Revista Heterotópica
Laboratório de Estudos Discursivos Foucaultianos - LEDIF
Instituto de Letras e Linguística da Universidade Federal de Uberlândia - ILEEL/UFU

Are the straights ok? Analyse multimodale de la resignification discursive de l'hétérosexualité sur Twitter et Instagram : défis et limites de la construction et préparation d'un corpus de données numériques

Are the straights ok? Análise multimodal da resignificação discursiva da heterossexualidade no *Twitter* e *Instagram*: desafios e limitas da construção e preparação de um corpus de dados digital

Are the straights ok? Multimodal analysis of the discursive resignification of heterosexuality on Twitter and Instagram: challenges and limits of the construction and preparation of a digital corpus

Alice Cesbron[1]
Université de Paris, Université de Greifswald
alice.cesbron@hotmail.fr

RÉSUMÉ : L'objectif de cet article est de présenter en détail les défis et limitations rencontrées lors de la création d'un corpus multilingues de données Twitter et Instagram, ayant pour objet la resignification de l'hétérosexualité, en vue d'une analyse lexicométrique et d'une analyse qualitative manuelle. Cette recherche articule la perspective de la *Queer Linguistics*, visant à mettre en lumière les idéologies de genre et sexualité normatives des discours (LEAP, 2015), la Critical Discourse Analysis à laquelle elle étroitement liée (FAIRCLOUGH, 1989), et la conception du discours issue de l'analyse du discours française (KRIEG-PLANQUE, 2017), à une approche *symétrique* des données numériques développée par Paveau (2013). En y abordant la problématique de la pertinence et de la représentativité d'une recherche thématique sur les réseaux sociaux numériques, les défis accompagnant la collecte de données ainsi que les choix entourant la préparation des données en vue d'une analyse quantitative et qualitative nous y montrons comment un corpus numérique de grande taille en analyse du discours vient contraindre les directions que pourra prendre le travail de recherche. L'article illustrera ensuite les choix méthodologiques développés, en présentant quelques résultats préliminaires d'analyses lexicométriques sur l'un des sous-corpus Instagram ainsi que de premières tendances thématiques des données.
Mots-clefs : Analyse du discours ; Réseaux sociaux numériques ; Linguistique de corpus ; Etudes de genre.

RESUMO: O objetivo deste trabalho é detalhar os desafios e limitações encontrados na criação de um *corpus* multilíngue de dados *Twitter* e *Instagram* sobre o tópico da ressignificação da heterossexualidade, em uma análise, ao mesmo tempo, lexicométrica quantitativa e qualitativa manual. Esta pesquisa articula a perspectiva da *Queer Linguistics*, que visa lançar luz sobre as ideologias normativas dominantes dos discursos (LEAP, 2015), a Análise Crítica do Discurso com a qual está intimamente relacionada (FAIRCLOUGH, 1989), e a concepção do discurso decorrente da análise do discurso francês (KRIEG-PLANQUE, 2017), com uma

[1] Doctorante en sciences du langage en cotutelle à l'Université de Paris (ED 622) et à l'Université de Greifswald (IFAA).

abordagem *simétrica* dos dados digitais desenvolvida por Paveau (2013). Ao abordar o problema da relevância e representatividade de uma pesquisa temática sobre redes sociais digitais, os desafios que acompanham a coleta de dados, bem como as escolhas em torno da preparação de dados para análise quantitativa e qualitativa, mostramos como um grande *corpus* digital na análise do discurso restringe as direções que o trabalho de pesquisa pode tomar. O artigo ilustra, então, as escolhas metodológicas desenvolvidas, apresentando alguns resultados preliminares de análises lexicométricas sobre um dos *sub-corpora* do Instagram, bem como algumas primeiras tendências temáticas dos dados.
Palavras-chave: Análise do discurso; Redes sociais digitais; Linguística de corpus; Estudos de gênero.

ABSTRACT: The purpose of this paper is to detail the challenges and limitations encountered in creating a multilingual corpus on Twitter and Instagram data about the resignification of heterosexuality, with the aim of performing a lexicometric analysis and a manual qualitative analysis. This research articulates the perspective of Queer Linguistics, which aims to shed light on the normative gender and sexuality ideologies of discourses (LEAP, 2015), the Critical Discourse Analysis to which Queer Linguistics is closely related (FAIRCLOUGH, 1989) and the conception of discourse stemming from French discourse analysis (KRIEG-PLANQUE, 2017), with a *symmetrical* approach to digital data developed by Paveau (2013). By addressing the problem of the relevance and representativeness of a thematic research on digital social networks, the challenges accompanying data collection, as well as the choices surrounding the preparation of data for quantitative and qualitative analysis, this article shows how the choice of a large digital corpus in discourse analysis itself constrains the directions that the research can take. The article will then illustrate the methodological choices developed, by presenting some preliminary results of lexicometric analyses on one of the Instagram sub-corpora as well as some of the data's first thematic trends.
Keywords: Discourse analysis; Social networks; Corpus linguistics; Gender studies.

Introduction

Dans la continuité d'un mouvement général de la libération de la parole sur les questions de genre et de sexualité initiés par MeToo en 2017 sur les réseaux sociaux numériques (RSN) dans les pays occidentaux, apparaissent en 2019 deux initiatives dédiées à la critique de l'hétérosexualité : le compte @heterocringe sur Instagram en février 2019 et le subreddit r/arethestraightsok? sur Reddit en octobre de la même année. Similaires aux comptes partageant des memes humoristiques, ces espaces sont dédiés à la publication de contenu qualifié d'*heterocringe* issu du mot anglais *cringe* à l'origine utilisé comme verbe signifiant *grimacer* face à quelque chose d'embarrassant. Le principe de ces pages est donc de poster du contenu trouvé sur internet et les RSN offrant une représentation jugée stéréotypique, problématique voire toxique de l'hétérosexualité ou de l'hétéronormativité et suscitant le malaise, afin de remettre en question le statut d'orientation sexuelle « naturelle et standard » dont l'hétérosexualité bénéficie. Le présent article découle donc de notre travail de recherche sur ces discours sur Twitter et Instagram, en anglais, français et allemand, visant à analyser les stratégies discursives et multimodales sur lesquelles reposent le processus de resignification ainsi que les nouvelles représentations de l'hétérosexualité et l'hétéronormativité qu'ils proposent.

Dans cet article, faisant suite à une communication donnée durant la journée d'étude *Analyse du discours numérique : enjeux épistémologiques et méthodologiques* organisée par l'ADAL, nous présentons principalement les défis survenus lors la création de notre corpus portant donc sur ces discours collectés sur Twitter et Instagram opérant une resignification de l'hétérosexualité, ainsi que les challenges que nous avons pu rencontrer dans le cadre de la préparation de ces données numériques en vue d'une analyse quantitative et qualitative. Nous parlerons tout d'abord de la manière dont le choix d'un terrain numérique et les attentes de pertinence et de représentativité accompagnant ce type de recherche, nous ont amené à faire évoluer notre objet de recherche d'origine. Nous continuerons en exposant les problématiques de tri de données liées à la collecte de tweets que nous avons pu rencontrer. Nous poursuivrons par une présentation des défis techniques que nous avons rencontrés durant le processus de préparation des données numériques dans l'optique d'une analyse lexicométrique quantitative, et des choix méthodologiques que cette étape nous a amenés à prendre. Ensuite, nous aborderons les limites et les difficultés de l'analyse qualitative en analyse du discours dans le cadre de notre recherche et face à un corpus de données numériques de grande taille en général.

Enfin, nous présenterons l'aboutissement de cette phase méthodologique à travers quelques analyses quantitatives et qualitatives préliminaires.

Cadre théorique

Ce travail s'inscrit tout d'abord dans le mouvement de la *Queer Linguistics*, lui-même fortement lié à la *Critical Discourse Analysis* (CDA) (FAIRCLOUGH, 1989) bien que cela ne soit pas toujours explicité (LEAP, 2015). A l'instar de la CDA, la *Queer Linguistics* s'intéresse en effet à la manière dont les rapports de pouvoir et les autorités normatives s'inscrivent dans les discours et les pratiques discursives, en croisant ici les notions de genres, de sexualités, de race, de classe et toute forme d'inégalité d'une manière générale (LEAP, 2012). S'appuyant en partie sur la théorie de la performativité établie par Butler (1999 [1990]) et selon laquelle les identités de genre et de sexualités sont construites de manière consciente et inconsciente à travers la reproduction et la répétition d'actes, notamment langagiers, produisant une illusion du naturel, ainsi que sur l'inscription des idéologies normatives dans le discours donnant l'impression d'être évidentes selon Althusser (1970), la *Queer Linguistics* s'intéresse à la manière dont les idéologies normatives en matière de genre et de sexualité se font passer pour évidentes et comment « the uncritical acceptance of those messages coincides with the conditions of difference, hierarchy, and exclusion. » (LEAP, 2015, p. 662). Nous allions également les objectifs de la *Queer Linguistics* en termes d'inscription des idéologies dans le discours, aux théories féministes issues du Mouvement de Libération des Femmes et principalement au concept d'*hétérosexualité obligatoire* (RICH, 1981) qui considère l'hétérosexualité entre autres comme une idéologie, une « pensée straight », omniprésente dans les relations humaines ainsi que dans l'inconscient (WITTIG, 1992 ; JACKSON, 1996). Enfin, au-delà de la manière dont les idéologies normatives sont inscrites dans le discours, nous nous positionnons dans le sillon de l'analyse du discours de tradition française pour qui le discours n'est pas un reflet de la réalité mais participe déjà en lui-même à constituer la réalité (KRIEG-PLANQUE, 2017).

Cette recherche se fonde également sur le concept de resignification et tout particulièrement sur l'apport qu'a pu y faire Marie-Anne Paveau (2019). Théorisée par Judith Butler en 1997 dans *Excitable Speech* et résumée par Marie-Anne Paveau, la resignification discursive « consiste à reprendre un élément langagier ressenti comme blessant et/ou injurieux et à en modifier la valeur axiologique négative pour le transformer en marque d'identité

habilitante » (2019, p. 1). La forme de resignification à laquelle nous nous intéressons diffère de cette définition sur plusieurs points. Nous n'avons tout d'abord ici pas à faire à la resignification d'un élément langagier mais bien d'une catégorie sociale, à savoir l'orientation hétérosexuelle. D'autre part nous sommes face à une configuration moins commune de resignification dans laquelle un groupe social en position de domination – les personnes hétérosexuelles – est resignifié par le groupe que lui-même stigmatise habituellement – la communauté LGBTQI+. Pour cette recherche nous proposons donc de nous situer dans la continuité des apports de Paveau (2019) mais aussi de Kunert (2012) à la notion de resignification. Kunert (2012) suggère en effet de ne pas seulement se concentrer sur l'unité lexicale mais considère aussi comme resignification la manipulation plus large de signes visant à opérer des changements dans les représentations. Paveau (2019), quant à elle, présente dans sa typologie des formes de resignification la catégorie de *resignification technodiscursive*, afin de prendre en compte les phénomènes de resignification au-delà de l'unité lexicale, notamment dans le cadre communication médiée par ordinateur et les potentialités multimodales qu'elle ouvre.

Nous avons également choisi d'inscrire notre travail analytique dans l'approche *symétrique* développée par Paveau (2013). En effet, nous ne souhaitons pas limiter notre analyse au niveau textuel dans la mesure où nos données sont entourées de tout un dispositif numérique. Dans cette approche Paveau (2013) suggère de penser la relation entre le verbal et le non-verbal comme un continuum et de « prendre en compte l'ensemble de l'écosystème des productions verbales, contenant l'ensemble des données qui nous entourent, qu'elles soient humaines ou non humaines » (p. 3).

Corpus

Cette recherche s'appuie sur un corpus multimodal et trilingue constitué de deux sous-corpus : un sous-corpus provenant de Twitter comprenant 163 819[2] tweets en anglais et français, et un sous-corpus provenant d'Instagram comprenant 2807 posts en anglais, allemand et français dans une moindre mesure. La collecte du sous-corpus Twitter s'est déroulée entre août 2020 et mars 2021 en utilisant le template de collecte de tweets TAGS[3], qui permet de collecter des tweets à partir d'une recherche de mots-clefs, hashtags ou syntagmes nominaux. Le flux

[2] Chiffre non définitif.
[3] En ligne sur : <https://tags.hawksey.info/>.

des données sur les deux plateformes ne permettant pas une délimitation chronologique ou événementielle du corpus, la fin de la collecte fut établie par rapport à la quantité de données acquises selon les deux plateformes. Les syntagmes utilisés pour la collecte furent déterminés grâce à des tests initiaux servant à estimer les combinaisons de termes permettant d'accéder à des quantités suffisantes de données. À la suite de ces tests les syntagmes « the straights », « straight people » et « les hétéros » furent retenus. Le sous-corpus Instagram se base lui sur les comptes @heterocringe [4] – dont le compte actuel rassemble 71K abonnés – et @german_hetero_cringe – totalisant 567 abonnés. Ces deux comptes furent sélectionnés car ce sont les deux principaux comptes directement centrés sur le contenu *heterocringe* dans une langue nous étant accessible[5]. Les posts furent collectés le 16 mars 2021 en utilisant un script Python permettant d'avoir accès à tous les posts des comptes jusqu'à cette date et aboutissant à un total de 2807 posts. Chaque set de donnés fut généré sous forme de fichier Excel rassemblant également un grand nombre de métadonnées que nous avons ensuite sélectionnées et gardées en fonction des usages prévus par rapport aux analyses. Dans le cas de Twitter nous avons conservé toutes les métadonnées permettant d'identifier les tweets et leur auteur. Nous avons également conservé les métadonnées sur la date de création, la localisation, si fournie, l'URL, et toutes les données dans les cas où les tweets étaient des réponses à d'autres tweets. Dans le cas d'Instagram, nous avons conservé les données indiquant si les posts contenaient des photos ou des vidéos, la date, le nombre de likes, de commentaires et de vues pour les posts avec vidéo, ainsi que l'URL et les hashtags si présents. Toutes les photos et vidéos n'étaient, elles, pas inclues dans les fichiers Excel mais ont été extraites dans des dossiers séparés.

Choix des critères de collecte : un défi de pertinence et représentativité

Le corpus que nous venons de présenter est l'aboutissement d'un long processus de réflexion et de tests, rendus nécessaires par le contexte de la recherche sur les RSN ainsi que par l'objet de recherche initialement choisi. En effet, cette recherche visait à l'origine à analyser la construction des féminités hétérosexuelles sur les RSN, ce qui, aux travers de nos observations et réflexions, s'est révélé être une entreprise difficilement réalisable dans le contexte de Twitter et d'Instagram. Cette incompatibilité entre notre objet de recherche initial

[4] A depuis été banni par la plateforme et recrée sous le nom <https://www.instagram.com/hetero.cringe/>.
[5] Il existe également des comptes mentionnant « heterocringe » dans leur titre et postant du contenu en polonais, hongrois ou espagnol.

et le terrain sur lequel nous voulions l'étudier s'est révélée durant le processus méthodologique préparant à la collecte de données au début de ce travail de recherche. En effet, il était pour nous tout d'abord crucial, d'identifier la méthode la plus adaptée dans le cadre d'une collecte de données sur les réseaux sociaux en partant d'un objet thématique mais également de déterminer comment la féminité hétérosexuelle s'indexe dans le discours de manière à pouvoir identifier des éléments discursifs qui serviraient de critères de collecte.

Nous avons donc tout d'abord entrepris une revue des méthodologies de collecte de données sur les RSN afin d'obtenir une vue d'ensemble des caractéristiques et fonctions des RSN pouvant être utilisées dans la construction d'un corpus. Pour répondre à cette question nous nous sommes donc appuyés sur la liste, établie par Rieder (2012), des méthodes envisageables lors d'une collecte de données provenant de Twitter. Dans cette liste, Rieder (2012) distingue six méthodes que nous avons chacune évaluées par rapport à notre objet et nos questions de recherche. Nous nous concentrerons sur la présentation des deux seules méthodes correspondant à notre type de question de recherche.

La première de ces options est *l'échantillon manuel* (RIEDER, 2012), dont le principe est de collecter un petit corpus localisé en sélectionnant par exemple des utilisateurs particuliers. Nous avons hypothétiquement considéré cette méthode pour construire un corpus basé sur une sélection d'utilisateur.ice.s dont l'hétérosexualité était rendue claire à travers leur compte. Cette méthode ne fut cependant pas retenue dans la mesure où peu d'utilisateur.ice.s rendent leur hétérosexualité visible d'une manière qui soit rapidement identifiable par exemple à travers l'utilisation dans leur « bio » d'un émoticône indexant l'hétérosexualité comme cela peut être le cas avec le drapeau multicolore chez certains membres de la communauté LGBTQI+. La deuxième méthode de collecte que nous avons étudiée est celle de *l'échantillon thématique* (RIEDER, 2012) basé sur des mots-clefs ou hashtags reliés à un sujet ou événement, qui selon Burgess et Bruns (2012) est la méthode la plus efficace lorsque l'on souhaite étudier un sujet précis comme c'est notre cas. Cette méthode présente cependant des limitations notamment en termes de représentativité comme le montrent Smyrnaios et Ratinaud (2013). Outre le fait que les utilisateur.ice.s de Twitter ne représentent qu'une frange spécifique d'une population observée – un objectif de représentativité que nous ne cherchons pas à atteindre car nous concentrons nos observations aux RSN eux-mêmes – la collecte thématique de données par mots-clefs, hashtags ou groupes nominaux reste limitée dans sa représentativité dans la mesure où il est extrêmement compliqué d'identifier la totalité des mots utilisés pour parler d'un sujet (SMYRNAIOS ; RATINAUD, 2013). Enfin, cette méthode implique également de prendre en compte des facteurs comme l'aspect polysémique des mots.

Malgré les limitations présentées, la collecte thématique nous a semblé rester la méthode la plus adaptée vis-à-vis de notre objet de recherche. Nous nous sommes donc ensuite intéressés à la façon dont la féminité hétérosexuelle s'indexait dans le discours de manière à définir des syntagmes nominaux pouvant servir de critères de collecte. Pour tenter de répondre à cette question nous nous sommes appuyés sur les études réalisées en *Language and Gender studies* qui l'abordent sous deux angles distincts. D'une part à travers les travaux fondateurs sur le *langage des femmes* de Lakoff (2004 [1975]), d'autre part à travers les recherches sur l'indexation de l'hétérosexualité dans le langage. Les recherches de Lakoff (2004), dans les premières années des LGS, constituent un apport fondateur pour la discipline et ont ainsi formé le point de départ de nos observations. Ces recherches portent sur certaines caractéristiques langagières considérées par la chercheuse comme plus présentes dans le langage des femmes et qui peuvent s'appliquer au langage écrit, que l'on retrouve sur Twitter. Parmi ces caractéristiques la chercheuse mentionne notamment, au niveau lexical, l'usage de diminutifs et d'euphémismes, de marqueurs d'atténuation comme *sort of* ou *I don't know* en anglais, et des adjectifs « vides » comme *fantastique* ou *divin* (LAKOFF, 2004). Elle observe également, au niveau phrastique, que les femmes ont tendance à être moins directes, plus polies et expriment des opinions de manière moins tranchée (LAKOFF, 2004). Enfin, au niveau conversationnel, que l'on peut dans une certaine mesure retrouver sur Twitter, elle remarque que les femmes sont plus collaboratives, utilisent davantage d'éléments non-verbaux, et ont tendance à employer une grammaire plutôt standard et une langue moins familière (LAKOFF, 2004). Bien que fondateurs, ces travaux n'en restent aujourd'hui pas moins critiqués, principalement car ils reposent sur une méthodologie introspective (BUCHOLTZ, 2014) et impliquent donc une conception hétérocentrée et blanche de la femme ne permettant pas d'identifier un langage correspondant à toutes les femmes (KIESLING, 2019).

Le travail de Herring (2000) sur les différences genrées dans la communication médiée par ordinateur nous a également aidé à cibler un certain langage féminin sur Twitter. Herring (2000) remarque en effet que les différences genrées habituellement observées dans le discours en face à face sont sensiblement reproduites dans la communication médiée par ordinateur, telles que les différences au niveau de la politesse, de l'assurance, du langage injurieux et de l'usage du langage non-verbal à travers les emojis. Cependant, bien que ces travaux permettent d'identifier un certain langage féminin, les caractéristiques relevées par Herring (2000) et Lakoff (2004) n'ont pas directement les féminités hétérosexuelles comme sujet et peuvent apparaître dans un grand nombre de contextes, faisant d'eux des éléments peut adaptés à une collecte de données thématique comme le décrivent Smyrnaios et Ratinaud (2013).

Nous nous sommes donc tournés vers les études sur l'indexation de l'hétérosexualité dans le langage. Une des études les plus importantes à ce sujet fut menée par Kitzinger (2005) et met en lumière l'omniprésence des références au couple hétérosexuel marié dans le langage. Bien que cette étude remonte à 2005 et repose sur le fait que la mention du mariage ne puisse faire référence qu'au mariage hétérosexuel, les analyses qu'elle apporte restent tout de même d'actualité. Kitzinger (2005) fait tout d'abord une distinction entre l'indexation explicite et implicite. Elle mentionne notamment les blagues et allusions sexuelles ainsi que les discours portant sur une activité hétérosexuelle ainsi que les discours sur la relation hétérosexuelle impliquant la thématique du mariage, en matière de références explicites. En ce qui concerne les références implicites, Kitzinger (2005) explique que celles-ci passent principalement par la référence de personne non-recognitionnelle, c'est-à-dire lorsqu'un locuteur fait référence à une personne inconnue des autres interlocuteurs via l'usage de pronom tels que *nous* et *il/elle/lui* en considérant que la personne à qui il fait référence va de soi. L'utilisation du pronom *nous* par un locuteur, lui-même dans un couple hétérosexuel, pour faire référence à lui-même et son ou sa partenaire auprès de personne ne le ou la connaissant pas comme dans la phrase « nous sommes partis en vacances cet été » en est un bon exemple. En plus de ces références de personne, nous trouvons également dans cette catégorie l'usage de termes de parenté tels que *mari* et *femme* ou les termes en rapport à la belle famille (KITZINGER, 2005). Les observations de cette étude nous ont ainsi permis de constater que l'hétérosexualité était indexée dans le langage à travers des termes directement en rapport avec cette orientation sexuelle et pouvant potentiellement servir de critères de collecte telles que les références explicites citées ci-dessus. De premiers tests de collecte à partir de syntagmes du champ lexical de la relation hétérosexuelle tels que *my husband, my wife, my boyfriend, my girlfriend* montrèrent cependant rapidement les trop importantes limites du choix de ces termes. Comme nous avions pu le lire chez Smyrnaios et Ratinaud (2013) dans la présentation des limites liées à cette méthodologie, les références explicites à l'hétérosexualité par Kitzinger (2005) étaient non seulement trop courantes pour ne donner accès qu'à des données en rapport avec les féminités hétérosexuelles, mais étaient également trop limitées pour prétendre à une quelconque représentativité sur le sujet des féminités hétérosexualités. Par ailleurs, couvrir le champ lexical de l'hétérosexualité féminine s'est rapidement révélée être une entreprise d'une trop grande envergure, qui en plus d'un profond travail de définition de la notion, aurait nécessité un très grand nombre de collectes différentes, elles-mêmes aboutissant à une importante quantité de données chacune.

Ces tests préliminaires nous ont donc poussé à resserrer notre objet de recherche, avec pour objectif d'identifier un phénomène moins polysémique. Une des manières d'atteindre cet

objectif fut de se concentrer sur des thématiques et phénomènes spécifiques aux RSN afin que l'objet d'étude puisse être délimitable à travers une petite poignée de syntagmes, mots-clefs ou hashtags. Ayant parallèlement continué à suivre les discours sur l'hétérosexualité et l'apparition de tendances et hashtags en lien avec l'hétérosexualité sur Twitter et Instagram nous avons finalement fait le rapprochement entre deux tendances similaires trouvées sur Reddit et Instagram venant du subreddit r/arethestraightsok et du compte Instagram @heterocringe, dont le but était de tourner l'hétérosexualité en dérision afin de lui retirer son image d'orientation sexuelle standard, « normale » et « saine ». Ce sujet se cristallisant notamment autour de la phrase « are the straights ok » et de l'expression « heterocringe », il nous a donc semblé plus adapté et pertinent pour une recherche thématique ayant les RSN pour objet.

Du corpus brut au corpus de référence : méthodologie de tri des données

Malgré les changements opérés afin d'effectuer une collecte de données la plus précise possible, nous avons tout de même dû passer par une phase de tri. En effet les APIs de Twitter, sur lesquels reposent les outils de collecte de données, ne font pas de distinction de contexte ou de forme et collectent donc tous les tweets contenant le ou les syntagmes recherchés, incluant aussi bien les retweets que les tweets comprenant les syntagmes recherchés mais n'appartenant pas au sujet ciblé. Par ailleurs, nous nous sommes également heurtés au manque de fiabilité de certaines fonctionnalités de ces outils conduisant à l'imprécision du set de données final. Nous avons donc dû traiter différents types de données ayant été obtenues à travers notre collecte mais ne faisant pas partie des données ciblées.

Le premier problème résultant en la collecte de données non-ciblées relève d'une limitation technique hors de notre portée car venant du moteur de recherche de l'outil TAGS et des conventions syntaxiques d'utilisation des moteurs de recherche. En effet, nous nous sommes aperçus après la collecte que l'utilisation des guillemets pour cibler un syntagme nominal figé n'avait pas fonctionné, entraînant la collecte de données contenant les deux mots recherchés sans que ceux-ci soient sous la forme des syntagmes figés. Ce problème s'est également croisé à la polysémie des mots recherchés. Le mot *straight* étant en effet polysémique en anglais, contrairement au mot *hétéro,* nous avons par exemple collecté des tweets contenant le syntagme recherché dans un contexte tout autre que celui ciblé, tel que celui de la Formule 1, où l'expression *by the straights* se trouvait régulièrement utilisée. Bien qu'il fut possible de repérer certaines de ces données non pertinentes en identifiant les expressions et contextes dans

lesquelles elles apparaissaient, comme dans le cas des tweets sur la Formule 1, et en utilisant l'outil de recherche sur la totalité du corpus, la taille du corpus et l'envergure du tri à effectuer a limité l'intervention manuelle.

Une seconde variable intervenant dans le tri des données concerne l'inclusion des retweets qui représentaient plus de la moitié des données brutes. Nous nous sommes ici recentrés sur notre question de recherche qui vise à analyser les différentes stratégies utilisées afin d'opérer une resignification de l'hétérosexualité dans les discours observés et s'intéresse donc moins au poids d'un discours à travers le nombre de retweet, information qui reste par ailleurs accessible via l'URL du tweet. A l'instar de Hardaker et McGlashan (2016) nous avons choisi de nous concentrer sur les formes directes de discours.

Le dernier aspect que nous avons dû prendre en compte dans le tri de ces données et qui représente un défi majeur des corpus numériques tient aux questions éthiques vis-à-vis de données provenant d'internet. Nous nous sommes en effet posé la question de l'inclusion au corpus des tweets provenant de profils privés ou ayant été supprimés après la collecte. Face à la tâche que la suppression manuelle des données provenant de comptes privés représentait nous avons ici choisi de faire une distinction entre l'inclusion dans le corpus de référence – le corpus contenant l'ensemble des données, qui sera utilisé pour l'analyse lexicométrique – et dans le corpus de travail – le corpus sur lequel reposeront les analyses qualitatives manuelles. Les données provenant de comptes privés ou ayant été supprimées seront donc uniquement intégrées dans l'analyse quantitative mais ne feront pas l'objet d'une analyse manuelle et a fortiori ne seront pas utilisées comme exemple illustratif.

Malgré les ajustements de notre objet de recherche afin de se rapprocher des attentes en termes de représentativité et de limiter les effets de polysémie dans le but d'obtenir un corpus le plus précis possible – chose que nous n'avons pas pu éviter – certaines variables sont intervenues dans la collecte, affectant l'exactitude du corpus brut. En effet nous avons non seulement dû faire face à des impasses techniques provenant des outils auxquels nous avons eu recours ainsi que des impasses liées à la taille conséquente du corpus rendant complexe toute opération manuelle à grande échelle. Nous avons également dû prendre en considération des standards éthiques, aspect parfois omis dans les méthodologies de collecte de données numériques.

Les données numériques multilingues face à l'analyse lexicométrique

Outre les défis méthodologiques liés à la création du corpus de référence nous nous sommes également heurtés à des obstacles durant la phase de préparation de l'analyse lexicométrique. Comme nous l'avons brièvement évoqué, nous avons fait le choix de réaliser une analyse quantitative lexicométrique sur la totalité du corpus de référence face à l'impossibilité de l'analyser manuellement dans sa totalité. Ce choix méthodologique a néanmoins comporté de nouveaux défis résultant de la complexe articulation entre la nature numérique et multilingue de nos données et les possibilités offertes par les outils d'analyse lexicométrique à notre disposition.

Une des premières limitations techniques que nous avons rencontrées provient des logiciels de lexicométrie et leur capacité d'une part à analyser la communication médiée par ordinateur (CMO) qui comprend entre autres des acronymes, des émoticônes et des répétitions de lettres ou de marques de ponctuation, et d'autre part à avoir le même niveau de prise en charge dans les trois langues avec lesquelles nous travaillons. En effet, les logiciels de lexicométrie sont souvent des initiatives de petites tailles, développées par des chercheurs, comme AntConc, ou des équipes de recherche. Leurs capacités techniques dans d'autres langues que l'anglais dépend donc de l'apport de la communauté d'utilisateurs comme c'est le cas avec les outils de lemmatisation et d'étiquetage morphosyntaxique fonctionnant par exemple sur AntConc. La préparation d'un corpus avec ces deux types d'outils est en effet une question incontournable dans le cadre d'une analyse lexicométrique (NEE, 2017). La lemmatisation et l'étiquetage morphosyntaxique sont généralement appliqués aux corpus afin de faciliter l'analyse lexicométrique et de la rendre plus précise. L'étiquetage morphosyntaxique consiste en effet à apposer une annotation sur chaque forme graphique indiquant entre autres la classe grammaticale, le genre, le nombre et le temps. La lemmatisation, quant à elle, consiste à regrouper « sous une forme canonique (telle qu'on la trouve dans les dictionnaires) tous les substantifs, les adjectifs, les verbes ainsi que les formes élidées » (NEE, 2017, p. 106) réduisant ainsi le « nombre de types d'unités pris en compte dans les calculs » (NEE, 2017). En plus de devoir être développés dans chaque langue du corpus cible, les lemmatiseurs et les étiqueteurs doivent également être entraînés pour le type de langage qu'ils vont devoir analyser, ce qui pose rapidement problème lorsqu'il s'agit de CMO. En effet, seuls quelques outils spécialisés sur la CMO en anglais et nécessitant de bonnes connaissances en programmation permettent d'effectuer les procédures dont nous venons de parler. Face à la nature multilingue et numérique de notre corpus, les limitations techniques que nous avons

rencontrées ne nous ont donc pas permis d'atteindre le niveau de précision pouvant être attendu de la préparation de données dans le cadre d'une analyse lexicométrique en linguistique du corpus.

Limites et difficultés de l'analyse qualitative des grands corpus numériques

La nature numérique de nos données a également posé des défis méthodologiques par rapport à la préparation de l'analyse qualitative. Comme nous le précisons dans notre cadre théorique, nous faisons le choix d'adopter l'approche symétrique proposée par Paveau (2013) impliquant une conception du contexte comprenant « l'ensemble du dispositif dans lequel [les énoncés] sont produits » (PAVEAU, 2013, p. 2), c'est-à-dire dans notre cas aussi bien les discussions dans lesquelles sont parfois insérées les données, les commentaires et réponses qui peuvent être présents, que le support numérique sur lequel elles sont publiées. Cette approche pose cependant la question du niveau d'intégration du contexte vis-à-vis du corpus de référence : faut-il collecter ces données langagières et non-langagières contextuelles ? et quelles sont les méthodes pour les collecter ? Bien qu'il existe quelques solutions pour répondre à ces questions, telles que les captures d'écran et la transcription des discours contextuels, celles-ci sont considérablement chronophages lorsque l'on travaille avec un corpus de grande taille. Nous avons donc fait le choix de nous en tenir à la consultation native des données, c'est-à-dire en utilisant leur URL pour les consulter sur la plateforme dont elles viennent, tout en capturant l'environnement natif des données lorsque celles-ci seront utilisées pour nos propos. Cette solution soulève cependant de nouvelles questions méthodologiques dans le contexte des corpus numériques : que faire lorsque les données natives n'existent plus ou qu'il n'est plus possible d'avoir accès à une partie du contexte les entourant ? Que faire par exemple lorsque la totalité d'un compte Instagram a été supprimé dans un acte de censure de la plateforme sur laquelle il se trouvait comme c'est le cas pour le compte @heterocringe ? Nous avons ici jugé que l'intérêt des données, pour les questions qu'elles posent en terme d'hétéronormativité et parce qu'elles semblent déranger – jusqu'à susciter la censure d'Instagram –, l'emportaient sur l'absence des données nativement.

Enfin le dernier défi que nous avons rencontré, dû à la nature de notre corpus et à sa taille conséquente, concerne le choix du point de départ de l'analyse qualitative, en particulier vis-à-vis de notre sous-corpus Twitter. Bien qu'il fut possible de lire la totalité du corpus Instagram en raison de sa plus petite taille, nous avons pris le parti, pour ce qui est du sous-

corpus Twitter, de baser nos premières observations sur plusieurs extraits de 1000 à 1500 posts prélevés à différents moments chronologiques du corpus, à la manière d'Ilbury (2022), afin de pouvoir repérer de premiers éléments saillants qu'il sera ensuite possible de rechercher et rassembler à plus grande échelle dans le corpus. A travers les éléments que nous venons de présenter nous voulions donc montrer dans quelle mesure un corpus numérique de grande taille provenant de Twitter et Instagram pouvait représenter des défis méthodologiques – notamment liés aux choix théoriques entrepris – impactant également des aspects cruciaux du processus analytique et de quelles manières nous avons choisi de contourner ces difficultés.

Analyse quantitative préliminaire

Les choix méthodologiques que nous venons de présenter peuvent être illustrés à travers les premiers résultats de nos analyses quantitatives et qualitatives. L'analyse quantitative fut réalisée sur le logiciel en ligne Sketch Engine qui a la capacité de prendre en charge la lemmatisation et l'annotation des discours numériques dans les trois langues sur lesquelles nous travaillons en plus de permettre la comparaison de nos données avec des corpus de données en ligne de français et d'anglais. Dans un premier temps, nous avons choisi d'observer la fréquence des mots et des n-grams ainsi que les mots et n-grams clefs en effectuant une recherche à partir des formes lemmatiques. Les résultats furent ensuite retranscrits dans des tableaux à 20 entrées dans lesquels nous n'avons retenus que les substantifs, adjectifs et verbes[6] afin de nous concentrer sur le contenu discursif du corpus observé (BAKER, 2006), aboutissant ainsi à quatre tableaux pour chaque set de données[7] et résultant en un total de 20 tableaux. Nous avons ensuite classé les entrées de chaque tableau dans des catégories thématiques afin de synthétiser l'analyse. Nous nous concentrerons ici sur la présentation des résultats quantitatifs du sous-corpus provenant du compte Instagram @heterocringe.

Mise à part la thématique du *cringe* – thème principal du compte observé – se retrouvant en première place du tableau et que nous avons classé dans une catégorie « autre » du fait de son caractère unique, trois thématiques principales se dégagent du tableau des mots les plus fréquents ci-dessous (Fig. 1): les mots en rapport avec la thématique du genre, de la sexualité

[6] Certains mots présents dans les tableaux peuvent appartenir à des catégories grammaticales autres que celles retenues. Nous avons également maintenu les acronymes de phrases dans le tableau de mots clefs.
[7] Nos sets de données représentent les deux sous-corpus Instagram pour le compte @heterocringe & @herman_hetero_cringe ainsi que les sous-corpus twitter « the straights », « straight people », « les hétéros »

et des individus en général, les verbes exprimant un processus mental ou sensoriel et les verbes d'action (voir Fig. 2). Sans surprise, nous notons que la majorité des termes en rapport avec le genre et la sexualité se trouvent dans les dix premières lignes du tableau. Nous remarquons tout d'abord que les lemmes *HETERO* et *MAN* ont des fréquences très similaires, laissant ainsi envisager que les termes sont peut-être utilisés conjointement ou du moins qu'ils revêtissent d'une importance similaire par la position de domination qu'ils dénotent. Nous observons également que la fréquence de ces deux lemmes représente presque le double de la fréquence du lemme *WOMAN*, illustrant la moindre présence des femmes comme thématique dans le sous-corpus. Nous précisons par ailleurs que le lemme *GUY* renferme ici principalement la forme *guys* utilisé dans la phrase *u guys* comme cela est montré dans le tableau des n-grams les plus fréquents (Fig. 3). Nous retrouvons également un important nombre de verbes exprimant un processus mental ou sensoriel que nous interprétons comme des verbes utilisés pour décrire ou donner une opinion sur le contenu partagé sur le compte. Enfin nous distinguons quelques lemmes représentant des actions parmi lesquels figure le lemme *POST* pouvant suggérer un métadiscours du compte sur sa propre pratique ou la pratique qu'il observe.

Figure 1 - Tableau des mots les plus fréquents du sous-corpus @heterocringe

Mot	Fréquence	Mot	Fréquence
1. Cringe[8]	130	**11. MEAN**	45
2. GUY	100	**12. THINK**	39
3. HETERO	84	**13. LOOK**	36
4. MAN	83	**14. Good**	36
5. GET	61	**15. STOP**	35
6. STRAIGHT	58	**16. FIND**	34
7. MAKE	55	**17. SEE**	33
8. KNOW	54	**18. LOVE**	30
9. POST	54	**19. NEED**	30
10. WOMAN	47	**20. People**	30

Figure 2 : Catégorisation thématique des mots fréquents

Catégories	Mots
Genre, sexualité et personnes	GUY, HETERO, MAN, STRAIGHT, WOMAN, People
Processus mentaux et sensoriels	KNOW, MEAN, THINK, LOOK, SEE, LOVE, NEED
Actions	GET, MAKE, POST, STOP, FIND
Autres	Cringe, good, POST

[8] Bien que les recherches aient été faites à partir de lemmes, nous n'avons pas inscrit la forme lemmatique lorsque les mots ne présentaient pas différentes formes.

Le tableau des n-grams les plus fréquents (Fig. 3) confirme certaines de nos interprétations tout en en suggérant de nouvelles. Nous remarquons tout d'abord que les n-grams sur le thème du genre, de la sexualité et des personnes sont beaucoup moins nombreux que dans le tableau des mots (Fig. 4). Parmi les trois entrées présentes nous retrouvons *hetero cringe* et *MAN BE*, confirmant que les thématiques de l'hétérosexualité, des hommes et de la masculinité semblent bien prendre plus de place dans les discours du compte Instagram. Par ailleurs, il est intéressant de noter la construction morphosyntaxique du lemme *MAN BE* qui appelle des mots venant qualifier *MAN*. Comme le mentionne Greco (2012) la construction *NP is/are X* joue un rôle clef dans la propagation des normes de genre notamment en raison de sa valeur de vérité générale. Cela nous indique donc la présence de ce type de discours généralisant mais pose également la question de la manière dont cette structure va être utilisée, dans le contexte d'un discours s'opposant à l'hétéronormativité, pour retourner les qualifications traditionnellement associées aux hommes. La thématique qui domine néanmoins ce tableau est celle des verbes exprimant un processus mental ou sensoriel. Nous retrouvons par exemple quatre de ces verbes précédés de la première personne du singulier confirmant ainsi la prédominance d'un discours centré sur l'avis de la créatrice. Nous percevons également la présence d'un interlocuteur à travers l'utilisation d'une forme raccourcie de *you* mais principalement de l'expression *u guys* désignant un groupe que nous supposons être les followers. Enfin nous retrouvons également de nouvelles marques de métadiscours venant appuyer l'existence d'une réflexion du compte sur sa propre pratique et qui viennent s'ajouter à *POST this* à travers le n-gram *this page* faisant référence au compte lui-même ou encore *FOUND by* qui semble qualifier le contenu partagé et pourrait être la trace d'une contribution de la part des followers.

Figure 3 - Tableau des n-grams les plus fréquents du sous-corpus @heterocringe

N-Gram	Fréquence	N-Gram	Fréquence
1. **I mean**	26	**11. POST this**	11
2. Found by	24	**12. THINK this**	11
3. hetero cringe	21	**13. LOOK at**	10
4. **I THINK**	20	**14. the fuck**	10
5. u know	18	**15. I wish**	10
6. u guys	16	**16. idk what**	9
7. MAKE me	15	**17. thank u**	9
8. WANT to	14	**18. I FEEL**	9
9. this page	12	**19. to make**	9
10. NEED to	12	**20. MAN BE**	8

Figure 4 - Catégorisation thématique des n-grams fréquents

Catégories	Mots
Genre, sexualité et personnes	Hetero cringe, u guys, MAN BE
Processus mentaux et sensoriels	I mean, I THINK, u know, WANT to, NEED to, THINK this, LOOK at, I wish, idk what, thank u, I FEEL
Actions	Found by, MAKE me, POST this, to make
Autres	This page

Contrairement aux tableaux de fréquences ceux de mots (Fig. 5) et n-grams clefs (Fig. 7) nous montrent les termes dont l'utilisation est plus élevée dans notre corpus comparé à leur apparition dans un corpus offrant une vue plus générale des discours numériques, ici le corpus English Web 2020 (enTenTen 2020). Typique des RSN, en partie en raison de leur utilisation sur mobile qui vient contraster avec le reste des données Web écrites par ordinateur, nous retrouvons quelques acronymes de phrases servant à exprimer une opinion telle que *idk* pour *I don't know* ou *tbh* pour *to be honest*. Également liée au contexte et aux pratiques spécifiques des RSN, sont les mentions du RSN Tiktok, de *MEME* ou de *tw*[9] et qui viennent ici souligner, dans ce sous-corpus, la centralité de Tiktok et des memes comme source de contenu qui est repartagé par le compte ainsi que la nature sensible de ce contenu qui nécessite régulièrement un trigger warning. Outre ces éléments, ce qui ressort principalement de ce tableau sont les entrées en rapport avec le genre, la sexualité et les personnes. Nous pouvons tout d'abord identifier plusieurs entrées impliquant des discriminations ou pratiques sexistes telles que

[9] Signifie « trigger warning ». Est communément utilisé sur les réseaux sociaux et inscrit en début de post pour indiquer la nature sensible du contenu publié.

incel[10], *terf*[11], *transphobic* voire *heteronormative*. D'autre part nous trouvons également des termes désignant les différents groupes sur lesquels le compte se concentre avec les mots *cis* et *hetero* – groupes dominants – face notamment aux *enbies*[12] et *wlw*[13] – groupes dominés. L'importance de cette thématique, mise en évidence à travers la précision et la diversité des entrées qui la compose, illustre en détail le contenu et les problématiques abordées par le compte Instagram ainsi que les groupes sociaux qu'il implique. Enfin nous notons, non sans un certain humour, la présence des termes *christmas* et *tshirt* signalant une prévalence notable du contenu en lien avec ces deux mots – le compte poste en effet beaucoup de photos de t-shirt contenant des messages sexistes – par rapport aux discours numérique en anglais en général.

Figure 5 - Tableau des mots clefs du sous-corpus @heterocringe

Mots clefs	
1. Hetero	**11. christmas**
2. Cringe	**12. weenies**
3. Idk	**13. ngl**
4. INCEL	**14. enbies**
5. Ppl	**15. MEME**
6. tbh	**16. terf**
7. queercringe	**17. tiktok**
8. tw	**18. heteronormative**
9. cis	**19. transphobic**
10. wlw	**20. tshirt**

Figure 6 - Catégorisation thématique des mots clefs

Catégories	**Mots**
Genre, sexualité et personnes	Hetero, incel,ppl, queercringe, cis, wlw, weenies, enbies, terf, heteronormative, transphobic
Pratiques des réseaux sociaux	Tw, MEME, tiktok
Acronymes d'opinion	Idk, tbh, ngl
Thématique du contenu partagé	Christmas, tshirt, cringe

Enfin, en ce qui concerne les n-grams clefs (Fig. 7.) il est intéressant de remarquer les similitudes des tendances notables (Fig. 8) par rapport à celles du tableau des n-grams fréquents. La domination numérique des n-grams comprenant des verbes exprimant un processus mental

[10] Signifie « involuntary celibate ».
[11] Acronyme pour « trans-exclusionary radical feminist ».
[12] Vient de la prononciation de « NB » en anglais et signifiant non-binary.
[13] Signifie « woman-loving-woman ».

ou sensoriel – ici à la première et deuxième personne du singulier – souligne à nouveau l'existence d'une conversation, au moins venant de la créatrice, entre celle-ci et les followers, que l'on distingue à travers d'autres entrées telles que *Thank u* ou *what DO it MEAN*. En revanche, nous constatons la présence de trois entrées mentionnant le mot *cringe*, ce qui souligne triplement la singularité, sur Internet, du phénomène dont le compte observé se fait le porte-parole. Ces premières analyses qualitatives confirment d'une part la centralité de la thématique de l'hétérosexualité et du cringe au-delà du nom du compte et permettent d'autre part de dégager des pistes d'analyses qu'il sera possible d'approfondir qualitativement notamment pour ce qui est de la place de la masculinité mais aussi des discours sur les attitudes discriminatoires et sexiste. Elle met également en évidence une certaine relation entre la créatrice du compte et les followers et laisse également apparaître des traces d'une réflexion métadiscursive sur cette pratique discursive.

Figure 7 - Tableau des N-grams clefs du sous-corpus @heterocringe

N-Gram clefs	
1. FIND by	11. MAKE me
2. hetero cringe	12. THINK this
3. u know	13. this page
4. u guys	14. POST it
5. I mean	15. the cringe
6. POST this	16. u need
7. the fuck	17. please stop
8. idk what	18. what DO it MEAN
9. thank u	19. I WISH
10. cringe BE	20. Warning slide

Figure 8 - Catégorisation thématique des n-grams clefs

Catégories	**N-gram**
Processus mentaux et sensoriels	u know, I mean, idk what, make me, think this, u need, I wish
Actions	FIND by, POST this, make me, POST it
Cringe	Heterocringe, cringe BE, the cringe
Politeness markers	Thank u, please stop, what DO it mean
Autres	U guys, the fuck, this page, warning slide

Analyse qualitative préliminaire

Nous évoquions plus haut la lecture d'extraits de données afin d'initier l'analyse qualitative face à la quantité importante de données présente. Suite à cette première lecture nous avons pu effectuer une classification thématique préliminaire des deux sous-corpus de données Instagram et Twitter. Bien qu'un très grand nombre de sujets soient abordés dans les données Instagram, nous avons fait le choix d'établir quatre grandes catégories à partir des thématiques les plus récurrences apparaissant dans les légendes des posts – et non à partir du contenu visuel partagé dans les posts. Une des thématiques ressortant le plus dans ces données est naturellement celle du genre et de la sexualité dans laquelle nous classons les posts parlant d'hétérosexualité et du couple hétérosexuel, de masculinités et féminités et des questions LGBTQI+ ainsi que des discriminations (Fig. 9). Nous avons également noté une récurrence de métadiscours faisant apparaître certaines dynamiques de la pratique discursive représentée par le compte notamment à travers des légendes à propos de la gestion du compte mais aussi soulignant des légendes interagissant avec les followers ou questionnant le sens du contenu partagé. Une troisième catégorie représente un métadiscours ayant cette fois-ci pour objet le contenu partagé dans les posts et pointant du doigts l'échec de certains types de mise en scène récurrentes de ce discours hétéronormatif. La créatrice souligne par exemple l'usage d'une mise en scène excessivement pompeuse, le recours à un humour de mauvais goût ainsi que les discours hyperboliques. Enfin dans une dernière catégorie nous avons regroupé les légendes exprimant des sentiments négatifs suscités par le contenu partagé tels que le désespoir, le choc ou le dégoût mais aussi le simple désaccord.

Figure 9 - Example du compte @heterocringe[14]

* **heterocringe:** Just get a divorce do something but please end these memes

En ce qui concerne le sous-corpus Twitter nous avons également pu observer quelques thématiques et stratégies discursives récurrentes. L'un des discours qui domine le corpus est tout d'abord l'expression, sous des formes diverses et variées, d'une aversion pour les personnes hétérosexuelles. De plus nous retrouvons beaucoup de discours opposant une culture hétérosexuelle à une culture LGBTQI+ et dénonçant la domination du regard hétéronormatif dans des aspects culturels ou encore l'appropriation par des personnes hétérosexuelles d'éléments LGBTQI+. Nous observons également la présence de discours opérant un retournement du stigmate en reprochant par exemple un manque de pudeur aux couples hétérosexuels s'embrassant dans la rue (Fig. 10). Par ailleurs, à l'image du sous-corpus Instagram, nous avons identifié des métadiscours ayant pour objet les discours sur les personnes hétérosexuelles, aussi bien d'un point de vue LGBTQI+ qu'hétérosexuel. Enfin parmi les discours récurrents, nous pouvons également mentionner les discours sur le rapport des personnes hétérosexuelles vis-à-vis de l'hétéronormativité et des comportements qui s'en affranchissent. A l'instar du sous-corpus Instagram ces catégories ne représentent que certains des discours récurrents présents dans les données.

[14] Il s'agit d'une recréation du post.

Figure 10 - Exemple de retournement du stigmate

Conclusion

Dans cet article nous avons voulu présenter les défis méthodologiques que nous avons rencontrés dans notre travail de recherche quantitatif et qualitatif en analyse du discours sur la resignification de l'hétérosexualité, liés à la nature numérique, multimodale et multilingue de notre corpus de données et à nos choix analytiques et comment ces défis nous ont contraint à faire des choix méthodologiques déterminants pour nos résultats finaux. Nous avons notamment montré comment, dans le cadre d'une recherche en analyse du discours, le choix d'un terrain numérique et les possibilités de collecte déterminées par la plateforme mais également les usages qu'en font les utilisateurs nous ont amené à faire évoluer notre objet de recherche initial vers une thématique facilement délimitable lexicalement, pour répondre d'une part à des attentes de pertinence de recherche sur les RSN et d'autres part pour atténuer certains des écueils de la collecte thématique de données sur les RSN répondant ainsi à des attentes en terme de représentativité. Nous avons également voulu mettre en avant la problématique de la collecte de données non pertinentes ainsi que l'intégration des considérations éthiques à l'échelle des grands corpus de données numériques. Nous avons ensuite mis en lumière la phase de préparation par laquelle peut passer un corpus numérique conséquent en vue d'une analyse quantitative lexicométrique et comment cette étape, à priori peu problématique lorsque l'on travaille avec un langage écrit standard, a soulevé un bon nombre de challenges méthodologiques et a nécessité des choix impliquant parfois une perte en précision d'analyse. Dans la suite de l'article nous avons également abordé les limites et difficultés rencontrées avec les corpus numériques lorsque l'on réalise une analyse qualitative en analyse du discours. Nous avons par exemple posé la question de l'intégration des données contextuelles si importantes à l'analyse, notamment face au caractère instable des données numériques et comment nous

avons été amenés à faire des concessions en termes d'intégration de ces données dans le corpus de référence compte tenu de leur disparition ou de limitations techniques en termes de collecte. Nous avons aussi évoqué la problématique de la sélection des données pour l'analyse qualitative finale lorsque l'on fait face à des centaines de milliers de tweets et comment nous avons choisi remédier à cette difficulté en sélectionnant des extraits du corpus permettant ainsi de repérer de premiers éléments pertinents sur une échelle réduite. L'article se termine enfin par la présentation de résultats et d'analyses quantitatives et qualitatives préliminaires dans lesquelles nous exposons notamment l'analyse lexicométrique effectuée sur le sous-corpus issu du compte @heterocringe ainsi que les premières tendances que nous avons observées en ce qui concerne les thématiques abordées sur la totalité du corpus.

Références bibliographiques

ALTHUSSER, Louis. Idéologie et appareils idéologiques d'État. (Notes pour une recherche) . **La Pensée**. Revue du rationalisme moderne, n. 151, p. 67-125, 1970.

BAKER, Paul. **Using Corpora in Discourse Analysis**. London; New York: Continuum, 2006. DOI: https://doi.org/10.5040/9781350933996

BUCHOLTZ, Mary. The Feminist Foundations of Language, Gender, and Sexuality Research. IN: EHRLICH, S.; MEYERHOFF, M.; HOLMES, J. (eds.). **The Handbook of Language, Gender, and Sexuality**. 2. ed. Chichester, West Sussex: Wiley Blackwell, 2014, p. 23-47. DOI: https://doi.org/10.1002/9781118584248.ch1

BURGESS, Jean; BRUNS, Axel. Twitter Archives and the Challenges of "Big Social Data" for Media and Communication Research. **M/C Journal**, v. 15, n. 5, 2012. Consulté sur : <https://journal.media-culture.org.au/index.php/mcjournal/article/view/561>. DOI: https://doi.org/10.5204/mcj.561

BUTLER, Judith. **Excitable speech**: a politics of the performative. New York: Routledge, 1997.

BUTLER, Judith. **Gender trouble**: feminism and the subversion of identity. New York; London: Routledge, 1999 [1990].

FAIRCLOUGH, Norman. **Language and power**. London ; New York: Longman, 1989.

GRECO, Luca. Production, circulation and deconstruction of gender norms in LGBTQ speech practices. **Discourse Studies**, v. 14, n. 5, p. 567-585, 2012. Consulté sur : <https://journals.sagepub.com/doi/10.1177/1461445612452229>. DOI: https://doi.org/10.1177/1461445612452229

HARDAKER, Claire; MCGLASHAN, Mark. "Real men don't hate women": Twitter rape threats and group identity. **Journal of Pragmatics**, v. 91, p. 80-93, 2016. Consulté sur :

<https://www.sciencedirect.com/science/article/pii/S0378216615003100?via%3Dihub>. DOI: https://doi.org/10.1016/j.pragma.2015.11.005

HERRING, Susan. Gender Differences in CMC: Findings and Implications. **CPSR Newsletter**, v. 18, n. 1, 2000.

JACKSON, Stevi. Récents débats sur l'hétérosexualité: une approche féministe matérialiste. **Nouvelles Questions Féministes**, v. 17, n. 3, p. 5-26, 1996.

KIESLING, Scott F. **Language, gender, and sexuality**: an introduction. London; New York: Routledge, Taylor & Francis Group, 2019. DOI: https://doi.org/10.4324/9781351042420

KITZINGER, Celia. "Speaking as a Heterosexual": (How) Does Sexuality Matter for Talk-in-Interaction? **Research on Language & Social Interaction**, v. 38, n. 3, p. 221-265, 2005. Consulté sur : <https://www.tandfonline.com/doi/abs/10.1207/s15327973rlsi3803_2>. DOI: https://doi.org/10.1207/s15327973rlsi3803_2

KRIEG-PLANQUE, Alice. **Analyser les discours institutionnels**. Malakoff: Armand Colin, 2017.

KUNERT, Stéphanie. Dégenrer les codes : une pratique sémiotique de défigement. **Semen**, n. 34, 2012. Consulté sur : <https://journals.openedition.org/semen/9770>. DOI: https://doi.org/10.4000/semen.9770

LAKOFF, Robin Tolmach. **Language and woman's place**: text and commentaries. Revised and expanded edition ed. Oxford: Oxford University Press, 2004 [1975].

LEAP, William L. Queer Linguistics as Critical Discourse Analysis. In: TANNEN, D.; HAMILTON, H. E.; SCHIFFRIN, D. (eds.). **The Handbook of Discourse Analysis**. 2. ed., v. 2. Chichester: Wiley Blackwell, 2015, p. 661-680. DOI: https://doi.org/10.1002/9781118584194.ch31

LEAP, William. L. Queer linguistics, sexuality and discourse analysis. In: GEE, J. P.; HANDFORD, M. (eds.). **The Routledge Handbook on Discourse Analysis**. New York; London: Routledge, 2012, p. 558-571.

PAVEAU, Marie-Anne. Genre de discours et technologie discursive. Tweet, twittécriture et twittérature. **Pratiques**, n. 158-157, p. 7-30, 2013. Consulté sur : <https://journals.openedition.org/pratiques/3533>. DOI : https://doi.org/10.4000/pratiques.3533

PAVEAU, Marie-Anne. La blessure et la salamandre. Théorie de la resignification discursive. v. hal-02003667, 2019. Consulté sur : <https://hal.archives-ouvertes.fr/hal-02003667>.

RICH, Adrienne. La contrainte à l'hétérosexualité et l'existence lesbienne. **Nouvelles Questions Féministes**, n. 1, p. 15-43, 1981.

RIEDER, Bernhard. The refraction chamber: Twitter as sphere and network. **First Monday**, 4 nov. 2012. DOI : https://doi.org/10.5210/fm.v17i11.4199

SMYRNAIOS, Nikos; RATINAUD, Pierre. Comment articuler analyse des réseaux et des discours sur Twitter: L'exemple du débat autour du pacte budgétaire européen. **Tic & société**, v. 7, n. 2, 2013. Consulté sur : <https://journals.openedition.org/ticetsociete/1578>. DOI : https://doi.org/10.4000/ticetsociete.1578

WITTIG, Monique. **The straight mind and other essays**. Boston: Beacon Press, 1992.

Recebido em: 2 de junho de 2022
Aceito em: 1 de agosto de 2022