Le site américain Reddit comme espace de variation de l'anglais

Marie Flesch

LE SITE AMÉRICAIN REDDIT COMME ESPACE DE VARIATION DE L'ANGLAIS[1]

Résumé de thèse[2]

Marie Flesch

Cette thèse en sciences du langage utilise des méthodes quantitatives pour étudier les relations entre les pratiques d'écriture de l'anglais en ligne et le genre dans un corpus de commentaires publiés sur les forums du site américain Reddit. Elle s'inscrit dans la tradition variationniste, et a pour ambition de briser la binarité avec laquelle le genre est le plus souvent étudié en sociolinguistique. Son corpus intègre ainsi les commentaires de femmes et d'hommes cisgenres, mais aussi de femmes et d'hommes transgenres et de personnes non binaires[3].

1. L'INTERSECTIONNALITÉ : UN ANCRAGE MÉTHODOLOGIQUE

Notre travail s'appuie sur l'approche intersectionnelle, qui a été décrite comme étant « la contribution théorique la plus importante des études féministes » (McCall, 2005 : 1771). Née aux États-Unis d'une critique de la recherche sur le genre, l'intersectionnalité postule que prendre en compte une seule catégorie sociale ne suffit pas pour étudier la diversité des comportements et des expériences (Levon, 2015). Cette approche, aujourd'hui dominante dans les études de genre aux États-Unis, a fait son entrée dans le domaine de la sociolinguistique à la fin des années 1990, essentiellement dans des travaux qualitatifs comme les études de Bucholtz sur les adolescentes et adolescents en Californie (2010), de Levon sur la prosodie en hébreu (2014), ou encore des études de Morgan sur les femmes afro-américaines (2004). Avec cette thèse, nous

---

[1] Thèse dirigée par Sophie Bailly (université de Lorraine, ATILF UMR 7118) et soutenue le 16 décembre 2020 à l'université de Lorraine, devant un jury composé de L. Greco (université de Lorraine), M. Candea (université Sorbonne Nouvelle), N. Kübler (université de Paris), et F. Baider (université de Chypre).

[2] Résumé de thèse paru dans L'Information Grammaticale, n°176, janvier 2023, http://www.informationgrammaticale.com/anneeencours.htm

[3] Les personnes cisgenres sont des personnes dont l'identité de genre correspond au genre qui leur a été assigné à la naissance. Les personnes transgenres ont une identité de genre différente de celle qui leur a été assignée à la naissance. Les personnes non binaires ne se reconnaissent pas dans une vision binaire du genre ; elles peuvent être par exemple agenre, *genderfluid* ou encore bigenre.

souhaitions intégrer cette approche dans une démarche quantitative, ce qui a rarement été mis en pratique en sociolinguistique. L'intersectionnalité a donc guidé les choix méthodologiques faits dans la construction du corpus et dans son analyse.

2. LE CORPUS

Le corpus a été réalisé à partir du site internet Reddit. Lancé en 2005 aux États-Unis, Reddit accueille des centaines de milliers de forums, appelés *subreddits*. Comme les internautes y commentent sous des pseudonymes, il est difficile d'y recueillir des données sociodémographiques de façon automatique. Nous avons donc opté pour une méthode manuelle : nous avons effectué des recherches dans les contenus des commentaires en utilisant les expressions *I am* et *I live in*. Cela nous a permis d'obtenir des informations riches sur de nombreux internautes : leur identité de genre, leur âge, leur lieu de résidence, et, dans certains cas, leur ethnicité[4]. Dans une optique intersectionnelle, nous souhaitions créer un corpus diversifié et richement annoté. Nous avons tiré parti de l'organisation de Reddit en communautés pour trouver des internautes aux profils variés, en surreprésentant dans le corpus les personnes transgenres et non binaires, ainsi que les internautes afro-américains et afro-américaines, hispaniques et asiatiques.

Créé entre avril et juillet 2017, le corpus comprend 460 707 commentaires, qui représentent 19 millions de tokens (ou mots). Les données ont été copiées manuellement, puis encodées pour être traitées avec le logiciel de textométrie TXM. Le corpus est composé des productions de 1044 internautes, tous anglophones, dont 78.83 % résident aux États-Unis : 372 femmes cisgenres, 372 hommes cisgenres, 100 femmes transgenres, 100 hommes transgenres, et 100 personnes non binaires. La majorité de ces personnes (49.52 %) a de 21 à 30 ans ; 14.08 % ont moins de 21 ans, et 36.40 % ont 31 ans ou plus.

3. LES VARIABLES LINGUISTIQUES

L'objectif de la thèse était d'étudier les pratiques linguistiques typiques de l'anglais d'internet. Onze phénomènes non standard ont été retenus. Ils ont été classés en deux catégories : les procédés d'ajout et les procédés de réduction. Les procédés d'ajout consistent à insérer davantage de caractères dans les messages que n'en nécessite l'orthographe standard de l'anglais, à utiliser des phénomènes paralinguistiques, ou à

4 Nous utilisons le terme « ethnicité » pour désigner le concept de « race » qui, comme le genre, est une construction sociale, non basée sur la biologie. Ce choix a été fait pour éviter l'utilisation du mot « race », qui est controversé en français.

employer des majuscules de façon excessive. Les six procédés d'ajout sont les émoticônes (**:-)**, **;)**), les émojis (😀, 💯, 💕) , les étirements de lettres (*giiiirrrllll*, *heyyyyy*), les étirements de ponctuation (*!!!!!!!!*), les mots en majuscules (*NO YOU DIDN'T*) et les interjections (*ah*, *ugh*). Les procédés de réduction répondent quant à eux au besoin d'économie qui caractérise souvent la communication sur internet. Nous en avons retenu cinq : les abréviations (acronymes - *lol*, *OMG* – mais aussi réductions – *bc*, *srs*), les graphies phonétiques (*wanna*, *tho*), les *g-droppings* ou omission du *-g* final (*fuckin*, *lookin*), les omissions d'apostrophe (*Im*, *cant*) et les omissions de la majuscule du pronom personnel *I*.

4. LES MÉTHODES STATISTIQUES

Les techniques statistiques ont été soigneusement choisies pour prendre en compte la spécificité des données linguistiques recueillies, pour offrir des résultats robustes, et pour respecter le projet intersectionnel de la thèse. Nous avons principalement utilisé la méthode de la régression multiple avec interactions, qui permet d'étudier les interactions entre nos trois variables d'intérêt (genre, âge et ethnicité), et qui a été décrite comme étant compatible avec l'approche intersectionnelle (Bowleg & Bauer, 2016). Nous avons opté pour des modèles de régression binomiale négative et *zero-inflated*, qui sont capables de gérer la sur-dispersion propre aux données de fréquence.

5. LES RÉSULTATS

5.1. Genre

Les analyses linguistiques explorent les relations entre le genre, l'âge et l'ethnicité, et l'utilisation de phénomènes caractéristiques de l'anglais d'internet. Certains de ces phénomènes ont déjà été abondamment étudiés ; c'est par exemple le cas des émoticônes, qui sont généralement considérées comme des marqueurs féminins. D'autres, comme les mots en majuscules ou les abréviations, ont moins attiré l'attention des chercheurs.

Dans un premier temps, les analyses ont permis de mettre en évidence des phénomènes « genrés ». Plusieurs phénomènes d'ajout, comme les émoticônes, les étirements de lettres et les étirements de ponctuation, sont utilisés plus fréquemment et de façon significative par les femmes cisgenres. Les hommes cisgenres, en revanche, emploient davantage certains procédés de réduction comme l'omission d'apostrophe, les *g-droppings* et l'omission de la majuscule de *I*.

Dans un deuxième temps, nous avons cherché à comprendre comment les internautes transgenres et non binaires se situent par rapport à ces variables genrées. Les hommes transgenres et les individus non binaires sont généralement dans une position médiane, c'est-à-dire qu'ils n'utilisent pas plus ou pas moins fréquemment les variables linguistiques que les femmes cisgenres ou les hommes cisgenres. Les femmes transgenres, par contraste, ont des usages peut-être inattendus : elles s'alignent parfois sur les usages des femmes cisgenres, et parfois sur ceux des hommes cisgenres. Par exemple, elles utilisent les émoticônes plus fréquemment que les hommes, mais elles

l'omission de la majuscule du pronom *I* aussi fréquemment qu'eux. Comme l'interaction entre genre et âge a été intégrée aux analyses quand elle était significative, les résultats sont souvent nuancés. Chez les personnes non binaires, le genre assigné à la naissance n'a pas d'impact sur la fréquence des graphies non standard.

5.2. Genre et âge

Nos analyses intègrent l'effet de l'âge, s'intéressant à son interaction avec le genre dans l'utilisation des variables linguistiques. Nous nous attendions à constater une corrélation négative entre âge et fréquence des variables. De nombreuses études (par exemple Tagliamonte, 2016) montrent en effet que les adolescents, adolescentes et jeunes adultes utilisent davantage de graphies non standard que les internautes plus âgés. Nos résultats révèlent un phénomène intéressant. Les hommes cisgenres plus âgés utilisent moins de graphies non standard que les plus jeunes, mais la corrélation négative est rarement présente chez les femmes cisgenres. Elle n'existe pas ou très peu chez les personnes transgenres et non binaires : les plus jeunes n'utilisent pas plus ou moins de phénomènes non standard que les plus âgées.

5.3. Genre et ethnicité

L'analyse de l'interaction entre genre et ethnicité a été réalisée sur un échantillon du corpus, composé de 347 personnes cisgenres. Elle permet de faire émerger deux résultats marquants. Tout d'abord, elle montre que l'intégration de l'ethnicité tempère les différences constatées dans les analyses portant uniquement sur le genre et l'âge. Par exemple, dans le cas du *g-dropping*, les différences entre femmes et hommes ne sont plus significatives quand on prend en compte l'ethnicité des internautes. L'effet du genre varie selon les groupes ethniques, avec des différences plus marquées entre femmes et hommes chez les internautes hispaniques, asiatiques et afro-américains que chez les internautes blancs. Nous remarquons également une distanciation entre les groupes asiatiques d'une part, et les groupes afro-américains et hispaniques de l'autre. Ces derniers utilisent plus fréquemment certains procédés non standard, ce qui suggère que ceux-ci sont une des stratégies utilisées pour indexer leur identité ethnique. Cela peut également indiquer que les internautes afro-américains et hispaniques jouent un rôle dans l'innovation linguistique en ligne, comme d'autres études le soulignent (Bamman et al., 2014, Eisenstein et al., 2010).

6. CONCLUSION

Un des apports principaux de notre thèse est son corpus singulier, qui intègre 300 personnes transgenres et non binaires et qui a une annotation démographique riche, rare dans les études de la langue d'internet. Le choix d'allier intersectionnalité et méthodes quantitatives s'est avéré pertinent : il apporte des nuances précieuses à l'étude des phénomènes linguistiques. L'analyse des interactions entre genre et âge, ou genre et ethnicité, permet de déconstruire les oppositions entre femmes et hommes que mettent en évidence de nombreuses études sur la langue et le genre.

Dans ce sens, notre thèse contribue à lutter contre les stéréotypes qui entourent le

langage et le genre. Elle montre qu'il n'existe pas une façon unique d'écrire sur internet « comme une femme » ou « comme un homme », mais que les usages des internautes sont influencés par un faisceau de variables, dont leur âge et leur ethnicité. Elle suggère ainsi que comparer systématiquement les femmes aux hommes sans prendre en compte d'autres facettes de leur identité occulte la réalité des usages. Et, pour la première fois en sociolinguistique, notre travail étudie les usages de centaines de personnes transgenres et non binaires, ce qui nous éclaire sur la façon dont elles utilisent le langage pour indexer, ou non, une identité de genre. Nos résultats suggèrent que la masculinité et la féminité ne sont pas des constructions homogènes, mais que les internautes font des choix qui les rapprochent ou les distancient des usages considérés comme typiquement féminins ou typiquement masculins.

Marie Flesch

UMR 7118 ATILF

Université de Lorraine-Nancy

RÉFÉRENCES BIBLIOGRAPHIQUES

BAMMAN D., EISENSTEIN J., SCHNOEBELEN, T. (2014), « Gender identity and lexical variation in social media », *Journal of Sociolinguistics*, 18/2, 135-160.

BOWLEG L., BAUER G. (2016), « Invited reflection: Quantifying intersectionality » *Psychology of Women Quarterly,* 40/3, 337–341.

BUCHOLTZ M. (2010), *White kids : Language, race, and styles of youth identity*. Cambridge University Press.

EISENSTEIN J., O'CONNOR B., SMITH, N. A., XING E. P. (2010), « A latent variable model for geographic lexical variation », *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 1277-1287.

LEVON E. (2014), « The politics of prosody : Language, sexuality and national belonging in Israel », *Queer Excursions: retheorizing binaries in language, gender and sexuality*, 101-128.

LEVON E. (2015), « Integrating intersectionality in language, gender, and sexuality research », *Language and Linguistics Compass,* 9/7, 295–308.

MCCALL L. (2005), « The complexity of intersectionality », *Signs: Journal of Women in Culture and Society*, 30/3, 1771-1800.

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*, Oxford University Press, 252-259.

TAGLIAMONTE S. A. (2016), « So sick or so cool? The language of youth on the internet », *Language in Society*, 45/01, 1-33.