UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LOL SUR TWITTER : UNE APPROCHE DU CONTACT DE LANGUES ET DE LA VARIATION PAR L'ANALYSE DES RÉSEAUX SOCIAUX

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN LINGUISTIQUE

PAR
JOSHUA MCNEILL

AOÛT 2018

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.10-2015). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Tout d'abord, je remercie tous ceux à l'Université du Québec à Montréal qui m'ont donné l'opportunité d'entreprendre ce travail. Si je regarde où j'en étais au début de mes études à l'UQÀM et mon ignorance, je me considère chanceux d'avoir obtenu cette opportunité. Je dois surtout remercier mon directeur, Philip Comeau, pour sa patience et ses commentaires judicieux, et Anne Rochette pour ses efforts réalisés dans le but d'obtenir les subventions nécessaires qui m'ont permis de m'inscrire dans le programme. Je dois également remercier mes lecteurs, John Lumsden et Elizabeth Allyn Smith, et tous les professeurs qui m'ont transmis leurs connaissances. J'ai beaucoup appris dans tous mes cours, et pour cela, je leur en suis reconnaissant.

On a more personal note, I want to thank my friends and family for their support when I've been stressed and their understanding when I've been too busy to give them the attention they deserve. My mom, Linda McNeill, has been especially patient when my weekly calls home became monthly calls. I wouldn't have been able to do this without your support. Last, but especially not least, I have to thank my favorite, Shari Ex, for managing to be all I've ever needed through this whole process despite being thousands of miles away and for supporting me so selflessly despite that distance. You're amazing.

TABLE DES MATIÈRES

INDEX DES FIGURES

INDEX DES TABLEAUX

LISTE DES FORMULES

RÉSUMÉ

Cette étude vise à effectuer une analyse de la façon dont la variable linguistique lexicale (lol), constituée de variantes d'origine française et anglaise, est réalisée sur Twitter. Nous employons des outils actuels de l'analyse des réseaux sociaux qui sont peu connus dans la sociolinguistique variationniste, surtout la méthode de Louvain (Blondel *et al.*, 2008), et ce, afin de détecter les communautés et le PageRank (Brin et Page, 1998) afin de quantifier la centralité des individus dans ces communautés. Nous nous interrogeons à savoir si la distribution des réalisations de variable (lol) change de communauté en communauté et si les individus présentent moins de diversité dans leurs réalisations de (lol) que l'ensemble de leurs communautés. Nos résultats nous permettent de répondre par l'affirmative pour la première question, mais nous hésitons à affirmer de fortes conclusions pour la dernière question en raison d'un manque de données. Des améliorations à la collecte de données sont donc proposées. Finalement, nous croyons que notre analyse joue bien le rôle d'une validation de principe pour ce qui est des outils actuels de l'analyse des réseaux sociaux et fait avancer du même coup l'étude du contact de langues.

MOTS-CLÉS : sociolinguistique, contact de langues, variation, Twitter, analyse des réseaux sociaux

INTRODUCTION

Les provinces maritimes au Canada sont constituées du Nouveau-Brunswick, de la Nouvelle-Écosse et de l'Île-du-Prince-Édouard. Selon le recensement de 2011, le français comme langue maternelle est parlé par environ 270 000 personnes habitant dans ces trois provinces, mais l'anglais est tout de même bien présent et en fait majoritaire dans plusieurs communautés, étant donné qu'il est la langue maternelle d'environ 1 440 000 personnes (Statistique Canada, 2016). Cette situation de langues en contact a entraîné l'apparition de variétés qui présentent assez d'éléments des deux langues à la fois, à un point tel que le parler d'un certain nombre de résidents de la ville de Moncton, au Nouveau-Brunswick, et parfois le parler de ceux habitant dans le sud-est du Nouveau-Brunswick en général, a hérité de son propre glossonyme unique : le chiac. Young (2002) a défini le chiac comme un bout d'un continuum de variétés, parlé par les jeunes citadins de Moncton, l'autre bout étant le français acadien, parlé par les aînés ruraux (p. 9-11), mais elle a également cité une locutrice de Moncton qui l'a décrit comme une variété générale, parlée dans le sud-est du Nouveau-Brunswick (p. 7-8). King (2008), pour sa part, était d'accord avec l'idée que le chiac peut être défini comme une variété urbaine strictement reléguée à Moncton (p. 137). Plus important encore, King (2008) a affirmé que le chiac de Moncton n'est pas nécessairement spécial par rapport à la présence de traits provenant de l'anglais et du français, puisque plusieurs de ces traits se retrouvent dans une grande part de ce qui s'appelle couramment le français acadien (p. 138-139).

Tout cela veut simplement dire qu'il est généralement admis qu'il existe une situation de contact de langues dans plusieurs zones des provinces maritimes dans lesquelles on s'attend à ce que les traits du français et de l'anglais s'entremêlent. Dans ce

mémoire, nous avons donc recueilli des tweets de Twitter, émanant des provinces maritimes, dans lesquelles on retrouve des tweets comme celui de Ben H.[1] dans l'Exemple 1, un utilisateur qui s'identifie comme venant de Frédéricton, au Nouveau-Brunswick :

Ben H. : @cyrillesimard Il faut que j'ai une petite talk avec Brian Gallant ou Francine Landry...lol..:P (Exemple 1)

Dans ce tweet, toute la structure est d'origine française sauf deux mots qui sont d'origine anglaise : *talk* et *lol*, ce dernier étant un sigle pour *laugh out loud* 'éclater de rire', adjoint qui signifie que l'on trouve quelque chose amusant. L'existence des mots d'origine française pour ces deux-ci, soit *conversation* et *mdr* respectivement, ce dernier étant un sigle pour *mort(e) de rire*, n'empêche pas l'utilisation de mots d'origine anglaise. Dans ce mémoire, nous ne nous intéressons donc pas tant à la manière dont on identifie le chiac, ni d'ailleurs le français et l'anglais, mais plutôt nous nous intéressons à ce que l'on peut apprendre en analysant la façon dont les unités lexicales, dans ce cas celles comprises dans la variable linguistique (lol)[2], sont réalisées dans une telle situation d'un point de vue sociolinguistique variationniste. Nous analysons plus précisément le comportement linguistique des personnes qui se servent de Twitter et qui séjournent ou habitent dans les provinces maritimes, compte tenu de deux objectifs :

1 Nous avons anonymisé tous les noms de tous nos sujets.

2 Il y a en fait de nombreuses variantes qui constituent la variable linguistique (lol), y compris *lol*, *lmao*, *rofl*, *kek*, *mdr*, *ptdr* et ainsi de suite. Pour plus de détails, voir la section 3.4.4 Variable linguistique (lol).

(O1) D'effectuer une analyse variationniste traditionnelle (p. ex. Eckert, 2000; Labov, 1966/2006; Milroy, 1980/1987; Nadasdi, Mougeon, et Rehner, 2004) d'une variable linguistique qui implique des variantes d'origine française et anglaise en mettant en œuvre des outils actuels de l'analyse des réseaux sociaux.

(O2) De mieux comprendre le rapport entre les variétés des individus, les variétés des groupes et la façon dont ceux-ci interagissent.

DÉFINITIONS

Avant d'exposer la problématique, il est important de présenter les termes qui seront utilisés tout au long de ce mémoire pour faire référence à différents types de langage. En dépit du fait que tous ces termes se retrouvent déjà dans la littérature sociolinguistique, les définitions antérieures ne sont pas nécessairement identiques à celles utilisées ici.

Deux extrêmes se distinguent en termes du type de données qui peuvent constituer la base empirique de la linguistique. Le premier est celui de Chomsky, qui est peut-être l'un des chercheurs les mieux connus de toute la linguistique théorique, et l'autre est celui de Labov, qui est peut-être l'un des chercheurs les mieux connus de toute la sociolinguistique. Pour Chomsky (1986), la linguistique est un sous-domaine de la psychologie (p. 3), c'est-à-dire que ce que les chercheurs devraient viser à comprendre est le système langagier sous-jacent de l'individu. Compte tenu de ceci, il a développé le concept de la langue-E et la langue-I. La langue-E se définit comme la langue que l'on peut percevoir, tandis que la langue-I se définit comme ce qui existe dans l'esprit d'un usager d'une langue et doit être inférée à partir de la langue-E (Chomsky, 1986, p. 20-22). En distinguant ces deux langues, Chomsky a effectivement écarté toute la variation dans les langues perceptibles comme du bruit statistique afin d'arriver à une langue invariante, la langue-I, qui pouvait constituer l'unité d'analyse principale de la linguistique théorique.

Labov a cependant adopté une approche différente de celle de Chomsky. Labov (1969) a presque rejeté les idiolectes, c'est-à-dire les langues des individus, en avançant l'idée qu'ils présentent peu de structure sans considérer le contexte social.

De plus, il a affirmé que les jugements de grammaticalité des chercheurs, souvent utilisés dans la linguistique théorique, produisent des données que l'on ne peut évaluer. Pour Labov, les langues des groupes sont bien plus systématiques que les idiolectes et doivent donc constituer la base empirique de la recherche linguistique (p. 757-759). L'approche de Labov a été raffinée de diverses façons dans la recherche sociolinguistique, et, selon Labov (1966/2006) en 2006, l'analyse des individus n'est jamais revenue avec vigueur (p. 157), mais il s'est trompé un peu. L'étude de la variation stylistique est au moins implicitement l'étude des individus, et en plus, il existe plusieurs études qui examinent en détail le comportement langagier de seulement un à quatre sujets (voir les sections 1.3.5 Individus et 2.5 Variation stylistique).

Ce que ces deux approches opposées se partagent est la tendance à confondre ce que l'on comprend couramment par les concepts tels que l'anglais ou le français et les structures empiriquement vérifiables qui concernent les linguistes. Les concepts tels que l'anglais et le français font référence à des entités socio-politiques, intersubjectivement constituées, essentiellement ce que Kloss (1978) a appelé l'*ausbausprache*, l'inverse étant les langues définies par leurs structures, ou l'*abstandsprache* (cité dans Brunstad, 2003, p. 58-59). Par exemple, et peut-être seulement pour des raisons pratiques[3], dans l'étude fondatrice de Labov (1966/2006) du Lower East Side de la ville de New York, il a effectué une analyse de la variation qui s'opère explicitement en « anglais » et non dans « la langue de *X* » où *X* est un groupe. Il a implicitement présumé que l'appellation *anglais* n'est pas seulement une étiquette socio-politique mais qu'elle réfère à quelque chose qui a une réalité

3 D'une certaine façon, c'est presque incontournable. Nous employons nous-mêmes les descriptions « d'origine anglaise » et « d'origine française » dans ce mémoire pour faire référence aux origines des unités lexicales, même si notre position affirme que les langues existent principalement comme des *ausbausprache*s, tandis que la réalité empirique de leur existence est très différente.

structurelle et dont fait partie la langue qu'eux, les New-Yorkais du Lower East Side, parlent. En 1983, Weiner et Labov ont plus explicitement affirmé qu'une entité structurelle monolithique qui s'appelle « l'anglais » existe en disant que « le besoin de variation stylistique conduit tous les locuteurs et écrivains de l'anglais à substituer un mot à un autre[4] » (p. 30). On peut supposer que la plupart des sociolinguistes reconnaissent cette confusion potentielle, mais celle-ci semble parfois resurgir, et nous voulons bien éviter cette confusion qu'ont aussi notée Le Page et Tabouret-Keller (1985) lorsqu'ils proposaient leurs quatre « sens de la langue » (p. 189-191). Des syntacticiens célèbres parlent également de « l'anglais » ou du « français » plutôt que de « la langue de *X* » où *X* est un individu (p. ex. Cinque, 2002; Kayne, 2007; Rizzi, 2013; etc.). Ils regroupent souvent même des données provenant des jugements de grammaticalité de plusieurs individus sous l'étiquette de « l'anglais » ou du « français » ou simplement n'identifient pas les sources des jugements, comme si une connaissance intuitive de la « langue » qu'ils étudient passe bien pour des données. Nous ne voulons pas dire que leurs travaux sont inutiles, mais plutôt qu'il est facile de simplifier ainsi les faits et, ce faisant, de perdre quelque chose.

Dans ce mémoire, nous voulons complètement éviter l'*ausbausprache*, et donc nous nous servirons d'une terminologie qui aidera à maintenir la distinction entre lui et l'*abstandsprache*, ainsi qu'une terminologie qui aidera à maintenir la distinction entre les *abstandsprache*s des groupes et les *abstandsprache*s des individus. Désormais, les termes *langue* et *dialecte* seront utilisés pour faire référence aux *ausbausprache*s, tandis que le terme *registre* sera utilisé pour faire référence aux *abstandsprache*s des groupes dans des contextes donnés[5], le terme *style* pour faire référence à un seul

4 Toutes les traductions dans ce mémoire sont les nôtres.

5 Nous nous concentrons sur l'identification des groupes dans ce mémoire, mais le contexte, soit une activité ou une localisation ou un temps, est également important, et c'est pour cette nature multidimensionnelle que nous choisissons le terme *registre*, à l'instar plus ou moins de l'usage de Halliday (1964/1968, p. 141).

abstandsprache d'un individu et le terme *répertoire* pour faire référence au système de styles d'un individu, c'est-à-dire tous les styles qu'un individu donné a dans son esprit. Finalement, le terme *variété* sera utilisé comme un terme général qui fait référence à n'importe quel type d'*abstandsprache*.

CHAPITRE I

PROBLÉMATIQUE

1.1 Introduction

Nos objectifs dans ce mémoire nécessitent de se pencher sur deux enjeux qui peuvent se retrouver dans la littérature : le contact de langues[6] et la délimitation des communautés, compte tenu de la position des individus et leurs systèmes langagiers dans ces communautés. Chaque enjeu sera examiné dans cette section-ci afin de développer des questions de recherche pertinentes qui découlent de nos objectifs.

1.2 Contact de langues

L'étude du contact de langues donne l'occasion de vérifier plus facilement l'existence de divers registres, divers styles et la façon dont ils sont liés que les études de la variation linguistique qui impliquent des variétés qui se ressemblent peut-être suffisamment au niveau structurel pour que l'on puisse les imaginer faire partie de la même langue. Par exemple, si deux locuteurs dans une étude ne se comprennent pas, on sait bien qu'il y a au moins deux variétés en jeu. Or, malgré cet avantage, le

6 Dans ce cas, le terme *langue* est maintenu non seulement parce qu'il est un terme technique dans la littérature, mais aussi parce qu'il distingue que nous nous intéressons aux variétés qui sont jugées assez disparates pour qu'elles prennent des glossonymes à *ausbausprache*. Un terme plus exact conforme à la position de ce mémoire serait le contact de variétés, puisque nous considérons toute la sociolinguistique variationniste comme l'étude du contact.

contact de langues ne figure pas en bonne place dans la sociolinguistique variationniste, même si les études du contact de langues peuvent souvent être décrites comme variationnistes. Cette section-ci discutera de l'interprétation des études du contact de langues comme des études de la variation et vice versa, des langues de base et enfin des emprunts contre les items alternés.

1.2.1 Études du contact de langues ou de la variation

Il existe plusieurs études de contact de langues qui peuvent être décrites comme variationnistes (p. ex. Brown, 2003; Ehresmann et Bousquette, 2015; Poplack, 1979/1980/2000; Poplack et Dion, 2012; Poplack, D. Sankoff, et Miller, 1988), mais on en trouve peu qui se servent des variables linguistiques constituées de variantes de différentes « langues », étudiées de la même façon dont on étudie la variable linguistique (ing) (voir Labov, 1966/2006; Trudgill, 1974; etc.), par exemple. En effet, Poplack *et al.* (1988) ont effectué une analyse des emprunts à l'anglais dans le français d'Ottawa et ses alentours, mais ils n'ont pas mis en œuvre des variables linguistiques pour le faire, dans le sens où deux formes ou plusieurs constituent une même variable de la même façon dont (ing) est constitué de [ɪŋ] et [ɪn] ; ils ont plutôt pris des mesures des taux d'usage de tous les emprunts à l'anglais à la fois. Ils ont donc constaté dans quelle catégorie grammaticale la majorité des emprunts apparaissent (Poplack *et al.*, 1988, p. 63), mais ils n'ont pas analysé l'usage de *snap* contre son équivalent d'origine française *bouton-pression*, par exemple. Quelques études ont employé des variables linguistiques qui sont constituées de variantes de plus d'une « langue », comme Mougeon (2007) et Perrot (2014), mais ces études-ci semblent être moins fréquentes dans la recherche sur le contact de langues. Nous

voulons donc ajouter une autre étude du contact de langues dans laquelle on se sert des variables linguistiques, dans ce cas-ci la variable (lol).

De même, les études typiques de la variation conceptualisent les phénomènes dont elles s'occupent peu souvent comme du contact. Une exception notable est Blom et Gumperz (1972/2000), qui ont étudié des variétés, appelées des dialectes dans leur étude, par rapport à l'alternance codique (p. 107), celle-ci pouvant se définir comme l'alternance successive d'un locuteur entre deux styles. Or, lorsque des variétés considérées comme distinctes de la variété en question se présentent dans les données des études variationnistes, elles sont parfois minimisées ou écartées comme des altérations de l'échantillon. Labov (1966/2006), par exemple, a explicitement écarté les sujets qui paraissent bilingues dans son étude des grands magasins (p. 45) ainsi que dans son étude du Lower East Side (p. 98). Plus récemment, Eckert (2000) a écarté les sujets qui sont déménagés dans la région qu'elle étudiait après l'âge de 8 ans, car on ne s'attend pas à ce que les locuteurs puissent complètement adopter de nouveaux patrons sonores après cet âge (p. 82). Eckert a cité le travail de Payne (1980), qui a en fait témoigné que l'adoption de nouveaux patrons sonores après l'âge de 8 ans est plus difficile, mais Bamman, Eisenstein et Schnoebelen (2014) ont également filtré tous les utilisateurs non-anglophones dans leur étude du genre et la variation lexicale sur Twitter sans savoir si la même restriction qu'a constatée Payne s'applique à ce qui est lexical (p. 139). Nous ne voulons pas dire que ces études ont été inutiles – leurs résultats ont fait bien avancer la théorie sociolinguistique – nous voulons simplement admettre que les chercheurs qui étudient la variation typique parlent rarement en termes de contact entre langues ou variétés. Pour notre part, nous visons à formuler une question de recherche qui réaffirme l'idée que la recherche sur la variation et la recherche sur le contact de langues sont une et même chose en

employant les outils typiques des études de la variation conceptualisée comme dans une seule langue, notamment les variables linguistiques.

1.2.2 Langues de base

Un concept fondamental du contact de langues est celui des langues de base, ou l'idée qu'un système distinct constitue la fondation dans laquelle des éléments d'un autre système distinct s'insèrent. Les méthodes pour identifier les langues de base sont allées de l'identification de la langue dont la majorité des traits phonologiques et morphologiques d'un discours sont tirés (Poplack, 1979/1980/2000, p. 210) à l'identification de la langue dont l'ordre des mots et les flexions sont issus (Poplack et Dion, 2012, p. 284) à la simple assertion qu'elle est la langue primaire dans le discours (Auer, 1988/2000, p. 165). Au vu des difficultés d'appliquer des critères qui identifient universellement les langues de base Gardner-Chloros et Edwards (2004) ont toutefois avancé que le concept est tout simplement inutile (p. 103/117-120). Essentiellement, on ne peut identifier une langue de base car des éléments des systèmes en question peuvent s'entremêler de n'importe quelle façon et donc on arrive à de nouveaux systèmes, tandis que l'idée d'une langue de base veut supposer que les systèmes originaux qui fournissent les éléments qui s'entremêlent demeurent plus ou moins intacts, à l'exception de quelques modifications des éléments insérés.

À titre d'exemple de nos données :

Andy B. : @bigbangs06HD Tu fais quoi avec ça un BBQ lol (Exemple 1.1)

Il y a trois façons générales de conceptualiser la phrase dans cet exemple :

- Les mots *BBQ* et *lol* s'insèrent dans le français.
- Les mots *Tu, fais, quoi, avec, ça* et *un* s'insèrent dans l'anglais.
- Les mots *BBQ* et *lol* de l'anglais et les mots *Tu, fais, quoi, avec, ça* et *un* du français s'unissent pour créer un nouveau système à part de l'anglais et du français.

Si l'on maintient l'idée des langues de base, il est difficile de justifier le troisième choix, car on doit présenter une explication pour la disparition des systèmes dont les éléments sont issus. Si l'on rejette l'idée des langues de base, les deux premiers choix sont plus difficiles à justifier, car on doit présenter une explication pour la persistance du français ou de l'anglais même malgré l'usage de mots considérés comme étrangers. Gardner-Chloros et Edwards semblent adopter cette dernière position.

Dans l'exemple donné, intuitivement, la tendance est peut-être de dire que le premier choix est le meilleur. Si l'on considère en revanche des cas plus extrêmes, tels que les créoles, le meilleur choix n'est plus évident. Par exemple, Klingler (2005, p. 359) a constaté la phrase suivante d'un sujet dans son étude de la démarcation entre l'anglais, le français et le créole louisianais en Louisiane :

RB : aoù to PART-ye ye, WELL (Exemple 1.2)
où 2P.POSS partie-PL COPULE bien

c'est une daube
c'est une daube

'Où que tes parties sont, bien c'est une daube'

Dans ce cas-ci, le choix de la langue de base n'est plus intuitivement clair. Il se peut que le français ou l'anglais ou une langue africaine ou autre chose se comporte comme la langue de base, mais peut-être que la réponse la plus simple est d'affirmer qu'il n'y a pas de langue de base, et que l'on prend des éléments de plusieurs systèmes pour former un nouveau système distinct. Selon nous, c'est un débat qui peut être mieux abordé dans la psycholinguistique ou la neurolinguistique, et nos questions de recherche ne s'en occuperont donc pas directement, mais nous partons de la supposition que la position de Gardner-Chloros et Edwards est justifiée.

1.2.3 Distinction entre les emprunts et les items alternés

En 2000, Poplack (1979/1980/2000) a décrit la question de la distinction entre les items empruntés, qui sont devenus des éléments de la langue de base, contre les items alternés, pris d'un autre système, qui demeurent donc étrangers par rapport à la langue de base, comme « fondamentale » à la recherche sur l'alternance codique (p. 205), un sentiment réaffirmé plus tard dans Poplack et Dion (2012, p. 279-280). En d'autres termes, quand un élément comme *BBQ* ou *lol* apparaît dans une phrase comme celle dans l'Exemple 1.1, on veut être en mesure d'identifier si le locuteur alterne vers un autre système avant de retourner au premier système ou si les items font tous partie du système en question, même si le locuteur semble alterner entre deux systèmes. Poplack (1979/1980/2000) a donc proposé que les emprunts sont morphologiquement et syntaxiquement intégrés dans la langue de base, tandis que les items alternés ne le sont pas (p. 205-206). En revanche, Myers-Scotton (1988/2000) a avancé l'idée que les items alternés déclenchent des sens sociaux, dans le sens où la forme est inattendue et alors rare. Là où il y a une absence de sens social dans l'usage d'une forme, on parle d'un emprunt (p. 133-134). Bref, il y a deux approches

générales à la distinction entre les emprunts et les items alternés : celle de Poplack qui avance que l'on peut les distinguer selon leur intégration linguistique et celle de Myers-Scotton qui avance que l'on peut les distinguer selon leur fréquence.

Les emprunts et les items alternés se retrouvent dans la littérature depuis longtemps, mais Poplack *et al.* (1988) ont introduit en plus l'idée des emprunts créés pour l'occasion. Un emprunt est un mot d'origine étrangère mais fréquent et intégré au niveau de la syntaxe et la morphologie, tandis qu'un emprunt créé pour l'occasion n'est pas fréquent, mais il est intégré au niveau de la syntaxe et la morphologie de la langue récipiendaire (Poplack *et al.*, 1988, p. 52-53). Ainsi, les emprunts créés pour l'occasion se situent sur le continuum des emprunts, tout en étant à part de celui des items alternés. Poplack *et al.* (1988) ont donc décrit l'alternance codique et le fait d'emprunter comme des « processus distincts » (p. 93). Cette distinction a été réaffirmée dans Poplack (1993, p. 255-256). Plus tard, Poplack et Dion (2012) ont présenté des preuves que tous les « seuls items d'une autre langue », un terme général qui fait référence aux emprunts et aux items alternés qui constituent un seul mot, sont des emprunts créés pour l'occasion et non des items alternés dans leurs corpus du français d'Ottawa et du Québec en démontrant que les seuls items d'une autre langue sont linguistiquement intégrés selon quatre critères :

- Ils prennent les flexions verbales du français.
- Ils pluralisent comme en français.
- Ils prennent les déterminants où ils sont pris en français.
- Ils s'accordent en genre comme en français. (p. 287-296)

Par exemple, ils présentent un seul locuteur qui prononce deux phrases avec un même verbe dans chacune (Poplack et Dion, 2012, p. 288) :

Locuteur :
- A few planes **crashed** in the World Trade Center.
  'Quelques avions ont percuté le World Trade Center.'
- Le building est en feu, il y a une- un avion qui **a crashé** dedans.
  'Le bâtiment est en feu, il y a une- un avion qui a percuté dedans.'

(Exemple 1.3)

Dans l'Exemple 1.3, le verbe *crash* prend la flexion verbale *-é* dans la phrase du français comme en français au lieu de *-ed* comme en anglais. Ce mot est apparu peu souvent dans les corpus, mais il était linguistiquement intégré quand même.

En revanche, Myers-Scotton (1988/2000) s'est appuyé sur la fréquence pour déterminer si un seul mot d'une autre langue est un emprunt ou un item alterné. Dans son modèle de marque, les mots qui déclenchent des sens sociaux dans un échange sont des items alternés (Myers-Scotton, 1988/2000, p. 133-134). Pour qu'un mot puisse déclencher un sens social, il doit violer les normes établies pour le genre d'échange en question, c'est-à-dire qu'il doit être inattendu en raison de sa rareté. Dans ce cadre, le mot *crashé* dans l'Exemple 1.3 serait un item alterné malgré sa flexion française, car il apparaît peu dans le corpus.

Or, les interprétations auxquelles les deux approches arrivent se chevauchent parfois naturellement, un fait qu'ont noté Deuchar et Stammers (2016). Pour notre part, nous ne visons pas à formuler une question de recherche qui nous permettra de résoudre ce débat, en partie parce que notre perspective est un peu différente. Les chercheurs mentionnés plus haut semblent effectuer leurs analyses à partir des données des groupes. Ces analyses ont permis de faire avancer nos connaissances en

sociolinguistique, bien sûr, mais notre supposition est que l'alternance codique est un phénomène qui se passe dans l'esprit d'un individu dans le sens où un individu peut utiliser un mot d'un de ses styles dans un autre. Si l'alternance codique se passe dans l'esprit d'un individu – et il se peut bien que les chercheurs mentionnés plus haut soient d'accord avec cette idée – alors la meilleure façon de l'étudier est d'analyser profondément un individu. Nous voulons donc fournir des informations pertinentes à la recherche à venir sur le phénomène en analysant le comportement linguistique des individus relatif aux communautés auxquelles ils appartiennent.

1.3 Délimitation des communautés

La délimitation des communautés est soulignée ici en raison de notre définition des registres : les *abstandsprache*s des groupes dans des contextes donnés. Ainsi, les communautés vont de pair avec les registres. Une synthèse de plusieurs développements clés en ce qui a trait à la délimitation des communautés sera donc exposée dans cette section. Nous trouvons ces méthodes antérieures efficaces en elles-mêmes, et elles ont bien contribué à faire avancer la sociolinguistique, mais nous croyons que les méthodes actuelles de l'analyse des réseaux sociaux peuvent identifier les mêmes communautés ou au moins des communautés aussi cohérentes tout en fournissant plus d'informations structurelles sur ces communautés. Nous ne nous interrogeons donc pas à savoir si les méthodes du passé fonctionnent, mais notre étude servira plutôt de validation de principe pour les méthodes de l'analyse des réseaux sociaux.

Cette section débutera par les communautés linguistiques, un concept employé par Labov lui-même, suivies de l'analyse des réseaux sociaux de Milroy (1980/1987), les

communautés de pratique d'Eckert (2000), le concept de communauté en général et enfin elle reviendra à la question de la place de l'individu et son répertoire de styles dans ce système.

1.3.1 Communautés linguistiques

Le concept des communautés linguistiques est bien utilisé dans la sociolinguistique variationniste jusqu'à présent – G. Sankoff (2015) les a toujours décrites comme « cruciales pour l'entreprise sociolinguistique » (p. 24) – mais diverses méthodes pour les identifier ont été utilisées. Par exemple, Labov (1966/2006) a justifié la caractérisation de la ville entière de New York comme une communauté linguistique selon trois critères : la différence entre le parler des natifs et celui des visiteurs, la cohérence de la façon dont les natifs ajustent leurs styles – les styles tels que définis par Labov[7] – et la conformité des évaluations des natifs envers les variables linguistiques, laquelle échoue lorsque des visiteurs sont ajoutés au groupe (p. 6). G. Sankoff (2015), dans un article récent où elle a plaidé en faveur de l'usage continu des communautés linguistiques, a décrit les Papouasiens d'une région de la Papouasie-Nouvelle-Guinée comme une communauté linguistique en raison de leur identité partagée, leur norme d'usage linguistique partagée et l'intelligibilité mutuelle de leurs parlers (p. 33-34). De même, elle a avancé que la ville de Montréal constitue une communauté linguistique car même les anglophones de Montréal qui apprennent « le français » acquièrent systématiquement la « variabilité de la langue dominante », c'est-à-dire le français montréalais (G. Sankoff, 2015, p. 38). Essentiellement, Labov et G. Sankoff ont identifié des communautés par la mise en œuvre d'indicateurs, en mettant souvent l'accent sur le parler, un accent qu'a aussi noté Davies (2005, p.

7 Pour plus de détails, voir la section 2.5 Variation stylistique.

559). Malgré les progrès dans la sociolinguistique en s'appuyant sur les communautés linguistiques identifiées par de tels indicateurs pendant des décennies, l'accent sur le parler entraîne une sorte de raisonnement circulaire où des gens constituent une communauté linguistique car ils se partagent des normes linguistiques, mais il se peut qu'ils se partagent des normes linguistiques parce qu'ils constituent une communauté linguistique. En effet, Milroy (1980/1987) a démontré qu'une communauté se comporte comme un mécanisme d'application des normes dans son étude de Belfast[8] (p. 162-163).

Quelques définitions des communautés linguistiques sont encore plus larges, telles que celle de Fishman (1967/2000), qui a affirmé qu'une communauté linguistique est une communauté qui n'exige aucun traducteur pour l'intracommunication (p. 77), c'est-à-dire la communication entre ses membres, et quelques-unes sont simplement différentes, telles que celle de Hymes (1967/1972), qui a dit qu'une communauté linguistique est l'intersection du champ de langue[9] et le champ de discours d'un locuteur, qui sont les portées dans lesquelles son répertoire de styles et ses normes de parler peuvent efficacement s'utiliser (p. 54-55). Hymes n'a pas donné une méthode pour délimiter le champ de langue et le champ de discours eux-mêmes, mais à l'époque, il était plus important d'établir des concepts sociolinguistiques de base que d'élaborer en détail ces concepts de base. La conceptualisation de Hymes est aussi notable pour son emphase sur l'individu.

Un problème avec l'usage des communautés linguistiques, c'est que l'on ne peut ni bien traiter les individus qui se trouvent à leurs périphéries ni bien expliquer leur

8 Pour plus de détails, voir la section suivante.

9 Le terme *champ de styles* serait peut-être plus approprié conforme à la terminologie de ce mémoire, mais le terme *champ de langue* est maintenu simplement car il est un terme technique dont le référent précis n'est pas tout à fait clair.

comportement langagier quand ils sont identifiés. Par exemple, Labov (1966/2006), dans son étude du Lower East Side, a écarté deux Afro-Américains qui sont déménagés dans un quartier de la Virginie à l'âge de 10 ans en raison de leur parler aberrant, selon les normes du quartier. Il a suggéré qu'ils parlaient un mélange de l'anglais du sud et de l'anglais new-yorkais (Labov, 1966/2006, p. 118), mais il n'avait pas les outils pour expliquer leurs positions dans la communauté afin de mieux analyser leur comportement.

Horvath et D. Sankoff (1987) ont remarqué, dans leur étude de « l'anglais australien » de la ville de Sydney, que les migrants sont donc d'habitude exclus dans de telles études (p. 202). Afin d'être capables de garder tous les sujets dans les échantillons, ils ont proposé l'analyse en composantes principales, qui est une technique statistique qui regroupe les sujets selon leurs réalisations des variables linguistiques en question avant de chercher des catégories sociales qui sont associées aux regroupements résultants. L'avantage, c'est que l'on peut rendre ainsi compte de tous les sujets, mais une autre façon d'accomplir cette tâche est l'analyse des réseaux sociaux.

1.3.2 Analyse des réseaux sociaux

En 1980, Milroy (1980/1987) a employé l'analyse des réseaux sociaux dans son étude de Belfast comme une réponse à des faiblesses perçues du concept des communautés linguistiques (p. 32-33). Elle a notamment présenté deux informateurs dans ses données qui se partageaient de nombreuses caractéristiques selon les variables sociales, variables sur lesquelles on s'appuie lorsque l'on emploie les communautés linguistiques, mais qui parlaient différemment (Milroy, 1980/1987, p. 131-134). Dans un cadre où un chercheur ne se sert que des variables sociales pour regrouper les

membres de la communauté dans des sous-ensembles, il n'est pas possible d'expliquer un tel résultat. Milroy a donc proposé de mesurer l'intégration des sujets dans leurs quartiers avec une mesure de centralité inspirée par l'analyse des réseaux sociaux.

L'analyse des réseaux sociaux exige que l'on élabore rigoureusement les réseaux sociaux des sujets sous forme de sociogrammes selon qui connaît qui (Milroy, 1980/1987, p. 46-51), comme dans la Figure 1.1. Une fois qu'un réseau est élaboré, on peut y appliquer des algorithmes pour détecter des agglomérats dans le réseau, autrement dit des communautés, et pour quantifier l'intégration des sujets dans les agglomérats/communautés, soit leurs centralités[10].

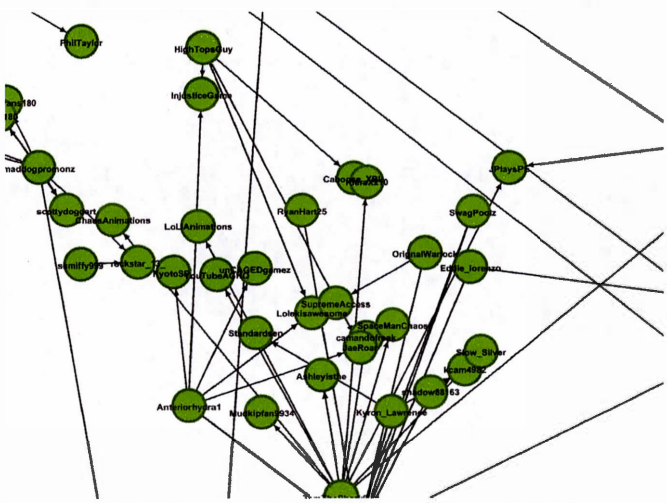

Figure 1.1: Une partie de la communauté 302, identifiée dans nos données

10 Pour plus de détails au sujet du cadre de l'analyse des réseaux sociaux, voir la section 2.4 Analyse des réseaux sociaux.

Pour mesurer l'intégration de ses sujets, Milroy (1980/1987) a créé son propre indice, qu'elle a appelé son échelle de force du réseau, dans lequel on accorde aux sujets un point pour chaque critère suivant qu'ils satisfont :

- Ils font partie d'un agglomérat territorial dense.
- Leurs parentés dans leurs quartiers sont fortes.
- Ils travaillent avec au moins deux autres personnes dans la zone.
- Ils travaillent avec au moins deux autres personnes du même sexe dans la zone.
- Ils passent volontairement du temps avec leurs collègues pendant leurs loisirs. (p. 141-142)

Elle a découvert que l'intégration dans un réseau, selon son indice, fonctionne comme un mécanisme d'application des normes et prédit la façon dont un locuteur parle (Milroy, 1980/1987, p. 162-163). Ce résultat a été ensuite corroborée par d'autres études après Milroy (p. ex. Auer, 1988/2000; Li, Milroy, et Ching, 1992/2000; Sharma, 2011).

En 2006, Labov (1966/2006) a caractérisé l'approche d'analyse des réseaux sociaux comme une approche qui sollicite des données des « individus étendus », car les locuteurs dans une même partie d'un réseau parleront probablement de façon très similaire, ce qui limite l'utilité de cette approche, selon Labov, lorsque l'on cherche de la variation (p. 366). Milroy (1980/1987) était elle-même, d'une certaine façon, d'accord avec Labov lorsqu'elle a remarqué, comme une faiblesse, que l'analyse des réseaux sociaux produit un échantillon qui n'est pas représentatif d'une

« communauté »[11] (p. 38). Ce que Labov et Milroy, les deux, ont caractérisé comme une faiblesse de l'analyse des réseaux sociaux s'avère au contraire, selon nous, une indication que la méthode réussit à identifier des unités cohésives qui représentent bien des communautés, et donc des registres.

La découverte de Milroy et son introduction de l'analyse des réseaux sociaux dans la sociolinguistique ont été des développements importants, mais les méthodes de l'analyse des réseaux sociaux utilisées dans la sociolinguistique sont demeurées plus ou moins désuètes, à quelques exceptions notables (p. ex. Dodsworth et Benton, 2017; Lev-Ari, 2018), en ce que les chercheurs se fient typiquement aux communautés linguistiques au lieu de détecter automatiquement les communautés. Ils emploient également de simples indices d'intégration malgré l'existence d'autres mesures de centralité qui sont plus rigoureuses, mathématiquement[12]. Nous voulons donc introduire quelques mesures dans ce mémoire et démontrer leur utilité.

1.3.3 Communautés de pratique

Une autre approche de la délimitation des communautés met en œuvre les communautés de pratique de Lave et Wenger (1991). Trois caractéristiques se présentent dans une communauté de pratique :

11 En effet, Milroy utilisait elle-même les communautés linguistiques. Ce qu'elle a employé de l'analyse des réseaux sociaux était l'idée des mesures de centralité. À part de ce développement important, elle a simplement fait connaître l'existence de l'analyse des réseaux sociaux aux autres sociolinguistes.

12 Pour plus de détails au sujet des méthodes plus récentes et des concepts dans l'analyse des réseaux sociaux en général, voir la section 2.4 Analyse des réseaux sociaux.

- L'engagement mutuel : les membres interagissent régulièrement
- Une entreprise commune : un processus dans lequel tous les membres s'engagent et négocient leurs rôles
- Un répertoire partagé : des ressources communes que les membres utilisent pour communiquer entre eux (Wenger, 1998, cité dans Holmes et Meyerhoff, 1999, p. 175-176)

Une communauté de pratique est donc une communauté centrée sur une activité. Il existe des précurseurs de ce concept dans la littérature sociolinguistique. Dans une étude des locuteurs aux Caraïbes de Le Page et Tabouret-Keller (1985), ils ont demandé aux sujets de raconter des histoires d'Anansi, qui ont par la suite produit des formes que Le Page et Tabouret-Keller ont caractérisé comme « archaïque », appartenant à un registre spécial (p. 102). Ce qu'ils ont démontré, finalement, c'est une indication que les registres se développent en tandem avec les activités, même si ce n'était pas leur objectif. Dans ce cas-ci, l'activité de raconter des histoires d'Anansi a entraîné des formes et des traits linguistiques qui n'apparaissent jamais durant d'autres activités ou dans d'autres contextes.

Eckert et McConnell-Ginet (1992, p. 34-35) ont introduit le concept de communautés de pratique dans la sociolinguistique, et Eckert (2000) a plus tard employé le concept dans son étude de Belten High afin d'analyser la façon dont la participation dans des activités peut prédire le parler de ses sujets. Elle a mesuré la participation avec un indice qui assigne des points selon le nombre d'activités auxquelles on participe et selon la force de l'engagement dans ces activités (Eckert, 2000, p. 156), qui ressemble bien à la façon dont Milroy (1980/1987) a quantifié l'intégration des sujets dans les réseaux avec son échelle de force du réseau (p. 139). En effet, Eckert a fini

par raffiner la méthode de Milroy, en soulignant simplement les activités comme des mécanismes qui entraînent la formation des communautés.

En dépit du fait que d'autres chercheurs ont essayé de distinguer les communautés de pratique des réseaux sociaux selon l'idée qu'il y a une différence idéologique entre les deux (p. ex. Davies, 2005), ceux-ci ne sont pas vraiment en concurrence. L'analyse des réseaux sociaux est en fait un ensemble d'outils qui peuvent être employés pour décrire la structure des communautés de pratique. Wenger (1998) a lui-même remarqué qu'une communauté de pratique pourrait être un agglomérat de liens forts si on l'examine comme faisant partie d'un réseau social (cité dans Schenkel, Teigland, et Borgatti, 2002, p. 6). C'est pour cette raison que Schenkel *et al.* (2002) ont déclaré que « toute communauté de pratique se constitue d'un réseau, mais tout réseau ne constitue pas une communauté de pratique. » C'est-à-dire que l'analyse des réseaux sociaux est un outil que l'on utilise pour présenter la structure de plusieurs genres de communautés, dont une communauté de pratique. Schenkel *et al.* (2002) ont donc proposé cinq caractéristiques structurelles des communautés de pratique dans la terminologie de l'analyse des réseaux sociaux :

- Une forte connectivité
- Des distances théoriques graphiques courtes
- Une forte densité
- Une structure avec un seul cœur et une périphérie
- Plus on participe, plus on fait partie du cœur de la communauté (p. 7-11)

Bref, une communauté de pratique n'a pas de sous-groupes, ce qui découle du quatrième caractéristique, et la plupart des membres d'une communauté de pratique

sont reliées directement, ce qui découle des trois premières caractéristiques[13]. Schenkel *et al.* sont arrivés à ces caractéristiques en supposant que les trois caractéristiques qui définissent les communautés de pratique selon Wenger (1998) – l'engagement mutuel, une entreprise commune et un répertoire partagé (cité dans Holmes et Meyerhoff, 1999, p. 175-176) – exigent une communauté dans laquelle les informations circulent facilement, ce qui entraîne une communauté qui est homogène à certains niveaux. Pour obtenir une telle communauté, la recherche sur l'analyse des réseaux sociaux et sur la théorie des graphes, les mathématiques sur laquelle l'analyse des réseaux sociaux est basée, suggère que ses membres doivent être bien reliés les uns avec les autres. En effet, c'est ce que Milroy (1980/1987) a déjà noté en citant des études dans la sociologique comme Bott (1972) et Boissevain (1974).

L'un des avantages de la représentation des communautés de pratique comme des réseaux sociaux, c'est que l'on est en mesure de quantifier les niveaux de participation des individus dans une communauté de pratique. Wenger (1998) a proposé qu'il y ait quatre niveaux de participation dans une communauté de pratique :

- La participation complète
- La participation périphérique
- La participation marginale
- Aucune participation (cité dans Schenkel *et al.*, 2002, p. 5)

Les niveaux de participation dans une communauté de pratique peuvent être facilement conceptualisés comme l'intégration – autrement dit les centralités – des individus dans la communauté, et si l'on peut mesurer la participation d'un individu

13 Pour plus de détails au sujet de cette terminologie, voir la section 2.4 Analyse des réseaux sociaux.

dans une communauté, on peut ensuite mieux expliquer leur comportement linguistique relatif à la communauté.

En résumé, le concept des communautés de pratique n'est pas une alternative à l'analyse des réseaux sociaux, car cette dernière est plutôt un ensemble d'outils qui nous permet d'élaborer plusieurs types de communautés, y compris les communautés de pratique, même si elle est parfois caractérisée comme une idéologie à part des communautés de pratique. Nous proposons que la sociolinguistique peut progresser de manière plus efficace en se servant davantage des outils de l'analyse des réseaux sociaux, et nous avons l'intention de démontrer les avantages que l'on en retire dans cette étude-ci.

1.3.4 Concept de communauté

Une communauté est finalement un concept et donc la question de comment l'opérationnaliser dépend de l'idée qu'a un chercheur de sa définition théorique. Dans les sections précédentes, nous avons exposé trois façons d'opérationnaliser les communautés qui semblent s'appuyer sur deux concepts différents de communauté. Dans le cas des communautés linguistiques, une communauté est un groupe dans lequel les membres peuvent différer considérablement et ne peuvent avoir que peu d'interaction directe, mais qui se partagent tout de même des traits linguistiques. Dans le cas des réseaux sociaux et des communautés de pratique, une communauté est un groupe dans lequel les membres diffèrent peu en raison de leur interaction considérable. Cette différence entre les définitions de communautés est importante, car là où un chercheur comme Labov (1966/2006) estime qu'une ville entière telle que New York constitue une communauté, avec toute sa variation linguistique, un

chercheur qui se sert de l'analyse des réseaux sociaux ne jugerait peut-être qu'un sous-ensemble d'individus dans cette ville constitue une communauté, comme tous les ouvriers employés à une même entreprise. Par conséquent, là où un chercheur trouve un sous-ensemble de ses sujets préfère une variante d'une variable linguistique et un autre sous-ensemble une autre variante, un chercheur comme Labov, qui s'appuie sur les communautés linguistiques, dirait peut-être que l'on regarde toujours une même communauté dans laquelle il y a de la variation, tandis qu'un chercheur qui s'appuie sur l'analyse des réseaux sociaux dirait peut-être que l'on a probablement deux communautés. Nous interprétons donc de telles situations de cette dernière façon.

1.3.5 Individus

Malgré la suggestion de Labov (1966/2006) en 2006 à l'effet que l'étude des individus ne soit jamais revenue avec vigueur dans la sociolinguistique (p. 157), les chercheurs qui s'intéressent à la variation stylistique les ont naturellement analysés. Dans des études notables, les chercheurs ne se sont concentrés que sur un seul individu, comme Coupland (1980) ou Rickford et McNair-Knox (1994). Ces études ont bien fait avancer la connaissance du comportement linguistique des individus, mais l'application des outils qui identifient précisément la position de ces individus dans leurs communautés nous permettrait d'en dire plus à ce propos.

Rickford et McNair-Knox (1994) ont analysé le changement dans le parler d'un seul individu, anonymisé comme Foxy Brown, de l'âge de 13 ans à l'âge de 18 ans. Ils ont

constaté que son usage du vernaculaire[14] était le plus élevé à l'âge de 13 ans et à l'âge de 18 ans qu'à l'âge de 15 ans (Rickford et McNair-Knox, 1994, p. 263-264). L'objectif de leur étude était d'examiner le cadre du design d'audience de Bell (1984). Dans ce cadre, un locuteur change d'un style à un autre pour réagir à l'audience actuelle ou pour influencer l'audience actuelle. Rickford et McNair-Knox (1994) ont donc mis de l'avant trois explications provisoires de l'usage du vernaculaire de Foxy :

- Foxy a commencé à étudier dans un lycée prestigieux où elle s'est aperçue de son vernaculaire marqué.
- L'entrevue avec Foxy a eu lieu dans un local différent des autres.
- Foxy affirmait son autorité intellectuelle. (p. 263-264)

Rickford et McNair-Knox n'avaient pas l'objectif de répondre à cette question dans l'étude, mais si l'on voulait déterminer quelle explication était la meilleure, avoir la capacité d'identifier précisément le changement de la position de Foxy par rapport à ses communautés donnerait des renseignements utiles. Par exemple, si Foxy était devenue bien centrale dans son nouveau lycée à l'âge de 15 ans puis n'occupait qu'une position périphérique dans cette communauté précise à l'âge de 18 ans, on aurait des indications que la première et la troisième explications sont meilleures que la deuxième explication. Avec l'analyse des réseaux sociaux, on peut facilement quantifier la centralité d'un individu comme Foxy dans ses communautés.

L'étude de Belten High d'Eckert (2000) présente une autre occasion où des mesures de centralité nous permettraient de vérifier un concept qu'elle a avancé, à savoir le

14 Ils n'ont pas donné une définition du vernaculaire, mais on peut supposer que c'est la définition de Labov. Pour plus de détails, voir la section 2.5 Variation stylistique.

concept des icônes sociolinguistiques. Ce sont les personnes qui emploient considérablement des variantes des variables linguistiques dont les fréquences s'avancent. De plus, la position centrale de ces icônes dans un agglomérat situé dans un réseau leur confère de l'autorité pour définir les sens sociaux des variables linguistiques (Eckert, 2000, p. 216-219). Eckert (2000) a dessiné le réseau social du lycée, mais elle a également admis que ce réseau n'a entraîné que des idées « suggestives », car elle n'a pas visé à analyser rigoureusement ce réseau (p. 177) ; toutefois, une analyse rigoureuse nous permettrait de quantifier les positions sociales de ses icônes. Si leurs positions sont uniques à certains égards, ce résultat nous permettrait ensuite de formaliser la définition des icônes sociolinguistiques puis de tester l'universalité du concept dans d'autres études.

Un autre concept que l'on pourrait formaliser par l'application de l'analyse des réseaux sociaux est celui du style maison[15] qu'a proposé Bell (1984), un style qu'adoptent les employés d'un établissement dans lequel ils parlent avec les clients d'une façon uniforme selon le parler des clients auquel s'attendent les employés en raison des interactions courtes et peu fréquentes (p. 170). En identifiant les communautés des employés et des clients et leurs positions dans ces communautés, on arriverait à un modèle de structure d'un réseau pour lequel on peut s'attendre à ce qu'un style maison soit utilisé. À partir d'une telle formalisation, on peut ensuite tester son universalité dans d'autres études.

Plusieurs concepts dans la sociolinguistique par rapport aux individus sont intéressants et utiles même sans formalisation, mais ils demeurent difficiles à tester et à vérifier si l'on ne peut les quantifier. Il s'agit en fait de l'essentiel de ce qu'a fait

15 Selon nos définitions, un registre maison serait peut-être un terme plus approprié, puisqu'il semble être conçu comme une variété partagée entre tous les employés.

Labov (1966/2006) dans son étude fondatrice : il n'a pas seulement proposé des concepts mais il a démontré des façons de les tester et de les vérifier à partir de méthodes rigoureuses et mathématiques. L'analyse des réseaux sociaux nous permet d'en faire autant en ce qui concerne des questions sur le comportement linguistique des individus.

1.4 Questions de recherche

Finalement, cette problématique présente quelques enjeux qui nous intéressent. L'idée selon laquelle ce qui se passe dans une étude typique de la variation dans un contexte dit monolingue est pareil à ce qui se passe là où les langues sont en contact nous amène à impliquer la variable linguistique (lol), contenant des variantes d'origine française et anglaise, dans nos questions de recherche, le choix de (lol) au lieu d'un autre item étant purement pratique (pour plus de détails, voir la section 3.2.2 Identification d'une variable linguistique lexicale). Nous croyons également qu'une étude comme la nôtre dans laquelle nous mettons un accent sur les individus va fournir des informations utiles pour la discussion au sujet de la distinction entre les emprunts et les items alternés. Enfin, là où des études antérieures ont démontré des résultats intéressants et utiles, nous croyons qu'il y a de meilleures méthodes pour analyser les communautés qui nous donnent à leur tour la capacité de quantifier les positions des individus dans leurs communautés et donc de mieux comprendre leurs comportements linguistiques. Ce mémoire est donc une démonstration de l'utilité des méthodes actuelles de l'analyse des réseaux sociaux. Nous posons alors deux questions de recherche :

(Q1) La distribution des réalisations de la variable linguistique lexicale (lol), constituée de mots d'origine française et anglaise, sera-t-elle la même pour chaque communauté identifiée sur Twitter à partir des tweets émanant des provinces maritimes au Canada ?

(Q2) Les individus bien intégrés dans leurs communautés réaliseront-ils la variable linguistique lexicale (lol) avec moins de diversité qu'au niveau de leurs communautés ?

CHAPITRE II

CADRE THÉORIQUE

2.1 Introduction

Afin de répondre aux questions de recherche que nous avons posées dans ce mémoire, il faudra présenter une synthèse du cadre théorique à partir duquel celles-ci vont être abordées. Cette section-ci exposera les études antérieures sur *lol*, les variables linguistiques, l'analyse des réseaux sociaux et enfin la variation stylistique. Quelques hypothèses développées à partir de ce cadre seront présentées à la fin du chapitre.

2.2 Variables linguistiques

Les variables linguistiques font partie intégrante de la sociolinguistique variationniste. Développées par Labov (1966/2006), elles sont des unités abstraites qui contiennent toutes les variantes possibles qui peuvent être réalisées pour un item linguistique donné (p. 32). Par exemple, (e) peut représenter une variable linguistique phonologique, laquelle contient [ɛ] et diverses variantes de [ɛ] qui diffèrent en hauteur dans un mot tel que *mais* prononcé soit comme [mɛ] soit comme [me] dans le français de Louisiane (Valdman et Rottet, 2010, p. 378). Les variables linguistiques peuvent également être des mots, tels que (on/nous) pour le pronom sujet de la première personne du pluriel du français, une variable que D. Sankoff et Laberge

(1978a) ont examinée dans leur étude du français montréalais (p. 120/126). De même, on peut avoir une variable linguistique lexicale comme (lol) dans ce mémoire.

Dans la recherche sociolinguistique, des associations entre les réalisations des variables linguistiques et des faits sociaux ou stylistiques sont souvent établies. Labov (1966/2006) a ainsi constaté que les locuteurs de New York dans ses données se différenciaient socialement par rapport à la variable linguistique (oh), c'est-à-dire que les locuteurs de différentes classes socio-économiques ont produit [ɔ] à différentes hauteurs (p. 165). En outre, il a constaté que la réalisation de (oh) changeait pour une même classe socio-économique lorsque le contexte était plus ou moins formel (Labov, 1966/2006, p. 164). De tels résultats se sont reproduits de nombreuses fois au fil du temps. Horvath et D. Sankoff (1987) ont remarqué que l'ethnie, la classe socio-économique et le genre ont entraîné différentes réalisations de leurs variables linguistiques à Sydney (p. 190-197), et Eckert (2000) a trouvé que le genre et l'affiliation avec les sportifs ou les burnouts à Belten High ont entraîné des réalisations spécifiques de ses variables linguistiques (p. 111-112).

Plus pertinent à l'étude actuelle, Milroy (1980/1987) a témoigné que le niveau d'intégration dans les communautés locales à Belfast a contraint les réalisations de ses variables linguistiques. De surcroît, l'étude de D. Sankoff et Laberge (1978b) est également pertinent, car ils ont incorporé une variable linguistique lexicale, (on/ils), qui représente le pronom sujet indéfini explétif humain au pluriel. Ils ont opérationnalisé l'idée du marché linguistique pour remplacer l'usage des classes socio-économiques. Dans ce cadre, des participants classent des descriptions des histoires personnelles des sujets selon le besoin de chaque sujet d'utiliser des variantes prestigieuses, autrement dit de participer au marché linguistique. Ils ont trouvé que la participation au marché linguistique était le facteur le plus important

pour prédire les réalisations des variables linguistiques qu'ils ont analysées à Montréal (D. Sankoff et Laberge, 1978b, p. 246).

Le contexte peut également entraîner des réalisations spécifiques des variables linguistiques. Par exemple, Milroy (1980/1987) a encore constaté à Belfast de la variation contrainte par la formalité du contexte qu'a observé Labov (1966/2006) à New York. Elle a sollicité trois styles dans son étude : un style spontané, où les locuteurs étaient détendus, un style d'entrevue, où la discussion était bien structurée, et un style de liste de mots, où les locuteurs ont lu des mots. Milroy (1980/1987) a remarqué que quelques variables linguistiques changeaient selon le style, telles que (a) et (th), qui représentent [ɑ] à différentes hauteurs et la présence ou l'absence de la fricative dentale sourde [θ] (p. 101-102). Coupland (1980) a témoigné que l'agente de voyage qu'il a étudiée à Cardiff changeait son parler selon l'interlocuteur, selon s'il s'agissait d'un collègue, d'un collègue à distance ou d'un client, et selon le thème du discours, soit au sujet du travail ou non.

Enfin, Tagliamonte (2006) a identifié cinq provenances possibles de la variation qu'observent les sociolinguistes dans les variables linguistiques, quelques-unes qui ne sont ni sociales ni stylistiques et quelques-unes qui le sont :

- La génération grammaticale des phrases
- Les processus de production et performance
- La physiologie de l'articulation
- Les décisions stylistiques conscientes des locuteurs
- Une construction analytique du linguiste (p. 134)

L'important en ce qui concerne l'étude actuelle, cependant, c'est que les variables linguistiques peuvent être des mots (D. Sankoff et Laberge, 1978b, 1978a), que leurs différentes réalisations peuvent être associées à des sous-ensembles sociaux (Eckert, 2000; Horvath et Sankoff, 1987; Labov, 1966/2006), y compris aux agglomérats dans les réseaux sociaux (Milroy, 1980/1987), et que leurs différentes réalisations peuvent être contraintes par le contexte du discours (Coupland, 1980). Cela suggère que nous pouvons nous attendre à ce que (lol) puisse servir de variable linguistique et que ses différentes réalisations puissent être contraintes par les communautés des sujets, mais avant de poursuivre à la section 2.3 *LOL* dans la littérature, il faudra se livrer à une petite digression pour aborder un problème qui se soulève surtout pour les variables linguistiques lexicales comme (lol).

2.2.1 Équivalence des variantes

Labov (1972b) a déclaré que l'existence de variation socialement et stylistiquement contrainte présume qu'il y a un sens référentiel partagé par les variantes de la variable linguistique (cité dans Lavandera, 1978, p. 174), c'est-à-dire que, pour identifier la variation sociale ou stylistique, on doit éliminer la possibilité que la variation soit due à une différence de référence entre les variantes concernées. Pour ce qui est des variables linguistiques phonologiques, il est facile d'établir que les variantes se partagent une référence, car elles n'ont pas de référence à elles seules. Par exemple, lorsqu'il s'agit d'une variable comme (e), les variantes [ɛ] et [e] n'ont pas de sens référentiels ; on ne peut les prononcer en dehors du contexte d'un mot puis être capable de dire qu'ils représentent un objet concret dans le monde ou un objet abstrait dans l'esprit. Or, pour ce qui est des variables linguistiques lexicales comme (on/nous) (D. Sankoff et Laberge, 1978a, p. 120/126) ou (lol), dont les variantes ont

des références tout seules, l'établissement d'une équivalence référentielle pose problème. Lavandera a exposé ce problème en 1978 (p. 175).

Weiner et Labov (1983) ont effectivement proposé que l'établissement d'une équivalence précise entre les variantes d'une variable linguistique n'était qu'un enjeu méthodique, non théorique, en affirmant que les études de la variation n'ont pas besoin de se tenir à des « façons alternatives de dire la même chose », mais qu'une telle étendue permet simplement de tirer de plus fortes conclusions (p. 31). En effet, Lavandera (1978) a fini par suggérer que l'on puisse assouplir la focalisation mise sur l'équivalence référentielle et la remplacer par une insistance sur une « comparabilité fonctionnelle » (p. 181). Elle n'a jamais bien défini cette comparabilité, mais on peut supposer qu'elle exige que l'effet discursif des variantes, dans le cas de (lol) du moins, soit identique. Si l'on écrit *lol* ou *mdr*, il s'agit toujours d'une sorte de commentaire humoristique, amical, ou peut-être moqueur envers le texte, et donc ces variantes peuvent être jugées comparables.

2.3 *LOL* dans la littérature

Notre choix de (lol) comme variable linguistique d'intérêt n'était que purement pratique : il varie avec au moins une forme d'origine française, et il était plus fréquent dans les données que les autres choix (pour plus de détails, voir la section 3.2.2 Identification d'une variable linguistique lexicale). Il sera néanmoins utile de parler de la recherche sur les variantes de (lol)[16].

16 En fait, nous ne connaissons pas d'étude sur (lol) comme une variable linguistique contenant des variantes d'origine anglaise et française, donc les études que nous pouvons citer sont celles qui ne traitent que les variantes isolément.

2.3.1 Définitions des variantes de (lol) dans la littérature

Le mot *lol*, qui représente *laugh out loud* 'éclater de rire', est défini similairement dans les quelques études qui le traitent dans la littérature. Baron (2004) l'a appelé « un mot de remplissage phatique, plus ou moins comparable à *OK*, *really* ou *yeah* dans le discours oral » en anglais (p. 416), qui ont les sens 'd'accord', 'vraiment' et 'ouais' respectivement. En accord avec Baron, Tagliamonte et Denis (2008) l'ont d'ailleurs défini comme « un signe d'engagement par l'interlocuteur, de la même façon dont on dirait mm-hm au cours d'une conversation » en anglais (p. 11). De plus, Cougnon et Ledegen (2008) n'ont pas seulement écrit que *lol* est un mot qui « ponctue[] les discours en indiquant l'émotion du scripteur », mais que l'équivalent d'origine française, *mdr*, qui représente *mort(e) de rire*, remplit cette même fonction (p. 9). Elles ont ajouté que même les émoticônes, ponctuations qui dessinent les visages comme :), ont cette fonction. L'essentiel, c'est qu'il y a peu de controverse par rapport à l'idée que ces mots sont phatiques, c'est-à-dire qu'ils réalisent des fonctions sociales, et qu'ils indiquent le point de vue du locuteur envers le discours, des interprétations avec lesquelles nous sommes également d'accord.

2.3.2 Variation des mots phatiques en ligne

Il n'existe pas d'études exactement comme la nôtre où on emploie (lol) comme une variable linguistique lexicale qui contient des variantes d'origine française et anglaise, mais il y a bien des études des mots phatiques en ligne par rapport à la variation. Quant à la variation impliquant *lol* ou *mdr*, Liénard (2014) a présenté un exemple de l'usage des deux mots dans un échange où les interlocuteurs utilisent d'autres éléments d'origine française et kibuchine (p. 158-159), une langue parlée à

Madagascar (p. 155), ce qui démontre que ces deux peuvent varier ensemble, mais cette étude n'a pas effectué une analyse de ces mots spécifiquement. En revanche, Tagliamonte et Denis (2008) ont visé à faire une analyse quantitative, mais puisque leur étude de la variation dans la messagerie instantanée était plus ou moins exploratoire à cette époque-là en raison du manque d'études similaires, ils ont seulement analysé la fréquence et la distribution par âge de *lol*, qui comptait comme 0,41 % de leur corpus, plus fréquent que d'autres mots semblables, mais moins fréquent que *haha* à 1,47 % du corpus, et qui était plus fréquent parmi ceux qui avaient entre 15 et 16 ans que ceux entre 17 et 18 ans ou 19 et 20 ans (p. 11-13).

Si l'on peut affirmer que les émoticônes ont les mêmes fonctions que les mots tels que *lol*, il est donc pertinent de parler de la recherche sur leur variation. En effet, Cougnon et Ledegen (2008), comme mentionné plus haut, ont avancé que les émoticônes ont bien les mêmes fonctions que *lol* (p. 9). Provine, Spencer, et Mandell (2007), dans leur étude sur les émoticônes sur les forums de discussion en ligne, ont remarqué que les émoticônes qu'ils ont classées dans la catégorie « rire » se déclenchent quand un utilisateur tape *LOL* (p. 302), ce qui suggère qu'il y a un lien généralement accepté entre les émoticônes et les mots phatiques. Provine *et al.* (2007) n'ont cependant pas étudié la variation des émoticônes, mais plutôt leurs dispositions relatives aux phrases. Ils ont témoigné que les émoticônes de rire sont fréquentes, constituant 20 % de toutes les émoticônes, et qu'elles se présentent entre les phrases (Provine *et al.*, 2007, p. 302-303), mais ils n'ont pas mentionné si les formes des émoticônes de rire varient selon le contexte social ou linguistique, qui serait plus pertinent à la présente étude.

En fait, nous n'avons pas réussi à trouver des études qui ont analysé la distribution de différentes formes des émoticônes qui signifient le rire. Les études sur les émoticônes

sont fréquentes, mais elles traitent d'habitude de la fréquence de l'usage de n'importe quelle émoticône ou, parfois, de la distribution des émoticônes comparativement aux autres façons d'exprimer une attitude envers le texte. À titre d'exemple de ce premier, Witmer et Katzman (1997) ont trouvé que les femmes ont plus fréquemment employé les émoticônes que les hommes sur les forums en ligne, tandis que Huffaker et Calvert (2017) ont trouvé l'inverse sur les blogs écrits par les adolescents. À titre d'exemple de ce dernier, sur les blogs japonais, Nishimura (2016) a plus récemment analysé la distribution des émoticônes, les symboles non-linguistiques comme ♪, les kanjis, qui sont des logogrammes en japonais, lorsqu'ils remplissent la même fonction et les émojis, qui sont similaires aux émoticônes mais dont les ponctuations utilisées pour les générer ne sont plus perceptibles (p. ex. ☺ est un émoji tandis :) est l'émoticône apparentée), par rapport à des facteurs extra-linguistiques. Or, encore une fois, il n'a pas analysé la variation des formes des émoticônes de rire, qui serait plus pertinente ici. L'important est simplement que ces études démontrent que les émoticônes peuvent être contrôlées par des facteurs extra-linguistiques, ce qui suggère que les mots phatiques contenus dans une variable linguistique comme (lol) peuvent avoir ces mêmes sortes de relations.

2.4 Analyse des réseaux sociaux

Milroy (1980/1987) a été la première à appliquer l'analyse des réseaux sociaux à la sociolinguistique variationniste dans son étude de Belfast en 1980, et donc son travail sert de point de départ pour nous dans ce mémoire-ci, mais plusieurs améliorations à la méthode pour détecter les communautés dans les réseaux et pour mesurer la centralité des individus ont été apportées dans la sociologie et d'autres domaines qui

sont quasiment absents dans la sociolinguistique. Cette absence, nous voudrions la voir changer.

2.4.1 Construction générale des réseaux

L'analyse des réseaux sociaux est généralement effectuée en élaborant les réseaux sociaux d'un échantillon de sujets avec lequel on travaille dans ce qui est connu comme un sociogramme à la Moreno (1934). Les sociogrammes sont constitués de points, lesquels représentent chacun une personne dans le réseau, et des liens, lesquels représentent la façon dont les personnes dans le réseau sont reliées. Le point dont on regarde le réseau est appelé l'ego (Milroy, 1980/1987, p. 46-47). Les sociogrammes des réseaux prennent une forme comme la suivante :

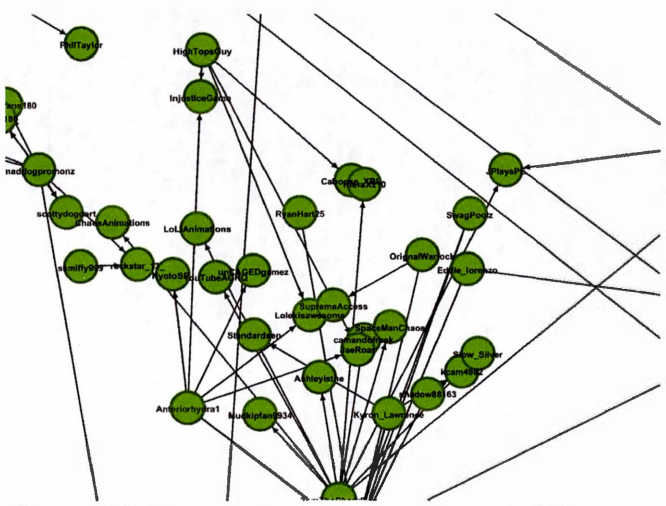

Figure 2.1: Une partie de la communauté 302, reproduction de la Figure 1.1

Pour caractériser les réseaux sociaux, Milroy (1980/1987) a introduit la densité et la multiplexité (p. 140). Un réseau dense correspond à un réseau dans lequel tous les points se relient à tous les points, calculé en divisant les connexions totales par les connexions totales possibles, où n'importe quel nombre de liens existant entre deux points constitue une seule connexion :

$$DR = \frac{connexions}{connexions\ possibles} = \frac{connexions}{\left(\frac{n(n-1)}{2}\right)} \qquad \text{(Formule 2.1)}$$

Un réseau multiplexe correspond à un réseau dans lequel les points se relient de plusieurs façons, telles que par parenté ainsi que par emploi, calculé en divisant les liens totaux par les liens totaux possibles, similaire à la densité mais prenant en compte le nombre de liens entre deux points.

Les liens entre les points d'un réseau peuvent représenter n'importe quoi. Milroy (1980/1987) était, elle, un peu vague en ce qui concerne la nature des liens. Elle semblait proposer que les relations familiales, professionnelles et d'amitié représentent des liens (Milroy, 1980/1987, p. 139-142), mais d'autres chercheurs ont construit les réseaux avec des liens qui représentent d'autres genres de contact. Li *et al.* (1992/2000), qui s'intéressaient principalement au caractère ethnique des réseaux de leurs sujets, en ont construit trois types qui s'appellent les liens passifs, les liens d'échange et les liens interactifs. Un lien passif est entre l'ego, un sujet dans ce cas-ci, et une personne avec qui l'ego interagit peu souvent mais qui influence et soutient l'ego de toute façon. Un lien d'échange est entre l'ego et un ami proche ou un

membre de sa famille. Un lien interactif est entre l'ego et une personne avec qui l'ego interagit souvent mais de qui l'ego ne dépend pas (Li *et al.*, 1992/2000, p. 177).

Twitter donne lui-même une occasion unique de construire des réseaux sociaux à partir des liens explicites, représentés par les abonnements des utilisateurs. Les utilisateurs de Twitter ont l'option de choisir d'autres utilisateurs de Twitter auxquels ils peuvent s'abonner pour que les tweets des utilisateurs abonnés apparaissent sur le fil d'actualité des autres utilisateurs qui s'y abonnent. Cependant, Kwak, Lee, Park, et Moon (2010) ont constaté que les abonnements ne sont réciproques qu'au taux de 22,1 % (p. 594), ce qui suggère que les réseaux construits à partir des abonnements ne représenteraient qu'une sorte de relation souvent unidirectionnelle. Bamman *et al.* (2014) se sont eux-mêmes servis des mentions-@ pour construire les réseaux sociaux (p. 140). Lorsqu'un utilisateur de Twitter veut envoyer un tweet directement à un autre utilisateur, il ajoute le symbole @ suivi du nom de l'autre utilisateur dans le tweet. Cette méthode pour construire les réseaux sociaux assure qu'une interaction considérable se passe dans le réseau résultant, et donc elle sert peut-être mieux une étude variationniste comme la nôtre puisque les interactions seraient suffisantes pour que les utilisateurs apprennent ou négocient des variantes appropriées pour le contexte.

2.4.2 Détection des communautés

Dans un réseau donné, on peut trouver des communautés, appelées des agglomérats par Milroy (1980/1987), ou des zones qui sont particulièrement denses et multiplexes relatives aux zones qui les entourent (p. 20-21/50-51). Cette définition des communautés est plus ou moins courante, bien que la méthode pour les identifier ait

bien changé. En fait, Milroy n'a proposé aucune méthode pour les identifier. Elle s'est servie de deux divisions administratives, les circonscriptions de Ballymacarrett et le Clonard, et d'un quartier, le Hammer (Milroy, 1980/1987, p. 43), supposant que les trois représentent des communautés dans le réseau social de la ville.

Or, si l'on veut détecter automatiquement les communautés, il existe des méthodes qui accomplissent cette tâche, telles que l'algorithme de Girvan-Newman (Girvan et Newman, 2002; Newman et Girvan, 2004) et la méthode de Louvain (Blondel, Guillaume, Lambiotte, et Lefebvre, 2008), cette dernière étant celle utilisée dans ce mémoire. La méthode de Louvain vise à maximiser la modularité Q du réseau. Q a été elle-même développée dans Newman et Girvan (2004), qui l'ont décrite comme « une mesure de la qualité d'une division particulière d'un réseau » (p. 8), c'est-à-dire que l'on divise un réseau en des communautés puis on calcule Q pour cette division. Leur calcul pour Q était le suivant :

$$Q = \sum_{\substack{\text{chaque} \\ \text{communauté}}} \left(\begin{array}{c} \text{connexions observées} \\ \text{entre les membres} \\ \text{de la communauté} \end{array} - \begin{array}{c} \text{connexions attendues} \\ \text{si elles se répartissaient} \\ \text{de façon aléatoire} \end{array} \right) \quad \text{(Formule 2.2)}$$

$$Q = \sum_i \left(e_{i,i} - a_i^2 \right)$$

Dans la formule, i représente une communauté, $e_{i,i}$ est un élément dans une matrice e qui représente la somme des connexions entre les membres d'une communauté et a_i est une colonne dans la matrice e qui, quand il est au carré, représente la somme des

connexions entre les membres d'une communauté à laquelle on s'attendrait si les connexions étaient réparties de façon aléatoire[17].

Un seul calcul de Q ne nous dit pas trop ; il faut la calculer plusieurs fois pour plusieurs divisions du réseau et comparer les résultats pour choisir la meilleure division. La méthode de Louvain est alors un algorithme qui effectue automatiquement cette étape. L'algorithme regroupe les points d'une façon, puis les regroupe encore d'une autre façon, jusqu'à ce que l'on trouve le regroupement qui maximise Q et donc identifie des communautés (Blondel *et al.*, 2008, p. 4-5).

Une façon courante d'évaluer les méthodes de détection des communautés est de les appliquer à des réseaux dont les structures des communautés sont déjà connues. L'un de ces réseaux est un club de karaté étudié par Zachary (1977). Dans ce club, il s'est produit une dispute qui a conduit à la dissolution du club et à la formation de deux nouveaux clubs. Zachary (1977) a démontré que la majorité de ceux qui étaient amis en dehors du club sont devenus membres du même club après la dissolution. On peut donc supposer indépendamment de l'analyse du réseau que deux communautés existaient tout avant la dissolution. Quand Q est maximisé pour ce réseau, sa valeur est 0,4198 et le réseau est divisé en quatre communautés, deux qui correspondent à un des clubs qui s'est formé après la dissolution et deux qui correspondent à l'autre (Aloise *et al.*, 2010, cité dans Waltman et Eck, 2013, p. 5-6), ce qui suggère que Q est une mesure exacte pour la division des réseaux en communautés. Newman et Girvan (2004) ont appliqué leurs propres algorithmes qui visent à maximiser Q, et ils ont obtenu des valeurs d'environ 0,4000 pour deux des algorithmes et de 0,1500 pour le troisième (p. 10-11), ce qui suggère que les algorithmes qui maximisent Q ne sont pas

17 Newman et Girvan (2004) donnent des justifications pour la formule, mais une discussion des mérites des mathématiques employées sort du cadre de ce mémoire.

tous égaux. Pour ce qui est de la méthode de Louvain, Waltman et Eck (2013) ont trouvé une valeur de 0,4151 pour Q (p. 471), qui s'avère presque la valeur la plus élevée possible pour le club de karaté. On peut donc supposer que la méthode de Louvain est fiable pour ce qui est de détecter de vraies communautés.

En utilisant un outil comme la méthode de Louvain pour détecter les communautés, les communautés que nous identifions seront denses et cohérentes, et il ne faudra pas nous fier à des indicateurs, comme pour les communautés linguistiques, ou simplement à des divisions administratives comme Milroy a fait à Belfast. Certes, on peut avancer qu'une communauté ne peut être définie comme une zone de densité dans un réseau, mais il est difficile d'imaginer une communauté dans laquelle les membres n'interagissent pas trop. De plus, Zachary (1977) a démontré que les communautés peuvent présenter des comportements qui diffèrent de façon significative. En fait, toute la sociolinguistique démontre ce fait, même si les communautés dont nous parlons ici sont décrites comme des sous-ensembles d'une seule communauté d'ailleurs[18]. Nous nous attendons donc à ce que la réalisation de (lol) différera de communauté en communauté dans l'étude actuelle.

2.4.3 Mesures de centralité

On peut analyser le comportement des individus relatif aux communautés où ils se trouvent simplement en identifiant où ils se trouvent dans un sociogramme donné, mais il vaut mieux quantifier leur centralité, autrement dit leur intégration, dans une communauté. Une façon de le faire est de se servir de l'échelle de force du réseau de Milroy (1980/1987), dont nous avons parlé dans la section 1.3.2 Analyse des réseaux

18 Un enjeu dont nous avons parlé dans la section 1.3.4 Concept de communauté.

sociaux, qui est un indice accordant un point pour chacun des cinq niveaux d'intégration, comme le fait de travailler avec au moins deux personnes dans le réseau ou non (p. 139-142). Or, il existe des mesures plus courantes dans l'analyse des réseaux sociaux qui sont aussi plus rigoureuses mathématiquement, dont le PageRank (Brin et Page, 1998), l'intermédiarité (Freeman, 1977, p. 37) ou même le degré.

Le degré d'un point dans un réseau indique le nombre de connexions entre lui et d'autres points ou, dans le cas du degré pondéré d'un point, le nombre de liens entre lui et d'autres points. En dépit du fait que le degré présente la mesure de centralité la plus simple de l'analyse des réseaux sociaux et encore plus de la théorie des graphes, Milroy ne l'a pas employée dans son étude de Belfast, peut-être parce qu'elle s'est servie des divisions administratives pour identifier les communautés au lieu de les élaborer sous forme de sociogrammes et donc ne savait pas les degrés de ses sujets. Cependant, nous ne nous servons pas seulement du degré mais aussi du degré entrant et du degré sortant. Dans ce cas-ci, ces deux mesures représentent les tweets dirigés vers l'individu et depuis l'individu respectivement.

Tandis que le degré est une mesure simple, le PageRank présente un outil plus robuste qui peut calculer la centralité d'un individu dans une communauté. Brin et Page (1998) l'ont développé pour classer les sites web dans les résultats de recherche de Google, mais la mesure peut être appliquée à n'importe quel genre de réseau, y compris les réseaux sociaux, sans modification. L'équation est la suivante :

$$PR(A) = (1 - d) + d\left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right) \qquad \text{(Formule 2.3)}$$

Dans cette équation, *PR(A)* représente le PageRank du point *A*, *PR(Tn)* représente le PageRank d'un énième point *T* relié à *A*, *C(Tn)* représente le nombre de points auxquels un énième point *T* est relié, y compris *A*, et *d* représente un facteur d'amortissement. À titre d'exemple, si l'on avait un point *A* auquel trois autres points, *B*, *C* et *D*, étaient reliés, ces trois autres points ayant des valeurs de *PR* de 0,75 chacun et deux, un et trois points auxquels ils se relient respectivement, eux, et enfin un facteur d'amortissement de 0,85, valeur typique selon Brin et Page (1998, p. 109) mais aussi plus ou moins arbitraire, on aurait le suivant :

$$PR(A) = (1 - d) + d\left(\frac{PR(B)}{C(B)} + \frac{PR(C)}{C(C)} + \frac{PR(D)}{C(D)}\right)$$

$$PR(A) = (1 - 0{,}85) + 0{,}85\left(\frac{0{,}75}{2} + \frac{0{,}75}{1} + \frac{0{,}75}{3}\right)$$

$$PR(A) = 1{,}31875$$

Si l'on augmentait le nombre de points reliés à *A*, la valeur de *PR* de *A* augmenterait. De même, si les points reliés à *A* avaient des valeurs de *PR* plus élevées ou s'ils étaient reliés à moins d'autres points, la valeur de *PR* de *A* augmenterait. De ce fait, *PR* est fonction des liens entrants, les tweets entrants dans ce cas, les sources de ces liens et l'unicité de ces liens. Les valeurs de *PR* initiales ne sont pas connues. Pour les connaître, on devine des valeurs initiales, on calcule toutes les valeurs de *PR* puis on les recalcule depuis ces valeurs-ci et on les recalcule encore, et ainsi de suite, jusqu'à ce que les valeurs ne changent qu'un petit peu.

L'intermédiarité est une autre mesure de centralité qui identifie théoriquement ce que Granovetter (1973) a appelé les « personnes intermédiaires » dans un réseau social,

autrement dit les points qui relient deux communautés d'un réseau avec des liens qui sont tous faibles (p. 1367-1368), quand on la calcule pour l'ensemble d'un réseau. Les intermédiaires se trouvent à chaque bout des ponts, les ponts étant les liens qui fournissent les seuls chemins entre deux points (Harary, Norman, et Cartwright, 1965, cité dans Granovetter, 1973, p. 1364). Les points avec les intermédiarités les plus prononcées devraient donc se trouver aux périphéries des communautés si l'on est d'accord avec les arguments mis de l'avant par Granovetter (1973). En effet, Zhao, Wu, et Xu (2010) ont produit des preuves que les liens faibles servent de ponts dans les réseaux sociaux en ligne, comme Twitter, par exemple. L'intermédiarité peut alors être décrite comme une mesure d'importance au lieu d'une mesure de centralité puisque les points qu'elle identifie ne sont pas bien intégrés dans leurs communautés, mais ils jouent néanmoins des rôles cruciaux dans la diffusion des informations à travers le réseau.

L'intermédiarité a été pour la première fois formalisée par Freeman (1977) et ainsi calculée (p. 37) :

$$C_B(p_k) = \sum_{i<j}^{n} b_{i,j}(p_k) \qquad \text{(Formule 2.4)}$$

Ici, la centralité d'intermédiarité C_B pour le point p_k est la somme de $b_{i,j}$ de p_k, qui représente les géodésiques entre les points p_i et p_j qui traversent p_k sur toutes les géodésiques entre p_i et p_j.

Ces mesures de centralité ainsi que les méthodes récentes pour détecter les communautés raffinent ce que Milroy (1980/1987) a introduit à la sociolinguistique

en 1980. Avec ses outils, elle a pu démontrer que la position des individus dans leurs communautés a un effet sur leurs parlers. On s'attend donc à ce que des outils plus précis, tels que la méthode de Louvain pour détecter les communautés (Blondel *et al.*, 2008) et le PageRank pour déterminer l'intégration des individus dans les communautés (Brin et Page, 1998), nous permettent également d'observer un effet.

2.5 Variation stylistique

La variation stylistique est essentiellement la variation au niveau de l'individu, autrement dit la façon dont un individu change son parler à mesure que le contexte change. Or, il y a plusieurs développements par rapport à sa conceptualisation au fil du temps, de Labov jusqu'à présent. Nous exposerons ces développements dans cette section-ci afin de situer nos résultats dans la théorie sociolinguistique.

La conceptualisation de style de Labov est fondamentale dans sa recherche ainsi que dans la recherche de ceux qui l'ont suivi. L'idée qu'un vernaculaire existe, défini comme le premier style acquis par un locuteur, parfaitement contrôlé et principalement parlé avec les amis et la famille (Labov, 1966/2006, p. 86), est à l'origine de sa conceptualisation. Labov (1972a) a défini les styles à partir de la combinaison de son principe vernaculaire, lequel dit que le vernaculaire est le style le plus stable et le plus constant, exigeant le moindre d'attention à produire, et son principe d'attention, lequel dit que les styles sont classés le long d'une seule dimension selon l'attention au parler du locuteur qu'ils exigent (p. 112). Cette perspective unidimensionnelle de la variation stylistique ne réduit effectivement les styles qu'à une mesure de formalité en ce que les contextes formels poussent un locuteur à faire plus attention à son parler afin de produire un style plus formel,

superposé sur son vernaculaire. Labov (1966/2006) croyait que l'on peut changer la formalité des entrevues en passant par quatre activités :

- L'activité de parler en dehors de l'entrevue
- L'activité de parler dans l'entrevue
- L'activité de lire des paragraphes
- L'activité de lire des listes de mots (p. 59-71)

Or, Mahl (1972) a effectué une expérience sur les variables linguistiques (dh) et (th) qui a remis en cause l'idée que l'attention au parler est à la base de la variation stylistique. Il a testé la capacité des sujets à produire des styles formels s'ils ne peuvent s'entendre comparativement à s'ils ne peuvent voir leur interlocuteur. Pour ce qui est de (dh), la perte de l'ouïe a entraîné des styles moins formels ainsi que le manque de vue de l'interlocuteur, mais pour ce qui est de (th), seul le manque de vue de l'interlocuteur a entraîné des styles moins formels (Mahl, 1972, cité dans Bell, 1984, p. 148-149). Bell (1984) a conclu que l'audience était plus importante en termes de variation stylistique (p. 149).

Bell (1984) a donc proposé le design d'audience. Dans ce cadre, un locuteur change d'un style à un autre selon la personne à qui il s'adresse. Bell (1984) a proposé trois faits possibles qui sont pertinents au locuteur dans sa décision d'alterner son style :

- Les caractéristiques personnelles de l'interlocuteur
- Le style général utilisé par l'interlocuteur
- Les valeurs des variables linguistiques utilisées par l'interlocuteur (p. 167-168)

D'autres personnes jouent également un rôle dans ce cadre, outre l'interlocuteur. Bell (1984) a nommé quatre « rôles d'audience » :

- Destinataire, reconnu, ratifié et auquel s'adresse directement le locuteur
- Auditeur, reconnu et ratifié par le locuteur
- Auditeur par hasard, reconnu par le locuteur
- Indiscret, non reconnu par le locuteur (p. 159)

Il y a enfin des groupes de référence qui sont pertinents pour le locuteur mais non présents (Bell, 1984, p. 161). Selon Bell, quand un locuteur parle, il détermine à qui il s'adresse vraiment, des faits de cette personne ou ces personnes, s'il y a une autre identité de référence qu'il veut exprimer, puis il choisit un style approprié. Cependant, Bell a réaffirmé l'idée que les styles se situent sur un continuum unidimensionnel, à l'instar de Labov (1966/2006).

En revanche, Eckert (2000) a affirmé qu'il n'y a aucune raison de croire qu'un style vernaculaire sur lequel tous les autres styles sont superposés existe, ni un style exigeant plus ou moins d'effort pour produire que n'importe quel autre style (p. 17-18). Ce changement de perspective libère le style d'être vu comme un phénomène multidimensionnel plutôt qu'un phénomène unidimensionnel. Eckert (2000) a poursuivi en suggérant que les styles soient construits à partir de l'appropriation créative de diverses ressources linguistiques dont on profite à nouveau (p. 213-216). On arrive donc à un répertoire de styles qu'un individu construit de façon créative en interagissant avec différentes communautés qui lui présentent différentes ressources.

Pour notre part, nos résultats ne contrediraient rien de ces théories antérieures. En effet, nous n'analysons pas l'alternance des styles, car la méthode de détection des communautés dont nous nous servirons classe chaque sujet dans une seule communauté, donc cette partie du contexte est fixe par rapport à chaque sujet. De plus, nous n'identifions pas les changements de thème de discours à l'intérieur des communautés. Nous sommes d'accord avec l'idée qu'un répertoire de styles est multidimensionnel, contraint par n'importe quel nombre de dimensions du contexte, mais puisque nos résultats ne décriront qu'un seul style pour chaque sujet, ils seront valides même si l'on part du principe qu'il existe un vernaculaire avec d'autres styles superposés dessus.

2.5.1 Diversité des réalisations des variables linguistiques des individus

Quant à la cohérence des réalisations des variables linguistiques des individus, la littérature ne présente pas trop d'indices. Premièrement, si l'on veut appliquer des mesures de dispersion à une variable linguistique, il faut savoir les fréquences de toutes ses variantes, mais elles sont peu souvent constatées pour un seul locuteur. Deuxièmement, il est courant de choisir une seule variante d'une variable linguistique, de regrouper les autres et d'analyser la variable comme une proportion. En effet, D. Sankoff et Laberge (1978a) ont avancé qu'une occurrence d'une variable linguistique est une épreuve de Bernoulli (p. 119), c'est-à-dire que les variables linguistiques sont binomiales. Les fréquences de chaque variante ne sont donc pas toujours constatées en général.

Il existe néanmoins des études utiles à cet égard, comme celle de Sharma (2011), qui a effectué une analyse de « l'anglais asiatique britannique ». Deux traits rendent cette

étude utile : elle n'a analysé que des variables linguistiques avec deux réalisations possibles et elle a constaté leurs réalisations dans plusieurs contextes pour quatre individus. Nous ne pouvons tirer de fortes conclusions à partir des données, mais elles sont intéressantes. Par exemple, Anwar, un homme plus âgé, a utilisé presque exclusivement des variantes britanniques pour toutes les variables linguistiques lorsqu'il parlait avec un mécanicien cockney, un avocat asiatique britannique distingué et des enfants asiatiques britanniques. Il a en revanche produit presque exclusivement des variantes asiatiques avec une femme de chambre sri-lankaise. Or, lorsqu'il parlait dans les entrevues, les fréquences entre les variantes britanniques et asiatiques étaient presque équivalentes (Sharma, 2011, p. 475). Si l'on peut supposer qu'il connaissait mieux le mécanicien, l'avocat, les enfants et la femme de chambre que les chercheurs qui ont effectué les entrevues, on peut également supposer que le manque d'intégration dans les réseaux des chercheurs a rendu difficile le fait de choisir des variantes appropriées. Ce genre de résultat s'est reproduit pour deux autres sujets, mais le quatrième a curieusement produit des variantes britanniques avec les chercheurs. En fait, ce sujet, jeune femme qui s'appelle Namrita, n'a produit des variantes asiatiques qu'avec ses parents à des taux presque équivalents aux variantes britanniques utilisées avec ses parents (Sharma, 2011, p. 480). Nous pouvons donc avancer que l'intégration d'un individu dans une communauté entraîne des réalisations cohérentes des variables linguistiques, mais seulement avec quelques hésitations.

Outre les observations dans la littérature, il existe aussi un argument logique qui nous suggère que les individus présenteront moins de diversité dans leurs réalisations des variables linguistiques que leurs communautés. Si un individu est assuré de sa place dans une communauté, il a eu le temps d'apprendre les normes de la communauté et d'apprendre comment la communauté répond à son style choisi, soit faisant partie de

ces normes ou non. Cet argument ressemble à ce que Le Page et Tabouret-Keller (1985) ont proposé pour expliquer la formation de nouvelles variétés. Ils ont avancé l'hypothèse que les variétés deviennent de plus en plus « nettes », comme une image, à mesure qu'une communauté se forme (Le Page et Tabouret-Keller, 1985, p. 115-116). Ce qu'ils entendaient par variétés nettes, c'était que la variation dans les variétés diminue. Le Page et Tabouret-Keller proposaient que les membres d'une nouvelle communauté se partagent leurs images de l'univers et qu'ils changent leurs images pour qu'elles soient plus semblables à mesure qu'ils deviennent plus proches. Essentiellement, c'est une supposition que les communautés deviennent homogènes car les membres apprennent à se connaître.

2.6 Hypothèses

Dans ce cadre théorique, nous avons donc présenté un argument en faveur de l'idée que la centralité ou l'intégration d'un individu dans une communauté conduit à moins de diversité dans ses réalisations des variables linguistiques que dans les réalisations de la communauté dans l'ensemble. Cet argument est en partie soutenu par ce que Sharma (2011) a trouvé dans son étude de l'anglais asiatique britannique. Nous avons également montré que la méthode de Louvain (Blondel *et al.*, 2008) réussit à détecter des communautés cohérentes qui correspondent aux communautés identifiées de façon indépendante, entre lesquelles les réalisations des variables linguistiques constituées de mots phatiques peuvent varier. Enfin, Milroy (1980/1987) a déjà démontré que l'intégration dans une communauté peut influencer les réalisations des variables linguistiques. À partir de ces faits, deux hypothèses peuvent être formulées pour répondre aux deux questions de recherche :

(H1) La distribution des réalisations de la variable linguistique lexicale (lol), constituée de mots d'origine française et anglaise, sera différente pour chaque communauté identifiée sur Twitter à partir des tweets émanant des provinces maritimes au Canada.

(H2) Les individus bien intégrés dans leurs communautés vont réaliser la variable linguistique lexicale (lol) avec moins de diversité qu'au niveau de leurs communautés.

Afin de clarifier H1, nous allons répondre à cette question par l'application du test exact d'indépendance de Fisher[19] (1922, 1925/1970), dans lequel on demande si la fréquence relative de chaque catégorie est égale pour toutes les communautés. Il y a donc de nombreuses façons dont l'hypothèse peut s'avérer exacte. Par exemple, si *lol* est plus fréquent dans cinq communautés, ***mdr*** est plus fréquent dans cinq autres et *ptdr* dans une autre, il se peut que le test nous dise que les distributions des réalisations de (lol) diffèrent entre les communautés à un niveau statistiquement significatif.

Enfin, pour que ces hypothèses s'avèrent exactes, une condition principale doit être remplie : les sujets que nous analysons doivent reconnaître les communautés auxquelles ils participent comme existantes. Comme a constaté Boissevain (1974), ce n'est pas l'existence objective d'une communauté qui influence le comportement des membres de ladite communauté, c'est plutôt la perception de son existence ou non qu'ont les membres (cité dans Milroy, 1980/1987, p. 61). De ce fait, on ne prévoit pas qu'une communauté se comporte comme un mécanisme d'application des normes si elle n'existe en fait pas dans l'esprit de ses membres, mais les hypothèses ci-dessus sont censées s'avérer exactes si cette condition est remplie.

19 Pour plus de détails, voir la section 3.5.2 Signification et indépendance.

CHAPITRE III

MÉTHODE

3.1 Introduction

Dans ce mémoire, nous nous intéressons aux patrons d'usage de la variable linguistique lexicale (lol) sur Twitter par les utilisateurs séjournant ou habitant dans les provinces maritimes du Canada, cette variable contenant des mots d'origine anglaise et française. Cette section-ci exposera la façon dont nous avons effectué la collecte de données et les avons analysées afin d'en savoir davantage sur les associations entre les communautés dans les réseaux sociaux et la réalisation de (lol) ainsi que sur le rapport entre les styles des individus et les registres des groupes. La collecte de données sera d'abord décrite, suivie de la construction du réseau social et la détection des communautés, le codage et enfin les analyses statistiques.

3.2 Collecte de données

La collecte de données a été effectuée deux fois : une fois pour identifier une variable linguistique lexicale appropriée et une deuxième fois pour les analyses principales. La même procédure a été suivie pour les deux, sauf quelques petits changements. Dans les deux cas, nous avons utilisé le site Netlytic (Gruzd, 2016) pour prendre les données. Ce site peut automatiquement prendre des données des sites de médias sociaux selon une recherche et une fréquence que l'on indique. Par exemple, on peut

lui demander de prendre tous les tweets de Twitter contenant le mot *chiac* toutes les 15 minutes pendant une semaine. Pour Twitter, la prise de données se restreint à 1 % des tweets du monde, qui est la même restriction qu'a n'importe quel utilisateur lorsqu'il fait une recherche sur le site web de Twitter. Les données extraites de Twitter représentent donc un échantillon et non pas un recensement. On peut ensuite exporter les données en format de fichier CSV, lequel peut être ouvert dans un tableur ou dans R. Pour chaque tweet, Netlytic prend plusieurs renseignements que nous avons inclus dans notre analyse le lien vers le tweet en tant que numéro d'identification, le nom d'utilisateur, le texte du tweet, le moyen par lequel le tweet a été envoyé, nommé la source, la localisation de l'utilisateur s'il la fournit et le fuseau horaire.

3.2.1 Paramètres de la recherche sur Netlytic

La recherche de base pour les deux collectes de données a été presque identique. Quant à la recherche initiale pour identifier une variable linguistique, nous nous sommes servis de la recherche suivante, qui a pris les données toutes les 15 minutes pendant un mois, du 11 janvier au 8 février 2017, et qui a pris 12 905 tweets :

- geocode:46.0878,-64.7782,200mi lang:fr exclude:retweets exclude:links

où chaque paramètre a le sens suivant :

- geocode:46.0878,-64.7782,200mi – Les coordonnées en latitude et longitude ainsi que le rayon
- lang:fr – La langue

- exclude:retweets – Exclure les tweets qui ne sont que des copies
- exclude:links – Exclure les tweets qui sont bien probables de ne contenir que des mèmes, des vidéos ou des articles

Cette recherche cible les tweets qui viennent d'un rayon de 200 miles (322 km) autour de Moncton, au Nouveau-Brunswick, qui sont jugés français par Twitter, qui ne sont pas des retweets et qui n'incluent pas d'hyperliens. La zone ciblée correspond plus ou moins aux provinces maritimes, tandis que les deux derniers paramètres réduisent la probabilité de prendre des tweets qui ne représentent pas des expressions originales, à l'instar de Pavalanathan et Eisenstein (2015, p. 199).

Le paramètre de langue assure que les tweets ne sont pas complètement d'origine anglaise, faute de quoi, il y aurait une forte possibilité que ce soit le cas puisque l'anglais est plus répandu que le français dans les provinces maritimes (Statistique Canada, 2016) et peut-être sur Twitter en général (Kim, Weber, Wei, et Oh, 2014, p. 246). Ce paramètre n'interdit pas les mots d'origine anglaise isolés, mais plutôt les tweets que Twitter juge complètement anglais, donc on peut toujours identifier des variables linguistiques lexicales qui contiennent des variantes d'origine anglaise dans ces données.

3.2.2 Identification d'une variable linguistique lexicale

Nous avons cherché la variable linguistique lexicale dans les données recueillies dans la recherche initiale en utilisant les nuages de mots que Netlytic construit à partir des tweets. Netlytic présente les mots les plus fréquents dans les tweets et permet de les éliminer un par un jusqu'à ce qu'un mot d'origine anglaise soit trouvé. Au lieu de

Tableau 3.1: Items lexicaux qui auraient pu servir de variantes de variables linguistiques lexicales et les raisons de leur exclusion ou non

Raison	Mots
1) beaucoup de fonctions, peu d'occurrences avec une seule fonction	nice, time
2) fait partie d'un nom propre ou composé d'habitude	night, game, snap, team, bro, project
3) (1) et (2)	top, power
4) seulement quelques utilisateurs responsables pour toutes les occurrences	guys, happy, mute
5) seul mot dans les tweets ou les tweets sont dépourvus d'autres éléments d'origine française	congrats, follow, close, jealous, savage, pass
6) sans une variante claire d'origine française ou peut-être de la troncation d'un mot d'origine française	wtf, cool, wow, phone, parking
7) catégorique (aucune occurrence des équivalents d'origine français)	fans, minions, jaywalking
8) approprié	lol, job, fun, stop, show, omg, condom

choisir le premier mot auquel nous sommes arrivés, nous avons décidé de continuer jusqu'à ce qu'une liste de mots avec au moins neuf occurrences chacun soit élaborée. Cette liste comprenait les mots dans le Tableau 3.1, classés selon la raison pour laquelle nous les avons trouvés non-appropriés.

Les mots que nous avons classés comme étant appropriés étaient les suivants :

- lol, job, fun, stop, show, omg, condom

Tableau 3.2: Fréquence des mots d'origine anglaise avec leurs synonymes et leurs quasi-synonymes d'origine française

Mot	Variantes	Fréquence
lol	mdr	255
job	emploi, travail, ouvrage, poste, boulot, devoir, responsabilité, tâche, métier	141
fun	amusant, amusement, sympa, sympathique, drôle, rigolo	134
stop	arrêt, cesser, empêcher, mettre fin, met fin	120
show	spectacle, programme, émission	100
omg	omd, mon dieu, oh seigneur, oh sainte, (oh) putain	99
condom	capote, préservatif	29

Nous avons cherché quelques synonymes ou quelques quasi-synonymes de ces mots sur Netlytic, qui sont d'origine française, afin d'obtenir une idée générale de leurs fréquences en tant que variables. Par exemple, nous avons cherché toutes les occurrences de *show* mais aussi de *spectacle*, *émission* et *programme*. Les variantes de chaque mot ainsi que leurs fréquences globales se présentent dans le Tableau 3.2. Puisque que ce n'était qu'un essai pour trouver une variable appropriée et fréquente, nous ne nous sommes pas trop occupés des questions de la validité des synonymes, alors on peut se demander si *rigolo* est vraiment un synonyme de *fun*, mais son inclusion n'a pas changé le fait que *fun* et ses synonymes possibles étaient moins fréquents que *lol*, le mot que nous avons finalement choisi, car ce dernier était beaucoup plus fréquent que les autres même en n'incluant qu'un seul synonyme.

3.2.3 Collecte de données principale

Quant à la collecte de données principale, également effectuée du 11 janvier au 7 février 2017, les mêmes démarches ont été suivies sans le paramètre de langue, c'est-à-dire que nous avons utilisé la recherche suivante :

- geocode:46.0878,-64.7782,200mi exclude:retweets exclude:links

Nous avons supprimé le paramètre de langue pour garantir de ne pas exclure des tweets dépourvus d'autres éléments d'origine française qui contiennent quand même des variantes d'origine française de la variable linguistique lexicale (lol), par exemple, un tweet hypothétique comme le suivant :

UtilisateurX : @UtilisateurY That was funny. mdr (Exemple 3.1)
'@UtilisateurY C'était drôle. mdr'

Si Twitter trouve la plupart des éléments dans ce tweet hypothétique anglais, il se peut qu'il l'exclue. Nous n'avons prévu ni un tweet comme dans l'Exemple 3.1, ni que Twitter l'exclurait si la recherche incluait le paramètre *lang:fr*, mais il vaut mieux ne pas présumer que ce ne soit que les tweets avec des éléments d'origine française qui permettent des éléments d'origine anglaise et non pas l'inverse et que Twitter fonctionne ainsi. De plus, ce choix nous a permis de déterminer si les tweets dépourvus d'autres éléments d'origine française interdisent les variantes d'origine française de (lol).

3.3 Construction du réseau social et mesures de centralité

Netlytic est en fait capable de dessiner automatiquement un réseau social à partir des tweets qu'il a recueillis et de détecter des communautés dans ce réseau en plus, mais il ne permet ni de contrôle sur le procès ni d'explication de la méthode de détection utilisée. Nous avons donc décidé d'importer les données dans l'application Gephi (Bastian, Heymann, et Jacomy, 2009), dans laquelle on peut construire les réseaux selon n'importe quels critères, choisir les paramètres de l'algorithme qui détecte les communautés et choisir les mesures de centralité.

3.3.1 Construction du réseau

La collecte de données principale a pris 1 265 789 tweets. Cependant, nous nous intéressions seulement aux tweets dirigés, c'est-à-dire les mentions-@, à partir desquelles nous avons construit le réseau social, comme Bamman *et al.* (2014) l'ont fait (p. 140). Chaque fois qu'un utilisateur envoie un tweet qui contient le symbole @ suivi du nom d'un autre utilisateur, cet autre utilisateur en est prévenu, sauf si le destinataire a bloqué l'envoyeur. Les mentions-@ représentent donc des communications directes. Les mentions-@ ont constitué 307 878 tweets, envoyés par 211 121 utilisateurs.

Ces mentions-@ ne servent pas seulement de méthode pour former le réseau initial, mais également de méthode pour mesurer l'activité entre toutes les paires d'individus du réseau en comptant chaque tweet à mention-@ comme un lien entre une paire. Le poids du rapport entre une paire donnée est le nombre de liens qui les relient, autrement dit le degré entre la paire. Sharma (2011) a mentionné le degré dans son

étude de la variation en anglais asiatique britannique (p. 471), mais elle ne l'a pas intégré dans son analyse. De notre connaissance, cette mesure du poids demeure peu utilisée dans la sociolinguistique variationniste. Compte tenu du poids des liens, nous nous sommes servis de la méthode de Louvain (Blondel *et al.*, 2008) dans Gephi pour détecter les communautés dans le réseau, une fois construit, en sélectionnant aussi l'option « aléatoire » et en tenant la résolution à 1,0, où une résolution plus élevée ou plus baissée détecterait moins ou plus de communautés, respectivement.

3.3.2 Mesures de centralité

Les mesures de centralité que nous avons employées agissent comme des facteurs. Pour chaque individu, plusieurs mesures ont été calculées dans Gephi : le degré, le degré pondéré, le degré sortant, le degré sortant pondéré, le degré entrant, le degré entrant pondéré, le PageRank et l'intermédiarité. Toutes ces mesures ont été calculées par rapport aux communautés sauf l'intermédiarité, c'est-à-dire que nous avons calculé les degrés et le PageRank en extrayant une communauté et en exécutant l'algorithme, ce qui veut dire que les liens qu'un individu avait avec des points hors de la communauté n'ont pas été considérés. Cependant, l'intermédiarité a été calculée en considérant les liens aux points hors de la communauté car elle indique principalement les individus qui relient des communautés.

Le PageRank a été choisi au lieu d'autres mesures de centralité similaires en raison de la nature des données et de son succès sur le web, mais le choix est forcément un peu arbitraire, aussi. Par exemple, le PageRank ressemble à la centralité de vecteur propre, mais ce premier est plus approprié lorsque l'on travaille avec des graphes qui peuvent indéfiniment s'étendre, comme un réseau social, grâce au facteur

d'amortissement que le PageRank introduit. Ce facteur, mis à 0,85 dans notre étude, assure que l'influence des points très lointains est réduite. D'autres mesures de centralité, comme la proximité (Bavelas, 1950; Freeman, 1978) et l'algorithme HITS (Kleinberg, 1999a, 1999b), demandent d'être plus analysées par rapport à leur application aux réseaux sociaux quand on veut spécifiquement étudier le comportement des traits linguistiques dans ces réseaux sociaux. Il se peut, par exemple, qu'un individu très central selon la proximité exerce une influence importante sur les normes de la mode et non sur les normes langagières. Faute de telles analyses, le succès évident du PageRank qui parvient au moins à identifier des sites web utiles dans les recherches sur Google a joué un rôle dans notre décision de le mettre en œuvre dans cette étude-ci.

Les degrés ne sont pas inclus dans cette analyse en tant que substituts de PageRank, mais afin d'augmenter ce que nous savons des points, c'est-à-dire la direction de leurs liens et leur nombre de liens, étant donné que le PageRank est plus basé sur l'importance de chacun de leurs liens. De même, l'intermédiarité nous indique quelque chose de distinct par rapport au PageRank : elle nous informe dans quelle mesure un individu relie deux communautés, ce qui va de pair avec sa position à la périphérie de sa communauté, selon l'argument de Granovetter (1973). L'intermédiarité a donc été employée comme un facteur et calculée dans Gephi avec l'algorithme de Brandes (2001) avec l'option « Normalise entre[0,1] » sélectionnée.

3.4 Codage

Dans les 307 878 tweets que nous avons sélectionnés, il y avait 4 733 occurrences de (lol). Puisque nous n'avions pas les moyens de lire tous les 307 878 tweets initiaux,

Tableau 3.3: Toutes les formes de base des variantes de (lol) et les mots-clés utilisés pour les extraire

Forme ciblée	Mots-clés
lol	lol, lool, looo
lawl	lawl, wwl, laa
lawlz	lz[20]
lel	lel, leel, leee
lul	lul, luul, luuu
l0l	l0l, l00l, l000, o0, 0o
kek	kek, keek, keee
lmao	lmao, maa, lmm
lmfao	lmfao, faa, mff, mmf
rofl	rofl, oof, ffl
mdr	mdr, mdd
ptdr	ptdr, tdd, ttd

nous sommes arrivés au nombre d'occurrences en les extrayant de l'ensemble des tweets selon des mots-clés qui ont saisi toutes les possibilités orthographiques des variantes, énumérés dans le Tableau 3.3. À titre d'exemple, le mot-clé *luuu* peut saisir *luuuul*, *luuuuuuul*, *luuullll*, et ainsi de suite. Ces mots-clés ont forcément pris de doubles tweets et des tweets dépourvus d'occurrences de (lol), mais nous les avons facilement supprimés pendant le codage à la main, où nous avons commencé par 16 583 tweets et fini par 4 733 occurrences de (lol). Chaque tweet a été finalement codé pour la communauté de l'utilisateur qui l'a envoyé, ainsi que les degrés, le PageRank et l'intermédiarité de cet utilisateur, et plusieurs autres facteurs sociaux et structurels ont été également considérés. Cette section détaillera ces variables-ci.

20 Contrairement aux autres ensembles de mots-clés, l'usage du mot-clé *lawlz* dans ce cas serait redondant, car *lawl* au-dessus capture déjà l'orthographe *lawlz*. Nous avons inclus *lz* afin de capturer les formes dans lesquelles il y a plus d'un L ou d'un Z, comme *lawlzzz*, par exemple.

Netlytic a automatiquement codé plusieurs facteurs pour chaque tweet, dont la source, la localisation et le fuseau horaire. La source fait référence au moyen par lequel le tweet a été envoyé, tel que par l'application Twitter d'un téléphone Android, le codage étant standardisé par nous comme *Android, Twitter*. Nous avons inclus cette variable simplement car elle a été déjà codée, mais elle ne semble pas pertinente *a priori*. Nous avons fini par identifier 28 niveaux du facteur *source*.

3.4.1 Localisation

La localisation de chaque tweet provient de ce que l'utilisateur qui l'a envoyé a entré dans son profil. Nous avons divisé cette entrée en trois variables : le pays, la province et la ville. Par exemple, si l'on a saisi « Fredericton » comme étant sa localisation, nous avons codé *Canada* pour le pays, *Nouveau-Brunswick* pour la province et *Frédéricton* pour la ville. En ce qui concerne les utilisateurs qui ont saisi une localisation moins précise, telle que « Nova Scotia », nous avons codé *Canada*, *Nouvelle-Écosse* et *indéfini* pour la ville. Parfois, deux localisations ont été énumérées dans les profils. Dans ces cas, nous avons indiqué les deux localisations pour chaque variable, séparées par un tiret. Naturellement, ces sujets n'ont pas pu être regroupés avec les sujets qui ont identifié une seule localisation qui correspondait à l'une des localisations indiquées par les sujets qui en ont identifié deux, mais ce groupe-ci ne constitue pas d'une grande part des cas. En effet, ce dernier constitue deux cas selon la variable *pays*, un cas selon la variable *province* et un cas selon la variable *ville*. Il est également à noter que le nombre de sujets qui ont indiqué leurs localisations diminue à mesure que l'on va du pays à la ville, c'est-à-dire que 242 tweets sont venus des utilisateurs qui n'ont pas indiqué leurs pays, mais 2 348 des utilisateurs qui n'ont pas indiqué leurs provinces et 3 232 des utilisateurs qui n'ont

pas indiqué leurs villes. Nous avons finalement identifié 9 niveaux pour les pays, 18 pour les provinces et 96 pour les villes[21].

Puisqu'il n'y aura pas d'interaction directe avec les sujets, il n'est pas possible de bien savoir leurs provenances. Twitter permet aux utilisateurs d'identifier leur localisation, mais elle ne nous indique pas leur provenance. De plus, le choix d'identifier la localisation est facultatif, et les sujets peuvent mentir, donc on ne peut trop se fier à cette auto-identification. Nous ne prévoyons pas qu'un nombre important de sujets aient menti, mais c'est une supposition à garder à l'esprit.

Le fuseau horaire a été laissé tel quel. Netlytic nous a donné des chiffres qui les représentent. Un grand nombre de tweets n'ont pas indiqué le fuseau horaire, soit 1 376. Ces tweets, nous les avons codés comme *indéfini*. Le reste n'a pas soulevé de problème.

3.4.2 Communautés et langues

En ce qui concerne les communautés, Gephi en a identifié un nombre énorme en utilisant la méthode de Louvain, 8 945 pour être précis, dont 19 sont d'intérêt en raison de l'usage de la variable linguistique (lol) dans des tweets constitués d'éléments d'origine française dans ces communautés. En effet, les variantes de (lol) d'origine française ne paraissent jamais dans les tweets qui ne sont pas constitués

21 Il faut ne pas confondre les localisations indiquées par les utilisateurs sur leurs profils et la région géographique ciblée par les paramètres de la collecte de données. Ces deux catégories ne sont pas du tout identiques. La région ciblée dans la collecte de données augmente la possibilité que nous trouverons des tweets envoyés depuis les provinces maritimes, mais il semble que les tweets passent simplement par des serveurs situés dans les provinces maritimes, qui peuvent différer des vraies localisations des utilisateurs (voir la section 3.5.3 Représentativité, ci-dessous).

d'autres éléments d'origine française. Les tweets qui s'en rapprochent le plus ne contiennent qu'un ou deux éléments d'origine anglaise, d'habitude, mais des éléments d'origine française sont toujours présents dans ces cas. Par exemple :

Bobby B. : @EldurSensei ptdr je m'en doutais tellement mais tkt jte follow parce que t'es un dieu (Exemple 3.2)

Hawk : @M_Gnangni mdrrr :D :D :D :D Quand tu seras dans ton mood faut revenir tu vas prendre pour toi (Exemple 3.3)

Le seul tweet que nous avons codé comme contenant une variante d'origine française de (lol) et aucun élément d'origine française ailleurs dans le tweet était le suivant :

Hawk : @EzechielDegny mdrrr yafoy bro (Exemple 3.4)

Nous n'avons pas jugé cet exemple un vrai exemple, toutefois. La langue a été codée comme *anglais-africain* à cause des termes *bro* et *yafoy*, respectivement. Cependant, *yafoy* est standard dans ce que l'on appelle souvent du français en Afrique. Nous aurions donc pu le coder également comme *français-anglais*, qui maintiendrait l'idée que les variantes d'origine française de (lol) ne paraissent jamais dans les tweets dépourvus d'autres éléments d'origine française. Nous avons finalement décidé que le codage *anglais-africain* était le choix plus prudent puisque son élément d'origine française était douteux et puisque nous nous intéressions principalement aux tweets contenant des éléments d'origine française.

Nous avons donc éliminé les tweets considérés non-français par Twitter, identifié la communauté de chacun et fini par 19 communautés. Tous les tweets qui contenaient des occurrences de la variable linguistique dans ces 19 communautés ont constitué le corpus. De plus, nous avons vérifié qu'il n'y avait pas de problèmes d'identification de langue par Twitter[22] en comparant les fréquences relatives de nos propres identifications du corpus final, les langues que nous avons jugées subjectivement[23], et les identifications de Twitter, arrondies au centième :

Tableau 3.4: Comparaison des fréquences relatives de l'anglais et du français selon Twitter et selon nous

$N = 4\,733$		
Langue	Selon Twitter	Selon nous
anglais	86,25 %	86,69 %
anglais-africain	s. o.	0,04 %
anglais-maori	s. o.	0,04 %
anglais-polynésien	s. o.	0,08 %
français	3,70 %	2,83 %
français-africain	s. o.	0,02 %
français-anglais	s. o.	0,70 %
français-italien	s. o.	0,02 %

Les fréquences relatives de l'anglais que nous avons calculées et que Twitter a calculées sont presque pareilles, mais les fréquences relatives du français diffèrent assez pour que nous devions offrir une explication de cette différence et son

22 Selon un membre du personnel, Twitter ne fournit pas son algorithme d'identification des langues (andypiper, 2016), mais @tm (2015), un de ses développeurs, a décrit la façon dont Twitter évalue lui-même ses algorithmes sur un billet de blog pour ceux qui veulent en savoir davantage.

23 Pour ce qui est des tweets classés comme *africain* et *polynésien*, nous avons réussi à les identifier comme venant de ces régions du monde, mais nous n'avions pas la compétence pour identifier les langues précises.

importance. Nous supposons que la différence vient du manque de précision dans les étiquettes de Twitter, qui identifie tous les tweets contenant même un seul élément d'origine française comme *français*, tandis que nous avons codé des tweets tels que ceux dans les exemples ci-dessus comme *français-anglais*. Le nombre de tweets que nous, nous avons identifiés comme simplement *français* est donc naturellement plus inférieur, mais si nous regroupons tous les tweets que nous avons identifiés comme *français* et une autre langue, nous arrivons à un taux de 3,70 % selon Twitter contre 3,57 % selon nous. De plus, si nous avons manqué quelques tweets d'origine française, ceci ne voudrait pas dire que nous avons également manqué des communautés d'intérêt. Tant que nous pouvions trouver au moins un tweet avec des éléments d'origine française dans une communauté donnée, nous incluions cette communauté-ci. Enfin, si nous avons en fait manqué des communautés d'intérêt, ceci ne rendrait pas du tout invalide notre analyse. Il vaut mieux avoir plus de communautés à analyser, mais ce n'est pas obligatoire.

Nous avons manuellement codé la langue de chaque tweet ainsi que la langue avant l'occurrence de la variable linguistique (lol) et la langue après. Pour ce qui est de la langue des tweets en général, nous avons fini par 10 niveaux. Quand les tweets avaient des éléments d'origine de deux langues ou plusieurs, peu importe le nombre de chacun des éléments, nous avons indiqué toutes les langues dans le codage, en les séparant par des tirets. Par exemple, il y avait deux tweets *anglais-maori*. Les tweets qui n'étaient constitués que d'un seul mot, c'est-à-dire (lol), ont été codés comme *indéfini*. L'anglais était la langue la plus fréquente pour les tweets à un taux de 4 103 sur 4 733, suivi par les tweets indéfinis à 451 et puis le français à 134. Ceux qui contenaient des éléments des deux langues d'intérêt, c'est-à-dire l'anglais et le français, comptaient comme 33 tweets sur 4 733. Enfin, il y avait 10 tweets qui ont montré une combinaison de l'anglais ou du français et d'une autre langue, et

seulement 2 tweets qui ne contenaient ni d'anglais ni de français ont été identifiés, l'un étant codé comme *allemand* et l'autre comme *arabe*.

3.4.3 Catégorie grammaticale

Nous avons identifié quatre valeurs pour ce qui est du codage de la catégorie grammaticale de la variable linguistique de chaque tweet et un seul tweet dont la variable linguistique a été codée comme *indéfini* car elle était d'origine arabe. Les catégories étaient *adjectif*, *adjoint*, *nom* et *verbe*, mais puisque les variantes d'origine française de la variable linguistique ne sont jamais apparues dans les tweets dépourvus d'autres éléments d'origine française, nous ne nous intéressions qu'aux catégories des tweets d'origine française. Ainsi, l'usage de (lol) est restreint à la catégorie *adjoint*, sauf l'occurrence suivante que nous avons codée comme *adjectif* :

Shelly J. : @Shelly J.[24] je suis mdrc quoi ca (Exemple 3.5)

Dans ce cas-ci, il semble que *mdr* soit plutôt un sigle de *morte de rire* et non lexicalisé. Cette sorte d'usage des variantes d'origine anglaise de (lol) se passe également dans les tweets d'origine anglaise, d'où le codage d'un nombre de ces cas-ci comme *adjectif* et *verbe*, mais cela est très rare parmi les usagers des éléments d'origine française dans nos données, donc nous avons laissé tomber la seule occurrence et relégué les analyses aux adjoints, qui constituaient le reste des occurrences.

24 Afin d'improviser un fil de commentaires, ce que Twitter ne permet pas en soi, les utilisateurs répondent à leurs propres tweets. D'autres utilisateurs pourront donc voir la suite de commentaires en ordre.

3.4.4 Variable linguistique (lol)

La variable linguistique (lol) a été strictement codée comme elle a été réalisée par les sujets. À titre d'exemple, dans le cas de Shelly J. ci-dessus, « mdrc » a été codé comme *mdrc*, malgré le fait qu'elle voulait probablement dire « ... mdr. C'est ... ». Ce choix nous a permis d'éviter de présumer la forme des lexèmes, mais il nous a également menés à proposer 139 variantes de (lol). Notre méthode statistique peut en fait traiter une variable avec tant de variantes, mais la majorité des variantes nous semblaient sans polémique des variantes orthographiques. Il est fort probable que « LOL » soit équivalent à « lol », et « mddddr » à « mdr », par exemple. Tagliamonte et Denis (2008) sont implicitement arrivés à la même interprétation dans leur étude de la variation dans la messagerie instantanée et l'ont considérée tellement évidente qu'ils ne l'ont mentionnée que dans une note de fin de document (p. 29). Un traitement de telles variantes serait donc un traitement de la variation orthographique[25] où nous, nous voulions traiter la variation lexicale. Nous avons alors regroupé les variantes que nous avons jugées orthographiques dans une même variante. Les regroupements sont élaborés dans le Tableau 3.5, ci-dessous.

Le codage de la variable linguistique (lol) et des langues, en général, avant et après, a été majoritairement effectué par Shari Ex et vérifié par moi. Tous les tweets codés comme *français* ou une combinaison de *français* et d'une autre langue ont été vérifiés, puisque ce sont les tweets dans lesquels (lol) peut être réalisé avec une variante d'origine française ou anglaise, ainsi que 5 % des autres tweets (214 tweets). Parmi ces tweets, j'ai dû corriger le codage de 10 tweets ou environ 2,6 % des tweets que j'ai vérifiés. Le codage de (lol) et des langues a donc été jugé fiable.

25 En effet, d'après nous, Twitter donne des occasions uniques d'analyser la variation orthographique similaire à la façon dont on analyse la variation phonologique. Pour des études récentes à ce sujet, voir Eisenstein (2015) et Tatman (2016).

Tableau 3.5: Les variantes lexicales de (lol) et toutes leurs variantes orthographiques

Variante lexicale	Variantes orthographiques
lol *laughing out loud* 'éclater de rire'	lol, #lol, #LOL, #Lols4Trolls, #lulz, #RootLol, lawl, lel, lels, Lels, llol, lLol, lOL, Lol, LoL, LOl, LOL, lol@, lolll, LOLLL, lollll, lollllll, lollllllll, lollol, LOLLOL, LOLLOLLOL, LOLLOLOLOLOL, lolno, LOLNOPE, Lolol, LOLOL, lololllol, lolololol, LOLOLOL, lolololo, lolololol, LOLOLOLOL, lolololololol, lolololololol, LOLOLOLOLOLOL, lolololololololol, LOLOLOLOLOLOLOLOLOL, LOLOLOLOLOLOLOLOLOLOLOLOL, LOLOLOOOOOOOOOOOOL, lols, Lols, LOLs, LOLUMAD, lolwut, lolz, LOLZzzz, Lool, Loollll, loool, looool, loooool, LOOOOOL, Looooool, loooooool, looooooool, Looooooool, looooooooooooool, loooooooooooooooooooooool, lul, LUL, lulz, rjlol, trololol, TROLOLOLOLOL, yolololololo
kek[26] 'éclater de rire'	kek, #FreeKekestain, #FreeKekistan, #RealFreeKekestain, Kek, keke, kekekeke, keks, Top kek
lmao *laughing my ass off* 'pété(e) de rire'	lmao, lmaaaaao, lmaaaao, LMAAAAO, lmaaao, lmaO, Lmao, lmao2, Lmaoo, lmaooo, lmaoooo, LMAOOOO, lmaooooo, lmaoooooo, lmaooooooo, LMAOOOOOOOOO, lmaooooooooooooooooooooooooooooo

26 Ce mot n'est pas un sigle. Il est issu du jeu vidéo World of Warcraft, dans lequel les joueurs créent des personnages et choisissent de jouer pour une de deux factions, la Horde ou l'Alliance. Les membres d'une faction ne peuvent parler avec les membres de l'autre. Le jeu remplace les lettres de ce que les membres de l'autre faction écrivent. Donc, quand un membre de la Horde écrit *lol*, les membres de l'Alliance voient ***kek***.

lmfao *laughing my fucking ass off* 'vraiment pété(e) de rire'	lmfao, #lmfao, lmafao, Lmfao, LMFAO, lmfaooo, LMFAOOOOO
rofl *rolling on the floor laughing* 'se rouler par terre de rire'	rofl, ROFL, rooofl
roflmao *rolling on the floor laughing my ass off* 'se rouler par terre et être pété(e) de rire'	ROFLMAO, ROTFLMAO
mdr *mort(e) de rire*	mdr, #MortDeRire, mdddr, mddr, mddrr, mdrr, Mdrr, mdrrr, Mdrrr, MDRRR, mdrrrr, MDRRRR, mdrrrrr, mdrrrrrr, MDRRRRRR, MDRRRRRRRRRRRRRRR
ptdr *pété(e) de rire*	ptdr, Ptdr, PTDR, PTDRRR

3.5 Analyses statistiques

Une fois codées, les données ont été soumises à des analyses statistiques. Dans un grand nombre d'études variationnistes, une analyse statistique implique la mise en application d'une analyse Varbrul, souvent dans le logiciel Goldvarb (D. Sankoff, Tagliamonte, et Smith, 2005) ou Rbrul (Johnson, 2009), qui est un ensemble d'opérations statistiques adaptées à l'étude de la variation linguistique. Malgré sa popularité, une analyse Varbrul traite les occurrences de la variable linguistique comme des épreuves de Bernoulli, c'est-à-dire des applications ou non d'une variante d'une variable qui est forcément binomiale. Puisque (lol) comprend huit variantes, et puisque le choix de l'une de ces variantes comme représentant un succès p, en accord

avec les termes des épreuves de Bernoulli, serait arbitraire, nous avons décidé d'utiliser d'autres statistiques qui s'appliquent spécifiquement aux données catégorielles et multinomiales.

3.5.1 Statistiques descriptives

Nous nous sommes servis du mode pour décrire les centres des distributions et de l'indice de Simpson (1949) pour décrire les dispersions des distributions, ce dernier demandant un peu d'explication comme il n'est pas trop répandu dans le domaine de la sociolinguistique variationniste. Nous avons calculé l'indice de Simpson en utilisant le package *vegan* dans R (Oksanen *et al.*, 2017), son équation étant la suivante :

$$D = \sum_{i=1}^{R} p_i^2 \qquad \text{(Formule 3.1)}$$

L'indice de Simpson *D*, autrement dit l'indice de diversité, est la somme des carrées des fréquences relatives *p* de chaque variante *i* de la variable, qui produit une valeur entre 0,0 et 1,0. Une valeur plus élevée indique une dispersion plus importante et donc une diversité plus importante. Cette statistique est standard dans l'écologie, mais elle apparaît sporadiquement dans la sociolinguistique et souvent pour décrire autre chose, telle que la diversité des langues dans une région (Greenberg, 1956), sur Twitter (Kim *et al.*, 2014) ou pour décrire la diversité ethnique d'un réseau social (Sharma, 2011).

3.5.2 Signification et indépendance

Afin de savoir s'il existe une association entre les facteurs et la variable linguistique (lol), nous avons employé le test exact d'indépendance de Fisher (1922, 1925/1970), calculé dans R avec la fonction *fisher.test* et l'attribut *simulate.p.value* mis à *TRUE*. Ce test d'indépendance est approprié pour n'importe quel tableau de contingence, tandis que le test plus populaire, le test du khi-deux, exige que la majorité des cellules dans le tableau de contingence aient des fréquences théoriques plus supérieures à 5, ce qui se passe peu fréquemment dans les données linguistiques et sûrement pas dans nos données. Enfin, pour savoir la taille d'effet des facteurs, nous avons employé le V de Cramér (1999), calculé dans R avec le package *lsr* (Navarro, 2015).

La question de l'indépendance des occurrences de la variable a été soulevée et traitée par Johnson (2009) lorsqu'il proposait son logiciel Rbrul. Johnson (2009) a remarqué que plusieurs occurrences d'une variable linguistique produites par un même locuteur ne sont pas indépendantes, mais Goldvarb, le logiciel courant à l'époque, les traitait telles quelles (p. 363). Son propre logiciel a mis en œuvre un modèle mixte qui tient compte de qui produit les occurrences de la variable linguistique pour améliorer les résultats statistiques. Puisque nous ne nous sommes pas servis de Rbrul, nous n'avons pas tenu compte de l'indépendance des occurrences. Cependant, notre source de données, Twitter, ne ressemble pas du tout aux sources de données dont Johnson parlait où, souvent, les chercheurs enregistrent de longues discussions avec relativement peu de sujets pour solliciter assez d'occurrences pour effectuer des analyses. Pour nous, la situation était l'inverse : nous avons recueilli peu d'occurrences de chaque sujet, mais nous avons observé beaucoup de sujets. En effet, nous avons codé 4 733 occurrences produites par 1 270 sujets. Le sujet qui a produit le nombre le plus important de (lol) était paultaylor47, qui l'a produit 115 fois, mais

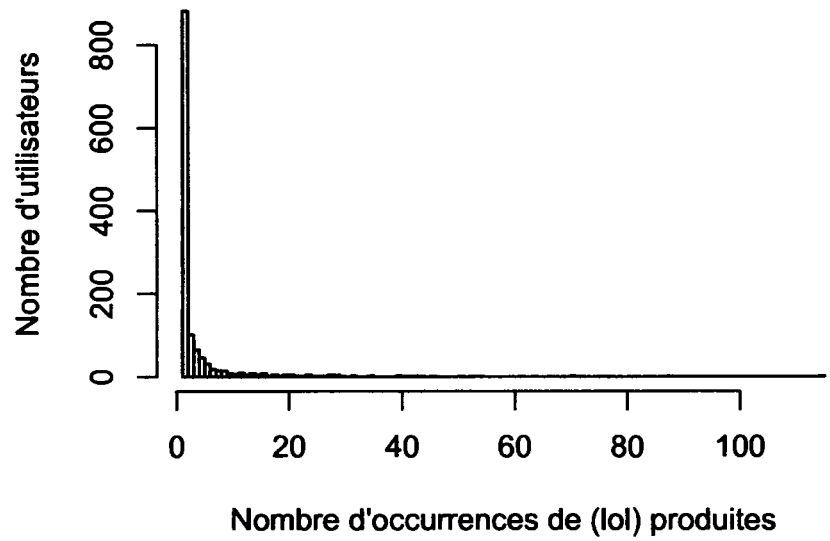

Figure 3.1: Fréquence des nombres d'occurrences de (lol) produits par les sujets

ce nombre était très rare, la plupart des sujets produisant un nombre inférieur de 10 (voir la Figure 3.1). Pour ce qui est des analyses des individus, ce manque de données constituait en fait une faiblesse de l'étude, mais pour ce qui est de l'établissement de l'indépendance des occurrences, il est une force.

3.5.3 Représentativité

Il est enfin à noter que nos données ne sont pas censées être représentatives des provinces maritimes, mais des communautés virtuelles. Nous avons ciblé les

Tableau 3.6: Nombre de tweets venant de chaque pays indiqué par les sujets

$N = 4\ 733$	
Pays	Tweets
Canada	2 324
Écosse	2
États-Unis	39
France	41
Australie-Nouvelle-Zélande	2
Nouvelle-Zélande	2 081
Pays-Bas	1
Royaume-Uni	1
indéfini	242

provinces maritimes non pas afin d'être en mesure de leur généraliser nos résultats, mais afin de nous assurer que nous trouverions des sujets ayant accès à des styles qui ressembleraient à ce que l'on appellerait du français et à ce que l'on appellerait de l'anglais. En effet, en dépit des efforts pour ne cibler que les utilisateurs séjournant ou habitant dans les provinces maritimes, le paramètre *geocode*, saisi dans la recherche sur Netlytic, a conduit à un corpus qui comprend beaucoup de tweets envoyés par des Néo-Zélandais, comme le montre le Tableau 3.6, qui suggère que les géocodes ciblent les tweets qui passent par un serveur dans la région indiquée et donc qui ne sont pas tout à fait équivalents aux tweets envoyés par les utilisateurs qui se trouvent physiquement dans la région indiquée. L'autre possibilité serait que les Néo-Zélandais se soient trouvés dans les données car les Canadiens leur ont envoyé des tweets, mais si nous comparons le degré entrant cumulatif au degré sortant cumulatif des Néo-Zélandais, nous voyons que parmi les 51 781 tweets qui les relient aux communautés d'intérêt, 3 536 étaient entrants et 48 245 étaient sortants. Ainsi, on voit bien que l'inverse est vrai : les Néo-Zélandais ont envoyé beaucoup plus de tweets

aux membres des communautés d'intérêt qu'ils en ont reçus. De toute façon, la présence de ces Néo-Zélandais dans les données n'invalide pas les analyses, car ils interagissent dans les communautés identifiées, mais il faut ne pas confondre ces communautés et les communautés physiques des provinces maritimes.

CHAPITRE IV

RÉSULTATS

4.1 Introduction

Dans ce chapitre, nous présenterons les résultats de notre analyse. Premièrement, nous donnerons des renseignements par rapport au caractère des communautés en général, suivis des analyses de l'association entre les communautés et la réalisation de (lol), puis des analyses de l'indice de diversité de Simpson des individus et leurs centralités et enfin des conclusions.

4.2 Caractère des communautés dans leurs ensembles

Il est premièrement important de présenter des caractéristiques générales des 19 communautés que nous avons analysées. Gephi a donné un numéro à chaque communauté, puis nous avons déterminé leurs tailles et leurs densités. Le Tableau 4.1 montre ces valeurs, ainsi que le mode et l'indice de Simpson, autrement dit la diversité *D,* pour tous les tweets des communautés, peu importe leurs langues ou leurs catégories grammaticales. *D* est une mesure de dispersion qui peut être considérée comme quantifiant la variation[27].

---

27 Pour plus de détails, voir la section 3.5.1 Statistiques descriptives.

Tableau 4.1: Caractéristiques générales des 19 communautés d'intérêt

Communauté	Points	Liens	Densité[28]	Mode	*D*[29]
173	2 480	2 846	0,000	lol	0,350
302	17 279	28 884	0,000	lol	0,272
322	19	18	0,053	mdr	0,000
572	3 601	4 152	0,000	lol	0,179
756	980	1 130	0,001	lol	0,694
799	33	32	0,030	mdr	0,180
1032	22 531	31 559	0,000	lol	0,188
1097	2 955	3 697	0,000	lol	0,152
1227	2 214	2 432	0,000	lol	0,153
1291	1 073	1 179	0,001	lol	0,321
1340	33	38	0,036	mdr	0,000
1782	2	1	0,500	ptdr	0,000
1917	4 432	5 849	0,000	lol	0,481
2067	44	44	0,023	mdr	0,616
2265	242	256	0,004	lol	0,440
2305	4	3	0,250	mdr	0,000
6445	2	1	0,500	mdr	0,000
6744	12	11	0,083	mdr	0,000
6817	592	641	0,002	lol	0,245

Ce qui est premièrement à noter, c'est que la taille des communautés va de 2 individus (la communauté 6445) à 22 531 individus (la communauté 1032). Quelques communautés sont si petites qu'il n'est pas possible d'en effectuer une analyse significative, comme les communautés 1782 et 6445, qui représentent plutôt des dyades. Ce qui est deuxièmement à noter, c'est que les ratios points-liens des

28 Calculée comme un graphe orienté.
29 Arrondie au millième.

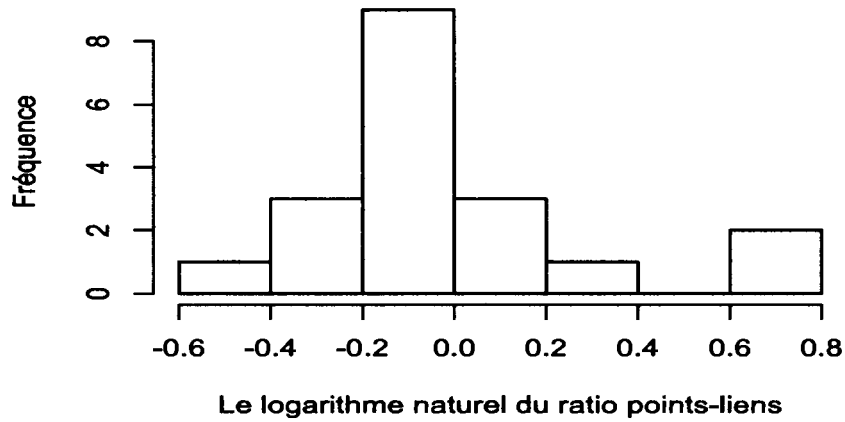

Figure 4.1: Fréquence des ratios points-liens des communautés

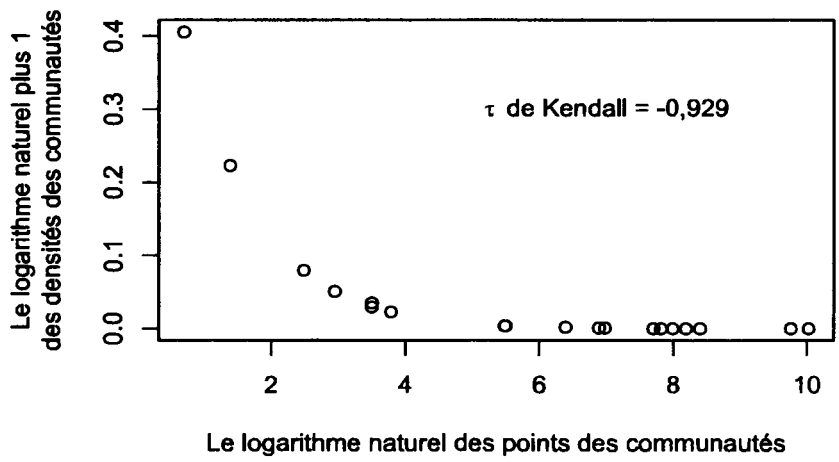

Figure 4.2: Nuage de dispersion des points et les densités

communautés sont presque d'un pour un, comme le montre la Figure 4.1. De ce fait, plus de points il y a dans une communauté, moins il y a de densité dans la communauté. C'est bien ce que suggère la relation monotone évidente dans la Figure 4.2.

Il existe également des patrons intéressants à remarquer dans le Tableau 4.1 en termes de mode et diversité de la variable linguistique (lol). Premièrement, les seules communautés dont le mode est une variante d'origine française sont les plus petites. La communauté la plus grande qui a une variante d'origine française comme le mode est la communauté 2067, constituée de 44 individus. C'est également la seule communauté avec un tel mode dont la diversité est relativement élevée à 0,616. Les autres communautés avec un mode d'origine française ne sont pas diverses du tout, six d'entre elles n'ayant aucune diversité, c'est-à-dire que D est 0,000. En ce qui concerne les communautés les plus grandes, et au vu du manque d'usage des variantes d'origine française dans les tweets dépourvus d'autres éléments d'origine française, nous pouvons supposer soit que (lol) est réalisé comme *lol* dans les tweets constitués d'éléments d'origine française dans ces communautés-ci ou que les usagers des éléments d'origine française sont peu nombreux dans ces communautés.

4.3 Signification des communautés

La première question de recherche que nous avons posée visait à déterminer si une association existe entre les communautés détectées par la méthode de Louvain et la réalisation de la variable linguistique (lol). La réponse brève est oui, mais ce qui complique la question, c'est que tous les facteurs ont eu un effet statistiquement significatif, comme le montre le Tableau 4.2.

Tableau 4.2: Signification et taille d'effet de tous les facteurs sur (lol) pour toutes les données

Facteur	Valeur de P[30]	Taille d'effet[31]	N[32]
ville	< 0,0005	0,4179309	1 501
communauté	< 0,0005	0,3650208	4 733
langue	< 0,0005	0,3550906	4 282
catégorie grammaticale	< 0,0005	0,2513084	4 732
province	< 0,0005	0,2205922	2 385
fuseau horaire	< 0,0005	0,1973552	3 357
pays	< 0,0005	0,1679545	4 491

Nous croyons que la taille de l'échantillon a entraîné ces valeurs de P significatives. Par exemple, nous avons également trouvé qu'il y avait une association significative pour deux autres facteurs que nous n'avons pas inclus sur le tableau : le codeur et la source. Le codeur est constitué de la personne qui a codé la valeur du tweet, et la réalisation de la variable linguistique (lol), mais il y a peu de raisons externes de croire que le codeur contrôle la réalisation de (lol), car la réalisation a déjà été produite par le sujet bien avant le codage. La source, autrement dit l'appareil à partir duquel l'utilisateur a envoyé le tweet, tel qu'un ordinateur ou un téléphone, a aussi présenté une valeur de P significative, mais on ne s'attend pas à ce que ce soit un facteur qui contrôle la réalisation de (lol). Si même ces facteurs ont fini par être significatifs, nous croyons que n'importe quel facteur serait significatif dans notre échantillon.

Ce que le codeur et la source ne présentent pas, cependant, ce sont des tailles d'effet importantes. C'est cette mesure qui nous intéresse finalement, et pour le codeur et la

30 À partir du test exact de Fisher.
31 À partir du V de Cramér.
32 Sans les tweets dont la valeur de la variable est indéfinie.

source, leurs tailles d'effet n'étaient qu'environ 0,11 chacune, ce qui est vraiment petit en comparaison avec les autres facteurs, comme montre le Tableau 4.2. Les tailles d'effet de la ville, de la communauté et de la langue, par exemple, étaient beaucoup plus élevées à environ 0,42, 0,37 et 0,36 respectivement.

Plusieurs variables qui ciblent des régions géographiques se chevauchent naturellement. Le fuseau horaire et le pays ont donc des tailles d'effet similaires, et la province et la ville ont des tailles d'effets même plus importantes, alors la région géographique semble jouer un rôle dans la réalisation de la variable linguistique (lol), mais il se peut qu'elle aille de pair avec la communauté en ce que ceux qui habitent physiquement ensemble finissent peut-être par se regrouper en ligne. De même, la langue témoigne d'une taille d'effet importante, mais il se peut que la langue aille également de pair avec la communauté. En effet, le Tableau 4.1 suggère déjà que les petites communautés sont dominées par les sujets qui préfèrent les éléments d'origine française, puisqu'elles ont toutes une variante d'origine française de (lol) comme mode, laquelle n'apparaît jamais dans les tweets dépourvus d'autres éléments d'origine française.

4.3.1 Communautés et langues

Quant aux éléments d'origine anglaise, si nous nous restreignons aux tweets que nous avons codés comme *anglais* ou une combinaison d'*anglais* et d'une autre langue qui n'est pas le français, effectivement tous les tweets dépourvus d'éléments d'origine française, nous trouvons qu'il y a peu de diversité dans ces données, une valeur de D d'environ 0,25 pour être précis, le mode étant *lol*. Cependant, la diversité des tweets codés comme *français* ou une combinaison de *français* et d'une autre langue est de

0,49, le mode étant *mdr*. De plus, comme nous l'avons déjà constaté, les tweets dépourvus d'autres éléments d'origine française ne permettent jamais des variantes d'origine française de (lol) (voir la section 3.4.2 Communautés et langues), mais c'est exactement cette sorte de mélange d'éléments couramment décrits comme provenant de différentes langues qui nous intéresse dans ce mémoire. Pour la plupart, si nous divisons les tweets dépourvus d'autres éléments d'origine française en communautés, la diversité diminue même davantage, comme le montre le Tableau 4.3, où seules trois communautés sur douze présentent une augmentation notable de diversité, les

Tableau 4.3: Mode et diversité des communautés en considérant seulement les tweets dépourvus d'autres éléments d'origine française

	$N = 4\ 111$		
Communauté	Mode	D[33]	N
173	lol	0,334	224
302	lol	0,270	1 272
572	lol	0,172	441
756	lol	0,268	33
799	lol/mdr[34]	0,500	2
1032	lol	0,184	1 248
1097	lol	0,150	172
1227	lol	0,136	235
1291	lol	0,191	103
1917	lol	0,458	344
2265	lol	0,354	32
6817	lol	0,000	5

33 Arrondie au millième.

34 Une occurrence de chacun, mais si le tweet contenant *mdr* est vraiment dépourvu d'autres éléments d'origine française est discutable (voir la section 3.4.2 Communautés et langues).

communautés 799, 1917 et 2265. On voit donc que les tweets dépourvus d'autres éléments d'origine française ne sont pas vraiment d'intérêt car ils ne présentent pas vraiment de variation à n'importe quel niveau. Le mode est toujours *lol* et la diversité commence déjà à un bas niveau et diminue à mesure que nous regardons toutes les communautés séparément. Nous pouvons donc restreindre l'analyse au sous-ensemble de tweets contenant des éléments d'origine française.

Pour nous assurer encore davantage que la variation qui se trouve dans les tweets dépourvus d'autres éléments d'origine française n'est pas aussi intéressante que la variation dans les tweets d'origine française, nous pouvons comparer séparément les tailles d'effet des deux catégories par rapport aux associations entre les communautés et les réalisations de (lol). Pour ce qui est des tweets dépourvus d'autres éléments d'origine française, la valeur de P est toujours significative ($P < 0{,}0005$), mais la taille d'effet diminue un peu, passant d'environ 0,365 à environ 0,316 pour tous les tweets, probablement parce que le mode est toujours *lol* et la diversité toujours

Tableau 4.4: Signification et taille d'effet de tous les facteurs sur (lol) pour les tweets d'origine française dans lesquels (lol) est un adjoint

Facteur	Valeur de P[35]	Taille d'effet[36]	N[37]
communauté	< 0,0005	0,6354218	168
ville	< 0,0005	0,5940664	113
fuseau horaire	< 0,0005	0,4926027	121
province	< 0,0005	0,3623589	132
pays	< 0,0355	0,1747749	155

35 À partir du test exact de Fisher.
36 À partir du V de Cramér.
37 Sans les tweets dont la valeur est indéfinie.

mineure, tandis que la taille d'effet pour ce qui est des tweets d'origine française augmente considérablement à environ 0,632. De même, la taille d'effet des tweets d'origine française par rapport aux communautés est d'environ 0,635 lorsque nous effaçons les tweets dans lesquels (lol) n'est pas un adjoint, afin de nous conformer au fait que (lol) se restreint à cette catégorie grammaticale dans les tweets contenant des éléments d'origine française (voir la section 3.4.3 Catégorie grammaticale). Ce résultat n'est pas étonnant puisque *N* ne passe que de 169 occurrences à 168 occurrences. Le Tableau 4.4 montre les valeurs de *P* et les tailles d'effet des associations entre (lol) et les communautés ainsi que les autres facteurs pour les tweets contenant des éléments d'origine française dans lesquels (lol) est un adjoint.

4.3.2 Communautés et régions géographiques

Nous voyons dans le Tableau 4.4 ci-dessus que l'association entre la communauté et (lol) témoigne de la taille d'effet la plus importante parmi les facteurs que nous avons analysés, mais il y en a d'autres qui impliquent des régions géographiques et dont les tailles d'effet sont aussi considérables, la ville et le fuseau horaire en particulier. La communauté est donc sans doute associée à la réalisation de (lol) dans nos données, mais il faut déterminer si les communautés sur Twitter sont associées aux régions géographiques. Une façon de le faire est de prendre un sous-ensemble de tweets d'une région et de regarder sa distribution à travers les communautés. Si nous arrivons à une distribution uniforme, c'est-à-dire que les occurrences sont uniformément réparties parmi les communautés, nous aurons des preuves que ceux dans la région donnée tendent à ne pas se regrouper sur Twitter. Si nous arrivons à une distribution non-uniforme, et/ou si le nombre de communautés diminue, nous aurons des preuves de l'inverse.

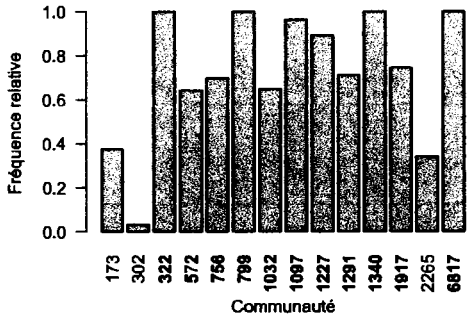

Figure 4.3: Distribution de ceux au Canada à travers les communautés sur Twitter

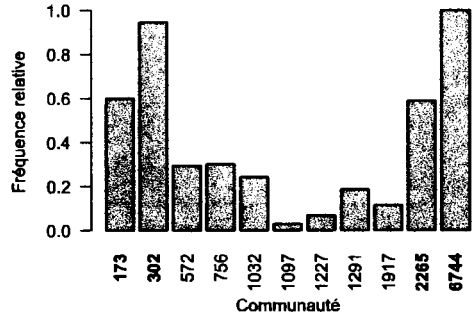

Figure 4.4: Distribution de ceux en Nouvelle-Zélande à travers les communautés sur Twitter

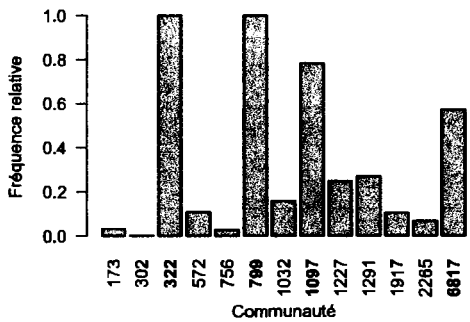

Figure 4.5: Distribution de ceux au Nouveau-Brunswick à travers les communautés sur Twitter

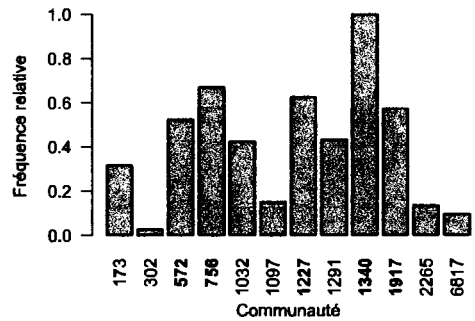

Figure 4.6: Distribution de ceux en Nouvelle-Écosse à travers les communautés sur Twitter

Selon ces critères, ceux au Canada, en Nouvelle-Zélande, au Nouveau-Brunswick et en Nouvelle-Écosse[38] tendent à se regrouper sur Twitter au moins un peu. Les

38 Nous avons choisi ces quatre régions car elles présentent les plus grands nombres d'occurrences dans nos données.

distributions des fréquences relatives des sujets dans ces régions à travers les communautés détectées sur Twitter ne sont pas trop uniformes, comme le montrent les Figures 4.3, 4.4, 4.5 et 4.6 ci-dessus. La distribution pour ceux au Canada (Figure 4.3) s'approche le plus clairement d'une distribution uniforme, mais seulement dans le sens où ceux au Canada font plus de 50 % des personnes dans la majorité des communautés où ils se trouvent, ce qui est indiqué par les communautés en gras, mais la différence entre une fréquence relative de 64 % dans la communauté 572 et de 96 % dans la communauté 1097 est toujours grande et non uniforme. Finalement, le nombre de communautés diminue dans tous les cas, de 19 à 14 pour le Canada, de 19 à 11 pour la Nouvelle-Zélande, de 19 à 13 pour le Nouveau-Brunswick et de 19 à 12 pour la Nouvelle-Écosse, qui soutient l'idée que les utilisateurs de Twitter dans nos données se regroupent dans des communautés dans une certaine mesure selon leurs régions géographiques.

Pour nous assurer encore davantage qu'il y a des raisons de croire que ceux dans une même région physique se regroupent sur Twitter dans nos données, nous pouvons

Tableau 4.5: Signification et taille d'effet des facteurs géographiques sur les communautés pour tous les tweets

Facteur	Valeur de P[39]	Taille d'effet[40]	N[41]
ville	< 0,0005	0,6332123	1 501
pays	< 0,0005	0,4379973	4 491
province	< 0,0005	0,4035079	2 385
fuseau horaire	< 0,0005	0,2806191	3 357

39 À partir du test exact de Fisher.
40 À partir du V de Cramér.
41 Sans les tweets dont la valeur est indéfinie.

employer le test exact de Fisher et le V de Cramér pour savoir si une association statistiquement significative existe entre les communautés et les régions. Le Tableau 4.5 témoigne bien de cette association. En général, nous voyons que la taille d'effet augmente à mesure que nous nous focalisons sur des localisations de plus en plus précises. Les habitants d'une même ville se regroupent donc le plus clairement sur Twitter dans nos données, et ceux qui résident dans le même fuseau horaire le moins clairement.

Ces tests ne sont pas aussi robustes que nous le voudrions, mais ils servent nos objectifs, faute d'analyses qui dépasseraient le cadre de ce mémoire. Nous ne sommes pas en mesure de dire si les gens dans une autre région géographique, hormis ceux que nous avons identifiés, se regroupent sur Twitter, par exemple. Encore plus important, nous n'avons codé que les tweets contenant la variable linguistique (lol), donc nous ne pouvons effectuer le test pour tous les membres des 19 communautés que nous étudions, qui constitueraient 58 528 personnes. De même, les cas représentent les tweets, et non les sujets. Parfois, une communauté est constituée d'un petit nombre de personnes qui ont produit un grand nombre d'occurrences de (lol). Par exemple, il y a 76 occurrences de (lol) dans la communauté 756, mais Shelly J. et Bobby B. en ont produit 21 et 16, respectivement. Cette situation ne se soulève pas souvent (voir la Figure 3.1 dans la section 3.5.2 Signification et indépendance), donc nous pouvons avoir confiance en les tests en ce qui nous concerne ici.

Nous pouvons finalement conclure qu'il existe une association statistiquement significative entre les communautés détectées sur Twitter par la méthode de Louvain et la réalisation de la variable linguistique (lol), surtout pour le sous-ensemble des tweets qui nous intéressent le plus : les tweets constitués d'éléments d'origine

Tableau 4.6: Caractéristiques des 19 communautés en considérant seulement les occurrences de (lol) qui sont des adjoints dans les tweets contenant des éléments d'origine française

	$N = 168$		
communauté	mode	D[42]	N
173	mdr/ptdr	0,500	2
302	lol	0,000	1
322	mdr	0,000	1
572	mdr	0,000	1
756	mdr	0,528	34
799	mdr	0,157	35
1032	lol	0,000	2
1097	lol	0,000	4
1227	lol	0,000	1
1291	mdr	0,000	11
1340	mdr	0,000	23
1782	ptdr	0,000	1
1917	mdr	0,278	6
2067	mdr	0,631	27
2265	mdr	0,560	5
2305	mdr	0,000	1
6445	mdr	0,000	1
6744	mdr	0,000	3
6817	lol	0,346	9

française dans lesquels (lol) est un adjoint. Le Tableau 4.6 présente un aperçu des caractéristiques de chaque communauté pour ce sous-ensemble-ci. Contrairement aux caractéristiques pour le sous-ensemble des tweets dépourvus d'autres éléments

42 Arrondie au millième.

d'origine française, le mode varie de communauté en communauté. Une autre différence entre les deux sous-ensembles, c'est qu'il y a peu d'occurrences de (lol) dans chaque communauté. En effet, sept des communautés n'en ont qu'une occurrence. Pour cette raison, il vaut peut-être mieux analyser les individus dans ce sous-ensemble séparément en comparaison avec leurs communautés en général.

4.4 Comparaisons entre les individus et leurs communautés

Les tweets qui contiennent des éléments d'origine française dans nos données sont peu nombreux en comparaison avec les tweets qui ne contiennent aucun élément d'origine française, 168 occurrences de (lol) comme un adjoint contre 4 033, mais un patron potentiel émerge si l'on analyse séparément les individus qui ont produit ces 168 occurrences compte tenu de leurs centralités dans leurs communautés. Le Tableau 4.7 montre que les sujets dont la production de (lol) présente plus de diversité que celle de leurs communautés ne sont pas trop centraux dans leurs communautés selon leurs PageRanks, ainsi que l'inverse.

Nous avons écarté les individus qui n'ont produit qu'une occurrence de (lol) car il n'est pas possible d'avoir de la diversité si le sujet n'en a produit qu'une occurrence. Nous avons également écarté quelques sujets qui ont produit toutes les occurrences de (lol) dans leurs communautés, ce qui empêche une comparaison entre eux et leurs communautés. Les individus écartés sont les suivants :

- Garland B., Laura P., Donna H., Audrey H., Norma J., Lucy M., Dr J., Jocelyn P., Hank J., Log L., Catherine M.

Tableau 4.7: Comparaison entre la diversité de la réalisation de (lol) pour des individus et leurs communautés

Individu			Communauté	
Nom	*PR* centile[43]	*D*[44]	Numéro	*D*[45]
Plus de diversité que sa communauté				
Leland P.	0,07	0,500	2265	0,448
Dale C.	0,09	0,667	173	0,291
Leo J.	0,52	0,480	2265	0,448
Andy B.	0,82	0,500	6817	0,245
Harry S.	0,98	0,338	302	0,276
Moins de diversité que sa communauté				
James H.	0,11	0,000	1291	0,321
Ed H.	0,25	0,408	1917	0,481
Ben H.	0,41	0,000	1097	0,154
Bobby B.	0,48	0,000	756	0,692
Dr H.	0,58	0,000	1097	0,154
Shelly J.	0,74	0,095	756	0,692
Pete M.	0,84	0,604	2067	0,616

Nous hésitons, bien sûr, à affirmer qu'un vrai patron existe dans les données présentées dans le Tableau 4.7 en raison du manque de données pour chaque individu, du manque d'individus adéquats pour l'analyse et des individus qui ne suivent pas le patron, mais nous croyons que nos résultats dans ce domaine suggèrent que l'on tirerait profit de l'exploration supplémentaire. Nous allons présenter ci-dessous des analyses plus approfondies des individus qui suivent le patron, suivies des individus qui ne le suivent pas.

43 Arrondi au centième.
44 Arrondie au millième.
45 Arrondie au millième.

4.4.1 Les individus qui suivent le patron

Parmi les individus qui suivent le patron, c'est-à-dire qu'ils présentent plus de diversité dans leurs usages de (lol) que leurs communautés s'ils ne sont pas centraux ou moins s'ils sont centraux, deux correspondent à la première catégorie et peut-être cinq à la dernière catégorie. Ce sont les sujets Leland P. et Dale C. et puis Ben H., Bobby B., Dr H., Shelly J. et Pete M., respectivement. Leurs modes et les modes de leurs communautés sont présentés dans le Tableau 4.8.

Les modes des sujets dans le Tableau 4.8 sont généralement ceux que nous avons prévus. Les deux individus qui présentent plus de diversité dans leurs réalisations de (lol) que leurs communautés n'ont pas vraiment de mode. Pour ce qui est des autres

Tableau 4.8: Modes des individus qui suivent le patron où elles et ils présentent plus de diversité que leurs communautés s'ils ne sont pas centraux dans leurs communautés et moins s'ils le sont

Individu		Communauté	
Nom	Mode	Numéro	Mode
Plus de diversité que sa communauté			
Leland P.	mdr/ptdr	2265	lol
Dale C.	lol/mdr/ptdr	173	lol
Moins de diversité que sa communauté			
Ben H.	lol	1097	lol
Bobby B.	ptdr	756	lol
Dr H.	lol	1097	lol
Shelly J.	mdr	756	lol
Pete M.	mdr	2067	mdr

sujets, Ben H., Dr H. et Pete M. ont des modes qui correspondent aux modes de leurs communautés, ce qui est un résultat que nous avons également prévu. Si un individu est bien intégré dans sa communauté, son comportement linguistique s'approche de sa communauté. Shelly J. et Bobby B. contredisent ce patron-ci : ils sont centraux dans leur communauté, mais leurs modes ne correspondent pas au mode de la communauté et ne correspondent même pas l'un à l'autre. La raison qu'ils produisent (lol) de cette façon n'est pas tout à fait claire, mais peut-être que leur proportion d'occurrences dans la communauté l'explique un peu. Les deux font ensemble 36 des occurrences sur 75, beaucoup plus que les autres membres de la communauté individuellement. Ils ne sont donc pas submergés de *lol* comme le sont les membres de presque toutes les communautés. Une autre possibilité est que ces deux individus soient équivalents aux icônes sociolinguistiques que Eckert (2000, p. 216-219) a décrit en ce que les deux individus sont centraux et ont confiance en leurs usages de (lol) dans le contexte de cette communauté, mais ils visent à se présenter comme innovateurs. Certes, ces explications ne sont finalement que provisoires, et des explications fiables dépassent le cadre de ce mémoire.

4.4.2 Les individus qui ne suivent pas le patron

Cinq individus sont des exceptions au patron et donc demandent d'être davantage examinés. Trois sont centraux dans leurs communautés mais présentent plus de diversité dans leurs réalisations de (lol) que leurs communautés, soit Leo J., Andy B. et Harry S., et deux ne sont pas centraux mais présentent moins de diversité que leurs communautés, soit James H. et Ed H. Nous pouvons examiner leurs autres mesures de centralité, présentées dans le Tableau 4.9, afin de chercher des raisons potentielles

pour lesquelles ces sujets ne suivent pas le patron, mais ces données soulèvent plus de questions qu'elles n'apportent de réponses.

L'intermédiarité ne semble pas expliquer les exceptions au patron, surtout parce que peu de sujets en général n'ont une intermédiarité, c'est-à-dire qu'ils ne se trouvent pas sur n'importe quelle géodésique. Parmi les exceptions, Andy B. et Harry S. ont des intermédiarités, les centiles étant de 0,99 et 0,98, respectivement. On pourrait

Tableau 4.9: Mesures de centralités à part du PageRank

Individu	Intermédiarité centile	Degré sortant et centile	Degré sortant pondéré et centile	Degré entrant et centile	Degré entrant pondéré et centile
Exceptions qui présentent plus de diversité que leurs communautés					
Leo J.	s. o.	20 (0,99)	75 (0,99)	1 (0,90)	3 (0,70)
Andy B.	0,99	66 (1,00)	124 (1,00)	1 (0,90)	1 (0,60)
Harry S.	0,98	2 (0,91)	63 (0,99)	1 (0,80)	7 (0,90)
Ceux qui suivent le patron					
Leland P.	s. o.	20 (0,99)	75 (0,99)	0	0
Dale C.	s. o.	14 (0,98)	45 (0,99)	0	0
Ben H.	0,98	10 (0,97)	15 (0,97)	1 (0,86)	1 (0,68)
Bobby B.	s. o.	206 (1,00)	1389 (1,00)	1 (0,85)	54 (1,00)
Dr H.	0,99	15 (0,99)	32 (0,99)	2 (0,94)	2 (0,82)
Shelly J.	s. o.	42 (1,00)	232 (1,00)	1 (0,85)	45 (0,99)
Pete M.	s. o.	40 (1,00)	190 (1,00)	1 (0,98)	11 (0,93)
Exceptions qui présentent moins de diversité que leurs communautés					
James H.	s. o.	49 (1,00)	327 (1,00)	0	0
Ed H.	s. o.	20 (0,99)	90 (1,00)	1 (0,79)	2 (0,76)

supposer que de telles intermédiarités élevées indiquent que ces individus interagissent fréquemment avec plus d'une communauté, et que cette interaction confond leurs réalisations de (lol), mais Ben H. et Dr H. ont également des intermédiarités élevées et suivent bien le patron. De plus, ceux dans ces deux groupes sont centraux selon leurs PageRanks, ce qui ne devrait pas être le cas selon l'argument de Granovetter (1973). Ils devraient se trouver aux périphéries de leurs communautés. Une analyse de l'intermédiarité ne nous donne donc ni de réponse à la question du comportement des exceptions au patron ni de confirmation de l'argument de Granovetter.

Une autre possibilité serait que les exceptions au patron qui sont centrales dans leurs communautés le soient centrales en raison de l'existence d'une connexion entrante d'un membre important dans la communauté. Le PageRank est finalement fonction des liens entrants et la centralité des points auxquels ces liens-ci se relient. Si un individu a alors un degré entrant de 1, par exemple, mais que cette connexion vient d'un individu avec un PageRank très élevé, ce premier aura également un PageRank élevé, même si leur interaction dans la communauté est rarement réciproque. Les trois sujets qui sont centraux mais qui présentent plus de diversité dans leurs usages de (lol) que leurs communautés ont aussi des degrés entrants de 1, donc il se peut qu'ils n'interagissent pas suffisamment avec les communautés en général pour apprendre les normes des communautés mais paraissent centraux à cause d'une seule relation importante.

En effet, les connexions entrantes d'Andy B. et Harry S. viennent de Windom E. et Maddy F., respectivement, qui ont des PageRank centiles de 1,00 et 0,99, respectivement. Leo J. est un peu différent : sa seule connexion entrante vient de Leo J., ce qui veut dire que sa seule connexion entrante est en fait une boucle, ou un lien

qui relie Leo J. à lui-même, et les boucles n'affectent pas les PageRanks. Leo J. semble donc contredire cette explication, mais les autres sujets qui suivent le patron la contredisent également. Par exemple, Ben H. est aussi central dans sa communauté selon le PageRank, et il suit le patron en ce qu'il présente beaucoup moins de diversité dans son usage de (lol) que sa communauté, mais son degré entrant n'est que 1, exactement comme Andy B. et Harry S., qui ne suivent pas le patron. Les degrés entrants ne semblent donc pas bien expliquer les exceptions qui montrent plus de diversité que leurs communautés.

En ce qui concerne les exceptions qui montrent moins de diversité que leurs communautés malgré leur manque d'intégration dans leurs communautés, les degrés entrants pourraient expliquer un individu, James H., mais non l'autre. James H. a un degré entrant de 0 et est également le seul membre de la communauté dont les tweets qui contiennent la variable linguistique (lol) contiennent aussi des éléments d'origine française. Ces faits pourraient suggérer que James H. est spammeur. Ce sujet n'a ni l'intention d'être accepté dans la communauté ni l'intention de se faire comprendre. Il n'est pas possible de prouver cette explication dans cette étude, cependant. On aurait besoin de plus de données et de focaliser davantage sur la réalisation des études de cas.

Similairement, l'explication de James H. pourrait s'appliquer à Ed H., mais nous n'avons pas de moyens de formuler une explication fiable à partir de nos données. Par exemple, nous avons jugé Ed H. non central selon son PageRank centile de 0,25, mais ce centile se trouve dans une position unique. Les autres individus jugés non centraux ont des PageRanks centiles de 0,11 ou moins et ceux jugés centraux de 0,41 ou plus. Nous avons choisi ces seuils principalement en raison du manque d'individus entre les deux valeurs, mais le choix était par ailleurs arbitraire. Nous pourrions également

dire qu'Ed H. est bien intégré dans sa communauté et que c'est pour cette raison qu'il présente moins de diversité que sa communauté. Une explication fiable nous élude donc toujours.

4.5 Conclusion

Ce qui rend notre analyse des exceptions au patron difficile est notre manque de données, mais ce manque est également ce qui rend impossible le fait d'affirmer avec certitude qu'un patron n'existe pas entre la diversité de la réalisation de (lol) des individus et leurs centralités dans leurs communautés. Les mesures de centralité dont nous nous sommes servis nous permettent de mettre de l'avant des explications possibles, mais une analyse plus concentrée sur quelques individus d'intérêt serait nécessaire pour avancer une réponse fiable. Par exemple, nous pourrions mieux connaître l'intention de James H., si ce sujet est un spammeur ou non, à partir d'une analyse du contenu de ses tweets, des autres communautés auxquelles ce sujet participe, peut-être d'autres variables linguistiques et ainsi de suite. De même, afin de savoir si les exceptions sont vraiment des exceptions, il faut plus d'individus appropriés. Comme le montre la Figure 4.7, il est difficile d'être certain que les exceptions représentent des observations aberrantes parce qu'il n'y a pas assez d'observations pour qu'un patron soit évident hors de doute. Il convient de répéter que le patron que nous avons identifié est potentiel. En effet, nous ne sommes pas en mesure de répondre à notre deuxième question de recherche faute de données, mais nous croyons que nos résultats à ce propos indiquent que l'on peut toujours tirer des bénéfices d'études supplémentaires dans ce domaine.

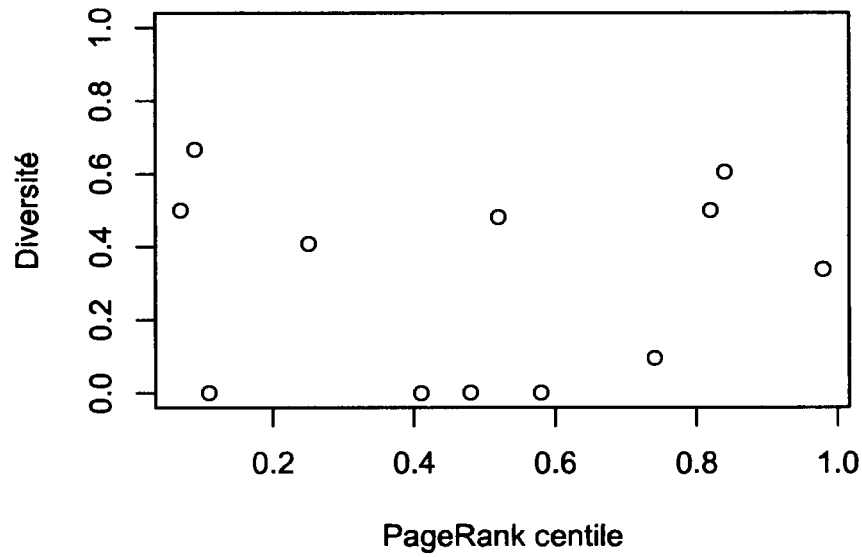

Figure 4.7: PageRank centile contre diversité de la réalisation de (lol) d'individus

Nous pourrions tirer de plus fortes conclusions sur la question des associations entre les communautés et la réalisation de (lol) si nous avions plus de données, aussi. Nos résultats à ce propos ici correspondent à ce que Milroy (1980/1987) a découvert et plusieurs études qui ont suivi qui ont mis en œuvre l'analyse des réseaux sociaux (p. ex. Auer, 1988/2000; Li *et al.*, 1992/2000; Sharma, 2011). Ils ont découvert que les communautés, autrement dit les agglomérats, servent de mécanismes d'application des normes linguistiques, comme le suggèrent nos résultats, même s'ils n'ont pas utilisé la méthode de Louvain pour détecter leurs communautés. Cette correspondance nous permet d'affirmer avec plus de confiance que nos résultats par rapport aux associations entre les communautés et la variable linguistique (lol) sont

fiables, mais effectuer plus de recherche sur Twitter et sur l'applicabilité de différentes méthodes de détection des communautés et de différents genres de liens en ce qui concerne les questions sociolinguistiques s'avérerait pertinent.

CHAPITRE V

DISCUSSION

Malgré le fait que nous pourrions mieux répondre à nos questions de recherche si nous disposions de plus de données, il reste plusieurs répercussions à noter à partir de ce que nous avons fait ici. Principalement, nous pouvons dire qu'il y a une association potentielle entre la centralité des individus et la diversité de leurs réalisations de (lol) et qu'il y a une association statistiquement significative entre les communautés et la réalisation de (lol). Là où nos données ne nous permettent pas de tirer des conclusions satisfaisantes, nous pouvons suggérer des améliorations à la collecte de données. Enfin, nos résultats témoignent de l'utilité de l'analyse des réseaux sociaux pour effectuer des analyses sociolinguistiques variationnistes et suggèrent des pistes de recherche à explorer davantage.

5.1 Améliorer la collecte de données

Ce qui aurait rendu notre analyse plus fiable aurait été un plus grand nombre d'occurrences de la variable linguistique (lol). D'autres études sociolinguistiques de Twitter ont tendance à examiner des vocabulaires entiers (p. ex. Bamman *et al.*, 2014; Pavalanathan et Eisenstein, 2015) ou l'écologie des langues sur Twitter (p. ex. Kim *et al.*, 2014) au lieu d'examiner en détail d'une à dix variables linguistiques comme nous avons essayé de le faire pour être conformes aux études variationnistes typiques (p. ex. Eckert, 2000; Labov, 1969, 1966/2006; Milroy, 1980/1987; Nadasdi *et al.*, 2004; Sharma, 2011). À cet égard, notre étude est assez unique, et nos commentaires

sur la façon d'améliorer la collecte de données sont donc instructifs. Nous proposons deux améliorations principales.

Premièrement, nous avons ciblé les provinces maritimes dans la collecte de données, mais d'autres régions pourraient être plus intéressantes, surtout puisque (lol) ne semble pas varier dans les tweets dépourvus d'autres éléments d'origine française (voir les sections 3.4.2 Communautés et langues 4.3.1 Communautés et langues). La collecte initiale de notre étude, où nous avons recueilli des tweets jugés français par Twitter afin d'identifier une bonne variable linguistique, a retrouvé 12 905 tweets, mais une recherche identique ciblant Montréal que nous avons démarrée durant la même période a retrouvé 299 400 tweets. Le mélange de langues est peut-être plus stéréotypé dans les provinces maritimes, mais il n'est pas du tout inédit à Montréal (p. ex. Friesner, 2009). Même une autre recherche identique qui a ciblé la Louisiane aux États-Unis, une région où un mélange de langues est bien constaté (p. ex. Blyth, 1997; Dajko et Carmichael, 2014; Klingler, Picone, et Valdman, 1997), a produit 86 804 tweets. En effet, un survol des tweets témoigne que *lol* et *mdr* se retrouvent tous les deux dans les tweets qui contiennent d'autres éléments d'origine française dans les tweets émanant de Montréal et de Louisiane. Prendre les données d'une autre région pourrait donc être plus efficace, même si l'on veut étudier la même variable linguistique lexicale que nous avons étudiée. On peut également choisir une autre variable qui apparaît plus fréquemment, comme un pronom sujet, mais ils n'impliquent pas des éléments d'origine anglaise et française dans nos données.

Deuxièmement, on pourrait faire la collecte de données sur une période plus longue, c'est-à-dire qui dépasse un mois, afin de prendre assez d'occurrences de la variable linguistique. Dans ce cas, tous les aspects de l'étude pourraient rester les mêmes, et on pourrait toujours finir par avoir assez de données pour effectuer une analyse plus

fiable, mais il n'est pas certain que l'on ne trouvera jamais des communautés où les membres qui écrivent des tweets contenant des éléments d'origine française sont majoritaires. Dans nos données, seule la communauté 6817 présente une proportion considérable d'individus qui ont employé des éléments d'origine française, 9 occurrences sur 21 pour être précis, mais au vu du manque de variation parmi les tweets dépourvus d'éléments d'origine française, on exigerait plus de telles communautés si l'on voulait mieux examiner les associations entre les communautés et la réalisation de (lol). Un examen d'une seule communauté dont les tweets contenant des éléments d'origine française sont majoritaires ne suffira pas si l'on veut être plus certain qu'une association existe. La collecte de plus de données de la communauté 6817 améliorerait finalement l'analyse des individus, bien sûr, mais le choix d'une autre région accomplirait la même chose et en plus améliorerait l'analyse des associations entre les communautés et la réalisation de (lol).

En effet, ce que nous voulions faire, c'était d'analyser non seulement les associations entre les communautés et la réalisation de (lol), mais aussi de mesurer le changement des diversités de la réalisation de (lol) pour des sous-ensembles des communautés, c'est-à-dire que nous voulions déterminer si la diversité diminue à mesure que l'on ne cible que les membres des communautés de plus en plus centraux, mais le manque de données a empêché une telle analyse. Le patron potentiel que nous avons trouvé en ce qui concerne les individus, leurs centralités et leurs diversités répond plus ou moins à la même question (voir la section 4.4 Comparaison entre les individus et leurs communautés), mais une analyse des sous-ensembles des communautés en présenterait une image plus claire. Ce serait une bonne piste de recherche à explorer davantage.

5.2 Identification des styles des individus

Dans la note 3 de bas de page 5 dans la section Définitions, nous avons expliqué la difficulté d'éviter l'usage des étiquettes couramment appliquées aux langues pour faire référence aux variétés qu'étudient les sociolinguistes. Nous nous servons donc souvent des descriptions « d'origine française » et « d'origine anglaise » dans ce mémoire, même si nous ne souscrivons pas à l'idée que ni « le français » ni « l'anglais » ni même une étiquette telle que « le français acadien » ne décrivent des entités structurelles cohérentes. Cette difficulté, d'après nous, provient de la nature des études sociolinguistiques variationnistes, dans lesquelles les chercheurs veulent typiquement décrire les variétés des groupes, plus ou moins équivalentes à ce que nous appelons les registres que nous définissons comme les variétés des groupes dans des contextes donnés (voir la section Définitions). Ces étiquettes servent donc de raccourcis où une analyse plus approfondie des réseaux sociaux des individus nous permettrait d'identifier les styles auxquels les individus ont accès, ce qui n'était pas possible dans la présente étude.

À titre d'exemple, un utilisateur de Twitter qui écrit toujours *lol* ou toujours *mdr* en interagissant dans une communauté donnée ne nous donne pas d'indication quant aux variantes qui existent dans son répertoire de styles. Il se peut que la seule variante que nous constatons soit en fait la seule variante dans ses styles, mais nous ne pouvons bien le savoir qu'à partir des observations seulement. Diane E. de la communauté 6817, par exemple, n'a produit que *lol* dans nos données, une variante que nous décrivons comme d'origine anglaise, bien que tous les autres éléments dans ses tweets codés seraient probablement décrits comme d'origine française, mais nous ignorons si Diane E. connaît les mots *mdr* et *ptdr* ou non à partir de ses renseignements, et une supposition là-dessus ne serait pas valide, c'est-à-dire qu'un

chercheur pourrait être tenté de supposer à partir de la présence des éléments d'origine française dans ses tweets que Diane E. connaît les mots *mdr* et *ptdr* dans une analyse superficielle où l'on emploie l'idée que « le français » est une entité structurelle monolithique, mais une analyse plus approfondie des styles de Diane E. pourrait témoigner que ce sujet a plusieurs styles, quelques-uns qui ressemblent à ce que l'on appellerait le français et quelques-uns qui ressemblent à ce que l'on appellerait l'anglais, mais dans tous ces styles, Diane E. ne produit que *lol* pour la variable linguistique (lol). Nous ne sommes simplement pas en mesure de dire si c'est le cas ou non sans une analyse plus approfondie de Diane E.

Dans le cas d'un individu qui ne produit jamais qu'une seule variante d'une variable linguistique, peu importe la communauté dans laquelle il est en train d'interagir, les étiquettes comme « le français » ou « l'anglais » sont peu utiles puisqu'elles supposent des structures qui sont inexactes pour cet individu. Il n'a pas accès à *mdr*, par exemple, même s'il a bien accès à d'autres éléments que l'on décrit couramment comme français. Un chercheur peut donc parler d'un registre donné, associé à une communauté, mais dès que l'on veut analyser plus en détail des individus, il faut identifier les vrais styles des individus. Dans ce cas, on peut parler des styles X_A et Y_A de l'individu A, les styles X_B et Y_B de l'individu B et la façon dont ces individus avec leurs styles se comportent dans un registre R. Pour ce faire, nous aurions dû nous servir d'une méthode de détection des communautés qui peut classer les individus dans plusieurs communautés. La méthode de Louvain n'est pas capable d'accomplir une telle tâche, mais il y en a d'autres dont nous parlerons dans la section 5.3.1.2 ainsi que d'autres mesures de centralité ci-dessous.

5.3 Répercussions des résultats

Jusqu'ici dans ce chapitre, nous parlons essentiellement des limitations de notre étude, et c'est en raison de sa nature : elle est une étude plus exploratoire qu'explicative. On peut donc en tirer plus de profit en considérant les limitations de l'étude qu'en essayant de confirmer des théories actuelles à partir des résultats. Il y a cependant des conclusions provisoires à tirer, notamment l'utilité des techniques actuelles de l'analyse des réseaux sociaux pour la sociolinguistique variationniste ainsi que le manque de distinction entre les études de la variation conceptualisée comme ayant lieu dans une même langue et les études du contact de langues.

5.3.1 Utilité des techniques actuelles de l'analyse des réseaux sociaux

Nos résultats dans ce mémoire équivalent à une validation de principe des techniques actuelles de l'analyse des réseaux sociaux pour les études de la variation ou du contact de langues. Bien que nous ayons démontré que les régions géographiques, qui sont souvent les unités que les chercheurs utilisent afin de délimiter leurs communautés linguistiques, sont associées aux communautés détectées par la méthode de Louvain sur Twitter, au moins dans nos données (voir la section 4.3.2 Communautés et régions géographiques), cette dernière s'avère une méthode plus précise. Il reste également d'autres méthodes à explorer, bien sûr, mais il devient de plus en plus difficile de s'en tenir seulement aux communautés linguistiques sauf dans le cas des études sommaires ou dans le cas où il n'est pas pratique d'élaborer un réseau social en raison des ressources limitées.

5.3.1.1 Communautés détectées contre communautés linguistiques

Nos résultats suggèrent que les régions géographiques vont de pair avec les communautés détectées à partir de la méthode de Louvain en ce que les résidents d'une région tendent à se regrouper sur Twitter dans nos données, et donc les communautés linguistiques, souvent définies selon des régions géographiques, demeurent utiles quand on veut donner un survol linguistique d'une région, surtout une région peu étudiée, mais des analyses plus détaillées et nuancées demandent des outils plus précis. De plus, l'usage des méthodes de l'analyse des réseaux sociaux nous permet d'éviter des problèmes éthiques qui se soulèvent en s'appuyant sur les variables sociales telles que l'ethnie, la religion, ou même le sexe, comme on fait souvent dans les études qui emploient les communautés linguistiques.

S'il est déterminé qu'une catégorie de personnes dans une communauté linguistique, soit celles venant d'un quartier ou celles faisant partie d'une ethnie particulière par exemple, préfèrent une variante spécifique d'une variable linguistique, il se peut qu'elles constituent également une communauté en termes d'analyse des réseaux sociaux, mais on ne peut ni le vérifier ni identifier les membres plus centraux et moins centraux à partir des variables sociales. Par exemple, il se peut qu'un membre du sous-ensemble des catholiques dans la communauté linguistique ne connaisse pas d'autres catholiques et donc soit mal classé dans ce sous-ensemble si un chercheur se sert seulement de communautés linguistiques, même si ce sujet s'identifie comme catholique. Ce n'est pas forcément le cas que tous ceux qui s'identifient à une classe sociale particulière se connaissent, et donc ils ne constituent pas forcément une communauté selon notre définition puisqu'ils n'interagissent pas[46]. Un chercheur

46 Il est vrai que ceux avec qui une personne n'interagit pas peuvent encore influencer cette personne-ci – par exemple, le parler des musiciens qu'elle aime peut l'influencer – mais on peut tenir compte de cette influence comme un lien avec un poids approprié, ce que l'usage des communautés linguistiques ne peut accomplir.

n'aurait aucune façon d'identifier ces individus mal classés qui rendent les résultats moins fiables. Un chercheur n'aurait également aucune façon d'étudier les membres plus centraux dans les communautés ou l'inverse. Dans notre étude, nous n'avions pas ces problèmes. Nous savions bien qui sont les membres centraux et non centraux, et le seul doute en ce qui concerne les personnes qui font partie d'une communauté dans notre analyse provient de la bonne ou mauvaise mise en application des outils de l'analyse des réseaux sociaux, et non d'un défaut des outils eux-mêmes.

Une dépendance seulement aux communautés linguistiques soulève aussi des problèmes éthiques en ce qui concerne les identités des communautés. Même si l'on peut se fier à des variables sociales pour identifier des sous-ensembles des communautés linguistiques qui sont plus ou moins équivalentes à des communautés détectées à partir des outils de l'analyse des réseaux sociaux, on le fait en invoquant des étiquettes qui ont tendance à simplifier les identités des peuples défavorisés. Par exemple, souvent dans les études de la variation, un sous-ensemble des sujets est identifié comme afro-américain (p. ex. Labov, 1969, 1966/2006; Nagy et Irwin, 2010; Strand, Wroblewski, et Good, 2010; Wolfram, 1969). Même si ce sous-ensemble constitue en fait une vraie communauté locale, l'usage de cette étiquette risque de les lier à d'autres communautés qui ont été identifiées comme afro-américaines. Il se peut que toutes ces communautés aient leurs propres caractéristiques locales, mais il est facile d'oublier cette possibilité si elles reçoivent toutes la même étiquette. Ce n'est certainement pas l'intention de la plupart des chercheurs, mais le risque existe. Dans le cas où l'analyse des réseaux sociaux peut identifier ces mêmes communautés à partir des interactions mesurées des sujets, on n'a pas besoin d'employer les étiquettes sociales, du moins si l'on ne s'intéresse pas à l'influence des constructions sociales sur les communautés, et donc on ne risque rien.

En effet, l'usage de la catégorie *afro-américain* a mené à la formulation d'un glossonyme pour un dialecte anglais : l'anglais afro-américain. Cette formulation a initialement eu un effet positif en ce qui concerne l'idée que le parler des personnes identifiées comme afro-américaines est un système régi par des règles, comme l'a avancé Bucholtz (2003, p. 402), mais l'usage d'un tel glossonyme risque aussi de faire en sorte, intentionnellement ou non, que toutes les personnes, ou au moins la plupart des personnes, que l'on peut identifier comme afro-américaines parlent une même variété monolithique qui s'appelle l'anglais afro-américain, ou que l'anglais afro-américain ne soit parlé que par les personnes que l'on peut identifier comme afro-américaines, des grosses simplifications qui sont inexactes selon toute vraisemblance. En effet, Benor (2010) a parlé d'un tel problème par rapport au peuple juif-américain comme une motivation pour le développement de son concept de répertoires ethnolinguistiques. Elle a remarqué qu'un chercheur qui s'intéresse à formuler un ethnolecte vise à décrire une variété que les membres d'une ethnie parlent mais doit finalement admettre que tous les membres ne le parlent pas (Benor, 2010, p. 159). Ces confusions se présentent facilement si l'on emploie les variables sociales comme l'ethnie, alors là où les mêmes communautés peuvent être détectées à partir d'une analyse du réseau des sujets, il vaut mieux s'appuyer sur l'analyse des réseaux sociaux, qui n'exige pas les variables sociales.

Dans notre étude, par exemple, nous n'avions pas besoin de considérer les classes sociales – et, certes, il aurait été difficile de les obtenir sur Twitter – car nous n'avons pas mis en œuvre les communautés linguistiques. Nous ne voulons pas dire que les classes sociales ne sont pas utiles, mais seulement que l'on ne devrait pas commencer par les invoquer si ce n'est pas nécessaire. Après la détection des communautés à partir d'une méthode d'analyse des réseaux sociaux, un chercheur peut chercher ce qui relie les membres de chaque communauté, soit une identité ethnique, soit leurs

genres, soit une activité, afin d'expliquer la nature de leurs communautés, mais l'application de ces explications avant la détection des communautés présume que ceux qui peuvent être décrits de la même façon constituent une communauté. De plus, on risque de lier la communauté locale à d'autres communautés si l'on utilise une étiquette courante, et ce faisant, on risque de simplifier l'identité de la communauté locale ainsi que les identités des autres communautés.

À ce propos, nous devons mentionner qu'il y a des chercheurs qui se fient aux auto-identifications des sujets à des classes sociales. À titre d'exemple, Eckert (2000) a classé ses sujets selon leurs propres identifications comme « sportif » ou « burnout ». De plus, Eckert (2000) n'a pas nommé ces deux catégories ; c'était les noms fournis par ses sujets (p. 82). Cela semble peut-être éviter le problème d'appliquer ses étiquettes aux sujets, mais ce n'est pas le choix d'étiquettes des chercheurs qui pose problème ; c'est plutôt que les problèmes se présentent quand on utilise n'importe quelle étiquette, peu importe qui l'a choisie. Cela est similaire à l'argument que nous avons mis de l'avant plus haut : un sujet qui se décrit lui-même comme catholique ne connaît pas forcément d'autres catholiques, mais on les regroupe quand même en utilisant les variables sociales telles que la religion. De même, les catholiques d'une région ne sont pas forcément équivalents aux catholiques d'une autre région. De même – et nous ne croyons pas que c'était l'intention d'Eckert de dire ceci, mais c'est une idée que l'on risque d'entraîner en tout cas – les sportifs de Belten High ne se connaissent forcément pas et ne sont pas forcément équivalents aux sportifs d'autres régions. Si l'on peut détecter ces mêmes communautés à partir de l'analyse des réseaux sociaux, c'est-à-dire sans s'appuyer sur les variables sociales, on peut éviter le risque de faire ces liens.

De plus, l'usage de l'auto-identification pour classer les sujets n'est pas un développement récent. Eckert (2000) était explicite par rapport à sa dépendance à l'auto-identification, mais Labov (1966/2006) s'en est lui-même servi en 1966 en ce qu'il n'a pas déterminé qui parmi ses sujets étaient juifs ou italiens, il l'a su en demandant à ses sujets de s'identifier. L'auto-identification permet d'éviter d'identifier un sujet comme noir qui ne se considère pas noir, par exemple, car cette identification peut se faire de façon erronée par un chercheur en regardant la couleur de la peau, mais ce n'est pas seulement une telle erreur que l'on veut éviter.

Or, nous voulons répéter que nous ne disons pas que les variables sociales ne sont jamais utiles. Si un chercheur s'intéresse à l'influence des constructions sociales sur ses sujets, les variables sociales sont bien pertinentes. Si un chercheur veut analyser la nature des communautés qu'il a détectées, c'est-à-dire ce qui lie les membres des communautés, les variables sociales sont bien pertinentes. Nous voulons seulement dire que les méthodes de détection des communautés de l'analyse des réseaux sociaux nous permettent d'identifier de vraies communautés, autrement dit des sous-ensembles d'une communauté linguistique, sans avoir besoin d'employer les variables sociales, et là où on peut éviter les risques de l'usage des variables sociales, il vaut mieux les éviter.

5.3.1.2 D'autres mesures de centralité et d'autres méthodes de détection des communautés

Dans ce mémoire, nous avons principalement employé le PageRank (Brin et Page, 1998) en tant que mesure de centralité et la méthode de Louvain (Blondel *et al.*, 2008) en tant que méthode de détection des communautés, mais d'autres outils

existent dans la littérature sur l'analyse des réseaux sociaux. Notre analyse a témoigné de l'utilité de ces outils-là, mais les sociolinguistes variationnistes pourraient tirer profit d'une comparaison entre ceux utilisés dans cette étude et d'autres à partir d'une perspective sociolinguistique.

L'objectif des mesures de centralité est d'identifier les personnes importantes d'un réseau ou d'une communauté, mais on peut être important de plusieurs façons. À titre d'exemple, les icônes sociolinguistiques d'Eckert (2000) sont centrales dans le sens où elles établissent de nouvelles modes, et il se peut qu'une mesure de centralité autre que le PageRank identifie mieux ces icônes et que le PageRank excelle dans l'identification des personnes qui suivent bien les icônes. Dans cette étude, en plus du PageRank, nous avons employé les degrés ainsi que l'intermédiarité, car ils nous fournissent d'autres renseignements des individus, mais de nombreuses autres mesures existent, telles que la proximité (Bavelas, 1950; Freeman, 1978), la proximité harmonique (Rochat, 2009), l'algorithme HITS (Kleinberg, 1999a, 1999b) et la centralité de vecteur propre. Il faut déterminer quel type d'individu est identifié par chacune de ces mesures en termes du comportement linguistique d'une communauté.

Cependant, les méthodes de détection des communautés ont pour but l'identification des communautés, un but qu'elles peuvent accomplir ou non. Si l'on peut dire qu'il y a plusieurs types de communautés, elles diffèrent dans la nature des liens entre les points, mais les chercheurs choisissent ce qui représente un lien avant d'appliquer les méthodes de détection des communautés. On n'a donc pas de raison de comparer les méthodes à partir d'une perspective sociolinguistique, mais les sociolinguistes feraient bien de se tenir informés des avancements dans ce domaine. À titre d'exemple, la méthode de Louvain utilisée dans ce mémoire est efficace et récente,

mais elle ne peut classer les individus dans plusieurs communautés à la fois. Cette méthode suppose qu'un individu est membre d'une seule communauté, une supposition qui s'avère exacte pour ce qui est des réseaux de neurones ou des réseaux informatiques, mais qui est peu probable d'être véridique dans le cas de réseaux sociaux. Une méthode qui résout ce problème en proposant des communautés qui se chevauchent est celle de Xie, Kelly, et Szymanski (2013), qui aiderait également à mieux analyser les répertoires de styles des individus dont nous avons parlé dans la section 5.2 Identification des styles des individus. Parés *et al.* (2018) ont proposé une autre méthode récente où ils ont affirmé que les périphéries des communautés sont fluides et visent à en tenir compte dans leur algorithme de détection. La recherche sur les méthodes de détection des communautés est bien active et les méthodes sont développées dans plusieurs domaines de recherche, ce qui la rend un peu difficile à suivre – par exemple, Xie *et al.* (2013) s'intéressent à la sociologie, tandis que Kleinberg (1999a, 1999b) s'intéresse à l'informatique, l'article de Yin, Chi, Y. Dong, et H. Dong (2017) se trouve dans un journal en physique et Schaub, Delvenne, Yaliraki, et Barahona (2012) s'intéressent aux neurosciences – mais les sociolinguistes peuvent sûrement en tirer profit.

5.3.2 Variation ou contact de langues

Dans notre étude, nous avons analysé (lol) de la même façon dont on étudierait (ing) ou l'effacement de (t/d). En dépit du fait que la recherche sur le contact de langues est variationniste dans plusieurs études (p. ex. Brown, 2003; Ehresmann et Bousquette, 2015; Poplack, 1979/1980/2000; Poplack et Dion, 2012; Poplack *et al.*, 1988), ces études n'emploient pas les variables linguistiques d'une façon qui implique des variantes de différentes variétés. Mougeon (2007) et Perrot (2014) sont des bons

exemples, mais ce n'est pas la norme. Notre réussite à analyser une telle variable, (lol), comme un cas typique de la variation réaffirme l'idée qu'il n'y a pas une grande différence entre l'étude de la variation typique, conceptualisée comme dans une seule langue, et l'étude du contact de langues, sauf peut-être que le contact de langues implique des variétés qui ont moins en commun que les variétés souvent considérées comme des parties d'une même langue. Nous voulons donc parler des répercussions pour ce qui est de la définition des variétés.

En partie, dans ce mémoire, nous nous occupons de la définition d'une unité d'analyse importante dans la linguistique : les variétés. Nous avons suggéré dans la section Définitions que les linguistes doivent parler de « la langue de *X* » où *X* est un groupe dans un contexte donné ou un individu au lieu des langues comme « le français » ou « l'anglais ». Comme chercheurs, nous cherchons des structures cohérentes et contraintes par des règles. En essayant d'expliquer la façon dont le monde fonctionne, on essaie d'expliquer que ce qui semble variable de façon aléatoire est en fait prévisible. Dans ce sens, même les variationnistes cherchent l'invariabilité. Ils veulent trouver des éléments linquistiques qui varient si l'on ne les considère que de façon large, mais qui présentent de fortes tendances envers une forme ou une autre lorsque l'on les analyse de façon restreinte, c'est-à-dire en les divisant en sous-ensembles. On n'arrivera jamais à de l'invariabilité totale dans les données du monde réel, bien sûr, mais les tendances qui intéressent les variationnistes sont envers l'invariabilité, du moins dans certains ensembles de leurs données.

En effet, en 2006, Labov (1966/2006) a finalement affirmé que l'analyse des réseaux sociaux produit des données des « individus étendus » et non des communautés (p. 399), en ce que les parlers des individus qui se trouvent dans un même coin d'un réseau social se ressemblent trop pour que ces individus soient jugés représentatifs

d'une communauté, on peut supposer ceci car une communauté doit comporter de la variation selon lui, mais ce manque de variation est une indication de l'utilité des outils de l'analyse des réseaux sociaux. On devrait chercher les contextes où la variation qui se présente hors contexte est réduite, car cette réduction indique que l'on a réussi à trouver un registre, suivant notre définition des registres (voir la section Définitions). Dans nos résultats, la réduction de la mesure de dispersion *D* en la calculant pour les communautés séparément comparativement au fait de la calculer pour tous les tweets en même temps témoigne que les communautés détectées à partir de l'analyse des réseaux sociaux prédisent où la variation sera réduite, et en fait l'affirmation au sujet des « individus étendus » de Labov implique la même conclusion.

Outre les communautés, il y a d'autres facteurs à considérer. Il serait raisonnable de supposer que d'autres facteurs contextuels, tels que la participation dans une activité ou le thème du discours, peuvent également prédire où la variation sera réduite. C'est bien ce que plusieurs sociolinguistes ont suggéré. Les communautés de pratique, proposées par des chercheurs comme Eckert (2000, p. 34-35/139) et Davies (2005), impliquent que la participation dans une activité a un effet sur la diversité de la réalisation des variables linguistiques, par exemple, de même que l'idée que le thème du discours peut influencer la variation existe dans la littérature au moins depuis Blom et Gumperz (1972/2000), qui ont constaté que les locuteurs à Hemnesberget, Norvège changeaient leurs parlers selon les changements du thème (p. 120-122), ce que Blom et Gumperz ont appelé de « l'alternance métaphorique » (p. 117).

La sociologie a probablement déjà largement étudié dans quelle mesure ces facteurs contextuels et les communautés se chevauchent, et ces découvertes devraient être intégrées dans la sociolinguistique. Weinreich (1953/1967) a proposé dix variables de

classe sociale générales auxquelles les langues maternelles peuvent être associées, y compris la région géographique, une résidence dans une zone urbaine ou rurale, le fait d'être indigène ou non, l'ethnie, la religion, la race, le sexe, l'âge, le statut social et la profession (p. 89-97), et c'est également ces variables qui se retrouvent dans les études de la variation encore aujourd'hui (p. ex. Carmichael, 2017; King, Martineau, et Mougeon, 2011; Lee, 2016; Morris, 2017; Sharma, 2011), mais il se peut que la détection des communautés en utilisant les méthodes de l'analyse des réseaux sociaux réduise le besoin de se fier à ces variables générales, surtout si ces communautés détectées et les sous-ensembles basés sur les variables sociales se chevauchent. Dans la présente analyse, ceci s'est avéré exact par rapport aux régions géographiques et les communautés (voir la section 4.3.2 Communautés et régions géographiques), mais nos données ne peuvent répondre en elles-mêmes à la question du chevauchement, ni à celle des autres variables et les communautés ou encore à celle de la même variable et d'autres communautés.

5.3.3 Unité d'analyse dans la linguistique théorique

Nous pouvons maintenant faire de petites remarques au sujet de ce qui constitue l'unité d'analyse principale dans la linguistique théorique afin de proposer un rapprochement entre elle et la sociolinguistique variationniste. Nos résultats suggèrent potentiellement que les individus très intégrés dans leurs communautés présentent moins de diversité que l'ensemble de leurs communautés (voir la section 4.4 Comparaison entre les individus et leurs communautés). De ce fait, on peut supposer que les styles des individus sont plus cohérents et invariables que les registres des groupes, ce qui les rendrait appropriés pour ce que font les linguistes théoriques. Or, les linguistes théoriques ont tendance à dépendre des données dérivées

des jugements de grammaticalité de différents locuteurs. Nous ne sommes pas en mesure d'avancer fortement des arguments pour ou contre différentes unités d'analyse dans la linguistique théorique, mais nous voudrions voir un rapprochement entre ce domaine-ci et la sociolinguistique, et nos résultats nous amènent à proposer que l'on peut atteindre un tel rapprochement si chaque étude linguistique théorique se restreint à un seul style d'un seul individu et si chaque étude sociolinguistique se restreint principalement à des registres des groupes et aux alternances dans les répertoires de styles des individus. De cette façon, la linguistique théorique serait effectivement un sous-domaine de la psychologie, comme Chomsky (1986, p. 3) l'a affirmé, tandis que le centre d'intérêt de la sociolinguistique serait ce qui est social.

CONCLUSION

Dans cette étude, nous avons effectué une analyse de la façon dont la variable linguistique lexicale (lol), constituée de variantes d'origine française et anglaise, est réalisée sur Twitter. Nos données sont venues des tweets émanant des provinces maritimes, mais les résultats représentent plutôt les communautés sur Twitter que nous avons détectées en utilisant la méthode de Louvain (Blondel *et al.*, 2008), un outil développé pour l'analyse des réseaux sociaux. Nous avons posé deux questions : si la réalisation de (lol) est statistiquement associée aux communautés et si les individus bien intégrés dans leurs communautés, selon leurs PageRanks (Brin et Page, 1998), présentent moins de diversité dans leurs réalisations de (lol) que leurs communautés. Nos résultats témoignent du fait que la réalisation de (lol) est en effet associée aux communautés et que les individus présentent moins de diversité dans leurs réalisations de (lol) que leurs communautés, même si nous ne pouvons fortement avancer ces conclusions en raison d'un manque de données. Nous avons donc offert des améliorations à la collecte de données. Finalement, nous croyons que cette étude témoigne de l'utilité des techniques actuelles de l'analyse des réseaux sociaux.

LISTE DES RÉFÉRENCES

andypiper. (2016 novembre). Is the Twitter language detection library publicly available? *Twitter Developers*. Récupéré de https://twittercommunity.com/t/is-the-twitter-language-detection-library-publicly-available/77573

Auer, P. (1988/2000). A conversation analytic approach to code-switching and transfer. Dans W. Li (dir.), *The Bilingualism Reader* (p. 154-174). Londres, R.-U. : Routledge.

Bamman, D., Eisenstein, J. et Schnoebelen, T. (2014 avril). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135-160. doi : 10.1111/josl.12080

Baron, N. S. (2004, 1 décembre). See You Online: Gender Issues in College Student Use of Instant Messaging. *Journal of Language and Social Psychology*, *23*(4), 397-423. doi : 10.1177/0261927X04269585

Bastian, M., Heymann, S. et Jacomy, M. (2009, mars). Gephi: An Open Source Software for Exploring and Manipulating Networks. Dans *Proceedings of the Third International ICWSM Conference*. International AAAI Conference on Web and Social Media (p. 361-362).

Bavelas, A. (1950, 1 novembre). Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, *22*(6), 725-730. doi : 10.1121/1.1906679

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145-204. doi : 10.1017/s004740450001037x

Benor, S. B. (2010 avril). Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics*, *14*(2), 159-183. doi : 10.1111/j.1467-9841.2010.00440.x

Blom, J.-P. et Gumperz, J. J. (1972/2000). Social meaning in linguistic structure: Code-switching in Norway. Dans L. Wei (dir.), *The Bilingualism Reader*. Londres, R.-U. : Routledge.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. et Lefebvre, E. (2008, 9 octobre). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 10008. doi : 10.1088/1742-5468/2008/10/P10008

Blyth, C. (1997). The Sociolinguistic Situation of Cajun French: The Effects of Language Shift and Language Loss. Dans A. Valdman (dir.), *French and Creole in Louisiana* (p. 47-70). New York, NY; Londres, R.-U. : Plenum Press.

Boissevain, J. (1974). *Friends of Friends: Networks, Manipulators and Coalitions.* St. Martin's Press.

Bott, E. (1972). *Family and Social Network* (2e éd.). Londres, R.-U. : Tavistock.

Brandes, U. (2001, 1 juin). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology, 25*(2), 163-177. doi : 10.1080/0022250X.2001.9990249

Brin, S. et Page, L. (1998, 1 avril). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1), 107-117. doi : 10.1016/S0169-7552(98)00110-X

Brown, B. (2003 février). Code–convergent borrowing in Louisiana French. *Journal of Sociolinguistics, 7*(1), 3-23. doi : 10.1111/1467-9481.00208

Brunstad, E. (2003 décembre). Standard language and linguistic purism. *Sociolinguistica, 17*(1), 52–70. doi : 10.1515/9783110245226.52

Bucholtz, M. (2003 août). Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics, 7*(3), 398-416. doi : 10.1111/1467-9481.00232

Carmichael, K. (2017, 1 novembre). Displacement and local linguistic practices: R-lessness in post-Katrina Greater New Orleans. *Journal of Sociolinguistics, 21*(5), 696-719. doi : 10.1111/josl.12253

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use.* New York, NY; Westport, CT; Londres, R.-U. : Praeger.

Cinque, G. (2002). A Note on « Restructuring » and Quantifier Climbing in French. *Linguistic Inquiry, 33*(4), 617-636. doi : 10.1162/002438902762731781

Cougnon, L.-A. et Ledegen, G. (2008, septembre). « c'est écrire comme je parle »: Une étude comparatiste de variétés de français dans l'écrit sms. Dans *Actes du Congrès annuel de L'AFLS* (p. 1-16). Oxford.

Coupland, N. (1980). Style-Shifting in a Cardiff Work-Setting. *Language in Society, 9*(1), 1-12. doi : 10.1017/s0047404500007752

Cramér, H. (1999, 23 mars). *Mathematical Methods of Statistics.* Princeton, NJ : Princeton University Press.

Dajko, N. et Carmichael, K. (2014 avril). But qui c'est la différence? Discourse markers in Louisiana French: The case of but vs. mais. *Language in Society, 43*(02), 159-183. doi : 10.1017/S0047404514000025

Davies, B. (2005 novembre). Communities of practice: Legitimacy not choice. *Journal of Sociolinguistics, 9*(4), 557-581. doi : 10.1111/j.1360-6441.2005.00306.x

Deuchar, M. et Stammers, J. R. (2016, 25 mai). English-Origin Verbs in Welsh: Adjudicating between Two Theoretical Approaches. *Languages, 1*(1), 7. doi : 10.3390/languages1010007

Dodsworth, R. et Benton, R. A. (2017 juin). Social network cohesion and the retreat from Southern vowels in Raleigh. *Language in Society*, *46*(3), 371-405. doi : 10.1017/S0047404517000185

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. (27). Madlen, MA : Blackwell Publishers, Inc.

Eckert, P. et McConnell-Ginet, S. (1992 octobre). Think Practically and Look Locally: Language and Gender as Community-Based Practice. *Annual Review of Anthropology*, *21*(1), 461-490. doi : 10.1146/annurev.an.21.100192.002333

Ehresmann, T. et Bousquette, J. (2015, 15 août, 15 août). Phonological non-integration of lexical borrowings in Wisconsin West Frisian. Dans J. Bondi Johannessen et J. C. Salmons (dir.), *Germanic Heritage Languages in North America: Acquisition, attrition and change.*, *18* (p. 234-255). Amsterdam; Philadelphie, PA : John Benjamins Publishing Company. 18 vol.

Eisenstein, J. (2015 avril). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, *19*(2), 161-188. doi : 10.1111/josl.12119

Fisher, R. A. (1922). On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87-94. doi : 10.2307/2340521

Fisher, R. A. (1925/1970). *Statistical Methods for Research Workers* (14th éd.). Édimbourg, R.-U. : Oliver and Boyd.

Fishman, J. A. (1967/2000). Bilingualism with and without diglossia; diglossia with and without bilingualism. Dans W. Li (dir.), *The Bilingualism Reader* (p. 74-81). Londres, R.-U. : Routledge.

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35-41. doi : 10.2307/3033543

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239. doi : 10.1016/0378-8733(78)90021-7

Friesner, M. (2009). *The social and linguistic predictors of the outcomes of borrowing in the speech community of Montréal*. (PhD). Philadelphie, PA : University of Pennsylvania.

Gardner-Chloros, P. et Edwards, M. (2004 avril). Assumptions Behind Grammatical Approaches To Code-Switching: When The Blueprint Is A Red Herring. *Transactions of the Philological Society*, *102*(1), 103-129. doi : 10.1111/j.0079-1636.2004.00131.x

Girvan, M. et Newman, M. E. J. (2002). Community Structure in Social and Biological Networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821-7826. doi : 10.1073/pnas.122653799

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360-1380.

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language*, *32*(1), 109-115. doi : 10.2307/410659

Gruzd, A. (2016). *Netlytic: Software for automated text and social network analysis.* Récupéré de https://netlytic.org/

Halliday, M. A. K. (1964/1968). The users and uses of language. Dans J. A. Fishman (dir.), *Readings in the Sociology of Language* (p. 139-169). Paris, FR : Mouton.

Holmes, J. et Meyerhoff, M. (1999 avril). The Community of Practice: Theories and methodologies in language and gender research. *Language in Society*, *28*(2), 173-183. doi : 10.1017/s004740459900202x

Horvath, B. et Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society*, *16*(2), 179-204. doi : 10.1017/s0047404500012252

Huffaker, D. A. et Calvert, S. L. (2017, 17 juillet). Gender, Identity, and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication*, *10*(2). doi : 10.1111/j.1083-6101.2005.tb00238.x

Hymes, D. H. (1967/1972). Models of the Interaction of Language and Social Life. Dans J. J. Gumperz et D. H. Hymes (dir.), *Directions in Sociolinguistics: The Ethnography of Communication* (p. 35-71). New York, NY : Holt, Rinehart and Winston, Inc.

Johnson, D. E. (2009, 1 janvier). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass*, *3*(1), 359-383. doi : 10.1111/j.1749-818X.2008.00108.x

Kayne, R. S. (2007, 1 mai). Several, few and many. *Lingua*, *117*(5), 832-858. doi : 10.1016/j.lingua.2006.03.005

Kim, S., Weber, I., Wei, L. et Oh, A. (2014). Sociolinguistic Analysis of Twitter in Multilingual Societies. Dans *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (p. 243–248). New York, NY. doi : 10.1145/2631775.2631824

King, R. (2008). Chiac in context: Overview and evaluation of Acadie's joual. Dans M. Meyerhoff et N. Nagy (dir.), *Social Lives in Language: Sociolinguistics and multilingual speech communities* (p. 138-178). Amsterdam; Philadelphie, PA : John Benjamins B.V. 24 vol.

King, R., Martineau, F. et Mougeon, R. (2011, 14 septembre). The Interplay of Internal and External Factors in Grammatical Change: First-person Plural Pronouns in French. *Language*, *87*(3), 470-509. doi : 10.1353/lan.2011.0072

Kleinberg, J. M. (1999 septembre). Authoritative Sources in a Hyperlinked Environment. *J. ACM*, *46*(5), 604–632. doi : 10.1145/324133.324140

Kleinberg, J. M. (1999 décembreb). Hubs, Authorities, and Communities. *ACM Computing Surveys*, *31*(4es), 1-3. doi : 10.1145/345966.345982

Klingler, T. A. (2005). Le problème de la démarcation des variétés de langues en Louisiane: étiquettes et usages linguistiques. Dans A. Valdman, J. Auger, et D. Piston-Hatlen (dir.), *Le français en Amérique du Nord: État-présent* (p. 349-367). Québec, QC : Presses de l'Université Laval.

Klingler, T. A., Picone, M. D. et Valdman, A. (1997). The Lexicon of Louisiana French. Dans A. Valdman (dir.), *French and Creole in Louisiana* (p. 47-70). New York, NY; Londres, R.-U. : Plenum Press.

Kwak, H., Lee, C., Park, H. et Moon, S. (2010, avril). What is Twitter, a Social Network or a News Media? Dans *Proceedings of the 19th International Conference on World Wide Web* (p. 591-600). New York, NY : ACM. doi : 10.1145/1772690.1772751

Labov, W. (1969). Contraction, Deletion, and Inherent Variability of the English Copula. *Language*, *45*(4), 715-762. doi : 10.2307/412333

Labov, W. (1972a). Some principles of linguistic methodology. *Language in Society*, *1*(1), 97-120. doi : 10.1017/s0047404500006576

Labov, W. (1966/2006). *The Social Stratification of English in New York City* (2e éd.). Cambridge, R.-U. : Cambridge University Press.

Lavandera, B. R. (1978). Where does the sociolinguistic variable stop? *Language in Society*, *7*(2), 171-182. doi : 10.1017/s0047404500005510

Lave, J. et Wenger, É. (1991). *Situated learning legitimate peripheral participation.* Cambridge, R.-U. : Cambridge University Press.

Le Page, R. B. et Tabouret-Keller, A. (1985). *Acts of Identity: Creole-Based Approaches to Language and Ethnicity*. Cambridge, R.-U. : Cambridge University Press.

Lee, J. (2016, 1 août). The Participation of a Northern New Jersey Korean American Community in Local and National Language Variation. *American Speech*, *91*(3), 327-360. doi : 10.1215/00031283-3701037

Lev-Ari, S. (2018 juillet). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, *176*, 31-39. doi : 10.1016/j.cognition.2018.03.003

Li, W., Milroy, L. et Sin Ching, P. (1992/2000). A two-step sociolinguistic analysis of code-switching and language choice: the example of a bilingual Chinese community in Britain. Dans W. Li (dir.), *The Bilingualism Reader* (p. 175-197). Londres, R.-U. : Routledge.

Liénard, F. (2014). Les communautés sociolinguistiques virtuelles. Le cas des pratiques scripturales numériques synchrones et asynchrones mahoraises. *Studii de lingvistică*, *4*, 145-163.

Milroy, L. (1980/1987). *Language and Social Networks*. (2) (2e éd.). Oxford : Blackwell.

Moreno, J. L. (1934). *Who Shall Survive: A New Approach to the Problem of Human Interrelations*. Washington, DC : Nervous and Mental Disease Publishing Co.

Morris, J. (2017, 1 avril). Sociophonetic variation in a long-term language contact situation: /l/-darkening in Welsh-English bilingual speech. *Journal of Sociolinguistics, 21*(2), 183-207. doi : 10.1111/josl.12231

Mougeon, R. (2007). Diversification du parler des adolescents franco-ontariens : Le cas des conjonctions et locutions de conséquence. *Cahiers Charlevoix : Études franco-ontariennes, 7*, 229-276. doi : 10.7202/1039327ar

Myers-Scotton, C. (1988/2000). Code-switching as indexical of social negotiations. Dans W. Li (dir.), *The Bilingualism Reader* (p. 127-153). Londres, R.-U. : Routledge.

Nadasdi, T., Mougeon, R. et Rehner, K. (2004, 16 juin). Expression de la notion de « vehicule automobile » dans le parler des adolescents de l'Ontario. *Francophonies d'Amerique, 17*(1), 91-106. doi : 10.1353/fda.2004.0017

Nagy, N. et Irwin, P. (2010 juillet). Boston (r): Neighbo(r)s nea(r) and fa(r). *Language Variation and Change, 22*(2), 241-278. doi : 10.1017/S0954394510000062

Navarro, D. (2015, 2 mars). lsr: Companion to « Learning Statistics with R » (Version 0.5). Récupéré de https://cran.r-project.org/web/packages/lsr/index.html

Newman, M. E. J. et Girvan, M. (2004, 26 février). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 1-16. doi : 10.1103/PhysRevE.69.026113

Nishimura, Y. (2016, décembre). A sociolinguistic analysis of emoticon usage in Japanese blogs: Variation by age, gender, and topic. Dans *AoIR Selected Papers of Internet Research*. The 16th Annual Meeting of the Association of Internet Researchers (Vol. 5). Phoenix, AZ. Récupéré de https://spir.aoir.org/index.php/spir/article/view/1137

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2017, 24 août). vegan: Community Ecology Package (Version 2.4-4). Récupéré de https://cran.r-project.org/web/packages/vegan/index.html

Parés, F., Garcia-Gasulla, D., Vilalta, A., Moreno, J., Ayguadé, E., Labarta, J., ... Suzumura, T. (2018, janvier). Fluid Communities: A Competitive, Scalable and Diverse Community Detection Algorithm. doi : 10.1007/978-3-319-72150-7_19

Pavalanathan, U. et Eisenstein, J. (2015 mai). Audience-Modulated Variation in Online Social Media. *American Speech, 90*(2), 187-213. doi : 10.1215/00031283-3130324

Payne, A. (1980). Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. Dans W. Labov (dir.), *Locating Language in Time and Space. 1.* New York, NY; Toronto, ON : Academic Press. 1 vol.

Perrot, M.-È. (2014). Le trajet linguistique des emprunts dans le chiac de Moncton: quelques observations. *Minorités linguistiques et société*, (4), 200-218. doi : 10.7202/1024698ar

Poplack, S. (1993, 1 juillet). Variation theory and language contact. Dans D. R. Preston (dir.), *American Dialect Research* (p. 251-263). Amsterdam; Philadelphie, PA : John Benjamins Pub Co.

Poplack, S. (1979/1980/2000). Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. Dans W. Li (dir.), *The Bilingualism Reader* (p. 205-240). Londres, R.-U. : Routledge.

Poplack, S. et Dion, N. (2012 octobre). Myths and facts about loanword development. *Language Variation and Change*, *24*(3), 279-315. doi : 10.1017/S095439451200018X

Poplack, S., Sankoff, D. et Miller, C. (1988). The social correlates and linguistics processes of lexical borrowing and assimilation. *Linguistics*, *26*(1), 47-104. doi : 10.1515/ling.1988.26.1.47

Provine, R. R., Spencer, R. J. et Mandell, D. L. (2007, 1 septembre). Emotional Expression Online: Emoticons Punctuate Website Text Messages. *Journal of Language and Social Psychology*, *26*(3), 299-307. doi : 10.1177/0261927X06303481

Rickford, J. et McNair-Knox, F. (1994). Addressee- and Topic-Influenced Style Shift: A Quantitative Sociolinguistic Study. Dans D. Biber et E. Finegan (dir.), *Sociolinguistic Perspectives on Register* (p. 235-276). New York, NY; Oxford : Oxford University Press.

Rizzi, L. (2013, 2 mai). Notes on cartography and further explanation. *Probus: International Journal of Latin & Romance Linguistics*, *25*(1), 197-226. doi : 10.1515/probus-2013-0010

Rochat, Y. (2009). Closeness centrality extended to unconnected graphs: The harmonic centrality index. Dans *ASNA* (p. 1-14). Zurich. Récupéré de https://infoscience.epfl.ch/record/200525/files/[EN]ASNA09.pdf

Sankoff, D. et Laberge, S. (1978a). Statistical Dependence among Successive Occurrences of a Variable in Discourse. Dans D. Sankoff (dir.), *Linguistic Variation: Models and Methods* (p. 119-126). New York, NY : Academic Press.

Sankoff, D. et Laberge, S. (1978b). The Linguistic Market and the Statistical Explanation of Variability. Dans D. Sankoff (dir.), *Linguistic Variation: Models and Methods* (p. 239-250). New York, NY : Academic Press.

Sankoff, D., Tagliamonte, S. A. et Smith, E. (2005). Goldvarb X: A variable rule application for Macintosh and Windows (Version 3.0b3). Récupéré de http://individual.utoronto.ca/tagliamonte/goldvarb.html

Sankoff, G. (2015, 1 janvier). The speech community as a social fact. *Asia-Pacific Language Variation*, *1*(1), 23-51. doi : 10.1075/aplv.1.1.02san

Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N. et Barahona, M. (2012). Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit. *PLoS ONE*, *7*(2), 1-11. doi : 10.1371/journal.pone.0032210

Schenkel, A., Teigland, R. et Borgatti, S. P. (2002, 6 décembre). Theorizing Structural Properties of Communities of Practice: A Social Network Approach. Dans *Communities of Practice or Communities of Discipline: Managing Deviations at the Oresund Bridge* (p. 1-31). Stockholm : The Economics Research Institute, Stockholm School of Economics.

Sharma, D. (2011 septembre). Style repertoire and social change in British Asian English. *Journal of Sociolinguistics*, *15*(4), 464-492. doi : 10.1111/j.1467-9841.2011.00503.x

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688. doi : 10.1038/163688a0

Statistique Canada. (2016, 23 novembre). *Population selon la langue maternelle et les groupes d'âge (total), chiffres de 2011, pour le Canada, les provinces et les territoires*. Récupéré de http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/hlt-fst/lang/Pages/highlight.cfm?TabID=1&Lang=F&Asc=1&PRCode=01&OrderBy=999&View=1&tableID=401&queryID=1&Age=1

Strand, T. R., Wroblewski, M. et Good, M. K. (2010, 1 septembre). Words, Woods, Woyds: Variation and Accommodation in Schwar Realization among African American, White, and Houma Men in Southern Louisiana. *Journal of English Linguistics*, *38*(3), 211-229. doi : 10.1177/0075424210373040

Tagliamonte, S. A. (2006). *Analysing Sociolinguistic Variation*. New York, NY : Cambridge University Press.

Tagliamonte, S. A. et Denis, D. (2008). Linguistic Ruin? Lol! Instant Messaging and Teen Language. *American Speech*, *83*(1), 3-34. doi : 10.1215/00031283-2008-001

Tatman, R. (2016, 1 décembre). "I'm a spawts guay": Comparing the Use of Sociophonetic Variables in Speech and Twitter. *University of Pennsylvania Working Papers in Linguistics*, *22*(2). Récupéré de http://repository.upenn.edu/pwpl/vol22/iss2/18

@tm. (2015, 16 novembre). Evaluating language identification performance. [Twitter]. Récupéré de

https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.html

Trudgill, P. (1974). *The Social Differentiation of English in Norwich*. Cambridge, R.-U. : Cambridge University Press.

Valdman, A. et Rottet, K. J. (2010). *Dictionary of Louisiana French: As Spoken in Cajun, Creole, and American Indian Communities*. Jackson, MS : University Press of Mississippi.

Waltman, L. et Eck, N. J. van. (2013, 1 novembre). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 471. doi : 10.1140/epjb/e2013-40829-0

Weiner, E. J. et Labov, W. (1983). Constraints on the Agentless Passive. *Journal of Linguistics*, *19*(1), 29-58. doi : 10.1017/s0022226700007441

Weinreich, U. (1953/1967). *Languages in Contact: Findings and Problems*. (1). La Haye : Mouton.

Witmer, D. F. et Katzman, S. L. (1997, 1 mars). On-Line Smiles: Does Gender Make a Difference in the Use of Graphic Accents? *Journal of Computer-Mediated Communication*, *2*(4). doi : 10.1111/j.1083-6101.1997.tb00192.x

Wolfram, W. (1969). *A Sociolinguistic Description of Detroit Negro Speech*. Washington, DC : Center for Applied Linguistics.

Xie, J., Kelley, S. et Szymanski, B. K. (2013 août). Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study. *ACM Computing Surveys*, *45*(4), 43-43:35. doi : 10.1145/2501654.2501657

Yin, G., Chi, K., Dong, Y. et Dong, H. (2017, 25 avril). An approach of community evolution based on gravitational relationship refactoring in dynamic networks. *Physics Letters A*, *381*(16), 1349-1355. doi : 10.1016/j.physleta.2017.01.059

Young, H. A. N. (2002 avril). *« C'est either que tu parles francais, c'est either que tu parles anglais »: A cognitive approach to Chiac as a contact language*. (PhD). Houston, TX: Rice University.

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, *33*(4), 452-473. doi : 10.1086/jar.33.4.3629752

Zhao, J., Wu, J. et Xu, K. (2010, 6 juillet). Weak ties: Subtle role of information diffusion in online social networks. *Physical Review E*, *82*(1). doi : 10.1103/PhysRevE.82.016105