

Parkins, R. (2012). Gender and emotional expressiveness : An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication*, *5*(1), 46-54.

Paunonen, H. (1994). The Finnish language in Helsinki (B. Nordberg, Éd.). In B. Nordberg (Éd.), *The sociolinguistics of urbanization : the case of the Nordic countries*. Walter de Gruyter.

Pavalanathan, U. & Eisenstein, J. (2015). *Emoticons vs. emojis on Twitter : A causal inference approach*. Récupérée 21 mars 2017, à partir de https://arxiv.org/abs/1510.08480

Pavalanathan, U. & Eisenstein, J. (2016). More emojis, less :) The competition for paralinguistic function in microblog writing. *First Monday*, *21*(11). Récupérée 21 janvier 2018, à partir de https://firstmonday.org/ojs/index.php/fm/article/view/6879

Peersman, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B. & Van Vaerenbergh, L. (2016). *The effects of age, gender and region on non-standard linguistic variation in online social networks*. Récupérée 19 mars 2017, à partir de https://arxiv.org/abs/1601.02431

Peters, P. (2004). *The Cambridge guide to English usage*. Cambridge University Press.

Phinney, J. S. & Onwughalu, M. (1996). Racial identity and perception of American ideals among African American and African students in the United States. *International Journal of Intercultural Relations*, *20*(2), 127-140.

pi's Yet Another Bloody List of Acronyms. (p. d.). Récupérée 20 avril 2020, à partir de http://piology.org/yabla.txt

Picone, M. D. (2016). Eye dialect and pronunciation respelling in the USA (V. Cook & D. Ryan, Éd.). In V. Cook & D. Ryan (Éd.), *The Routledge handbook of the English writing system*. Routledge.

Pilcher, P. (2009). *Talking young femininities*. London, Palgrave Macmillan.

Plag, I. (2003). *Word-formation in English*. Cambridge, Cambridge University Press.

Podesva, R. J. & Van Hofwegen, J. (2014). How conservatism and normative gender constrain variation in inland California : The case of/s. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 129-137. Récupérée 1 août 2017, à partir de http://repository.upenn.edu/pwpl/vol20/iss2/15/

Poplack, S. (1978). Dialect acquisition among Puerto Rican bilinguals. *Language in society*, *7*(1), 89-103.

Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B. & Gaspar, R. (2018). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics*, *35*(7), 1925-1934. https://doi.org/10.1016/j.tele.2018.06.005

Pratt-Johnson, Y. (2005). The growing use of uptalk in the United States : Language trend or shift, In *Language and global communication conference*.

Preston, D. R. (2000). Mowr and mowr bayud spellin' : Confessions of a sociolinguist. *Journal of Sociolinguistics*, *4*(4), 615-621. https://doi.org/10.1111/1467-9481.00132

Provine, R. R., Spencer, R. J. & Mandell, D. L. (2007). Emotional expression online : Emoticons punctuate website text messages. *Journal of Language and Social Psychology*, *26*(3), 299-307. https://doi.org/10.1177/0261927X06303481

Pyke, K. & Dang, T. (2003). "FOB" and "Whitewashed" : Identity and internalized racism among second generation Asian Americans. *Qualitative Sociology*, *26*(2), 147-172. https://doi.org/10.1023/A:1022957011866

Quarantined Subreddits. (p. d.). Reddit Help. Récupérée 7 juillet 2020, à partir de https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits

R Core Team. (2018). *R : A language and environment for statistical computing*. manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R FAQ. (p. d.). The Comprehensive R Archive Network. Récupérée 2 janvier 2020, à partir de https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f

Race and ethnicity. (p. d.). United States Census Bureau. Récupérée 3 août 2020, à partir de https://data.census.gov/cedsci/profile?q=United%20States&g=0100000US

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B. & Daelemans, W. (2015). Overview of the 3rd author profiling task at PAN 2015 (L.

Cappellato, N. Ferro, G. Jones & E. San Juan, Éd.). In L. Cappellato, N. Ferro, G. Jones & E. San Juan (Éd.), *CLEF 2015 Labs and Workshops, Notebook Papers*.

Rankin, S. & Beemyn, G. (2012). Beyond a binary : The lives of gender-nonconforming youth. *About Campus*, *17*(4), 2-10. https://doi.org/10.1002/abc.21086

Rao, D., Yarowsky, D., Shreevats, A. & Gupta, M. (2010). Classifying latent user attributes in Twitter, In *In 2nd International Workshop on Search and Mining UserGenerated Content. ACM*.

Rayson, P., Leech, G. N. & Hodges, M. (1997). Social differentiation in the use of English vocabulary : some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, *2*(1), 133-152.

Reddiquette. (p. d.). reddit. Récupérée 29 décembre 2019, à partir de https://www.reddit.com/wiki/reddiquette/

reddit. (2012, février 12). *A necessary change in policy : blog*. reddit. Récupérée 7 juillet 2020, à partir de https://www.reddit.com/r/blog/comments/pmj7f/a_necessary_change_in_policy/?st=jk1aqjk0&sh=747b9941

Reddit Content Policy. (p. d.). Reddit. Récupérée 30 avril 2020, à partir de https://www.redditinc.com/policies/content-policy

Reddit rolls out user profiles amid site makeover [newspaper]. (2017). *Reuters*. Récupérée 20 janvier 2020, à partir de https://www.reuters.com/article/us-reddit-profiles-idUSKBN16S1QO

Reese, S., Boleda, G., Cuadros, M. & Rigau, G. (2010). Wikicorpus : A word-sense disambiguated multilingual Wikipedia corpus.

Resolution Concerning Updating the International Standard Classification of Occupations. (p. d.).

Reyes, A. (2005). Appropriation of African American slang by Asian American youth. *Journal of Sociolinguistics*, *9*(4), 509-532. Récupérée 10 avril 2017, à partir de http://onlinelibrary.wiley.com/doi/10.1111/j.1360-6441.2005.00304.x/full

Rezabeck, L. L. & Cochenour, J. J. (1995). Emoticons : Visual Cues for Computer-Mediated Communication. Récupérée 21 mars 2017, à partir de http://eric.ed.gov/?id=ED380096

RFC Editor Abbreviations List. (2020, avril). Récupérée 20 avril 2020, à partir de https://www.rfc-editor.org/materials/abbrev.expansion.txt

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S. & Zannettou, S. (2020). *The evolution of the Manosphere across the web*. Récupérée 7 octobre 2020, à partir de http://arxiv.org/abs/2001.07600

Richards, C., Bouman, W. P. & Barker, M.-J. (Éd.). (2017). *Genderqueer and non-binary genders*. London, Palgrave Macmillan.

Richterich, A. (2014). 'Karma, precious karma!' Karmawhoring on Reddit and the Front Page's econometrisation. *Journal of Peer Produc-*

tion, *4*(1). Récupérée 26 septembre 2017, à partir de http://www.academia.edu/download/42669175/Richterich_2014_Reddit.pdf

Rinker, T. W. (2018). *lexicon : Lexicon Data* (Version 1.2.1). http://github.com/trinker/lexicon

Rintel, E. S., Mulholland, J. & Pittam, J. (2001). First things first : Internet Relay Chat openings. *Journal of Computer-Mediated Communication*, *6*(3). https://doi.org/10.1111/j.1083-6101.2001.tb00125.x

Riordan, M. A. & Kreuz, R. J. (2010). Cues in computer-mediated communication : A corpus analysis. *Computers in Human Behavior*, *26*(6), 1806-1817. https://doi.org/10.1016/j.chb.2010.07.008

Rivard, L. (2014, septembre 18). *America's Hidden History : The Eugenics Movement | Learn Science at Scitable*. nature.com. Récupérée 20 juillet 2020, à partir de https://www.nature.com/scitable/forums/genetics-generation/america-s-hidden-history-the-eugenics-movement-123919444/

Ronkin, M. & Karn, H. E. (2002). Mock Ebonics : Linguistic racism in parodies of Ebonics on the Internet. *Journal of Sociolinguistics*, *3*(3), 360-380. https://doi.org/10.1111/1467-9481.00083

Rose, N. A. (2020, juillet 22). *AAVE Is More Than Just Your Internet Slang*. dear dark skinned girl. Récupérée 11 septembre 2020, à partir de https://deardarkskinnedgirl.com/2020/07/22/aave-is-more-than-just-your-internet-slang/

Rosen, L. D., Chang, J., Erwin, L., Carrier, L. M. & Cheever, N. A. (2010). The relationship between "textisms" and formal and informal writing among young adults. *Communication Research*, *37*(3), 420-440. https://doi.org/10.1177/0093650210362465

Rosenthal, S. & McKeown, K. (2011). Age prediction in blogs : A study of style, content, and online behavior in pre-and post-social media generations, In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, Association for Computational Linguistics. Récupérée 19 mars 2017, à partir de http://dl.acm.org/citation.cfm?id=2002569

Rouhani, S. (2015). *Intersectionality-informed quantitative research : A primer*. The Institute for Intersectionality Research & Policy.

RStudio Team. (2016). *RStudio : Integrated development environment for r*. manual. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Rubin, D. L. & Greene, K. (1992). Gender-typical style in written language. *Research in the Teaching of English*, *26*(1), jstor 40171293, 7-40.

Rust, J., Golombok, S., Hines, M., Johnston, K., Golding, J., Team, A. S. Et al. (2000). The role of brothers and sisters in the gender development of preschool children. *Journal of experimental child psychology*, *77*(4), 292-303.

s. (p. d.). Urban Dictionary. Récupérée 15 mai 2020, à partir de https://www.urbandictionary.com/define.php?term=%5Cs

Sacked Reddit employee speaks out [newspaper]. (2015). *BBC News : Technology*. Récupérée 7 juillet 2020, à partir de https://www.bbc.com/news/technology-33787004

Saisuwan, P. (2016). Kathoey and the linguistic construction of gender identity in Thailand (E. Levon & R. B. Mendes, Éd.). In E. Levon & R. B. Mendes (Éd.), *Language, sexuality, and power : Studies in intersectional sociolinguistics*. Oxford University Press.

Salinas Jr., C. (2019). Mapping and recontextualizing the evolution of the term latinx : an environmental scanning in higher education. *Journal of Latinos and Education*, *18*(4), 302-315. Récupérée 24 août 2020, à partir de https://www.tandfonline.com/doi/full/10.1080/15348431.2017.1390464

Salles, D. (2009). L'analyse factorielle des correspondances simples. In *Analyse factorielle simple en sociologie méthodes d'interprétation et études de cas* (p. 129-239). Bruxelles, De Boeck
OCLC : 495330588.

Sankoff, G. (2006). *Age : Apparent time and real time* (K. Brown, Éd.). In K. Brown (Éd.), *Encyclopedia of language and linguistics*. Elsevier.

Sankoff, G. & Thibault, P. (1977). L'alternance entre les auxiliaires "avoir" et "être" en francais parlé à Montréal. *Langue française*, (34), 81-108.

Santa Ana, O. & Bayley, R. (2004). Chicano english : Phonology. *A handbook of varieties of English*, *1*, 417-434.

sballens. (2018, mai 10). *Why doesn't anyone use emojis on reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/NoStupidQuestions/comments/8iazpv/why_doesnt_anyone_use_emojis_on_reddit/

Schilling-Estes, N. (2004). Constructing ethnicity in interaction. *Journal of Sociolinguistics*, *8*(2), 163-195.

Schilt, K. (2010). *Just one of the guys ? : Transgender men and the persistence of gender inequality*. Chicago, University of Chicago Press.

Schmid, H. J. (2003). Do women and men really live in different cultures ? Evidence from the BNC (A. Wilson, R. Rayson & T. McEnery, Éd.). In A. Wilson, R. Rayson & T. McEnery (Éd.), *Corpus linguistics by the Lune*. Lódz Studies in Language.

Schmid, H. (p. d.). *TreeTagger* (Version 3.2). Institute for Computational Linguistics of the University of Stuttgart. https://www.ims.uni-stuttgart.de/en/research/resources/tools/treetagger/

Schnoebelen, J. (2012). *Emotions are relational : Positioning and the use of affective linguistic resources*. Stanford University. Stanford. Récupérée 2 janvier 2019, à partir de https://stacks.stanford.edu/file/druid:fm335ct1355/Dissertation_Schnoebelen_final_8-29-12-augmented.pdf

Schnoebelen, T. (2012). Do you smile with your nose ? Stylistic variation in twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, *18*(2), 118-125.

Schulman, K. A., Berlin, J. A., Harless, W., Kerner, J. F., Sistrunk, S., Gersh, B. J., Dube, R., Taleghani, C. K., Burke, J. E., Williams, S. Et al. (1999). The effect of race and sex on physicians' recom-

mendations for cardiac catheterization. *New England Journal of Medicine*, *340*(8), 618-626.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P. & Ungar, L. H. (2013). Personality, gender, and age in the language of social media : The open-vocabulary approach (T. Preis, Éd.). *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Scott, M. (2016). *WordSmith Tools version 7*.

Scott, N. A. & Siltanen, J. (2017). Intersectionality and quantitative methods : assessing regression from a feminist perspective. *International Journal of Social Research Methodology*, *20*(4), 373-385. https://doi.org/10.1080/13645579.2016.1201328

Sebba, M. (2007). *Spelling and society*. Cambridge, Cambridge University Press.

Seethaler, I. (2013). "Big bad Chinese mama" : How internet humor subverts stereotypes about Asian American women. *Studies in American Humor*, (27), jstor 23823982, 117-138.

Sehulster, J. R. (2006). Things we talk about, how frequently, and to whom : Frequency of topics in everyday conversation as a function of gender, age, and marital status. *The American Journal of Psychology*, *119*(3), jstor 20445351, 407-432. https://doi.org/10.2307/20445351

SenpaiThrowMeAway. (2019, juillet 7). *What does it REALLY mean if she adds multiple letters at the end of the words like 'heyyy'?* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/dating_advice/comments/ca38e3/what_does_it_really_mean_if_she_adds_multiple/et5lh47/

Settanni, M. & Marengo, D. (2015). Sharing feelings online : studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01045

Sexyoldmann. (2014, juillet 26). *Girls, why do you add extra letters to words like hey or hi?* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/teenagers/comments/2br3gu/discussion_girls_why_do_you_add_extra_letters_to/cj82lxr/

Shapiro, E. (2004). 'Trans' cending barriers : Transgender organizing on the Internet. *Journal of Gay & Lesbian Social Services*, *16*(3-4), 165-179. https://doi.org/10.1300/J041v16n03_11

Shortis, T. (2009). Spelling, vernacular orthography, and "unregimented writing" (S. Wheeler, Éd. ; IAT). In S. Wheeler (Éd.), *Connected Minds, Emerging Cultures : Cybercultures in Online Learning* (IAT).

Siegler. (2010, juin 9). *Reddit finally becomes usable on the mobile web — cause they hate the App Store*. TechCrunch. Récupérée 12 septembre 2020, à partir de https://social.techcrunch.com/2010/06/09/reddit-mobile/

Significant other. (2020, avril 9). In *Wikipedia*. Récupérée 30 juin 2020, à partir de https://en.wikipedia.org/w/index.php?title=Significant_

other&oldid=950032829
Page Version ID : 950032829

simbawulf. (2017, février 15). *Introducing r/popular*. redd. Récupérée 7 juillet 2020, à partir de https://www.reddit.com/r/announcements/comments/5u9pl5/introducing_rpopular/

simply-hopeless. (2019). *Can someone help me understand Reddit's... lingo, jargon, language, whatever you want to call it ?* Reddit. Récupérée 14 mai 2020, à partir de https://www.reddit.com/r/help/comments/9jzsyp/can_someone_help_me_understand_reddits_lingo/

Sklar, E. S. (1976). The possessive apostrophe : The development and decline of a crooked mark. *College English*, *38*(2), jstor 376342, 175-183. https://doi.org/10.2307/376342

Skovholt, K., Grønning, A. & Kankaanranta, A. (2014). The communicative functions of emoticons in workplace e-Mails : :-). *Journal of Computer-Mediated Communication*, *19*(4), 780-797. https://doi.org/10.1111/jcc4.12063

Sleeper, C. B. (1930). *Samplings of leisure-time conversations to find sex differences in drives*. Manuscrit non publié.

Smith, A. (2014, janvier 6). *African Americans and technology use*. Pew Research Center : Internet, Science & Tech. Récupérée 12 août 2020, à partir de https://www.pewresearch.org/internet/2014/01/06/african-americans-and-technology-use/

Social justice warrior. (2020, mai 13). In *Wikipedia*. Récupérée 14 mai 2020, à partir de https://en.wikipedia.org/w/index.php?title=Social_justice_warrior&oldid=956403580
Page Version ID : 956403580

Soyboy. (p. d.). Dictionary.com. Récupérée 7 juillet 2020, à partir de https://www.dictionary.com/e/slang/soyboy/

Spack, N. P., Edwards-Leeper, L., Feldman, H. A., Leibowitz, S., Mandel, F., Diamond, D. A. & Vance, S. R. (2012). Children and adolescents with gender identity disorder referred to a pediatric medical center. *PEDIATRICS*, *129*(3), 418-425. https://doi.org/10.1542/peds.2011-0907

spez. (p. d.). *Out with 2016, in with 2017*. reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/announcements/comments/5q4qmg/out_with_2016_in_with_2017/

Squires, L. (2007). *Whats the use of apostrophes ? Gender difference and linguistic variation in instant messaging*. American University TESOL Working Papers. Récupérée 21 mars 2017, à partir de http://www.leadership.american.edu/cas/tesol/pdf/upload/WP-2007-Squires-Instant-Messaging.pdf

Squires, L. (2012). Whos punctuating what ? Sociolinguistic variation in instant messaging (Jaffe), In *Orthography as Social Action : Scripts, Spelling, Identity and Power* (Jaffe). Mouton de Gruyter.

Steinmetz, K. (2015, novembre 16). *Oxford's 2015 Word of the Year Is This Emoji*. Time. Récupérée 18 juin 2020, à partir de https://time.com/4114886/oxford-word-of-the-year-2015-emoji/

Stirratt, M., Meyer, I. H., Ouellette, S. C. & Gara, M. A. (2008). Measuring identity multiplicity and intersectionality : Hierarchical Classes Analysis (HICLAS) of sexual, racial, and gender identities. *Self and Identity*, (7), 89-111.

Stoeffel, K. (2013, février 19). *How to tweet like a girl*. The Cut. Récupérée 1 août 2020, à partir de https://www.thecut.com/2013/02/how-to-tweet-like-a-girl.html

Stoke, S. M. & West, E. D. (1930). The conversational interests of college students. *School and Society*, *32*, 567-570.

Stommel, W. (2007). *Mein Nick bin ich !* Nicknames in a German forum on eating disorders. *Journal of Computer-Mediated Communication*, *13*(1), 141-162. https://doi.org/10.1111/j.1083-6101.2007.00390.x

Stryker, S. (2017). *Transgender history. The roots of today's revolution* (Second). New York, Seal Press.

Submit to Reddit. (p. d.). Récupérée 20 janvier 2020, à partir de https://new.reddit.com/submit

Subrahmanyam, K., Greenfield, P. M. & Tynes, B. (2004). Constructing sexuality and identity in an online teen chat room. *Journal of Applied Developmental Psychology*, *25*(6), 651-666. https://doi.org/10.1016/j.appdev.2004.09.007

subredditstats. (p. d.). *r/france*. subredditstats.com. Récupérée 13 septembre 2020, à partir de https://subredditstats.com/r/france

Sulfruous. (2017, mars 31). *Why is there an unspoken agreement to never use emojis on Reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/AskReddit/comments/62l8q2/why_is_there_an_unspoken_agreement_to_never_use/

SwillFish. (2017, janvier 26). *Mobile Reddit is terrible ! I need to rant.* reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/mobileweb/comments/5q8mek/mobile_reddit_is_terrible_i_need_to_rant/

Tagg, C. (2012). *Discourse of text messaging : Analysis of SMS communication*. London, Continuum.

Tagg, C., Baron, A. & Rayson, P. (2012). Analysis and normalisation of SMS spelling variation. *Lingvisticae Investigations*, *35*(2), 367-388.

Tagliamonte, S. A. (2016a). So sick or so cool ? The language of youth on the internet. *Language in Society*, *45*(01), 1-32. https://doi.org/10.1017/S0047404515000780

Tagliamonte, S. A. (2016b). *Teen talk : The language of adolescents*. Cambridge University Press.

Tagliamonte, S. A. & Baayen, R. H. (2012). Models, forests, and trees of York English : *Was/were* variation as a case study for statistical practice. *Language Variation and Change*, *24*(2), 135-178. https://doi.org/10.1017/S0954394512000129

Tagliamonte, S. A. & Denis, D. (2008). Linguistic ruin ? Lol ! Instant messaging and teen language. *American Speech*, *83*(1), 3-34. https://doi.org/10.1215/00031283-2008-001

Tannen, D. (1990). *You just don't understand : Women and men in conversation*. New York, Ballantine.

Tenbarge, K. (2020, janvier 26). *Internet slang like 'periodt' is rooted in cultural appropriation - Insider*. Insider. Récupérée 11 septembre 2020, à partir de https://www.insider.com/internet-slang-origin-i-oop-meaning-sksk-vsco-girls-stans-2020-1

Thelwall, M. (2008). Fk yea I swear : cursing and gender in MySpace. *Corpora*, *3*(1), 83-107.

Thelwall, M. & Stuart, E. (2018). *She's Reddit : A source of statistically significant gendered interest information ?* Récupérée 15 mars 2019, à partir de http://arxiv.org/abs/1810.08091

thendofthebeginning. (2018, août 29). *Which do you use more : mobile reddit or desktop reddit ?* reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/AskReddit/comments/9bbd57/which_do_you_use_more_mobile_reddit_or_desktop/

Thibault, P. (1997). Âge (M.-L. Moreau, Éd.). In M.-L. Moreau (Éd.), *Sociolinguistique : les concepts de base*. Mardaga.

ThirdEyeTrippyShit. (2014, décembre 12). *ELI5 : Why don't people use emojis on reddit ?* reddit. Récupérée 18 juin 2020, à partir de https://www.reddit.com/r/explainlikeimfive/comments/2p3btw/eli5_why_dont_people_use_emojis_on_reddit/

Thompson, C. A., Craig, H. K. & Washington, J. A. (2004). Variable production of African American English across oracy and literacy contexts. *Language, Speech, and Hearing Services in Schools*, *35*, 269-282.

Thompson, D. & Filik, R. (2016). Sarcasm in written communication : Emoticons are efficient markers of intention. *Journal of Computer-Mediated Communication*, *21*(2), 105-120. https://doi.org/10.1111/jcc4.12156

Thompson, R. & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology*, *40*, 193-208.

Thurlow, C. (2003). Generation txt ? The sociolinguistics of young people's text-messaging. https://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html?report=reader#

Thurlow, C. (2014). Determined creativity : Language play in new media discourse (R. Jones, Éd. ; Routledge). In R. Jones (Éd.), *Discourse and creativity* (Routledge). London ; New York.

TIOBE Index. (p. d.). TIOBE - The Software Quality Company. Récupérée 2 janvier 2020, à partir de https://www.tiobe.com/tiobe-index/

Tirado, F. (2019, juin 10). *'Queer eye"s Jonathan Van Ness : "I'm nonbinary"*. Récupérée 25 août 2020, à partir de https://www.out.com/lifestyle/2019/6/10/queer-eyes-jonathan-van-ness-im-nonbinary

Tognini-Bonelli, E. (2001). *Corpus linguistics at work* (T. 6). John Benjamins Publishing.

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H., Rahmati, A. & Zhong, L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior*, *28*(2), 659-663. https://doi.org/10.1016/j.chb.2011.11.012

Tousignant, C. (1987). *La variation sociolinguistique : modèle québécois et méthode d'analyse*. Presses de l'Université du aQuébec.

Transgender acronyms. (2019, avril 16). Transgender Map. Récupérée 20 avril 2020, à partir de https://www.transgendermap.com/resources/words/acronyms/

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge, Cambridge University Press. https://doi.org/10.1007/978-1-349-25582-5_15

Tsou, A. (2016). How does the front page of the Internet behave ? Readability, emoticon use, and links on Reddit. *First Monday*, *21*(11). Récupérée 28 octobre 2017, à partir de http://journals.uic.edu/ojs/index.php/fm/article/view/7013

Tuan, M. (1998). *Forever foreigners or honorary whites ? : the Asian ethnic experience today*. Rutgers University Press.

Turnage, A. K. (2007). Email flaming behaviors and organizational conflict. *Journal of Computer-Mediated Communication*, *13*(1), 43-59. https://doi.org/10.1111/j.1083-6101.2007.00385.x

The TV Corpus. (p. d.). English-Corpora. Récupérée 30 juin 2020, à partir de https://www.english-corpora.org/tv/

Tyler, J. C. (2015). Expanding and mapping the indexical field : Rising pitch, the uptalk stereotype, and perceptual variation. *Journal of English Linguistics*, *43*(4), 284-310.

U.S. Census Bureau QuickFacts. (p. d.). United States Census Bureau. Récupérée 14 mai 2020, à partir de https://www.census.gov/quickfacts/fact/table/US/PST045219#PST045219

Uchida, A. (1992). When "difference" is "dominance" : A critique of the "anti-power-based" cultural approach to sex differences. *Language in Society*, *21*(4), jstor 4168392, 547-568.

Ungratefulpanda. (2018, août 31). *Is Reddit making its mobile site worse on purpose to force people to use the app ?* reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/NoStupidQuestions/comments/9bs5bz/is_reddit_making_its_mobile_site_worse_on_purpose/

Van Hofwegen, J. & Wolfram, W. (2010). Coming of age in African American English : A longitudinal study. *Journal of Sociolinguistics*, *14*(4), 427-455. https://doi.org/10.1111/j.1467-9841.2010.00452.x

Varma, R. (2007). Women in computing : The role of geek culture. *Science as Culture*, *16*(4), 359-376. Récupérée 6 juillet 2018, à partir de http://www.tandfonline.com/doi/abs/10.1080/09505430701706707

Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H. & Kwong, T. E. (2010). Lol : new language and spelling in instant messaging. *Reading and Writing*, *23*(6), 719-733. https://doi.org/10.1007/s11145-009-9181-y

Veenstra, G. (2011). Race, gender, class, and sexual orientation : intersecting axes of inequality and self-rated health in Canada. *International Journal for Equity in Health*, *10*pmid 21241506, 3. https://doi.org/10.1186/1475-9276-10-3

Venables, W. & Ripley, B. (2002). *Modern Applied Statistics with S* (Fourth edition). New York, Springer.

Verheijen, L. (2017). WhatsApp with social media slang ? Youth language use in Dutch written computer-mediated communication (D. Fišer & M. Beißwenger, Éd.). In D. Fišer & M. Beißwenger (Éd.), *Investigating Computer-Mediated Communication : Corpus-based Approaches to Language in the Digital World*. Ljubljana University Press.

Wang, Y.-C., Burke, M. & Kraut, R. (2013). Gender, topic, and audience response : An analysis of user-generated content on Facebook. ACM Conference on Human Factors in Computing Systems (CHI).

Warner, D. F. & Brown, T. H. (2011). Understanding how race/ethnicity and gender define age-trajectories of disability : An intersectionality approach. *Social Science & Medicine*, *72*(8), 1236-1248. https://doi.org/10.1016/j.socscimed.2011.02.034

Warren, P. (2016). *Uptalk : The phenomenon of rising intonation*. Cambridge University Press.

Watson, J., Breed, W. & Posman, H. (1948). A study in urban conversation : Sample of 1,001 remarks overheard in Manhattan. *The Journal of social psychology*, *28*(1), 121-133.

West, C. & Zimmerman, D. H. (1983). Small insults : A study of interruptions in cross-sex conversations between unacquainted persons. (B. Thorne, C. Kramarae & N. Henley, Éd.). In B. Thorne, C. Kramarae & N. Henley (Éd.), *Language, Gender and Society*. Rowley, MA, Newbury House.

What do all these acronyms mean ? (p. d.). Reddit Help. Récupérée 15 mai 2020, à partir de https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/what-do-all-these-acronyms-mean

What is karma ? | Reddit Help. (p. d.). Récupérée 18 mars 2020, à partir de https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/what-karma

Whilchins, R. (2004). *Queer theory, gender theory : An instant primer*. Los Angeles, Alyson Publications.

Whittle, S. (1998). The trans-cyberian mail way. *Social & Legal Studies*, *7*(3), 389-408.

Wickham, H. (2016). *ggplot2 : Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

William, L., Encrevé, P. & Kihm, A. (1976). *Sociolinguistique / William Labov ; présentation de Pierre Encrevé traduit de l'anglais par Alain Kihm*. Paris, Les Éditions de Minuit.

Witmer, D. F. & Katzman, S. L. (1997). On-line smiles : Does gender make a difference in the use of graphic accents ? *Journal of Computer-Mediated Communication*, *2*(4). Récupérée 8 novembre 2017, à partir de http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00192.x/full

Wolf, A. (2000). Emotional expression online : Gender differences in emoticon use. *CyberPsychology & Behavior*, *3*(5), 827-833. Récupérée 30 avril 2017, à partir de http://online.liebertpub.com/doi/abs/10.1089/10949310050191809

Wolfram, W. (1974). *Sociolinguistic aspects of assimilation : Puerto Rican english in New York City*. Washington, DC, Center for applied linguistics.

Wolfram, W. & Schilling, N. (2016). *American English. Dialects and variation* (3ème). Wiley Blackwell.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Urban language series, no. 5.

Wong, A. (2007). Two vernacular features in the English of four American-born Chinese. *University of Pennsylvania Working Papers in Linguistics*, 216-230.

Woodfield, R. (2000). *Women, work and computing*. Cambridge University Press.

Woods, H. B. (1979). *A socio-dialectology survey of the English spoken in Ottawa : a study of sociological and stylistic variation in Canadian English*. University of British Columbia. https://doi.org/10.14288/1.0100260

xauxiheo. (2019, février 14). *Why is Reddit mobile a piece of shit ?* reddit. Récupérée 15 mai 2020, à partir de https://www.reddit.com/r/AskReddit/comments/aqfolt/why_is_reddit_mobile_a_piece_of_shit/

Yaguello, M. (1979). *Les mots et les femmes : essai d'approche socio-linguistique de la condition féminine* (T. 75). Payot.

Yelland, P. (2010). An introduction to correspondence analysis. *The Mathematica Journal*, *12*. https://doi.org/10.3888/tmj.12-4

yishan. (2014, septembre 4). *Every man is responsible for his own soul*. Upvoted. Récupérée 8 juillet 2020, à partir de https://redditblog.com/2014/09/06/every-man-is-responsible-for-his-own-soul/

Yuan, J. & Liberman, M. (2011). Automatic detection of "g-dropping" in American English using forced alignment, In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. Understanding (ASRU), Waikoloa, HI, USA, IEEE. https://doi.org/10.1109/ASRU.2011.6163980

Zammuner, V. L. (1987). Children's sex-role stereotypes : a cross-cultural analysis (P. Shaver & C. Hendrick, Éd.). In P. Shaver & C. Hendrick (Éd.), *Sex and Gender*. Newbury Park, Sage.

Zeileis, A. & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7-10.

Zentella, A. C. (1997). Growing up bilingual : Puerto Rican children in New York. *Lingua*, *1*(103), 59-74.

Zimman, Lal. (2014). The discursive construction of sex : Remaking and reclaiming the gendered body in talk about genitals among trans men (L. Zimman, J. Davis & J. Raclaw, Éd.). In L. Zimman, J. Davis & J. Raclaw (Éd.), *Queer Excursions : retheorizing binaries in language, gender and sexuality*. Oxford University Press.

Zimman, L. (2015). Facebook, the gender binary, and third-person pronouns, In *The OUPblog Tenth Anniversary Book*. New York, Oxford University Press.

Zimman, L. (2016). Sociolinguistics agency and the gendered voice : Metalinguistic negotiations of vocal masculinization among female-to-male transgender speakers (A. Babel, Éd.). In A. Babel (Éd.), *Awareness and control in sociolinguistic research*. NY et Cambridge, Cambridge University Press.

Zimman, L. (2017). Gender as stylistic bricolage : Transmasculine voices and the relationship between fundamental frequency and /s/. *Language in Society*, *46*(3), 339-370.

Zimman, L. (2018). Transgender voices : Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, *12*(8), e12284. https://doi.org/10.1111/lnc3.12284

zombreness. (2015, février 3). *Can you lovely ladies please confirm or deny if this is true ? Is there such code hidden in your texts ? We tend to just be too happy you texted at all to notice normally.* reddit. Récupérée 5 août 2020, à partir de https://www.reddit.com/r/TrollXChromosomes/comments/2uolol/can_you_lovely_ladies_please_confirm_or_deny_if/coa8x9z/

Zuberi, T. (2001). *Thicker than blood : How racial statistics lie*. Minneapolis, University of Minnesota Press.

Zucker, K. & Cohen-Kettenis, P. T. (2008). Gender identity disorder in children and adolescents, In *Handbook of sexual and gender identity disorders*. New York, John Wiley & Sons.

Zuur, A. F., Hilbe, J. M. & Ieno, E. N. (2015). *A beginner's guide to GLM and GLMM with R : A frequentist and Bayesian perspective for ecologists*. Newburgh, Highland Statistics.

Zweben, S. & Bizot, B. (2018). *Another year of record undergrad enrollment ; Doctoral degree production steady while Master's production rises again*. Computer Research Association.

Index

Table des figures

Liste des tableaux

Annexe A

Annexes de la partie II

Tableau A.1 – Catégories socioprofessionnelles dans RedditGender, classification de l'ISCO. Nous avons ajouté 4 catégories (étudiant·es, lycéen·nes, sans emploi, parents au foyer)

Catégorie	Exemples ou traductions	Eff.
Professionals	Médecins, enseignants, architectes, ingénieurs, etc.	352
Services and sales workers	Cuisinier·ères, serveur·ses, caissier·ères, aides-maternelles, etc.	93
Technicians and associate professionals	Technicien·nes	26
Managers	Managers, cadres	22
Clerical support workers	Secrétaires, employé·es de bureau, caissier·ères de banque	14
Craft and related trades workers	Employé·es du bâtiment, électricien·nes, etc.	11
Plant and machine operators, and assemblers	Ouvrier·ères, conducteur·trices de machines, etc.	4
Skilled agricultural, forestry and fishery workers	Agriculteur·trices, technicien·nes forestier·ères, etc.	4
Armed forces occupations	Métiers de l'armée	3
Elementary occupations	Agents de nettoyage, éboueurs, etc.	3
College students	Étudiant·es	229
High-school students	Lycéen·nes, collégien·nes	58
Unemployed	Sans emploi	48
Stay-at-home parents	Parents au foyer	13
Inconnue		164

Sideways Latin-only emoticons

Icon	Emoji	Meaning
:-) :) :] :-] :3 :-3 :> :-> 8-) 8) :-} :} :o) :c) :^) =] =)	😊🙂😊 😀😀	Smiley or happy face.[4][5][6]
:-D :D 8-D 8D x-D xD X-D XD =D =3 B^D	😃😄😆 😍	Laughing,[4] big grin,[5][6] laugh with glasses,[7] or wide-eyed surprise[8]
:-))		Very happy or double chin[7]
:-(:(:-c :c :-< :< :-[:[:-|| >:[:{ :@ :(; (	☹🙁😠 😡😞😔 😟😖	Frown,[4][5][6] sad,[9] angry,[7] pouting
:'-(:'(	😭😭	Crying[9]
:'-) :')	😂	Tears of happiness[9]
D-': D:< D: D8 D; D= DX	😰😧😨 😱😫😩	Horror, disgust, sadness, great dismay[5][6] (right to left)
:-O :O :-o :o :-0 8-0 >:O	😮😯😲	Surprise,[3] shock,[4][10] yawn[11]
:-* :* :x	😗😙😚 😘😍	Kiss[4]
;-) ;) *-) *) ;-] ;] ;^) :-, ;D	😉😜😘	Wink,[4][5][6] smirk[10][11]
:-P :P X-P XP x-p xp :-p :p :-Þ :Þ :-þ :þ :-b :b d: =p >:P	😛😝😜 🤪	Tongue sticking out, cheeky/playful,[4] blowing a raspberry
:-/ :/ :-. >:\ >:/ :\ =/ =\ :L =L :S	🤔😕😟	Skeptical, annoyed, undecided, uneasy, hesitant[4]
:-| :|	😐😑	Straight face[5] no expression, indecision[9]
:$://) ://3	😳😞😖	Embarrassed,[6] blushing[7]
:-X :X :-# :# :-& :&	😬🤐	Sealed lips or wearing braces,[4] tongue-tied[9]
O:-) O:) 0:-3 0:3 0:-) 0:) 0;^)	😇🙏	Angel,[4][5][10] saint,[9] innocent
>:-) >:) }:-) }:) 3:-) 3:) >;) >:3 >;3	😈	Evil,[5] devilish[9]
|;-) |-O	😎🛀	Cool,[9] bored/yawning[10]
:-J	😏😏	Tongue-in-cheek[12]
#-)	—	Partied all night[9]
%-) %)	😵😵🤕	Drunk,[9] confused
:-###.. :###..	🤒🤕🤢	Being sick[9]
<:-|	—	Dumb, dunce-like[10]
',:-| ',:-|	😐	Scepticism, disbelief, or disapproval[13][14]

FIGURE A.1 – Liste des émoticônes de Wikipedia (« List of emoticons », 2020) utilisée pour extraire les émoticônes du corpus

TABLEAU A.2 – Liste des interjections du package `lexicon` (Rinker, 2018), utilisée pour extraire les interjections du corpus

ah	eeeek	hup	oooh	uh-oh
a-ha	eek	hurrah	oops	uhh
aaaahh	eep	ich	ouch	uhhuh
aaah	eh	ick	ow	uhm
aah	er	jeez	oww	uhoh
aha	err	lahdedah	oy	uhuh
ahem	ew	meh	oyh	um
ahh	eww	mhm	pff	umm
ahhh	ewwww	mm	pffh	umph
argh	feh	mmh	pfft	vavavoom
aw	gak	mmhm	phew	whee
aww	grr	mmhmm	poof	whew
awww	grrrr	mmm	pooh	whoa
aye	ha	mwah	pshaw	whoopdedoo
ba-dump	haha	neenerneener	pssh	woo
bada-zing	hamanahamana	now	psst	wow
bah	hardyharhar	oh	rah	yay
blah	heehee	oh-lala	sheesh	yikes
blech	hey	oh-oh	shh	yohoho
boo	hist	ohoh	sis	yoohoo
booh	hm	oi	tchah	yow
boohoo	hmm	ok	tish	yuck
booya	hmmmm	okay	tsk-tsk	yuk
brr	hmph	ol	tsktsk	yum
brrrr	hoho	ooh	tut-tut	zing
bwahhahhah	hohum	ooh-la-la	ugh	zoinks
doh	hubbahubba	oohlala	uh	zowie
duh	huh	oomph	uh-hu	

TABLEAU A.3 – Script utilisé pour créer les tableaux de régression avec le package `stargazer` (MC808, 2013). « m » est ici le nom du modèle.

```
OR.vector <- exp(m$coef)
CI.vector <- exp(confint(m))
p.values <- summary(m)$coefficients[, 4]
stargazer(m, coef = list(OR.vector), ci = T, ci.custom =
list(CI.vector), p = list(p.values), single.row = T)
```

Annexe B

Annexes de la partie III

Tableau B.1 – Tableau de contingence utilisé pour réaliser le diagramme en mosaïque présentant les corrélations entre âge Reddit et genre

	Hommes cis	Femmes cis	MTF	FTM	NB
0-2 ans	98	120	60	70	40
2-3 ans	112	130	24	12	24
4 ans et +	162	122	16	18	36

Tableau B.2 – Tableau de contingence utilisé pour réaliser le diagramme mosaïque (résultats p. 177) présentant les corrélations entre âge Reddit et ethnicité

	Blancs	Afro-Am.	Asiatiques	Hispaniques
0-2 ans	40	27	21	22
2-3 ans	41	37	20	20
4 ans et +	47	25	22	25

Tableau B.3 – Données utilisées pour réaliser les modèles de régression pour l'étude du karma (résultats p. 182)

	Sous-corpus	Effectifs
Genre	Hommes cisgenres	332
	Femmes cis	333
	Hommes trans	92
	Femmes trans	93
	Non-binaires	91
Âge	14-20 ans	131
	21-30 ans	463
	31 ans et +	347
Total		941

Tableau B.4 – Karma de post, effet de l'ethnicité (résultats p. 188)

	Variable dépendante :
	Karma de post
Intercept	1,559.620** (806.004, 3,501.576)
Femmes cisgenres	0.628 (0.366, 1.079)
Afro-américain·es	1.123 (0.566, 2.270)
Asiatiques	1.408 (0.684, 3.067)
Hispaniques	0.621 (0.311, 1.310)
21-30 ans	0.931 (0.380, 2.089)
31 ans et +	1.131 (0.466, 2.531)
Modérateur·trices	1.656 (0.848, 3.605)
Observations	311
Log Likelihood	−2,124.645
θ	0.200** (0.014)
Akaike Inf. Crit.	4,265.289
Note :	*p<0.05 ; **p<0.01

Tableau B.5 – Karma de commentaire, effet de l'ethnicité (résultats p. 188)

	Variable dépendante :
	Karma de commentaire
Intercept	9,753.369** (5,534.288, 18,830.050)
Femmes cisgenres	0.939 (0.625, 1.410)
Afro-américain·es	0.858 (0.502, 1.481)
Asiatiques	0.830 (0.472, 1.510)
Hispaniques	0.910 (0.516, 1.649)
21-30 ans	0.660 (0.337, 1.224)
31 ans et +	1.702 (0.845, 3.266)
Modérateur·trices	0.980 (0.583, 1.763)
Observations	311
Log Likelihood	−2,929.027
θ	0.327** (0.022)
Akaike Inf. Crit.	5,874.054
Note :	*p<0.05 ; **p<0.01

TABLEAU B.6 – Tableau de contingence utilisé pour réaliser l'analyse factorielle des correspondances simples présentées dans la section 8.4.2

	Male	Female	MTF	FTM	NB
Education & science	1857.37	1508.18	357.57	259.25	235.08
Entertainment	2451.11	2415.95	777.91	262.16	186.91
Gaming	2938.36	1287.39	609.67	315.63	650.63
General	9934.18	8261.21	2151.60	1311.28	857.87
Humor	1976.02	2126.35	782.03	408.15	871.58
News & politics	2750.26	1344.82	628.12	261.55	365.39
Random/others	503.82	453.21	117.99	109.49	29.64
Sports & fitness	2165.16	1125.43	161.61	175.17	74.85
Technology	1215.89	286.98	220.79	49.44	111.32
Nobbies	1801.31	1703.76	332.02	81.94	124.07
X-rated	898.82	195.12	161.66	60.94	98.54
Personal advice	8707.57	16491.31	3698.92	6705.03	6394.09

Annexe C

Annexes de la partie IV

Tableau C.1 – 50 émoticônes les plus fréquentes dans RedditGender

Rang	Émoticône	Fréq.	Rang	Émoticône	Fréq.
1	:)	8382	26	:c	69
2	:(	3113	27	:o	65
3	:/	1409	28	=D	53
4	<3	1307	29	;D	53
5	;)	1299	30	:<	49
6	:P	1046	31	> :(	39
7	:-)	715	32	:S	35
8	:p	497	33	:S	34
9	XD	462	34	=P	34
10	xD	367	35	;P	34
11	:3	361	36	8)	30
12	D :	218	37	:>	26
13	:-(	141	38	:x	26
14	;-)	139	39	=]	25
15	:\	136	40	=\	21
16	:’(	109	41	:^)	21
17	:=)	105	42	> :)	12
18	:-\	100	43	;p	11
19	:O	90	44	:-|	10
20	=/	87	45	:X	8
21	:0	87	46	:$	6
22	:]	84	47	> :O	5
23	:|	83	48	XP	5
24	:))	78	49	:o)	4
25	:’)	69	50	D ;	4

TABLEAU C.2 – 100 étirements de lettres les plus fréquents du corpus

Rang	Type	Fréq.	%	Rang	Type	Fréq.	%
1	soooo	245	2.40	51	shhhh	22	0.22
2	sooo	243	2.38	52	wayyy	22	0.22
3	hmmm	223	2.18	53	yasss	22	0.22
4	awww	218	2.13	54	awwwww	21	0.21
5	sooooo	146	1.43	55	ewwww	21	0.21
6	oooh	143	1.40	56	yesssss	21	0.21
7	mmm	130	1.27	57	ewww	20	0.20
8	ahhh	126	1.23	58	ayyyy	19	0.19
9	ohhh	80	0.78	59	loooong	19	0.19
10	awwww	73	0.71	60	ohhhhh	19	0.19
11	ooooh	70	0.69	61	oooooh	19	0.19
12	ooo	67	0.66	62	aaah	18	0.18
13	uhhh	67	0.66	63	suuuper	18	0.18
14	ahhhh	64	0.63	64	suuuuper	18	0.18
15	mmmm	63	0.62	65	uhhhh	18	0.18
16	soooooo	62	0.61	66	ehhhh	17	0.17
17	yessss	58	0.57	67	aaand	16	0.16
18	waaaay	57	0.56	68	buuuut	16	0.16
19	ummm	55	0.54	69	hmmmmm	16	0.16
20	xxx	51	0.50	70	soooooooo	16	0.16
21	aaa	49	0.48	71	xxxx	16	0.16
22	nooooo	49	0.48	72	aaaaaand	15	0.15
23	ohhhh	48	0.47	73	fuuuuck	15	0.15
24	yesss	46	0.45	74	looong	15	0.15
25	nooo	43	0.42	75	ughhhh	15	0.15
26	oooo	43	0.42	76	uhhhhh	15	0.15
27	shhh	39	0.38	77	wooo	15	0.15
28	noooo	37	0.36	78	yesssssss	15	0.15
29	waaay	37	0.36	79	allll	14	0.14
30	soooooooo	35	0.34	80	alllll	14	0.14
31	hmmmm	34	0.33	81	ooohhh	14	0.14
32	loooove	34	0.33	82	oooooh	14	0.14
33	waaaaay	34	0.33	83	sss	14	0.14
34	aaaand	33	0.32	84	fuuuuuck	13	0.13
35	nooooooo	32	0.31	85	hhh	13	0.13
36	ehhh	30	0.29	86	huuuuge	13	0.13
37	ggg	30	0.29	87	looooong	13	0.13
38	mmmmm	30	0.29	88	ohhhhhh	13	0.13
39	ummmm	29	0.28	89	roxxxy	13	0.13
40	ughhh	28	0.27	90	viii	13	0.13
41	waaaaaay	28	0.27	91	yaaaas	13	0.13
42	aaaaand	27	0.26	92	yaaay	13	0.13
43	ooooo	27	0.26	93	yassss	13	0.13
44	ahhhhh	26	0.25	94	yeahhh	13	0.13
45	ayyy	26	0.25	95	yeahhhh	13	0.13
46	traaaaaaannnnnnnnnns	24	0.23	96	riiiight	12	0.12
47	pffft	23	0.23	97	shhhhh	12	0.12
48	wayyyy	23	0.23	98	ughhhhh	12	0.12
49	looooove	22	0.22	99	waaaaaaay	12	0.12
50	looove	22	0.22	100	wayyyyy	12	0.12

Tableau C.3 – Liste des interjections recensées dans le corpus

Rang	Interjections	Fréq.	%	Rang	Interjections	Fréq.	
1	oh	9225	31.55	45	oy	36	0.12
2	haha	3814	13.05	46	argh	31	0.11
3	hey	3134	10.72	47	pfft	26	0.09
4	ah	1396	4.78	48	eek	24	0.08
5	ugh	1280	4.38	49	hurrah	23	0.08
6	ha	784	2.68	50	shh	23	0.08
7	um	778	2.66	51	blech	21	0.07
8	eh	747	2.56	52	oi	20	0.07
9	yay	718	2.46	53	aaah	18	0.06
10	huh	694	2.37	54	mhm	17	0.06
11	hmm	631	2.16	55	ick	16	0.05
12	uh	630	2.15	56	grr	14	0.05
13	aww	452	1.55	57	doh	13	0.04
14	meh	434	1.48	58	boohoo	12	0.04
15	blah	380	1.30	59	zing	9	0.03
16	aw	368	1.26	60	aah	8	0.03
17	er	318	1.09	61	grrrr	8	0.03
18	ooh	278	0.95	62	pff	8	0.03
19	ahh	242	0.83	63	pooh	7	0.02
20	ol	221	0.76	64	psst	7	0.02
21	oops	207	0.71	65	a-ha	6	0.02
22	ew	206	0.70	66	eep	6	0.02
23	yikes	184	0.63	67	oomph	6	0.02
24	duh	162	0.55	68	whee	6	0.02
25	ouch	139	0.48	69	hmph	5	0.02
26	hm	135	0.46	70	pssh	5	0.02
27	jeez	128	0.44	71	rah	5	0.02
28	umm	127	0.43	72	yoohoo	4	0.01
29	mm	126	0.43	73	yuk	4	0.01
30	ahhh	122	0.42	74	heehee	3	0.01
31	aye	76	0.26	75	mmhmm	3	0.01
32	uhh	75	0.26	76	oww	3	0.01
33	whew	73	0.25	77	brr	2	0.01
34	yum	64	0.22	78	feh	2	0.01
35	ahem	62	0.21	79	hist	2	0.01
36	ow	58	0.20	80	pshaw	2	0.01
37	uhm	56	0.19	81	uh-oh	2	0.01
38	eww	52	0.18	82	gak	1	0.005
39	yuck	49	0.17	83	mmhm	1	0.005
40	phew	42	0.14	84	mwah	1	0.005
41	bah	40	0.14	85	tut-tut	1	0.005
42	sheesh	40	0.14	86	uhoh	1	0.005
43	poof	38	0.13	87	uhuh	1	0.005
44	err	36	0.12	88	umph	1	0.005