An investigation into contact-induced semantic shifts in Quebec English : conciliating corpus-based vector models and variationist sociolinguistic inquiry

Filip Miletic

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 2 - Jean Jaurès

Présentée et soutenue par

Filip MILETIC

Le 20 juin 2022

An investigation into contact-induced semantic shifts in Quebec English: conciliating corpus-based vector models and variationist sociolinguistic inquiry

---

Ecole doctorale : **CLESCO - Comportement, Langage, Education, Socialisation, Cognition**

Spécialité : **Sciences du langage**

Unité de recherche :
CLLE - Unité Cognition, Langues, Langage, Ergonomie

Thèse dirigée par
Anne PRZEWOZNY-DESRIAUX et Ludovic TANGUY

Jury

M. Stefan DOLLINGER, Rapporteur
Mme Sabine SCHULTE IM WALDE, Rapporteure
M. Kris HEYLEN, Examinateur
Mme Amélie JOSSELIN-LERAY, Examinatrice
Mme Anne PRZEWOZNY-DESRIAUX, Directrice de thèse
M. Ludovic TANGUY, Co-directeur de thèse

Le monde n'est pas exactement le même
quand chaque objet a deux noms différents ;
c'est bizarre de penser à ça.

Nancy Huston, *Lignes de faille*, 2006

Abstract

This dissertation investigates contact-induced semantic shifts in Quebec English, i.e., preexisting English words which are used with a different meaning due to the potential influence of French. This sociolinguistic phenomenon has been described in several studies, but its diffusion, the constraints on its use, and the social meaning that it conveys remain poorly understood. I therefore propose a novel approach at the intersection of natural language processing and variationist sociolinguistics, aiming to provide a more comprehensive descriptive account as well as assess the contributions of the implemented methods.

In order to conduct computational analyses of semantic variation, I created a corpus containing 78.8 million tweets published by 196,000 speakers from Montreal, Toronto, and Vancouver. It was used to implement different types of vector space models, i.e., computational representations of word meaning. Type-level models were used to identify new semantic shifts based on the semantic differences between Montreal and the other two cities. Token-level models were used in finer-grained analyses and allowed to further characterize their use. Despite promising results, extensive qualitative analyses suggest that these methods are hampered by noise related to their inherent characteristics as well as corpus structure. This is corroborated by a systematic quantitative evaluation on a custom-built 80-item test set, demonstrating that SOTA-like performance on a standard semantic change detection task does not directly translate to practical value in discovering new semantic shifts.

These large-scale approaches were complemented with finer-grained data collected through sociolinguistic interviews with 15 speakers living in Montreal. I used a standard sociophonological protocol, ensuring comparable and reliable results, as well as a novel perception test examining the acceptability of 40 semantic shifts attested in the Twitter corpus. Varying correlations between lexical items and a range of sociodemographic factors, coupled with qualitative remarks on their use, point to four distinct patterns of synchronic variation; these in turn reflect potential diachronic processes. Moreover, interspeaker variability suggests that the use of semantic shifts is driven by speakers who tend to be younger and proficient in both English and French. Finally, the acceptability ratings are weakly correlated with computational variation measures, suggesting that they capture different dimensions of semantic variation.

Overall, this dissertation has provided the first systematic description of contact-induced semantic shifts in Quebec English, based on corpus analyses and face-to-face interviews. It has highlighted the complementarity of approaches used in different disciplines: the sociolinguistic object of study determined the setup of the computational experiments, which in turn provided the stimuli used in the sociolinguistic interviews, which in turn constituted further evaluation of the computational methods. These considerations have provided a pathway towards a better-informed use of corpus-based computational methods in studies of sociolinguistic phenomena.

Résumé

Cette thèse étudie les glissements de sens induits par le contact de langues en anglais québécois, à savoir des mots anglais préexistants utilisés avec un sens différent en raison d'une influence potentielle du français. Ce phénomène sociolinguistique est décrit dans plusieurs études antérieures, mais il reste de nombreuses inconnues quant à sa diffusion, les contraintes sur ses usages et la valeur sociale qu'il véhicule. Nous proposons une approche novatrice à l'intersection du traitement automatique des langues et de la sociolinguistique variationniste, afin de fournir une description exhaustive de ce phénomène ainsi que d'évaluer les contributions des approches sur corpus mises en œuvre ici.

Afin d'effectuer des analyses computationnelles de variation sémantique, nous avons constitué un corpus composé de 78,8 millions de tweets, publiés par 196 000 locuteurs de Montréal, Toronto et Vancouver. Le corpus a été utilisé pour mettre en œuvre différents types de modèles vectoriels, à savoir des représentations computationnelles du sens des mots. Les modèles statiques ont permis d'identifier de nouveaux glissements de sens (en identifiant des différences entre les locuteurs de Montréal par rapport aux deux autres villes), alors que les modèles contextuels ont permis de caractériser plus finement leurs utilisations. Malgré des résultats prometteurs, les analyses qualitatives indiquent que ces méthodes sont limitées par le bruit lié à leurs caractéristiques intrinsèques et à la structure du corpus. Ceci est corroboré par une évaluation quantitative systématique effectuée sur un jeu de données composé de 80 items. Celle-ci a montré que des résultats comparables à l'état de l'art sur une tâche classique de détection de changement sémantique ne se traduisent pas directement par la capacité pratique à repérer de nouveaux glissements de sens.

Ces approches à grande échelle ont été complétées par des données plus fines recueillies au moyen d'entretiens sociolinguistiques avec 15 locuteurs vivant à Montréal. Nous avons utilisé un protocole sociophonologique classique, garantissant des résultats comparables et fiables, ainsi qu'un nouveau test de perception portant sur l'acceptabilité de 40 glissements de sens attestés dans le corpus de tweets. Les corrélations entre ces variables linguistiques et différents facteurs sociodémographiques, ainsi que les remarques qualitatives sur leur utilisation, indiquent quatre patterns de variation synchronique ; ceux-ci pourraient à leur tour refléter des processus diachroniques. Par ailleurs, la variabilité inter-locuteurs suggère un rôle important des locuteurs bilingues et plus jeunes dans l'utilisation des glissements de sens. Enfin, les scores d'acceptabilité sont faiblement corrélés avec les mesures computationnelles, ce qui suggère que ceux-ci reflètent d'autres dimensions de variation sémantique.

Dans l'ensemble, cette thèse a fourni la première description systématique, menée sur corpus et au moyen d'entretiens, des glissements de sens en anglais québécois induits par le contact avec le français. Elle a également mis en évidence la complémentarité des approches développées dans des disciplines différentes : notre objet d'étude sociolinguistique a orienté la mise en place des expériences computationnelles ; celles-ci ont fourni les stimuli utilisés dans les entretiens sociolinguistiques ; ces derniers ont apporté une évaluation supplémentaire des méthodes computationnelles. Ces considérations ouvrent la voie à une utilisation plus avisée des méthodes computationnelles basées sur corpus dans des études de phénomènes sociolinguistiques.

Acknowledgments

This work would not exist without my advisors, Anne Przewozny-Desriaux and Ludovic Tanguy, who provided the perfect balance between steadfast scientific guidance and openness to explorations in all sorts of directions. The knowledge and skills that they so readily shared permeate all the chapters of this dissertation, and will stay with me well beyond it. Above all, thank you for your unfailing support, kindness, and encouragement through all manner of challenges and tight deadlines.

I am grateful to Stefan Dollinger, Kris Heylen, Amélie Josselin-Leray, and Sabine Schulte im Walde for accepting to evaluate my work and dedicating their time to it. Special thanks are due to Amélie and Kris, whose insights at an intermediate assessment strongly informed the final direction of my dissertation.

This work directly builds on the interests that I developed as a student at the University of Genoa. I am particularly grateful to Cristiano Broccias and Anna Giaufret, who supervised my initial research on Canadian English. I would also like to thank Wim Remysen, whose course at the University of Sherbrooke sparked my ongoing interest in variationist sociolinguistics. I am grateful for his continued encouragement, as well as his help in organizing my research stay in Quebec and valuable advice on conducting sociolinguistic interviews (in the midst of a pandemic surge, no less). I am also deeply indebted to the Montrealers who took the time to participate in the interviews. The wealth of linguistic knowledge and personal experiences that they shared constitutes the best conclusion to this work I could have hoped for.

I had the good fortune to conduct my PhD in the CLLE research lab, where despite being a newcomer to Toulouse I immediately felt at home. Particular thanks go to the NLP group, and especially Cécile Fabre, Nabil Hathout, and Mai Ho-Dac, for stimulating questions and discussions.

This experience would have been much less enjoyable without the company of all the other PhD students. A special shoutout to Marine, Chiara, Daniele, Julie H., Bénédicte, Lison, Camilla, Claire, Mariame, Killyam, Victoria, and all the others, for all the laughs, coffee breaks, and raclette parties. I would also like to thank Julie R., not only for her friendship, but also for the many discussions on Quebec, as well as for her help in analyzing the recordings for this dissertation. And I am particularly grateful to Silvia, who has been the definition of a friend. Thank you for your immense kindness, encouragement, and all the weekend walks.

Finally, a special thank you to Nevena, for your boundless reserves of optimism and a friendship that continues to defy thousands of kilometers. Thank you to my parents Slađana and Borivoj, for your unconditional support. And thank you to my sister Aleksandra, for always leading the way.

Contents

List of Tables

List of Figures

Introduction

When walking around Montreal, or reading a newspaper article published in the city, or scrolling through the Twitter profile of a Montrealer, it is not unusual to come across an utterance such as this one:

(1) I really want to go to an art museum or an art **exposition**

In this example, taken from the corpus of tweets created in this dissertation, the lexical item *exposition* is used to refer to what is usually known as an *art exhibition*. It is not conventionally used in this way in English; the sense attested here is instead associated with the homographhous French lexical item *exposition*. A number of existing sociolinguistic studies describe this linguistic practice – whereby an existing English lexical item is used with a sense associated with a phonologically and/or semantically similar French lexical item – as typical of the way in which English is spoken in Quebec. The prevalence of this phenomenon, usually termed *semantic shift*, is explained by the ongoing contact between English and French, the latter being spoken by a large majority of Quebecers. But although various descriptive sources provide evidence of its existence, limited systematic information is available on this sociolinguistic behavior. We know very little about its diffusion within the speech community, the linguistic and social constraints on its use, and the social meaning that it conveys. This is the gap that the present dissertation aims to address, specifically from a variationist sociolinguistic perspective.

But any attempt to pursue this description is faced with a series of challenges. From a theoretical standpoint, variationist sociolinguistics can draw on decades of research to investigate phonological and morphosyntactic phenomena. On the other hand, its treatment of the lexicon, and lexical semantic issues in particular, is considerably less well-established. This has implications from the methodological standpoint as well. Standard data collection methods, such as the sociolinguistic interview, entail practical constraints which lead to corpora that are too limited in size to systematically study lexical variation. Other approaches, such as written dialect surveys, circumvent this issue by eliciting directly comparable information from a larger number of informants. However, they provide more limited sociodemographic background, are disconnected from spontaneous communication, and are limited to predefined sets of lexical items.

A potential solution comes from the field of natural language processing, where vector space models – computational representations of word meaning – have been used to study semantic change. They allow for a systematic, quantitative assessment of the evolution of word meaning over time or across other dimensions. These analyses can be extended to the entire vocabulary, potentially allowing for a bottom-up detection of previously unknown cases of the

phenomenon under study. However, they come with methodological challenges of their own. First, there is the issue of data: in order for these analyses to be meaningful, they are conducted on very large corpora, at least two orders of magnitude larger than those usually created through sociolinguistic interviews. And then, there is the choice of the model architectures, hyperparameters, and variation measures to be implemented, as well as the as yet uncertain reliability of the produced results. The first problem could be solved by using the vast amounts of publicly available, geolocated social media data, but this in turn entails additional uncertainty over the extent to which the resulting descriptions reflect real-life communication. The second problem may be addressed through a systematic evaluation of semantic change detection methods, but no readily available benchmarks exist for contact-induced semantic shifts in Quebec English.

Since there appears to be no simple solution to the problem at hand, I adopted an interdisciplinary perspective. My aim is to produce a comprehensive descriptive account by drawing on the complementary aspects of the two types of approaches, all the while circumventing their shortcomings. This in turn provides an opportunity to evaluate the descriptive contributions of the implemented computational methods both quantitatively and qualitatively, as well as to assess the reliability of social media corpora in sociolinguistic descriptions. More specifically, I used vector space models created from a custom-built corpus of tweets to obtain a systematic, large-scale overview and an initial characterization of contact-induced semantic shifts in Quebec English. The set of lexical items identified through these analyses was then examined more closely through face-to-face sociolinguistic interviews conducted with 15 Montrealers. The joint outcome of these two approaches clarified the factors behind the use of semantic shifts, the representations that are associated with them, and provided a systematic analysis of their diffusion within the speech community. It also demonstrated the promising role of large-scale computational methods in facilitating descriptive work, while also highlighting important shortcomings and the continued importance of linguistically informed analyses.

In the remainder of the introduction, I will briefly discuss my scientific position with regard to the language community under study. I will then present the publications produced as part of this work, and outline the structure of the dissertation.

Scientific position

As the opening chapters of the dissertation will show more clearly, the sociohistorical context in which Quebec English is spoken is far from neutral. The use of English and French is more closely associated with identity in Quebec than arguably any other Canadian province. Given this situation, it is relevant to state that my interest in contact-induced phenomena in Quebec English is principally driven by the linguistic side of the sociolinguistic continuum. In other words, rather than analyzing the underlying societal structures, my main focus is a description of the way in which this specific variety of English is used. That being said, this use must necessarily be interpreted against the backdrop of the social context in which it occurs. Although I am external to the community under study, I am able to draw on several months of my own lived experience in Quebec, which involved first-hand participation in the kinds of

bilingual interactions that I aim to describe. More generally, my background in sociolinguistics initially developed with a focus on Quebec French, which was then complemented with work on Quebec English. This, I hope, provides sufficient safeguards against hasty conclusions which would not do justice to the complex social reality of Quebec.

Publications and presentations

The work conducted as part of this dissertation led to the following publications:

- Miletic, F., Przewozny-Desriaux, A., Tanguy, L. (2021). Detecting contact-induced semantic shifts: What can embedding-based methods do in practice? *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10852–10865.
- Miletic, F., Przewozny-Desriaux, A., Tanguy, L. (2020). Collecting tweets to investigate regional variation in Canadian English. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6255–6264.

This work also led to the following presentations:

- Miletic, F., Przewozny-Desriaux, A., Tanguy, L. (2021). The status and representation of contact-induced semantic shifts in Quebec English: From Twitter users to sociolinguistic informants. Poster presented at *New Ways of Analyzing Variation 49 (NWAV 49)*, Austin, TX, USA (online).
- Miletic, F. (2021). Bridging across datasets and disciplines: The contribution of corpus phonology to the study of lexical semantic variation. Paper presented at *PAC 2021 – Spoken English varieties: redefining and representing realities, communities and norms*, Toulouse, France.
- Miletic, F., Przewozny-Desriaux, A., Tanguy, L. (2021). Modeling fine-grained sociolinguistic variation: The promises and pitfalls of Twitter corpora and neural word embeddings. Paper presented at *Corpus Pitfalls: Dealing with Messy Data (and Other Traps for the Unwary) (Workshop at ICAME42)*, Dortmund, Germany (online).
- Miletic, F., Przewozny-Desriaux, A., Tanguy, L. (2020). Methodological issues in using word embeddings in a sociolinguistic perspective: The case of contact-induced semantic variation across Canadian Twitter corpora. Poster presented at *Empirical Studies of Word Sense Divergences across Language Varieties (Workshop at DGfS 2020)*, Hamburg, Germany.
- Miletic, F. (2019). Contact-induced lexical variation in Quebec English: An accountable description. Poster presented at *RJC2019 – 22èmes Rencontres des jeunes chercheurs en Sciences du Langage*, Paris, France.

Outline of the dissertation

The remainder of this dissertation is organized in four main parts. They are followed by a general conclusion.

Part I provides a broad background on the linguistic and sociohistorical context of this dissertation, focusing on a wide range of general mechanisms which underpin my view of the descriptive object of study. Chapter 1 discusses bilingualism from the standpoint of individual speakers as well as wider communities, highlighting the aspects which provide support for cross-linguistic influence on the lexical semantic level as well as factors which may influence their use in the context of language contact. Chapter 2 provides a background for the specific situation of language contact under study: it outlines the sociohistorical context of Quebec, as well as key features of Quebec French and Quebec English. Building on a discussion of the existing descriptions of semantic shifts in Quebec English, Chapter 3 presents the theoretical view of this issue adopted in this dissertation.

Part II discusses the bases of the interdisciplinary approach that I adopted. It specifically addresses data collection methods in Chapter 4, modeling of lexical semantic variation in Chapter 5, and ways of accounting for observed patterns of language variation in Chapter 6. Each chapter reviews complementary practices developed in variationist sociolinguistics and natural language processing. This leads to a general overview of the proposed method, outlined in Chapter 7.

Part III presents the corpus-based analyses conducted in this dissertation. Chapter 8 describes the creation of a large corpus of tweets allowing for an analysis of regional patterns of lexical semantic variation. Chapter 9 presents an exploratory analysis of the collected data, providing an initial confirmation of the presence of regional trends in the data, as well as highlighting likely shortcomings of the implemented methods. Chapter 10 examines these trends more thoroughly, aiming for a better understanding of the object of study, the collected data, and the implemented methods; these include type-level and token-level vector space models. Chapter 11 addresses the observed issues more systematically: it introduces a test set for semantic shift detection, which is subsequently used to evaluate the performance of vector space models. This provides a clearer view of their descriptive contributions, and leads to the formulation of finer-grained hypotheses regarding the use of contact-induced semantic shifts.

Part IV introduces the sociolinguistic interviews conducted in this dissertation. The protocol used for this study, including a novel semantic perception task, as well as the recruitment procedure are presented in Chapter 12. The sociodemographic and attitudinal characteristics of the participant sample are discussed in Chapter 13. An analysis of contact-induced semantic shifts, focusing on their status and diffusion within the community, is described in Chapter 14. Finally, Chapter 15 outlines a comparative assessment of the two methodological approaches, highlighting their complementarity and providing a promising pathway for future studies of other phenomena.

Part I

Semantic effects of language contact in Quebec English: an overview

The chapters presented in this part of the dissertation provide a general overview of the mechanisms which may explain the emergence of contact-induced semantic shifts, the specific context in which they are used, and the theoretical view adopted in describing them. Chapter 1 introduces the central notions of bilingualism and language contact. It focuses both on the characteristics of individual bilingual use, as well as their implications for community-level linguistic practices. It moreover discusses the linguistic effects of individual bilingualism, including the mechanisms which may facilitate lexical semantic influence in a situation of language contact. Chapter 2 presents the sociohistorical context and language varieties which are at the center of this dissertation. It describes the historical development of Quebec society and its current demolinguistic profile, which provide a clear basis for ongoing language contact. It then outlines some of the main characteristics of Quebec French and Quebec English, focusing on those that are implicated in contact-related processes or are otherwise directly relevant for this dissertation. This includes a summary of previous descriptions of semantic shifts, which are further addressed in Chapter 3. It provides a more precise definition of this object of study and outlines key theoretical principles taken into consideration in its analysis. This sets the stage for the development of the methodology proposed in this dissertation.

Chapter 1

Bilingualism and language contact

It is traditionally considered that "two or more languages will be said to be *in contact* if they are used alternatively by the same persons" (Weinreich, 1953, p. 1). Already in Weinreich's work, the interest in language contact is motivated by the study of *interferences*, or "deviations from the norms of either language which occur in the speech of bilinguals as a result of their familiarity with more than one language" (p. 1). I will similarly view contact-induced language change as "the product of innovations that individual multilingual speakers introduce into discourse in a multilingual setting" (Matras, 2009, p. 5). That is why, in order to understand the mechanisms underpinning this type of language change, I will begin by discussing the specifics of bilingual language use.

This chapter first addresses bilingual language acquisition and use from the perspective of the individual speaker (Section 1.1). It then presents the development of bilingual communities and their link with identity (Section 1.2). Finally, it outlines the main manifestations of bilingualism in the speech of individual speakers, underscoring the link between these patterns and language change at the community level (Section 1.3). Bearing in mind my general focus on contact-induced semantic shifts, this chapter will provide evidence for the mechanisms facilitating their emergence, as well as highlight the factors which may condition their use.

1.1 Individual bilingualism

This section sets the ground for an overview of individual bilingualism by providing a definition of bilingualism and outlining the most common ways in which it is classified. This is followed by an overview of bilingual language acquisition, distinguishing between simultaneous and successive acquisition.

1.1.1 Defining bilingualism

In analyzing bilingualism, I will adopt François Grosjean's view that "bilinguals are those who use two or more languages (or dialects) in their everyday lives" (Grosjean, 2010, p. 4).[1] This

[1]Following this definition, I will use the term *bilingual* to refer to both bilingual and multilingual speakers. The forthcoming discussion of the use of two languages should be taken to apply to three or more languages as well.

definition brings to the fore the regular use of multiple languages, without restricting the notion based on factors such as proficiency or age of acquisition. While this view is now widely accepted, it is also relevant to examine how this focus evolved over time.

As Grosjean (2010) points out, most early analyses of bilingualism focused on language fluency rather than language use (p. 19). For instance, Bloomfield (1933) defines bilingualism as "native-like control of two languages" (p. 56). He argues that it is more frequent in children than in adult speakers, despite conceding that it is difficult to establish the precise threshold at which proficiency is native-like (pp. 56–57). Haugen (1969) similarly focuses on proficiency in formulating the minimum requirement for bilingualism as the ability to "produce *complete, meaningful utterances* in the other language" (p. 7). Unlike Bloomfield, however, he suggests that bilingualism applies to different degrees of proficiency, the highest of which corresponds to the ability "to pass as a native in more than one linguistic environment" (p. 7). The idea of a scale of bilingualism is further developed by Diebold's (1961) notion of *incipient bilingualism*. It extends to passive knowledge of another language, thereby removing strict requirements in terms of proficiency at the lower end of the scale. The various points are linked into a continuum by Kachru (1965), who introduces a description based on a *cline of bilingualism*. This is a scale which ranges from absolute monolingualism to absolute ambilingualism (native-like control in both languages), and passes through various degrees of bilingual proficiency. While the definitions adopted by different researchers underscore the evolution of the notion of bilingualism over time, they are also closely associated with the specific contexts that each of them studied. In describing the characteristics of World Englishes, it is important to note that English-language interactions routinely take place between bilinguals with varying degrees of proficiency, who constitute an integral part of their communities (Kachru, 2008); this view is reflected by the definition of Quebec English that I introduce in Chapter 2.

Moreover, Mackey (1962) argues that determining whether a speaker is bilingual based on their degree of proficiency is inherently arbitrary. He therefore defines bilingualism as "the alternate use of two or more languages by the same individual" (p. 52). Although Weinreich's (1953) earlier definition of bilingualism similarly references "the practice of alternately using two languages" (p. 1), the shift in focus to language use is more limited in his work. For instance, in discussing the behavior of the supposed "ideal bilingual", he precludes practices such as intrasentential codeswitching (p. 73), which has since been shown to represent an integral part of bilingual language use (see Section 1.3.1). More generally, the historical focus on "ideal" or "balanced" bilingualism has had the detrimental effect of supporting the idea that there are other, less valuable kinds of bilingualism (Romaine, 1995, p. 6). The very notion of balanced bilingualism is in fact largely a reflection of a monolingual bias in describing bilinguals (Romaine, 1995, p. 19).

One way of tackling this bias is to approach bilingualism as "a language user's competence which cannot be described in a single grammar" (Wald, 1974, p. 307). Put differently,

> the bilingual is an integrated whole which cannot easily be decomposed into two separate parts. The bilingual is *not* the sum of two complete or incomplete monolinguals; rather, he or she has a unique and specific linguistic configuration. The

> co-existence and constant interaction of the two languages in the bilingual has produced a different but complete language system. (Grosjean, 2008, pp. 13–14)

This is what Grosjean terms *the wholistic view of bilingualism*. As we shall see in the coming sections, this perspective allows us to account for different characteristics of bilingual language use. For now, let us note that bilinguals use their languages for different purposes and in different contexts. As a result, they tend to develop different linguistic competencies in the languages they speak. This in turn means that, contrary to popular belief, their fluency is rarely balanced across their linguistic repertoire (Grosjean, 2008, pp. 13–14).

1.1.1.1 Types of bilingualism

As mentioned in the previous section, I will consider all individuals who regularly use two or more languages as bilingual. This deliberately broad definition encompasses a wide range of profiles of bilingual speakers, so it is important to identify the characteristics which may help to differentiate them.

Butler (2012, pp. 112–115) provides an overview of classifications of bilingualism. They involve a variety of dimensions:

- relative language proficiency: *balanced* bilinguals are equally proficient in all their languages, *dominant* bilinguals are not;
- functional ability: *productive* bilingualism involves the active knowledge of another language, *receptive* bilingualism is limited to passive knowledge;
- age of acquisition: *simultaneous* bilinguals acquire their languages in parallel from birth, *sequential* bilinguals acquire them in succession, *late* bilinguals acquire them as adults;
- organization of linguistic codes: linguistic structures are shared in *compound* bilingualism, they are independent in *coordinate* bilingualism, and the use of one language is mediated by the other in *subordinate* bilingualism;
- language status and learning environments: *elite* (or *elective*) bilingualism involves the acquisition of another language by choice due to its perceived value, whereas *folk* (or *circumstantial*) bilingualism arises out of necessity, for example as a result of immigration;
- effect of L2 learning on L1 retention: *additive* bilingualism implies the preservation of the L1, *subtractive* bilingualism implies that the L1 is negatively affected by the acquisition of another language;
- cultural identity: *L1 monocultural* bilinguals maintain their initial cultural identity, *L2 accultural* bilinguals acquire the cultural identity associated with the newly acquired language, *bicultural* bilinguals develop a cultural identity shaped by both cultures, *decultured* bilinguals lose both cultures.

While classifications such as these illustrate the overall diversity of bilingual speakers, Butler (2012, pp. 112–116) also argues that they do not fully capture the complexity of bilingualism or its evolution over time. He points out, first, that many classifications are continuous rather

than categorical. They therefore require identifying cut-off points for the categories that they entail, which can be arbitrary. Second, he underscores that the classifications do not take into account the role of context, which is central to the way bilinguals use their languages. This is particularly important for studies of language variation, because speakers are known to adapt their language use to the specific communicative situation (this phenomenon, known as style shifting, is further addressed in the discussion of data collection in Chapter 4). Finally, these classifications of bilingualism do not account for the fact that bilingual profiles change over time. Since many classifications involve binary oppositions between two languages, they are additionally no longer mutually exclusive when applied to multilingual speakers.

These concerns echo Grosjean's (2008) caution against classifying bilingual speakers into discrete categories in a definitive manner (p. 269). One of the reasons for this is the functional specialization of a bilingual's languages, which he formally defines as the Complementarity Principle: "Bilinguals usually acquire and use their languages for different purposes, in different domains of life, with different people. Different aspects of life often require different languages" (Grosjean, 1997, p. 165). This principle affects a person's use of a language, their fluency in it, and the domains in which they employ it. Incidentally, these factors are taken to be indicative of language dominance, which is often invoked in classifications of bilinguals, including the ones I have presented. But as these factors evolve over time, so does language dominance (Grosjean, 2013, pp. 11–14). This once again points to the instability of categories that are routinely presented as immutable.

These critical observations highlight important limits of classifications of bilingualism such as the ones presented above. However, they can be useful in guiding the description of bilingual language use, as they point to a variety of factors which may play an important role in conjunction with the wider context and the speaker's individual trajectory. Another key aspect in analyzing bilinguals, referenced in some of the classifications, is the degree of bilingualism; I will focus on it next.

1.1.1.2 Degree of bilingualism

Estimating the extent to which a speaker is bilingual is essential in accounting for their linguistic behaviors, as well as the way in which they interact with other speakers. A prominent approach to describing bilingual ability was proposed by Mackey (1962). It takes into account four characteristics:

- *degree*, which is based on testing the bilingual's skills in all of their languages;
- *function*, which involves describing the use of different languages in terms of *external functions*, roughly corresponding to the communicative contexts in which they are used, and *internal functions*, focusing on the use of language outside of interaction (e.g. for internal speech) or as a reflection of the speaker's demographic or cognitive characteristics;
- *alternation* between the languages, examined in terms of the quantitative patterning of codeswitching and its conditioning on contextual factors;
- *interference*, understood here as the introduction of features from one language while

using another.

Overall, this approach brings together (i) measures of monolingual proficiency; (ii) contextual, demographic, and cognitive factors; and (iii) bilingual usage patterns.

While a variety of other classifications have been proposed, Pienemann and Keßler (2007) draw attention to some of their shortcomings. Language-external factors similar to the ones used by Mackey (1962) are often employed to construct straightforward taxonomies, which, the authors argue, provide little insight into the social or cognitive specifics of bilingual language use. In other cases, these factors are linked to measures of language proficiency; as we will see later on, these come with an additional set of difficulties (pp. 249–251).

Romaine (1995) furthermore notes that methods such as the one proposed by Mackey respond to the need to assess bilingual ability based on different linguistic skills. However, the approaches attempting to quantify bilingualism remain imperfect. First of all, they fail to capture finer qualitative differences in bilingual language use. Moreover, they often involve self-assessment, which is unreliable as it may be affected by a variety of factors. These include (i) the speaker's attitude towards the languages they speak and their perceived prestige; (ii) culturally specific understanding of what it means to be a competent speaker of a language; (iii) differences in literacy across the languages; and (iv) varying patterns of language use due to external factors (e.g. interlocutor, setting, topic) (pp. 12–17). Similar problems also affect census statistics, which are often used in bilingualism research due to the large coverage they provide, but are known to suffer from conceptual ambiguity (e.g. in defining key notions such as *mother tongue*) in addition to the general shortcomings related to self-assessment (pp. 26–30).

There are also issues that specifically affect the measurement of proficiency. A central problem emphasized by Pienemann and Keßler (2007) is the fact that many proficiency measures were originally designed for monolingual competence, making the cross-linguistic comparisons that rely on them questionable at best. Examples include rating scales (where samples of speech production are evaluated by trained raters) and word naming tests (where the subject is asked to provide as many words as possible related to a given domain). Other methods directly estimate differences in the subject's behavior in different languages, using methods such as verbal association tests (where the subject is asked to provide as many associations as possible to a given word, in the same language as the word) and reaction time measures (which examine the difference in the time taken to complete the same task in different languages). However, these are not general measures of bilingualism, and it is unclear which specific aspect of bilingual ability they address. Another, more general problem is the fact that different disciplines often adopt different measures, many of which present inherent limitations (Pienemann and Keßler, 2007). This lack of comparability has motivated a recent proposal of a "bilingualism quotient", a single quantitative estimate of bilingualism modeled after the intelligence quotient, but it remains unclear how it should be computed (Marian and Hayakawa, 2021).

In addition to these methodological issues, the very idea of quantifying proficiency is somewhat at odds with its complex and dynamic nature. Grosjean (2010) insists on the interaction between language fluency and language use, arguing that it is the patterning of these features that provides a full understanding of a person's bilingual ability. Contrary to popular represen-

tations of bilingualism, a person may, for example, have native-like fluency in one language and use it rarely, and have intermediate fluency in another language and use it daily. This interaction may further evolve over time, hence the importance of studying the language history of bilinguals: time, place, and manner of acquisition, and patterns of fluency and use over time (pp. 23–27). When such information is not available, corpus-based estimates of language dominance can be produced, using measures such as lexical richness (Treffers-Daller, 2011).

In summary, most approaches to estimating the degree of bilingualism combine proficiency measures with information on language-external factors. Different methods involve different types of limitations, including a lack of clarity on the evaluated linguistic skill, and a limited reliability of some sources of information. A comprehensive estimate of bilingualism should therefore incorporate different types of data, as well as account for the dynamic nature of bilingual ability as reflected by personal language history. A key aspect in this regard is the way a bilingual speaker acquires their languages, which is explored in the following section.

1.1.2 Bilingual language acquisition

The acquisition of multiple languages can take place under different circumstances. Bilingualism research usually draws a distinction between simultaneous and successive acquisition. Simultaneous acquisition implies that a bilingual's languages are acquired at the same time, hence there is no "first" language as such. This is often the case with childhood bilingualism. Successive acquisition, as its name suggests, involves a differentiation between the acquisition of the first and the second language. This can occur at any stage in a person's life.

Both processes can make a person bilingual, and they can lead to the same degree of bilingualism (Grosjean, 2010, p. 178). However, they imply distinct mechanisms and contexts of acquisition, and are associated with different common preconceptions. I will address these two types of acquisition in turns, discussing the key stages and factors for each of them. I will then provide a general overview of the differences and similarities in the resulting bilingual ability.

1.1.2.1 Simultaneous acquisition

Simultaneous acquisition is defined by Yip (2013) as "the concurrent acquisition of two languages in a child who is exposed to them from birth and uses both regularly in early childhood" (p. 120). Early childhood is understood here to continue up to around the age of five, so the definition targets linguistic ability preceding the effect of school. It also underscores language use, thereby excluding passive bilingualism.

Apart from the obvious difference in the number of acquired languages, simultaneous and monolingual acquisition are remarkably similar. They involve the same processes in language development, i.e. babbling, one-word, and two-word stages. In simultaneous acquisition, the stages occur at the same age as in monolinguals, but not necessarily at the same time in both languages. While in principle all languages acquired in this manner are equally important, patterns of dominance often emerge, mainly due to the functional differentiation of languages (Yip, 2013, pp. 119–122).

In addition to following similar developmental patterns as monolingual children, simultaneous bilinguals very quickly acquire the communicative strategies typical of bilingual adults. Already at an early age, bilingual children differentiate their languages in both perception and production; they are able to choose the language appropriate to the context; and they deploy strategies such as borrowing and codeswitching under similar conditions as adult bilinguals (Yip, 2013, pp. 126–137).

While these trends hold true in a general way, considerable differences in bilingual ability have been reported between different speakers. This may be explained by a variety of factors that can affect childhood bilingual acquisition. Grosjean (2010, pp. 171–177) argues that the foremost among them is the need for the language in question: if a bilingual is strongly compelled to use a language, and other factors are favorable, that language is likely to be acquired; otherwise, its use is likely to cease and it may be lost entirely. The other factors that Grosjean references include input, both in terms of its amount (e.g. provided in varied situations, by interlocutors important to the speaker) and type (e.g. natural input provided by a monolingual speaker, written input obtained through reading). In terms of the wider context, he underscores the role of the family as it ensures the use of the home language, which is particularly important for minority languages. The school and the wider community are also instrumental in lending importance to the minority language. Finally, the attitudes towards the specific language and the associated culture, as well as towards bilingualism in general, are crucial. They are acutely perceived by children and may strongly influence their linguistic behaviors.

From a theoretical standpoint, Grosjean (2010, pp. 181–183) outlines two contrasting positions on the development of languages in simultaneous bilinguals. On the one hand, the one-system view advocates that children first develop a single linguistic system, with a differentiation occurring at a later stage. Supporting evidence includes early mixing of the two languages, the use of a linguistic rule specific to one language in the other, and the limited overlap of the two vocabularies. On the other hand, the differentiated system view posits that linguistic systems develop separately from the outset. This is supported by early ability to differentiate languages (as evidenced by the use of the interlocutor's language), as well as the ability to differentiate grammatical systems (as when using language-appropriate inflectional morphemes and word order). The mixing of different languages may simply represent codeswitching. On balance, Grosjean argues that a bilingual's languages are not in fusion, but in some kind of contact. As we will see in more detail, it is this contact that leads to cross-linguistic influence.

1.1.2.2 Successive acquisition

Whereas simultaneous acquisition is defined with respect to a precise set of circumstances – exposure to multiple languages at the earliest stages of childhood – successive bilingualism applies to a wider range of scenarios. The acquisition of a new language, in addition to those already spoken, can occur at any point in life. By this definition, the study of successive bilingualism is closely related to second language acquisition (Li, 2013, p. 145).

It is commonly observed that children acquire additional languages with more ease and success than adults; that is why a key issue in successive acquisition is that of age effects. This was

traditionally formulated under the critical period hypothesis (Lenneberg, 1967), which claims that native-like language acquisition could only occur between early infancy and puberty. The argument is based on a supposed link between language acquisition and brain maturation, and specifically the development of hemispheric specialization (p. 179). This is in turn predicated on the notion of brain plasticity, which can be traced back to Penfield and Roberts (1959). They argue that the brain of a child is "plastic", i.e. more easily adaptable, and that it is this adaptability that explains children's superior learning skills compared to adults, including in particular in language acquisition (p. 240).

While subsequent work has confirmed the presence of an age effect, it appears to be more complex than a clearly defined threshold for native-like language acquisition. In an influential study, Johnson and Newport (1989) find that linguistic performance declines linearly as age of acquisition increases, but only until a cut-off point at age 16. The performance is thereafter overall lower, but uncorrelated to age and highly variable across individual speakers. Potential influence of typological distance on these conclusions is underscored by Birdsong and Molis (2001), who use the same experimental setup to examine a typologically closer language pair. Unlike in the initial study, performance is uncorrelated to age in the pre-16 group due to a ceiling effect (it is universally high); in the post-16 group, it declines linearly as age increases. The older group again presents overall lower performance, but interindividual variability is comparatively more limited; moreover, occasional native-like performance is observed. Taken together, observations such as these constitute "evidence against the existence of a simple, clearly bounded, and monotonically developing, critical period" (Li, 2013, p. 149).

It has also been argued that age does not affect the general ability to acquire another language, but rather specific processes associated with it. Liu et al. (1992) observe age effects on the choice of processing strategies in bilingual speakers. These effects arise in interaction with other factors, such as language use at home, and in a nonlinear manner. For instance, in some circumstances early exposure to another language facilitates backward transfer (the use of L2 strategies in L1 processing), despite the received view that it leads to balanced bilingualism. Grosjean (2010) suggests that differences between early and late bilinguals might result from different learning strategies adopted by children and adults. Pointing out that the only clear advantage in early bilinguals is pronunciation, he argues that children are unsophisticated learners, and that any inherent advantages they may have are likely offset by the more complex cognitive mechanisms used by adults. He moreover underscores that simultaneous and successive acquisition are influenced by the same factors: the need to use a language, the amount and type of input, the role of family and school, and attitude (pp. 185–186).

More generally, Birdsong (2018) draws attention to the evidence commonly presented in support of age effects. He argues that the insistence on the supposed inability of late bilinguals to attain native-like performance is biased by the use of monolingual standards to evaluate them. Native-like performance in specific areas of linguistic ability has in fact been reported in some late bilinguals. What neither early nor late bilinguals can do is behave exactly like monolinguals in all regards. This is because all languages a bilingual speaks influence one another, most crucially due to the phenomenon of coactivation, addressed in the next section.

Overall, we have seen that simultaneous and successive bilingualism occur in different

contexts, involve different learning processes, and may result in some differences in the attained linguistic ability. However, other factors, shared among all bilinguals, influence the outcome far more than age of acquisition taken in isolation. Moreover, a large body of research questions the received view that an inherent difference exists between early and late bilinguals, making the former somehow more bilingual than the latter. It is in fact the case that no bilingual can demonstrate monolingual native-like performance in all aspects of linguistic knowledge, as they cannot simply "turn off" a language they are not using. This has implications in terms of language choice and control, as well as cross-linguistic influence. I now turn to these issues.

1.1.3 Language choice and control

A central aspect of bilingual interaction is the choice of the language to be used. According to Grosjean (2013), the way this issue is resolved depends on whether both interlocutors are bilingual or not. A bilingual speaker who interacts with another bilingual faces a complex decision: they can use any one of the languages that they share with the other speaker, or a combination of those. This depends on factors including the participants (their relative proficiency in the shared languages, the language history between them, attitude, demographic factors); situation (the place of the interaction, the presence of monolinguals); topic (as per the Complementarity Principle discussed above); and the function of the interaction (including or excluding an interlocutor, formulating a request, and so on). By contrast, a bilingual speaker who interacts with a monolingual is left with no choice but to use the monolingual's language. However, the interaction may still present traces of common bilingual strategies, such as minimal codeswitching (e.g. if the bilingual speaker is unable to find the appropriate word in the language used in the interaction). Moreover, bilinguals are likely to present cross-linguistic influence at all levels of linguistic structure, whatever the communicative situation (pp. 17–21).

The distinction opposing bilingual and monolingual interactions is paralleled by more general principles of language control. Grosjean (2013) bases his analysis on the notion of language mode, which he defines as "the state of activation of the bilingual's languages and language processing mechanisms at a given point in time" (p. 14). He argues that language mode depends on two decisions taken by the speaker: first, the language to be used, which becomes what he terms the base language; second, whether another language should be brought in. The resulting choice can be conceptualized as a continuum ranging from the monolingual mode, where only one language is activated (e.g. conversation with a monolingual interlocutor), to the bilingual mode, where both are fully activated (e.g. conference interpreting). Various other points exist between the two extremes (e.g. using a single language with a bilingual interlocutor). Bilinguals constantly move along the continuum, including to change the base language. The language not being used as the base language is by definition active in the bilingual mode, but it appears to remain active to some extent even in the monolingual mode (pp. 14–17).

One effect of language coactivation is crosslinguistic influence on different levels of linguistic structure. Serratrice (2012) discusses the interaction observed in bilingual children, showing for example that the syntax of one language can be used as a model in the other. This leads to patterns such as more frequent null realizations of a constituent, the use of constructions

which do not ordinarily exist, or pragmatically or semantically inappropriate use of existing constructions. Cross-linguistic influence has similarly been observed in successive bilinguals. Li (2013) suggests that this may be related to the fact that adults are unable to radically alter an already learned system or develop new categories. This is particularly clear in phonology, where the development of new categories is based on similarity with the native phonological system. In lexical acquisition, a key issue is crosslinguistic discrepancy between corresponding lexical items on the phonological, morphosyntactic, and fine-grained semantic levels. Relatedly, different languages may categorize the world in different ways, which is reflected by an incomplete overlap in the polysemic structure of corresponding lexical items (pp. 151–156).

As for the direction of cross-linguistic influence, it is often exerted by the L1 on the L2, but the reverse is also true. A possible outcome in the latter case is language attrition, or the loss of a language, including in late language learners (Li, 2013, pp. 157–160). More generally, crosslinguistic influence has been reported in both simultaneous and successive bilinguals, and its effect does not appear to diminish in size with age. To this extent, it appears to represent a feature of bilingual ability rather than a developmental phenomenon (Van Dijk et al., 2021).

As we have seen, a bilingual speaker decides which language to use depending on a variety of situational factors, the foremost among them being the linguistic ability of their interlocutor. They may moreover introduce elements from their other languages to varying degrees, as formalized by the notion of language mode. Whatever the specific situation, however, all languages spoken by a bilingual remain active to some extent, which leads to different types of crosslinguistic influence. A specific way that this is reflected in the linguistic ability of bilingual speakers is the mental organization of the bilingual lexicon. This is what I turn to next.

1.1.4 The bilingual lexicon

Following early work which posited the existence of a mental switch which would entirely activate or deactivate a given language (e.g. Penfield and Roberts, 1959; Macnamara and Kushnir, 1971), it is now widely accepted that bilingual communication involves the activation of lexical information from all of a bilingual's languages. This is true across different activities, including reading, listening to spoken language, and speech planning (Kroll and Ma, 2018, p. 295). Different theoretical models have been put forward to explain the mechanisms at play. I will briefly review some of the key positions, as they provide potential explanations for semantic interference in bilinguals.

An influential theoretical proposal from the standpoint of language production is the Revised Hierarchical Model (RHM; Kroll and Stewart, 1994). It posits multiple levels of representation in a bilingual's mind: concepts are stored in a language-independent abstract memory system, and words are stored in language-specific lexical memory systems. Bilingual memory is organized using (i) conceptual links, which associate concepts with lexical items in either of the languages, and (ii) lexical links, which directly relate lexical items between the different languages, without concept mediation. The strength of the associations depends on the speaker's proficiency in the languages they speak and their relative dominance. It is assumed that the L1 mostly accesses meaning directly (using conceptual links), whereas the L2 does so

via L1 translation equivalents (using lexical links), at least at lower levels of L2 proficiency.

But this does not explain the way in which control over languages is ensured: how does a bilingual choose the word in the right language rather than its translation equivalent? One theoretical account of this issue is the inhibitory control model (Green, 1998). Similarly to RHM, it presupposes a separation of conceptual and lexical levels in the bilingual memory: conceptual representations are associated with lemmas, which are in turn specified using a language tag. Relying on general attentional mechanisms, this model argues that the goal of producing an utterance in a given language alters the activation levels of different representations. This results in the inhibition of lemmas with the wrong language tag, ensuring the correct output.

Further evidence of semantic representations shared across languages comes from research on bilingual visual word recognition. For instance, the Bilingual Interactive Activation model (BIA+; Dijkstra and van Heuven, 2002) argues that word recognition proceeds in a bottom-up fashion. It starts with the activation of sublexical and then lexical orthographic representations. These are word candidates competing for selection, which in turn activate the related phonological and semantic associations. This process is language non-selective: the activation of word candidates is based on their similarity with the input string rather than the language to which they belong. Language identification occurs at a later stage.

While theoretical models differ in the adopted perspective and the specifics of the described mechanisms, they all suggest that semantic representations are shared across languages, and that all of a bilingual's languages are always activated to some extent. These claims are also broadly supported by experimental evidence. In a review focusing on bilingual semantic representations, Francis (2005) shows that research on episodic memory is indicative of the use of shared memory stores to represent words from all of a bilingual's languages. This is supported by observations such as the fact that recalling a word in a specific language, seen in a mixed-language word series, is more difficult than recalling a word from a single-language series. Similarly, experimental research into semantic and conceptual systems has repeatedly highlighted the existence of shared semantic representations. This is illustrated by trends such as comparable lexical decision times in semantic comparisons within and across languages.

Coactivation of lexical information is also supported experimentally, as discussed by Kroll and Ma (2018). Bilinguals do not appear to be able to ignore the language not being used, be it in processing (of both isolated words and words in sentence context) or in production. This points to an integrated bilingual lexicon, as well as a language selection mechanism which does not appear to be context-based or to operate in a top-down manner. The specific way in which language control is ensured varies depending on the speaker's relative proficiency in different languages. However, crosslinguistic influence persists even among highly proficient bilinguals.

It should be noted that methodological reservations have been expressed regarding some of these conclusions. For instance, it has been argued that some experimental evidence may simply be consistent with coactivation rather than explicitly supporting it (Costa et al., 2006). Attention has also been drawn to the fact that many studies do not put bilingual speakers in a fully monolingual mode. This is the condition which would demonstrate the full extent of inherent coactivation, as opposed to that arising from the simultaneous use of multiple languages in bilingual mode (Grosjean, 2013, pp. 16–17). However, this debate is beyond the scope of the

present discussion, which has nevertheless underscored this key point: semantic representations appear to be shared across languages, and this provides a basis for crosslinguistic influence.

1.2 Social dimensions of bilingualism

I have so far discussed how individual speakers acquire and use multiple languages. On the societal level, this bilingual ability intersects with community membership. As Romaine (2012) points out, all speakers belong to multiple communities, many of which are related to language use. These are known as language communities, which she defines, following Baker and Prys Jones (1998, p. 96), as "those who use a given language for part, most, or all of their daily existence" (p. 446). By virtue of using multiple languages in their everyday life, bilingual speakers belong to a variety of language communities. Some of these communities are likely monolingual, and others may themselves be bilingual (pp. 446–447).

The participation of bilingual speakers in different communities depends on a variety of factors. These include the reasons behind the development of bilingualism on the societal level, the way in which different language communities come into contact, and the implications bilingualism has in terms of identity. All of these aspects have the potential to influence processes of language variation and change. I will therefore briefly review each of them in turn.

1.2.1 Development of societal bilingualism

Bilingualism is a remarkably widespread phenomenon. This is true both on the individual and on the societal level: the majority of the world's population speaks multiple languages, and bilingualism is present in practically every country (Grosjean, 2010, p. 13). But considerable differences exist in the role of bilingualism in different communities. A first step in understanding this consists in looking at how different language communities come into contact.

Different sociohistorical reasons may give rise to the development of societal bilingualism. Edwards (2012, pp. 7-8) identifies several frequent pathways:

- immigration (of settlers or invaders), including in limited numbers, as in the case of colonial expansion;
- political unions among different groups of speakers, such as between English-speaking and French-speaking Canadians;
- cultural and educational motivations.

The contribution of immigration to language contact is further discussed by Sankoff (2002, pp. 4-5), who argues that it usually results in a rapid assimilation of the immigrant group. Its language is particularly affected by this process: short periods of contact suffice for the integration of borrowings, whereas structural language change is usually observed if contact lasts over multiple generations. The language of the immigrant community can also exert influence on the locally spoken one, although it is usually limited unless the incoming population is demographically or socially dominant.

A speech community only exists to the extent that people continue to speak its language, so it is also worth examining the community factors impacting the development of individual bilingualism. For example, Pearson (2007) underscores the importance of a cohesive community of heritage language speakers, as it provides motivation for acquisition as well as a concrete opportunity to use the language in question. She also argues that education in the immigrant language plays an important role, particularly as it can offset a potential lack of input in the family setting.

If conditions such as these are not provided, the loss of the immigrant language can occur in what is a relatively swift process. It was traditionally considered that the first generation of immigrants would remain strongly dominant in their native language, the second generation would be bilingual, and the third generation would be monolingual in the new language (Veltman, 1988). Subsequent studies have suggested that, rather than reflecting a linear generational trend, sustained knowledge of the immigrant language was better explained by more complex factors such as the use of the immigrant language at home and the distance from the social network in the country of origin (Hakuta and D'Andrea, 1992).

Despite differences in the specifics, all of these findings underscore the dynamic and often precarious nature of language communities in which bilingual speakers participate. This leads me to look at what kinds of language communities exist, and what indices can be used to describe their status.

1.2.2 Status of language communities

Sociohistorical factors such as the ones discussed above may lead to different expressions of bilingualism on the societal level. Wei (2012, pp. 30–31) identifies three main types:

- territorial bilingualism involves multiple languages whose speakers are mostly confined to their respective territories. The territories are defined both geographically and politically, and provide an official status to the language spoken by their community. Other languages may also be used, but without an official status. This is the case of Canada;
- diglossia refers to the coexistence of multiple languages within a single community of speakers. However, the languages are used in a complementary way and have a different social status. The relationship between the languages may evolve over time;
- widespread multilingualism corresponds to the situation where numerous languages are used by different groups of speakers and coexist with one or more languages of wider communication. Most members of these communities are highly bilingual.

A more specific type of bilingualism, particularly relevant for sociolinguistics, is that of minority language communities or linguistic minorities. This notion may refer to demographic as well as social or political status. A further general distinction is frequently made between indigenous (autochtonous) and non-indigenous (immigrant or migrant) minorities. However, this is often contentious because it is debatable how long it takes for a community to be considered as indigenous after its first arrival in a territory (Romaine, 2012, pp. 450–452). Studying the issues surrounding language use in these communities can be especially insightful given that

"conflicts involving language are not really about language, but about fundamental inequalities between groups who happen to speak different languages" (Romaine, 2012, p. 463).

The place a linguistic minority occupies in a society can be gleaned from the geographical distribution of its members. Edwards (2012, p. 8) outlines a typology based on a series of distinctions first discussed by White (1991):

- unique minorities (unique to one state, e.g. the Breton minority in France), non-unique minorities (present in multiple states, but subordinate in all of them, e.g. the Basque minority in Spain and France), and local-only minorities (a minority in a specific setting, but a majority elsewhere, e.g. the French community respectively in Canada and in France);
- if the same language community exists in different states, it can be adjoining (geographically contiguous, as in the case of the Basque minority) or non-adjoining (e.g. the French community in Canada and France);
- in terms of spatial cohesion within a single state, a language community can be cohesive (e.g. the Cree community in Canada) or non-cohesive (e.g. the Spanish community in the United States).

However, Edwards (2012, p. 9) also argues that these dimensions only provide a general overview of the situation. This can be complemented by analyzing the intersections of three key categories of variables – speakers, language, and setting – with various disciplinary perspectives – demography, sociology, linguistics, psychology, and so on.

One such approach, developed by Giles et al. (1977), consists in studying the ethnolinguistic vitality of language communities, defined as "that which makes a group likely to behave as a distinctive and active collective entity in intergroup situations" (p. 308). Vitality is analyzed based on three groups of variables:

- status variables, which are reflective of the group's prestige in the intergroup context (e.g. the group's economic influence, the representations associated with its language);
- demographic variables, which directly indicate how many members the group has (e.g. birth rate) and how they are distributed geographically (e.g. the proportion of speakers relative to the outgroup);
- institutional support variables, which account for the group's representation across a range of institutions (government, industry, media etc.).

Broadly speaking, ethnolinguistic vitality is positively associated with high prestige, favorable demographic trends, and strong institutional support. A high degree of vitality is in turn associated with the group's survival as a collective entity; conversely, if vitality is limited, the group may cease to exist (Giles et al., 1977, pp. 308–318).

While these approaches provide a global overview of language communities, they are limited in describing the position of the individual speakers who comprise them. This is what I turn to next, focusing specifically on potential relationships between bilingualism and identity.

1.2.3 Bilingualism and identity

A key aspect of the definition of language community is "the sense of perceived solidarity and interaction based on reference to a particular language and the relationships among people who identify themselves as members of that community" (Romaine, 2012, p. 447). Similarly, Edwards (2012) associates language with the notions of allegiance and belonging. This is due to the role that language plays in transmitting a group's tradition and culture, which in turn has ramifications for identity (p. 19).

To be sure, bilingualism does not outright determine the identity of all bilingual speakers. However, it may have an impact, particularly for speakers who are more deeply linguistically and culturally integrated into multiple groups. This usually translates to a coherent identity which reflects the different groups, even though in most cases one language and culture remain dominant on the psychological and emotional level. Moreover, the relationship between language and identity is rarely problematic for monolingual majority groups, as the communicative and symbolic value of language coincide: the same language is used to, say, communicate in everyday contexts and to transmit cultural traditions. But the reverse is usually true for minority language communities. Although bilinguals do not inherently constitute minorities, many are in similar situations as members of groups which are not socially dominant (Edwards, 2012, pp. 19–23). That is why "a link will often exist between bilingualism and a heightened awareness of, and concern for, identity" (p. 23).

The association of identity with the use of multiple languages and membership of multiple cultures is also related to the notion of biculturalism. Bicultural individuals can be defined based on the following criteria:

> Firstly, they take part, to varying degrees, in the life of two or more cultures. Secondly, they adapt, at least in part, their attitudes, behaviours, values, languages, etc., to these cultures. Thirdly, they combine and blend aspects of the cultures involved. (Grosjean, 2015, p. 575)

Biculturalism develops through contact with the different cultures, in childhood or in later life. It often, but not always, develops at the same time as bilingualism. Much like bilingualism, it can evolve over time, with the relative dominance of cultures changing depending on a variety of factors (e.g. immigration, work, romantic partners). Bicultural behavior is similar to language modes: biculturals are situated at different points on a continuum, ranging from a monocultural extreme (exclusive use of a single culture) to a bicultural extreme (e.g. interaction with other biculturals with switches from one culture to another). It is more difficult to deactivate a culture than a language; some culture blending may therefore persist even in the monocultural mode (e.g. discrepancies in terms of eye contact, distance between interlocutors etc.). While identity is not central to biculturalism, it is related to it, as biculturals often have trouble accepting their belonging to multiple cultures. In the case of bilingualism, this trend is less pronounced, but remains present: for instance, some people avoid labeling themselves as bilinguals because of an outdated perception of what bilingualism is (Grosjean, 2015).

Furthermore, bilingualism is an object of attitudes on the part of both bilingual and monolingual speakers, as outlined by Grosjean (2010, ch. 9). Bilinguals associate bilingualism with

a range of advantages. These include evident communicative benefits – interacting with people from different countries and cultures, having access to different literary traditions – as well as a social and cultural dimension, both instrumental (e.g. more job opportunities) and symbolic (e.g. a different perspective on life). The disadvantages bilinguals most often perceive are related to difficulties in language use: for instance, in the case of dominant bilingualism, it can be tiring to use the non-dominant language, frustrating to make mistakes, and so on. Another commonly perceived disadvantage has to do with identity, and in particular the feeling of not belonging to any cultural group. As for monolingual speakers, their views of bilingualism range from very positive to very negative. A key factor is the socioeconomic status of the bilingual speaker: positive attitudes are associated with speakers of higher status, and negative attitudes, with speakers of lower status. The latter is particularly true of groups such as immigrants and language minorities, especially when speaking the majority language with an accent (Grosjean, 2010, pp. 97–105).

Different methodological approaches have been proposed to examine the link between bilingualism and identity. As Wei (2012) notes, sociolinguistics generally investigates it by analyzing bilingual speakers as social actors, and bilingualism itself as a socially constructed phenomenon. Here, linguistic means of expressing identity, such as language choice, are taken to position the speaker within a broader sociohistorical context. A closely related idea is that of negotiating identities, which can be traced back to research in social pscychology. This includes the previously discussed ethnolinguistic approach (Giles et al., 1977), which provides a clear overview of some aspects of bilingualism at the societal level. However, a key shortcoming of views such as these is the introduction of rigid links between language and identity. They moreover often embed a monolingual, monocultural bias, as exemplified by the focus on the opposition between ingroup and outgroup members (Wei, 2012, pp. 43–44).

One more recent approach, attempting to overcome these issues, has been formulated by Pavlenko and Blackledge (2004). In investigating how identities are negotiated, they emphasize the fact that speakers have multiple identities which are moreover dynamic. The process of identity negotiation is analyzed with respect to power relations in a wider sociohistorical context, and it is understood to involve the appropriation of languages (p. 10).

In addition to highlighting societal aspects, this approach points to the fact that bilingualism is associated with specific linguistic behaviors. The next section focuses on two such types of behavior: codeswitching and lexical borrowing. It then analyzes how individual behavior can lead to community-level language change.

1.3 Linguistic manifestations of bilingualism

Bilingual speakers have at their disposal a variety of linguistic structures, which are not necessarily limited to the use of elements from a single language. Like other speakers, bilinguals are linguistically socialized to choose context-appropriate forms based on a range of factors, including interlocutors, topics, and institutional settings. In some contexts, the use of elements from multiple languages is allowed or even desirable; if such a pattern of language use acquires

currency in the linguistic community, it may lead to language change (Matras, 2009, p. 4).

Bilingual speakers are particularly likely to introduce elements from one language into another when communicating with other bilinguals. There are two main ways in which they can go about this: codeswitching and borrowing (Grosjean, 2010, p. 51). This can be distinguished from another related phenomenon, which occurs in monolingual communication and in which the language being spoken is influenced by a deactivated language. Two types of influence can be identified: transfer (permanent effects of one language on another, e.g. pronunciation influenced by the dominant language) and interference (occasional effects of one language on another, e.g. a slip leading to a semantically inappropriate use of a lexical item) (Grosjean, 2012).

In this section, I will address the linguistic behaviors specific to bilingual communication: codeswitching and borrowing. They are particularly relevant in studying lexical manifestations of bilingualism, and will provide a basis for a discussion of how individual behaviors can translate to community-level language change. This will moreover clarify the theoretical distinction between bilingualism and language contact. The finer-grained effects mainly observed in monolingual communication, such as interference, will be of particular importance in our subsequent analyses focusing on lexical semantics. They will be addressed in more detail in Chapter 3.

1.3.1 Codeswitching

Grosjean (2010, pp. 51–52) defines codeswitching as "the alternate use of two languages, that is, the speaker makes a complete shift to another language for a word, phrase, or sentence and then reverts back to the base language". Codeswitched elements of an utterance are linked together on the prosodic, syntactic, semantic, and pragmatic level (Romaine, 1995, p. 121). Although negative attitudes to codeswitching are often expressed by both monolinguals and bilinguals (Grosjean, 2010, p. 52), it occurs frequently and naturally in bilingual discourse (Romaine, 1995, p. 121).

The importance of codeswitching is related to the range of functions it fulfills. In conversation with other bilinguals, it often responds to a linguistic need, such as finding the most adequate way of expressing a notion or a concept, or reporting speech originally heard in the other language. It can also be used to fulfill a communicative or social goal, such as positioning the speaker as a member of a community or excluding somebody from the conversation (Grosjean, 2010, pp. 53–55). Although it may appear counterintuitive, codeswitching also occurs in communication with monolinguals. Frequent reasons include introducing a proper noun from the other language, filling a lexical gap (especially for highly dominant bilinguals), and addressing a topic usually discussed in the other language (pp. 66–67).

In discussing the communicative goals alluded to above, Romaine (1995) follows Blom and Gumperz (1972) in drawing a distinction between transactional (non-situational) switching, which is controlled by components of the speech event, such as topic and participants; and metaphorical (situational) switching, which is related to the desired communicative effect. Following Gumperz (1982), she also identifies specific discourse functions with which

codeswitching can be used: delimiting direct and reported speech; introducing interjections or sentence-fillers; qualifying the message, e.g. by introducing a topic; specifying an addressee (switching to a monolingual's language, drawing a bilingual's attention); marking personalization or objectivization (e.g. lending more authority to one's words) (Romaine, 1995, pp. 161–165).

From a structural standpoint, three types of codeswitches are usually described:

- tag switches, where an other-language element such as an interjection is inserted in the sentence without affecting its syntactic structure;
- intrasentential switches, consisting in the insertion of other-language elements within the structure of another sentence or a constituent;
- intersentential switches, where an entire clause or a major sentence constituent is produced in the other language (Poplack, 2015, p. 918).

It has been suggested that these types of codeswitches are also functionally different. For instance, Poplack (1980) contends that interactional effects like those posited by Gumperz are brought about by what she terms emblematic codeswitching, which involves strategies such as tag switching and hence does not require much bilingual skill. By contrast, intrasentential switching, which requires a high degree of bilingual ability, represents a discourse mode in its own right and thus potentially constitutes a specific part of the bilingual repertoire (pp. 613–614). Nevertheless, all three types of codeswitching may occur within the same discourse. They can moreover be analyzed in terms of a continuum, ranging from whole sentences to isolated words, with larger spans of discourse between the two endpoints (Romaine, 1995, pp. 123–124).

In analyzing structural constraints of codeswitching, much attention has been dedicated to intrasentential codeswitching. A crucial observation is that it does not occur at random, but is rule-governed. Moreover, it can take two major forms, involving respectively the introduction of a lone content word or of a multiword fragment; this does not generally extend to lone grammatical elements (Poplack, 2015, pp. 918–919). A range of theoretical approaches have been proposed to account for these observations.

One well-known view is formulated under the Equivalence Constraint (e.g. Poplack, 1980). Focusing on the linear structure of the codeswitched utterance, it predicts that intrasentential codeswitching will occur at a point where the introduction of elements from another language does not violate the surface syntactic structure of either language, i.e. where the two languages structurally correspond to one another. The resulting utterance is grammatical by the standards of both involved languages (pp. 586–588). One important criticism of this approach is that it assumes equivalence of grammatical categories between languages. That is not always the case, particularly for typologically distant language pairs (Romaine, 1995, pp. 128–129).

Another influential approach is the Matrix Language Frame model (e.g. Myers-Scotton, 2002). Under this view, codeswitched utterances are produced through an interaction between a matrix language, which determines word order and provides grammatical elements, and an embedded language, which provides content elements. It is assumed that surface morpheme order, as well as all system morphemes with grammatical relations external to their head constituent, come from the matrix language; this is used as a criterion in identifying it. However,

Romaine (1995, pp. 134–137) argues that the identification of the matrix language is often not as straightforward, with a potential solution involving an extension of the analysis to surrounding utterances.

A view closer to variationist sociolinguistics, Poplack (2015) argues, is the one advocated for example by Schindler et al. (2008) and grounded in the the Optimality-Theoretic approach to codeswitching (Bhatt, 1997). It analyzes codeswitching as a resolution of conflicting constraints; these are defined in a small set, universal, and allow for different violations depending on the language combination. This position is related to variationism because it posits that, for general constraints, applicability is universal, instantiation varies depending on the language pair, and implementation is variable (Poplack, 2015, p. 920).

Theories on codeswitching abound and consensus remains elusive, but some takeaways are undisputed. Codeswitching is part and parcel of bilingual behavior. Socially and discursively, it can be motivated by different factors and serves a variety of purposes. Structurally, it is not random but rule-governed. However, a major stumbling block in the analysis of these rules is the treatment of lone other-language items due to the inherent similarity with the notion of borrowing; this is the focus of the next section.

1.3.2 Borrowing

According to Grosjean (2010), "unlike code-switching, which is the alternate use of two languages, borrowing is the integration of one language into another" (p. 58). As we will see, the surrounding theoretical discussion is far more complex, but this definition is a good starting point. Two types of borrowing can be distinguished:

- loanwords or nonce borrowings, where both form and content come from one language, and are integrated into another. The most easily borrowed elements are nouns, followed by verbs and then adjectives, while other parts of speech are borrowed much less frequently;
- loanshifts, which involve (i) extending a word's meaning so that it corresponds to the meaning of a word in the other language, or (ii) reproducing a surface pattern from another language in order to convey a new meaning (also known as calque or loan translation) (Grosjean, 2010, pp. 58–60).

The term borrowing has also been used to describe cross-linguistic influence on other levels of linguistic structure, such as phonology and morphosyntax (e.g. Matras, 2009, ch. 8). This is related to a more general observation that lexical borrowing may in turn trigger other types of language change, most notably on the phonological level (Sankoff, 2002, p. 658). However, our focus will remain on lexical borrowing. The case of loanshifts directly involves a lexical semantic dimension; it will be discussed in more detail in Chapter 3, and will constitute the core of this dissertation. Here, I turn to loanwords.

Just like codeswitching, borrowing can be motivated by linguistic need, such as finding the most precise word or discussing a domain which is usually addressed in the other language. Referential effects can also be observed, particularly in the case of immigrants whose L1 words

cannot adequately denote the realities in their new country (Grosjean, 2010, pp. 60–61). In addition to these, Matras (2009) discusses two other borrowing scenarios. Some borrowings are related to prestige, like when a word is borrowed despite there being an equivalent in the recipient language. This process has little to do with denotative needs; rather, the borrowed element conveys social value, such as reflecting the social position of the speakers of the donor language or distancing itself from another community. Other borrowings are related to cognitive pressures of bilingual processing. It is specifically hypothesized that the production of cognitively demanding linguistic structures may reduce the ability to keep the other language inhibited, which in turn facilitates borrowing (pp. 149–152).

All of the described cases involve the presence of lone other-language items, i.e. isolated lexical items from one language appearing in a span of speech produced in another language. The analysis of these items constitutes a central theoretical and empirical issue in research on language contact: it is often the adopted position that determines if an item is considered as an instance of codeswitching or borrowing. This is further related to the distinction between the general process of borrowing, including when it occurs on the spot, and the development of established loanwords. This in turn has to do with the status of other-language items in bilingual speech, as well as the link between synchronic variation and diachronic change (see Section 1.3.3).

On the distinction between codeswitching and borrowing, an influential position is the one first developed by Poplack et al. (1988) and refined in subsequent studies (e.g. Poplack and Meechan, 1995; Sankoff et al., 1990). A central tenet of this approach is

> the clearcut conceptual distinction between borrowing, in which an L_2 lexical item submits to L_1 morphological and syntactic rules in L_1 discourse, and code switching, in which each monolingual fragment is lexically, morphologically, and syntactically grammatical in one language. (Poplack et al., 1988, p. 93)

Put otherwise, the key characteristic distinguishing borrowing from codeswitching is the morphosyntactic integration of the borrowed item into to the recipient language. It may be accompanied by some degree of phonological integration, but this is not seen as a defining feature of borrowing (Poplack et al., 1988, p. 96).

This line of research moreover claims that the integration-based distinction between borrowing and codeswitching holds whether an other-language item is borrowed on the spot or is an established loanword. Known as the Nonce Borrowing Hypothesis, this view has been summarized as

> captur[ing] the empirical observation that speakers not only code-switch spontaneously, but may also *borrow* spontaneously, and these spontaneous borrowings assume the morphological and syntactic identity of the recipient language even *prior* to achieving the social characteristics of established loanwords (recurrence in the speech of the individual, and dispersion across the community). (Poplack, 2012, p. 645)

In other words, if an other-language lexical item is used for the nonce, it is considered as a borrowing so long as it is morphosyntactically integrated into the recipient language; otherwise,

it is considered as a single-word codeswitch. A small proportion of nonce borrowings may over time become established in the speech of the initial speaker, of other bilingual speakers, and eventually even monolinguals. These cases correspond to established loanwords. Their status relative to nonce borrowings can be determined from data including historical attestation, frequency of use, and phonological integration (Poplack et al., 1988, p. 96). More generally, the distinction between nonce borrowings and established loanwords reflects the difference between individual bilingualism and lexical transmission at the social level (Poplack et al., 1988, p. 93–94).

Various aspects of the approach developed by Poplack and her associates have been questioned. As concerns the distinction between codeswitching and borrowing, a key issue is related to its reliance on morphosyntactic integration. While it can be readily operationalized in languages with rich inflectional morphology, in languages where that is not the case this type of analysis is exceptionally challenging (Romaine, 1995, p. 151). Poplack (2015) herself states that it is difficult to establish this distinction in "inherently language-neutral constructions", but she also underscores the importance of word order as an indicator (pp. 922–923).

Another criticism leveled at this approach is related to the distinction between nonce borrowings and established loanwords. A case in point is the study by Stammers and Deuchar (2012), which investigates the integration of listed (attested) and non-listed (nonce borrowed) English verbs into Welsh, focusing on the realization of the morphosyntactic phenomenon known as soft mutation. They find that it is associated with frequency, occurring significantly more often in listed verbs. They interpret this finding as a refutal of the Nonce Borrowing Hypothesis, which predicts that morphosyntactic integration would occur even at the nonce stage, i.e. independently of frequency. However, Poplack (2012) argues that these results in fact confirm the Nonce Borrowing Hypothesis, to the extent that the same patterning was reported in monolingual and bilingual data, albeit at a different rate. This references the idea that the relevant benchmark in determining if an item is morphosyntactically integrated is comparison with the distribution of the same morphosyntactic pattern in the language in question in the absence of other-language content (Poplack, 2015, p. 922).

While the theoretical approach outlined so far, and the debate surrounding it, has dominated variationist sociolinguistic work on code-switching, other standpoints have also been defended. For instance, Myers-Scotton (2002) argues that within the Matrix Frame Language model it is not strictly necessary to describe borrowings as distinct from codeswitching. Just like longer stretches of other-language material, established loanwords as well as non-integrated borrowings can be viewed as codeswitches within the morphosyntactic frame of the matrix language. Crucially, morphosyntactically integrated other-language items, which Poplack would analyze as borrowings, are seen here as involving codeswitching within a single constituent (e.g. the base belonging to one language, and the affix to the other) (Myers-Scotton, 2002, pp. 153–155).

A less clear-cut position is taken by Matras (2009), who suggests that codeswitching and borrowing are best analyzed as a continuum along multiple dimensions. These include the speaker's degree of bilingualism, the linguistic composition of other-language material, the regularity of occurrence, and so on. A prototypical case of borrowing is represented by the

"regular occurrence of a structurally integrated, single lexical item that is used as a default expression, often a designation for a unique referent or a grammatical marker, in a monolingual context" (Matras, 2009, p. 113). A prototypical example of codeswitching is an "alternational switch at the utterance level, produced by a bilingual consciously and by choice, as a single occurrence, for special stylistic effects" (pp. 113–114). All other cases, situated between these two extremes, are fuzzy to some extent (pp. 110–114).

Echoing the notion of continuum, Romaine (1995) raises the issue of compromise forms. These are lexical items whose partial phonetic similarity across languages makes it difficult to judge from which language the attested form comes. An example is the English preposition *of* and its Dutch homonym *of* 'or'. The Dutch element is attested with the English meaning, but it is difficult, if not impossible, to determine if this involves a lexical or a semantic transfer from English, as first described by Clyne (1987, p. 755). This problem is compounded by the fact that forms seen as unacceptable in isolation may be seamlessly integrated in conversational context (Romaine, 1995, pp. 151–152).

In summary, we have seen that bilingual speakers frequently introduce lone other-language items in communication. This can be motivated by different factors, including linguistic need, prestige, and cognitive load in bilingual processing. In addition to studying the process by which an initial integration of an element can spread and lead to the formation of an established loanword, an impressive amount of work has gone into determining whether such items constitute instances of borrowing or of codeswitching. It has been argued (e.g. Poplack, 2012) that this is not a theoretical but an empirical question, which can be settled by evaluating morphosyntactic integration into the recipient language using a diagnostic appropriate for the language pair under study. But when push comes to shove – when there is insufficient inflectional morphology to evaluate integration, as in the case of English-French bilingualism – this does seem to be a theoretical issue, to the extent that it is settled differently by different theories.

We have so far seen how individual speakers can introduce elements from one language into the other. But the impact of these practices in a larger scheme of things remains unclear; this is what I turn to next.

1.3.3 From bilingualism to language contact

A hallmark of bilingual communicative behavior is the ability to combine elements from multiple languages in a single utterance in different ways. In addition to representing a specific type of communication typical of bilingual speakers, these phenomena have a broader interest. As mentioned at the beginning of this chapter, it is behaviors such as these that can give rise to contact-induced language change.

The link between individual and community-level bilingual patterns has long been underscored. For instance, Mackey (1962) considers that language contact is related to "the direct or indirect influence of one language on another resulting in changes in 'langue' which become the permanent property of monolinguals and enter into the historical development of the language". By contrast, "bilingualism is not a phenomenon of language; it is a characteristic of its use. [...] It does not belong to the domain of 'langue' but of 'parole'" (p. 51). A similarly

Saussurean distinction is present in the opposition between speech-level interferences, which appear in the utterances of a bilingual speaker as a result of that speaker's personal knowledge of multiple languages; and language-level interferences, which originate in the same type of behaviors, but become established in the wider speech community and are no longer dependent on individual bilingualism (Weinreich, 1953, p. 11).

While this is reminiscent of the distinction that Grosjean (2012) draws between static and dynamic interferences, the dimension at play is not the same. Grosjean's focus is on whether cross-linguistic influence is related to temporary or permanent effects that one language has on another, but in both cases this is observed on the level of the individual speaker. The speech-level and language-level phenomena described by Weinreich (1953) respectively correspond to the start and end points of the process of contact-induced change described by Matras (2009). As he additionally points out, in this context, individual behaviors can be seen as synchronic, and structural patterns as diachronic phenomena (p. 1).

Moreover, Sankoff (2002) notes that contact-induced change is different from the change occurring within monolingual communities. In order to describe these patterns, research on individual speakers must be complemented with analyses of community-level mechanisms, accounting for both sociohistorical and language-internal constraints (pp. 638–641). One way of doing so is to adopt a variationist sociolingusitic perspective, in which the analysis of language contact

> involves the study of linguistic processes by which forms from two or more languages may be combined as a result of their common use, the linguistic constraints on such combination, and its consequences for the structure of the languages involved. We have also sought to ascertain the social meaning of language choice as exemplified by speaker 1) behavior, 2) attitudes, and 3) perceptions. (Poplack, 1993, p. 254)

This approach is mainly concerned with conventional bilingual interaction. Sociolingusitic studies therefore tend to focus on speakers whose bilingualism is stable, in order to limit the influence of processes such as language acquisition and attrition. Speakers of varying degrees of bilingualism are included in these studies so long as they are considered to be well-integrated members of the speech community. That being said, their bilingual ability remains an important explanatory variable (Poplack, 1993, p. 255). Further discussion of the variationist sociolinguistic approach to language contact, and particularly the way it informs the methodology developed in this work, will be presented in Part II.

1.4 Summary

On a general level, bilingualism is understood as the use of two or more languages in everyday life; in the Quebec context, this corresponds to the speakers who regularly use (at least) English and French. In the present study, no specific type of bilingualism, degree of bilingualism, or manner of bilingual acquisition will constitute an inclusion requirement. Rather, all bilinguals are taken to be valuable members of speech communities; moreover, we have seen that both

simultaneous and successive bilinguals, for example, can attain comparable degrees of linguistic ability and present similar patterns of cross-linguistic influence. This is the view adopted in the definition of Quebec's speech communities, presented in Chapter 2. That being said, all aspects of a bilingual's language history will be described in detail wherever possible and will constitute a set of potential explanatory variables in the coming analyses. Particular attention will also be directed to the contextual factors which may influence bilingual interaction. The precise way in which I implement this approach is presented in Chapter 8, for corpus-based analyses, and in Chapter 12, for the sociolinguistic interviews.

On the societal level, we have seen that the development of bilingualism is related to a range of sociohistorical factors, and that the status of speech communities in which bilinguals participate is dynamic and often precarious. Because of this and other reasons, bilingualism is closely related to identity. These aspects will be taken into account in defining the status of the Quebec English community on the whole (see next chapter). They will also be used in interpreting the communicative practices of the individual speakers who constitute the community; with respect to the sociolinguistic interviews, a general overview of the reported identities is provided in Chapter 13.

In addition to being associated with sociolinguistic factors, bilingual communicative practices manifest themselves in a variety of linguistic patterns. In particular, I have discussed the ways in which elements from different languages can be combined in a single utterance, focusing particularly on codeswitching and borrowing. I have also suggested that individual occurrences of these linguistic practices, produced by individual speakers, can spread in the speech community and constitute patterns of language variation and change. These issues are discussed more extensively throughout Part II.

The remainder of this work will focus on one particular linguistic practice, underpinned by cross-linguistic semantic influence. My focus on this issue is grounded in the previously discussed fact that the languages in a bilingual's brain are in constant interaction. The bilingual's lexicon is therefore not considered to be language-independent; rather, semantic representations are shared across languages. Consequently, the meanings associated with lexical items in different languages can interact in the mental lexicon of a bilingual speaker. In Chapter 3, I will address this phenomenon from the standpoint of community-level patterns, described by sociolinguistics and other related disciplines, as I aim to define the notion of contact-induced semantic shifts. But first, let us turn to the more general background in which this behavior takes place, and which can help us better understand its importance.

Chapter 2

Language contact in Quebec

In the last chapter, we saw that key aspects of bilingual communication are firmly rooted in the specific context of the language community under study. This dissertation focuses on the use of English in Quebec; it is therefore essential to better understand its sociohistorical profile as well as the main characteristics of the languages that are spoken there. This chapter begins with a brief overview of Quebec's history, linking its key stages to the development of its linguistic communities, and then illustrating its present-day demolinguistic composition (Section 2.1). It then draws on existing sociolinguistic research to present the main features of Quebec French (Section 2.2) and Quebec English (Section 2.3), in the latter case also summarizing the existing accounts of contact-induced semantic shifts in Quebec English. It finally concludes with a brief summary (Section 2.4).

This chapter will point to defining aspects of contact-related semantic influence in Quebec English, which will be theoretically refined in Chapter 3. It will also provide important background for both internal and external factors which might account for these patterns of language variation, and which will be explored in more detail in Chapter 6. More immediately, it will highlight a range of sociodemographic characteristics enabling me to propose a definition of Quebec's language communities. This will guide data collection and analysis implemented using both computational (Part III) and variationist sociolinguistic (Part IV) approaches. Note finally that in the forthcoming discussion I will use the term Francophone to denote native French speakers, Anglophone to denote native English speakers, and Allophone to denote native speakers of a language other than English and French (cf. Lepage, 2020, p. 5).

2.1 Sociohistorical context

Quebec is one of Canada's thirteen provinces and territories. With a surface of around 1.5 million km^2, it is the second largest in size, behind Nunavut (Statistics Canada, 2016); with 8.5 million inhabitants, it is the second most populous, behind Ontario (Statistics Canada, 2022). To put this into a European context, Quebec is three times the size of mainland France, but it only has an eighth of its population (Insee, 2022). Like elsewhere in Canada, the vast majority of that population lives in the southernmost belt bordering the United States; around half of it is concentrated in the area surrounding Montreal, the largest city in Quebec and the second

FIGURE 2.1: Map of Canada indicating the position of Quebec and main population centers.[1]

largest in Canada, behind Toronto (Statistics Canada, 2022).

Quebec must be understood within its wider North American context (Figure 2.1), where it is enveloped in the cultural and linguistic influence exerted both by the rest of Canada and by its more powerful and populous southern neighbor. Quebec stands out in this picture in one important way: a large majority of its inhabitants are native speakers of French, a linguistic island surviving – even thriving – in a continent dominated by 300-odd million English speakers. But within this enclave of sorts lies another one. Quebec's English-speaking community constitutes a demographic minority within the province, and is in intense everyday contact with French. A minority situation such as this one is of particular relevance for sociolinguistics, as it can be expected to facilitate contact-induced linguistic behaviors like those discussed in Chapter 1.

However, the context is more complex than it may first seem, since the provincial French-speaking majority constitutes a country-level and continent-level minority. This has led to persistent tensions reported by both linguistic communities, which partly condition sociolinguistic behaviors in Quebec and confer them social meaning. They are also the result of longstanding historical trends, which are addressed in the next section; this will be followed by a more detailed overview of the current demolinguistic profile of Quebec.

[1]Map adapted from https://commons.wikimedia.org/wiki/File:Quebec_in_Canada_2.svg.

2.1.1 History of Quebec

As alluded to above, Quebec constitutes an exception to the general linguistic profile of most other Canadian provinces and territories. This makes it a prime example of the fact that

> Canada is not a culmination of centuries of a common history, language, and culture; it is a new nation of disparate and diverse geographic and economic, cultural and linguistic communities. (Saywell, 1996, p. 3)

It is in this context of diversity between different Canadian regions, as well as within the province itself, that Quebec's social and historical characteristics take on significance. This section provides a brief overview of the history of Quebec, starting with precolonial Indigenous populations and leading to today's complex multicultural society.

2.1.1.1 Indigenous peoples

Before the arrival of Europeans settlers, North America was already inhabited by an important Indigenous population.[2] Archaeological evidence suggests that its presence goes back 12,000 years, if not more. In particular, the north-eastern area of North America, including what is now Quebec, was home to Algonquian peoples, who were mainly nomadic hunters-gatherers; and, further south in that area, Iroquoian peoples, who led a more sedentary life sustained by agriculture (Lackenbauer et al., 2010, p. 3). These populations were fully capable of satisfying their material as well as spiritual needs using the resources provided by their natural surroundings. They also formed clearly structured, complex societies, including ones based on democratic systems of government (Canada, 2013).

Indigenous peoples came into contact with Europeans starting in the 11th century, with the presence of Norse explorers and subsequently that of fishermen from Western Europe; they formed trading relationships with these populations. When the French arrived, they played a vital role in enabling the survival of their settlements. As before, they established commercial exchanges, crucially providing the settlers with access to furs, which constituted their main export. Indigenous peoples also played a military role, forming longstanding alliances with the British as well as the French, and participating in North American conflicts between the two colonial powers. However, as Canada became more institutionally organized in the 19th century, Indigenous populations were increasingly seen as subjects rather than allies. The results of this situation include land treaties leading to the creation of reservations, as well as notoriously violent attempts at cultural assimilation (Canada, 2013).

Despite these tensions, Indigenous populations remain present in Canadian society. An important way in which they define it is by contributing to its linguistic diversity, including by providing loanwords to English and French; we will come back to this later on in the chapter.

[2] In present-day Canada, three specific Indigenous groups are recognized: First Nations, which include over 50 distinct Nations living across the country; Inuit, who traditionally inhabit the Arctic regions; and Métis, whose communities historically descend from unions of Indigenous and European populations. The three groups are culturally distinct, and are subject to differences in legal recognition. See https://www.rcaanc-cirnac.gc.ca/eng/1100100013785/1529102490303.

But let us first take a look at the historical events arising from the contact and confrontation between the two later arriving communities, starting with the French settlers.

2.1.1.2 New France

French colonial explorations in what is now Quebec started in the 16th century. Initial contact was established through three voyages led by Jacques Cartier between 1534 and 1541. He claimed the land for the French king, and explored the territory by following the St. Lawrence River upstream from the Gaspé peninsula, past present-day Quebec City, and onto what would later become Montreal. Beyond their declared religious goal of converting the Indigenous populations to Christianity, these voyages were motivated by finding a passage to Asia, as well as bringing back minerals such as gold; this, at least initially, did not materialize. And while occasional contacts related to fishing and fur trade continued in the late 1500s, it is at the turn of the century that first colonial settlements were established; most notably, Quebec City was founded by Samuel de Champlain in 1608 (Mathieu, 2021).

The rate of settlement remained extremely slow due to difficult conditions, including harsh winters and infertile land. The first generations of settlers nevertheless succeeded in exploring the extensive waterways and established trading alliances with Indigenous peoples. Moreover, the territory controlled by France expanded over the course of the 17th century, with explorers reaching the Mississippi River and following it down to the Gulf of Mexico, claiming the land for the French king (Saywell, 1996, pp. 19–21). The early decades of French presence were also marked by the founding of Montreal in 1642 by Paul de Chomedey de Maisonneuve and Jeanne Mance. Located some 200 km upstream from Quebec City, it was similarly created as a missionary colony, but its *raison d'être* soon became fur trade (Linteau, 2017, pp. 26–43).

The development of New France was particularly pronounced in the second half of the 17th century. Following a period of control by merchant companies, it was officially designated a French province in 1664. This translated to a more direct exertion of royal power, including further immigration supported by the European mother country. This, coupled with high local birth rates, led to a population of thousands towards the end of the century. In terms of social structures, the multiple roles played by the Catholic church should be noted, as its focus included education and charitable activities. With the conversion rate of Indigenous populations remaining limited, it is these other activities that ultimately produced long-lasting effects (Durand, 2002, pp. 18–24).

By the beginning of the 18th century, New France had reached the limits of its territorial expansion, covering much of western North America (Figure 2.2). But this century was also marked by two major conflicts, which arose from European tensions and ultimately led to the demise of the French rule on the continent. First, as a result of the War of the Spanish Succession (1701–1715), the Treaty of Utrecht signed in 1713 led to France ceding some parts of its North American territory to Great Britain. The 30 years that followed nevertheless represented a period of peace and development. New France constructed fortifications at key locations and forged alliances with Indigenous nations. Its society also began taking a more complex shape with the formation of an upper class; the vast majority of the population continued to farm, but

FIGURE 2.2: New France around 1750.[3]

each generation also cleared and settled additional swaths of land (Mathieu, 2021). The number of settlers rose to 60,000 by the 1760s (Saywell, 1996, p. 19), but this should be contrasted with the 1.5 million people living in the Thirteen British Colonies around the same time (p. 17).

The definitive end to the French presence in North America came with the Seven Year's War (1756–1763), which pitted Great Britain against France. In 1758, the conflict spilled over from Europe into North America, where the population of New France was far outnumbered by that of the Thirteen Colonies. British forces swiftly captured Quebec City (1759) and Montreal (1760). A three-year transition period followed, during which around 3,500 British troops were left in charge of the entire French territory. Given their numerical disadvantage, they attempted to foster a climate of collaboration, maintaining many of the existing rights and customs for the French population. But this situation came to a close with the Treaty of Paris, which marked the official end of the conflict. France ceded all of its possessions in North America, with the exception of the small island territory of Saint Pierre and Miquelon. The whole of French Canada came under British rule (Durand, 2002, pp. 40–47).

The two or so centuries of French presence in North America – bookended by Cartier's arrival in 1534 and the Treaty of Paris in 1763 – were marked by a focus on trade relations, varying and overall limited interest in the development of the settlements, and comparatively moderate success in populating them. Nevertheless, traces of this origin – particularly the French language and the Catholic faith – have persisted in Quebec to this day; they characterize much of its society and distinguish it from the rest of North America. Let us now see how these characteristics evolved under British rule.

[3]Map source: https://commons.wikimedia.org/wiki/File:Nouvelle-France_map-en.svg.

2.1.1.3 Quebec under British rule

In the years following the Treaty of Paris, the British attempted to culturally assimilate the French Canadian population. However, immigration into the new colony never took off: fewer than 2,000 Britons settled in Quebec between 1760 and 1776. These attempts were additionally complicated by tensions between Great Britain and its remaining North American colonies (Durand, 2002, pp. 49–53). It is within this context that the Quebec Act (1774) was enacted and should be interpreted. It enlarged the territory of the colony by establishing that the Ohio–Mississippi valley, a fur trading area, would continue to be governed from Quebec; this was in effect a way to limit the westward expansion of the Thirteen Colonies. The Act also provided guarantees regarding the use of French and the maintenance of the Catholic faith. But the political situation changed once more with the American Revolutionary War (1775–1783). Like Nova Scotia, another British territory, Quebec did not join the other colonies in seeking independence; despite being partly invaded by them, it never fell. Quebec remained under British control after the war, but it was definitively cut off from the Ohio–Mississippi territory, ceded to the United States (Saywell, 1996, pp. 24–25).

The aftermath of the American Revolution was important in multiple ways. From a demographic standpoint, over 40,000 inhabitants of the former Thirteen Colonies who had remained loyal to Britain moved north. Known as the United Empire Loyalists, most of them settled in Nova Scotia, but around 7,000 moved to Quebec. They were followed by the so-called late Loyalists, who claimed loyalty to Britain simply to obtain free land, and ordinary Americans moving north without realizing the extent of political boundaries. From a territorial standpoint, Quebec was once again altered by the 1791 Constitutional Act. It was split into the colonies of Upper Canada (current Ontario) and Lower Canada (current Quebec); as a result, the territory inhabited by French Canadians was reduced (Saywell, 1996, p. 26). However, the act also introduced a legislative assembly in each colony, which inadvertently provided French Canadians with some political power. As for societal trends, commercial activities were taken over by the English-speaking elites, whereas the French population remained mostly rural, persisting in large part due to a high birth rate (Durand, 2002, pp. 52–57).

Political tensions over self-government led to rebellions in 1837 and 1838, first in Lower and then in Upper Canada. A military repression ensued; in London, Lord Durham was commissioned to devise a way of putting the situation under control. His 1839 report recommended that the principle of self government be applied, but it also argued for the unification of the two colonies in order to assimilate French Canadians. This was the effect of the Union Act, enacted in 1841, which created a single province of Canada; English was made its only language (Couture, 2021). Despite these affronts, a new balance of power emerged over the following decades. The British retained control over politics, business, and key projects such as the construction of the railway, but they gave up on assimilating French Canadians. This allowed for a largely peaceful coexistence in the period leading up to the 1867 Confederation (Durand, 2002, p. 63).

In summary, during the century of the British colonial rule – from the Treaty of Paris to Confederation – the fate of Quebec was often determined indirectly, within the broader con-

text of power relations between Britain and its other North American colonies. This period resembles a series of back-and-forth movements: for example, decisions such as the reduction of Quebec's territory were often counterbalanced by gains in terms of political rights. But a clear societal trend also emerged: that of the minority English-speaking population occupying positions of political and economic power, at the expense of the majority French-speaking population. This would have far-reaching consequences for the subsequent structure of Quebec's society and the tensions that continue to permeate it.

2.1.1.4 Confederation and economic growth

Towards the 1860s, the British territories in North America were increasingly considering some form of union. At the time, these included the province of Canada; the colonies of Nova Scotia and New Brunswick; Newfoundland, Prince Edward Island, British Columbia, Rupert's Land (privately owned by the Hudson's Bay Company), and the North-Western Territory. The impetus for a union mainly came from fears of American expansion, which arose as a result of the Civil War, trading difficulties with the United States, and uncertainty over the British commitment to the defense of North American territories. Following a series of gradual steps, the Dominion of Canada was created through the British North America Act, adopted in London in 1867. The resulting confederation initially included Quebec (once again separate), Ontario, Nova Scotia, and New Brunswick. It reached the Pacific in 1871, with the addition of British Columbia; it continued to evolve until 1999, with the addition of Nunavut (Waite, 2021).

The new constitution – the British North America Act – granted the federal government power over key national policies, while the provinces retained control over many other important issues. This crucially meant that Quebec was in charge of its linguistic and religious specificity. However, this failed to defuse the tensions between English and French Canadians. Linguistically, the rest of Canada was strongly English-speaking and the rights to education in French, for instance, were scarcely respected; politically, French Canadians may well have retained control over their province, but their will was repeatedly overpowered in Ottawa by the English-speaking rest of Canada (Saywell, 1996, pp. 87–89). More broadly, these tensions reflected a different understanding of the Confederation: French Canadians mainly interpreted it as a federation of nations – the British and the French; most English Canadians understood it to result in a homogeneous nation (Couture, 2021).

On the demographic front, there had been sustained immigration from the British Isles between 1815 and 1860, with a quarter of Quebec's 1.2 million people coming from Britain at the time of Confederation. In parallel, the number of French Canadians was rising due to a high birth rate, but they also started emigrating to the United States in the late 1800s. The second half of the 19th century was also marked by more general social evolutions. This was a period of increasing industrialization, which in turn accelerated urbanization, particularly benefiting the development of Montreal. A new bourgeoisie formed, mainly composed of Montreal-based business owners of English and Scottish descent. French Canadians had less economic clout, but they shared some political power and controlled French-speaking businesses. Industrialization also gave rise to the working class, mostly comprising low-skilled and poorly paid French

Canadians. The Catholic Church continued to exercise strong influence, controlling Quebec's healthcare and education, as well as peoples' world views (Linteau, 2021).

The same trend towards industrialization continued in the first decades of the 20th century. It was driven by capital coming from outside of Quebec, with the new industries managed by English-speaking Canadian, American, or British owners. French Canadians occupied low-paying jobs, becoming the most poorly paid workers in Quebec by the middle of the century. To put it more vividly, "in their own province, French Canadians became the hewers of wood and the drawers of water in an urban and industrial society dominated by others" (Saywell, 1996, p. 90). Demographically, this period was marked by a new wave of immigration, led by Eastern European Jews, followed by Italians. As for economic growth, it came to a halt with the Great Depression and the Second World War. The slowdown in international trade was most acutely felt in Montreal, Canada's main port. But this period had positive social consequences as well: Quebecers who served in Europe came into contact with other cultures; rural Quebecers were increasingly involved in Canadian industry; and many women were employed during this period, opening up new perspectives (Linteau, 2021).

The postwar period brought about renewed economic growth, as well as immigration, this time from the British Isles and southern Europe, particularly Italy and Greece. It was during this time that a new middle class formed, comprised mainly of highly skilled workers, with an important effect on the subsequent social and political evolution of Quebec. From the political standpoint, this period was dominated by the rule of Maurice Duplessis (1944–1959). His views were economically liberal and strongly socially conservative, emphasizing the Catholic faith, French language, and rural traditions. However, these positions discounted the evolution of Quebec's society, as reflected by the term *la grande noirceur* (the Great Darkness) which was applied to them by young French Canadian intellectuals of the era. This disconnect set the stage for transformative social change in the decades that followed (Linteau, 2021).

Summarizing, the century between Confederation in 1867 and the beginning of the Quiet Revolution in 1960 set the foundations of modern-day Quebec's society. Economically, the province became industrialized, and fully integrated into Canadian and North American trends; socially, it started shifting away from traditional values. However, this period also exacerbated some of the persisting conflicts in the province. It reinforced the economic and political importance of the English-speaking minority, leading to a striking power imbalance relative to the French-speaking majority, and ultimately giving rise to a renewed feeling of French Canadian nationalism.

2.1.1.5 Quiet Revolution and beyond: social modernity and political tensions

Starting in 1960, a wide-ranging set of reforms was put in place under the government of Jean Lesage (1960–1966), and pursued to a lesser extent over the two decades that followed his term in office. This period of change is known as the Quiet Revolution. One of its central components was an education reform, which introduced full provincial control over schools, created junior colleges, and founded the Université du Québec. Taken together, these measures led to a dramatic increase in the level of education among Quebec's Francophones.Other re-

forms included the development of the welfare state, with provincial control of hospitals and social services; nationalization of private electricity companies; and an infrastructure construction program. These reforms brought about important societal changes over a short period of time, but they also fit into the preexisting trends in the evolution of Quebec's society, with a continued rise in industrialization, urbanization, and standard of living (Linteau, 2021).

From a political perspective, the second half of the 20th century was marked by an increasing affirmation of Quebec's Francophones. An important step in this process was a series of language laws which came into force between 1969 and 1977 (Linteau, 2021); their effects on the demographic and linguistic dynamics of Quebec will be discussed in the next section. In more strictly political terms, these laws reflected a rise of French Canadian nationalism. A key role in this regard was played by the Parti québécois (PQ), formed under René Lévesque in 1966 and elected to power in 1976. This was a highly symbolic turning point, marking a new period in the political control of Quebec and, more broadly, the beginning of constitutional struggles over the position of Canada's provinces. In particular, Canadian Constitution was amended in 1982 through the Constitution Act, but Quebec never formally approved it. Instead, negotiations between all provinces and the federal government ensued, one of the central issues being the recognition of Quebec as a distinct society. Despite multiple attempts, no constitutional accord has been adopted (Saywell, 1996, pp. 95–112). The repercussions of this debate are perhaps best illustrated by the referendums addressing Quebec's sovereignty, conducted in 1980 and 1995. Both were rejected; however, the margin in the second referendum stood at 1%, and the Anglophone vote played a deciding role both times (Linteau, 2021).

In demographic terms, the second half of the 20th century was characterized by a decrease in the birth rate of Francophone Quebecers, and a significant out-migration of the Anglophone population to other provinces, especially following the adoption of language laws mentioned above. Since that period, population renewal has largely depended on immigration. It has included a broader range of origins since the 1960s, especially those from French-speaking regions such as France, North African countries, and Haiti. Economically, this was a time of successive periods of growth and decline, with the the end of the century marked by a strong shift to services and technology (Linteau, 2021).

On the whole, the last 60 years of Quebec's history can be seen as a break away from the earlier trends, with an affirmation of the French-speaking majority in the political life of the province and concrete expressions of the nationalist sentiment, opposing Quebec to the rest of Canada. However, this is also a result of the issues observed since the British conquest of New France, namely a growing accumulation of political and economic power by parts of the English-speaking minority. Let us now take a closer look at the way in which the historical context discussed so far translates to the demolinguistic composition of Quebec.

2.1.2 Demolinguistic profile of Quebec

Drawing on the main historical events outlined so far, this section will more explicitly link them to the evolution of the populations speaking Quebec's languages. It will then discuss the structure and impact of linguistic legislation introduced in the 20th century, and finally provide

an overview of the current demolinguistic structure of Quebec.

2.1.2.1 Historical demographic trends

As we have seen, the first European settlers in today's Quebec were French speakers. Although there is evidence of isolated contact with English speakers during the New France period, such as captives and fugitives, it is assumed that they fully integrated into the French-speaking community (Dickinson, 2007, pp. 11–12). The first notable Anglophone population formed after the British conquest in 1760, but this was a slow process. By 1776, there were around 70,000 inhabitants in the province of Quebec (current Ontario and Quebec), only a few hundred of whom were English speakers (Walker, 2015, p. 43).

A more important influx came in the wake of the American Revolutionary War, with the arrival of the United Empire Loyalists to the then-province of Quebec. The precise estimates of their numbers vary; Walker (2015, p. 47) cites around 6,000 arrivals between 1779 and 1784, whereas Dickinson (2007, p. 14) puts the number at 10,000 for the entire existence of the province of Quebec in its form at the time (1774–1791). The fact remains that these populations mainly settled in western regions that would go on to constitute the province of Upper Canada, later Ontario, in 1791. Although this means that French speakers remained numerically dominant in present-day Quebec, the English-speaking minority was highly influential. Initially composed of military personnel, it soon expanded through the arrival of merchants, and it also included farmers with the arrival of Loyalists. It wielded political power, as all government positions were occupied by British subjects until the Quebec Act (1774); while the Act nominally opened these positions to French-speaking Catholics, its concrete effects were limited. English speakers also held economic power, with particular importance of Scottish traders in the fur business in Montreal (Dickinson, 2007, pp. 13–15).

The 19th century saw considerably more intense migratory movements. They were in large part encouraged by Great Britain in order to shore up its presence in Canada, following the invasion by the United States during the War of 1812. Between 1815 and 1867, more than 1 million people are estimated to have moved from the British Isles to British North America. Most of them settled in Upper Canada, i.e. present-day Ontario (Walker, 2015, pp. 48–50). But some remained in what is now Quebec, where English speakers represented a quarter of the total population by mid-century, corresponding to more than a quarter million people. In contrast to their earlier demographic status, they constituted a majority in regions including the Eastern Townships (the area bordering the United States), the Ottawa Valley, and Montreal. And unlike the first English-speaking settlers in Quebec, most of them held no better jobs than the French-speaking majority (Donovan, 2019). Although the Anglophone community later diversified, the basis of the English language spoken in Quebec was formed by this British core – composed of English, Scottish, and Irish input dialects – which was to some extent influenced by American features introduced by Loyalists. As a result, Quebec English at its outset closely resembled the variety spoken in Ontario, from which it was mainly distinguished by a more limited impact of the Loyalist population (Boberg, 2014, p. 57).

Following Canada's Confederation in 1867, significant numbers of English-speaking Que-

becers living in rural areas started moving west due to economic hardship (Dickinson, 2007, p. 14). As a result, the largest part of the English-speaking population was concentrated in Montreal, a trend that continues to this day (Donovan, 2019). The composition of the city's English-speaking community diversified in the early 20th century, as its rising economic importance attracted populations from other parts of Canada and from abroad (Dickinson, 2007, p. 15). Of particular note were Jewish populations coming from Eastern Europe in the late 19th century, as well as subsequent Italian arrivals; both groups tended to join the English-speaking community. In the following decades, newly arriving English speakers also included Black, Chinese, and South Asian migrants. Today, most English-speaking Quebecers are of non-British origin (Donovan, 2019). In terms of general demographic trends, the proportion of English speakers in Quebec peaked in the 1860s, but it continued to grow in absolute terms until the 1970s. The subsequent decline is mainly linked to a wave of out-migration to other provinces, starting in the 1960s. It was driven by the rising economic importance of Toronto and the introduction of language laws reinforcing the position of French in the province (Dickinson, 2007, p. 15); this issue is addressed in the next section.

To summarize, the majority of Quebec's population is of French origin, in many cases with direct links to the settlers arriving during the New France period. The English-speaking presence is the result of more recent, successive waves of immigration. These include a very limited presence following the British conquest in 1763; a more important influx of Loyalists towards the end of the 18th century; considerable immigration from the British Isles over the course of the 19th century; and the arrival of more diverse groups joining the English-speaking community, starting in the late 1800s. This evolution was also associated with a changing social status of the English-speaking community. As Dickinson (2007, pp. 15–17) notes, the initial arrivals after British invasion mostly maintained good relationships with the French-speaking majority. However, the rise of industrialization in the 19th century conferred overwhelming economic importance to English-speaking elites, even though most Anglophones were in no better position than the French-speaking majority. The well-to-do gradually distanced themselves from the rest of society in a trend suggesting that, by the early 20th century, "the Montreal bourgeoisie [had grown] perhaps complacent with its God-given right to wealth and influence" (Dickinson, 2007, p. 17). This position was undermined by the political events of the later decades, including laws aiming to strengthen the use of French in the province.

2.1.2.2 Language planning and evolution of linguistic groups

The use of languages in Canada is partly regulated by federal and provincial legislation, which have tended to operate in different directions:

> while provincial language laws and regulations often eroded the vitality of Francophones in the ROC [Rest of Canada] and Anglophones in Quebec, federal language laws in the last decades sought to equalize and protect the status of official language minorities as a way of maintaining Canadian unity. (Bourhis and Landry, 2012, p. 32)

On the federal level, English–French bilingualism is enshrined in the Official Languages

Act (Canada, 1985, first adopted in 1969). It establishes that Parliament, federal courts, and certain federal institutions must use both languages. Bilingualism is also protected under the Canadian Charter of Rights and Freedoms (Canada, 1982), which defines the equal status of English and French as Canada's official languages and sets forth the educational rights of official language minorities. Supportive institutional norms such as these likely contribute to the overall favorable conditions of contact between Anglophones and Francophones (Adsett and Morin, 2005), with the nation-wide rate of bilingualism rising from 12% in 1961 to 18% in 2016 (Lepage, 2017b, p. 1).[4]

But in Quebec, the use of English and French had become diglossic by the middle of the 20th century. Despite its minority demographic status, the social prestige of English was particularly evident in the workplace, especially in highly bilingual areas such as Montreal and the regions bordering Ontario and the United States (Bourhis, 2001, p. 109). Although the intergenerational transmission of French was largely unimpeded in Quebec, the French-speaking population perceived other potential threats: its decline in the rest of Canada; its decreasing birth rate; the tendency for immigrant children to enroll into English schools in Quebec; and the economic domination of English-speaking Quebecers (Bourhis, 2001, p. 113).

Beginning in the late 1960s, a series of provincial laws were adopted in Quebec to address these issues. The most impactful was the Charter of the French Language (1977), also known as Bill 101. Its aim was to ensure the maintenance of French in the province and to improve its status relative to English (Bourhis, 2001, p. 114). Its most consequential provisions include the fact that it made French the only official language of Quebec; it considerably limited access to English-language schools in Quebec; it required that businesses with over 50 employees provide guarantees for the use of French in the workplace; it made public signage French-only; it introduced the right for consumers to demand to be served in French; and it prescribed that Quebec institutions could only be referred to using their French names (Quebec, 1977).

The most immediately visible effects of Bill 101 came from the requirement for public signage and commercial advertising to be exclusively in French, as shown in an overview of these provisions by Bourhis and Landry (2002). This requirement, like the bill more generally, was received positively by French-speaking Quebecers; however, the opposite was true for the English-speaking minority. This is hardly surprising: linguistic landscape has both an informative and a symbolic function, and it influences the way in which a linguistic community perceives its own vitality. The public signage provisions amounted to a symbolic exclusion of Anglophones from institutional and commercial life, leading to years of public outcry on their part. These provisions were also challenged legally, with the Supreme Court of Canada deeming them unconstitutional in 1988. Following several amendments, the use of languages other than French is now allowed in most cases, providing that French is also used and that it is visually predominant. Surveys conducted in Montreal in the late 1990s suggest that the resulting policy found a reasonable balance, enjoying broad support across language groups.

The debate surrounding public signage more generally reflects the fact that Bill 101 overtly

[4] In this analysis produced by Statistics Canada, a person is considered bilingual if they report the ability to conduct a conversation in both English and French. See below for further discussion of the language variables collected in the Canadian census.

elevated French to a symbol of Quebec identity. Although it also referenced the need to respect linguistic minorities, Anglophone Quebecers perceived its potential demographic consequences as threatening. As a result, and somewhat ironically, it was the promulgation of the bill that led Quebec's English-speaking communities to view *their* language as an identity symbol (Bourhis and Landry, 2002). Nevertheless, by the beginning of the 21st century the initial tensions had mostly eased in everyday interactions between Anglophones and Francophones. One way in which this is exemplified is the willingness to switch to an unknown interlocutor's language. A series of surveys conducted in downtown Montreal has shown that this rate of convergence increased over time: in 1977, 95% of Francophones converged to English, and 60% of Anglophones converged to French; in 1997, the rate was up at 99% and 88%, respectively (Bourhis et al., 2007, p. 208). A more general indication of improved relations is the fact that there is broad support across linguistic groups for using Bill 101 to protect the vitality of French, despite the persistence of often divergent concerns regarding its scope (Bernard Barbeau, 2018).

But these symbolic issues are also related to other, more tangible consequences. In demographic terms, a net of 148,000 Anglophones outmigrated to other provinces between 1976 and 1986, i.e. in the decade following the election of the Parti Québécois and the adoption of Bill 101. This was a continuation of a trend spurred by two earlier language laws, which had already led to the net outmigration of 102,000 Anglophones in the decade prior (Bourhis, 2012, p. 323). Longer-term consequences were produced by regulating the language of instruction. Faced with declining fertility rates, the viability of both Quebec's official language communities increasingly depended on the assimilation of Allophone immigrants (Dickinson, 2007, pp. 12, 21). In order to reinforce the demographic position of the Francophone community, Bill 101 made French-language instruction mandatory for most children attending public or subsidized private schools, through the end of secondary education. Specific exceptions were granted; they currently include children who previously attended English-language school in Canada, or whose one parent or sibling did so (OQLF, 2019, p. 38). In a confirmation of this measure's effectiveness, the rate of enrollment of Allophone students in French primary and secondary schools increased from 20% in 1976 to 89% in 2015. A weaker but similar trend is observed for Anglophone students, with the rate rising from 8% to 28% (p. 39).

These demographic trends are further reflected by the evolution of Quebec's composition in terms of linguistic groups, illustrated in Figure 2.3. The solid lines indicate the proportion of the population by mother tongue, defined as "the first language a person learned at home in childhood and still understood" at the time of the Census (Lepage, 2020, p. 30). Between the most distant points in time – 1971 and 2016 – the proportion of French mother tongue speakers declined slightly, from 80.7% to 78%. This trend was more pronounced for the English-speaking population, dropping by more than a third from 13.1% to 8.1%. This points to a decreasing relative importance of both official language communities in Quebec, as the proportion of Allophone speakers more than doubled from 6.2% to 13.8%. In absolute terms, the number of French mother tongue speakers linearly increased from just over 6 million in 1971 to just over 8 million in 2016. By contrast, the Anglophone population declined from 789,000 to 657,000 over the same period; however, it reached its lowest level in 2001 (591,000), after which it started growing again. Finally, the absolute number of Allophone speakers tripled

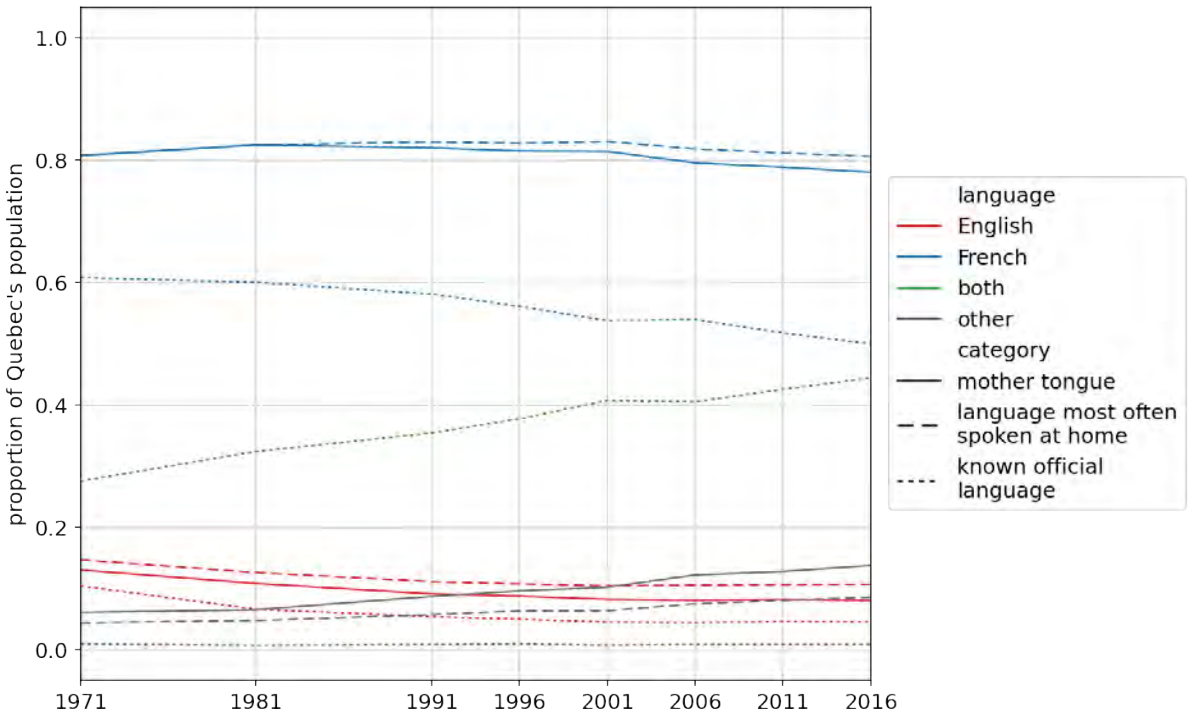


FIGURE 2.3: Historical linguistic trends in Quebec.
Data source: Census 1971–2016 (Statistics Canada, 2019a,b, 2020).

from 372,000 in 1971 to 1.1 million 2016; they now outnumber the English mother tongue population by nearly two to one (Statistics Canada, 2019a).

This transformative demographic change should be contextualized in two ways. First, the data on the language most often spoken at home (indicated as a dashed line in Figure 2.3) show that the proportion of speakers who use English or French at home is higher than the corresponding mother tongue population. This difference comes from the Allophone speakers assimilated into the official language communities; while it was present for English in 1971, it only developed for French later on. This confirms the effectiveness of Bill 101, as well as the importance of Allophone speakers for the vitality of the official language communities. The second important trend concerns the reported knowledge of official languages (represented as a dotted line). This Census variable reflects a broader view of language use, corresponding to the ability to conduct a conversation in English and/or French (Lepage, 2020, p. 29). The main tendency to be noted is a decrease over time in monolingual knowledge of both English and French. This is in fact compensated by a marked increase in English–French bilingualism, which rose from 27.6% in 1971 to 44.5% in 2016 (Statistics Canada, 2019b).

As we have seen, Bill 101 signaled a dramatic turning point for Quebec's language communities. For its proponents, it righted historical wrongs because of which the language of the majority occupied a lower social position, its survival potentially threatened. For its detractors, the bill encroached upon the rights of another language group, crystallizing oppositions within the province. While these positions are difficult to reconcile, the bill indisputably triggered a transformation of Quebec's society and the place that languages occupy in it: symbolically, through changes such as the very visible shift from English to French in public signage; more

profoundly, by altering the demographic trends in the province. To conclude this discussion, let us take a closer look at the structure of Quebec's language communities and the issues affecting them today.

2.1.2.3 Quebec's language communities today

As a result of historical demographic trends as well as language planning, Quebec is a majority French-speaking province (Figure 2.4). As of 2016, 77.9% of its inhabitants – close to 6.3 million people – report that their mother tongue is French. Ten times fewer Quebecers – 7.8% of the population, or just under 630,000 individuals – are native speakers of English. Compare this with Canada outside of Quebec, where nearly three quarters of inhabitants (72.5%, 19.4 million) are native speakers of English; French is the mother tongue of close to 1 million (3.6%) non-Quebecers (Statistics Canada, 2017e).

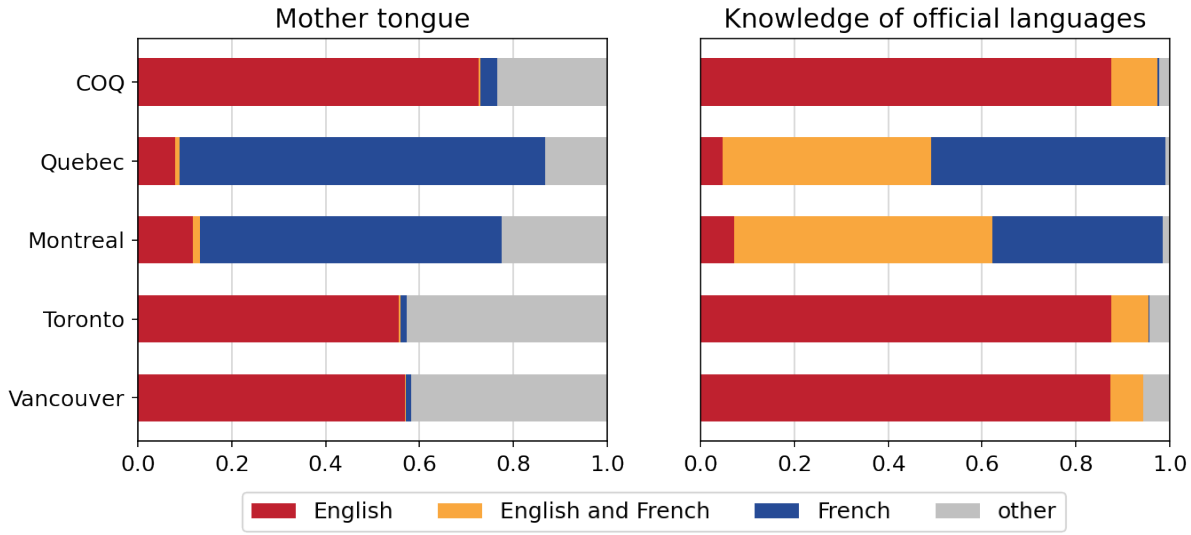


FIGURE 2.4: Present-day demolinguistic profile of key regions: Canada outside of Quebec (COQ), Quebec, Montreal, and – for comparison – Toronto and Vancouver. The values are expressed as a proportion of the overall population. For mother tongue, all language categories except for "other" include joint knowledge of a non-official language (e.g. "English" includes monolingual English speakers, as well as bilingual speakers of English and a language other than French). Data source: 2016 Census (Statistics Canada, 2017e).

These trends should also be contextualized with regard to knowledge of official languages, i.e. the reported ability to carry a conversation in English and/or French. From this standpoint, 50% of Quebec's population (slightly over 4 million inhabitants) is monolingual French; 4.6% (372,000) is monolingual English. It has the highest rate of official language bilingualism among all Canadian provinces, at 44.5% (just under 3.6 million speakers). In Canada outside of Quebec, the rate stands at 9.8% (2.6 million speakers). Within Quebec, official language bilingualism is in large part driven by Montreal, where over half of the inhabitants (55.1%, 2.2 million) report speaking both English and French. The rate drops in Quebec outside of Montreal (33.8%, 1.4 million), but even there it remains over three times higher than elsewhere in Canada. The rate of bilingualism is higher among mother-tongue English speakers (68.8%) than among mother-tongue French speakers (40.2%), but the difference between the two groups is smaller in Montreal (70.6% and 52.2%, respectively) (Statistics Canada, 2017c).

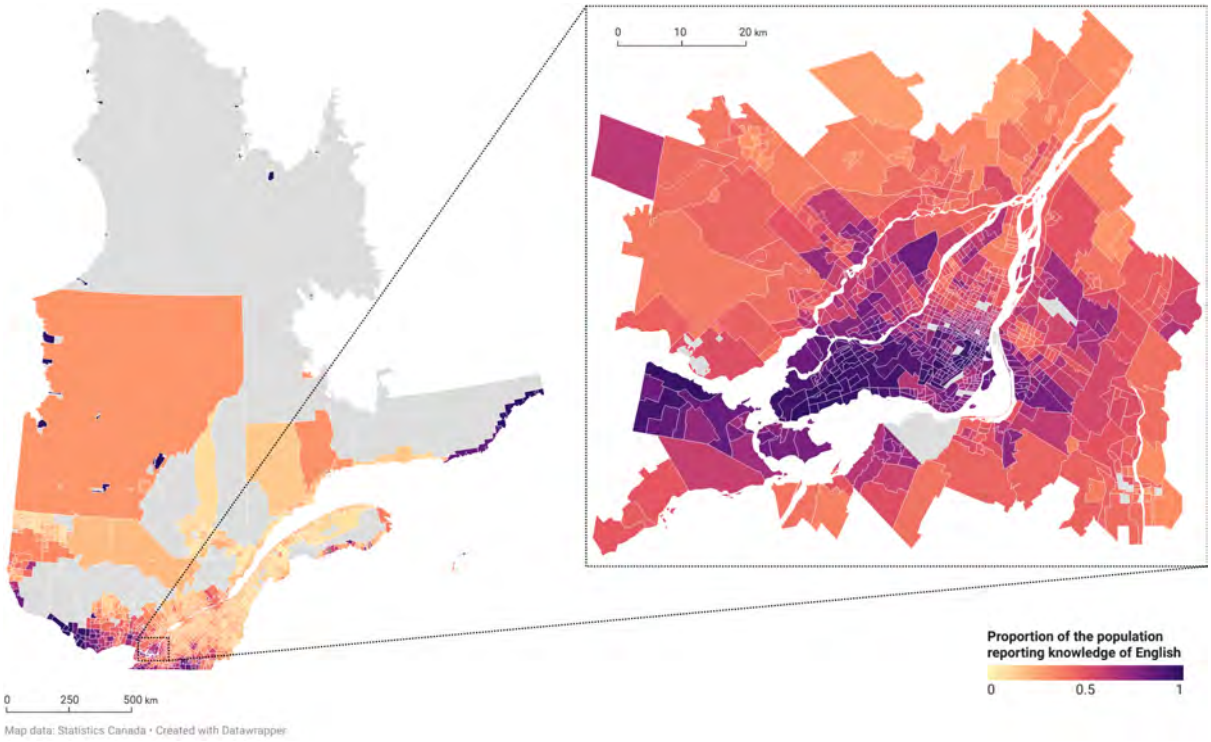


FIGURE 2.5: Geographical distribution of the population of Quebec (left) and Greater Montreal (right) based on the knowledge of English. Color coding reflects the proportion of the population reporting knowledge of English, including jointly with French. The total number of inhabitants does not include those who speak neither English nor French. Map created based on the data from the 2016 Census (Statistics Canada, 2017b,d). The left-hand map shows census subdivisions, which correspond to municipalities or equivalent areas. The right-hand map shows census tracts, smaller and relatively stable areas whose population usually does not exceed 10,000 persons.[5]

Regional differences in reported knowledge of English are mapped in Figure 2.5. In addition to the broad distinction between Montreal and the rest of Quebec, the map also points to finer-grained trends. First, it shows other significant English-speaking communities, in particular those bordering Ontario to the southwest and the United States to the south. It also indicates the presence of additional smaller communities in the province, especially in the eastern Gaspé peninsula and the sparsely populated, vast northern regions of Nord-du-Québec and Côte-Nord. While these communities present a particular interest due to their isolated nature and potentially more intense contact with French, they are comparatively very small, accounting for 2% of the entire English-speaking population of Quebec (Statistics Canada, 2017a).

The map also highlights the fact that, although Montreal constitutes the core bilingual area of Quebec, the use of languages is unevenly spread on its local level, too. The rate of English knowledge is higher on the Island of Montreal, the boomerang-shaped structure in the center of the map, than in the surrounding areas.[6] Within the Island, it is the highest – close to 100% – in the traditionally Anglophone neighborhoods of the West Island, as well as in the

[5]For definitions of Census geographic units, see https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/az1-eng.cfm.

[6]The St. Lawrence River was historically considered to flow in a west–east direction. As a result, the interpretation of the cardinal directions in Montreal is conventionally rotated by around 45 degrees compared to their actual position, represented in Figure 2.5. The commonly accepted "west" corresponds to the point of the Island that is geographically located to the southwest, the "north" points to the northwest, and so forth.

downtown area, home to more recent English-speaking immigrant communities and the highly international campuses of Concordia and McGill universities. The rate of English knowledge is the lowest – around 35% – in neighborhoods located in the east end, which has historically housed French-speaking communities (Statistics Canada, 2017d).

Moreover, socioeconomic differences align broadly with the geographic distribution of language communities, with neighborhoods in the west part of the Island generally exhibiting a higher median household income.[7] However, finer-grained trends are highly variable, similarly to the characteristics of the English-speaking population of Quebec as a whole: even though its mean income is higher than that of the Francophone population, its median income is lower. This suggests that, with the exception of a minority of high earners, most Anglophones have a lower socioeconomic status than their French-speaking counterparts (Donovan, 2019).

More generally, the demographic fact that English speakers constitute a minority in Quebec must be interpreted together with the place of French speakers in the rest of Canada. A key difference lies in the fact that most Francophone communities outside of Quebec are undergoing language shift – adopting a new home language – at a high rate. That is not the case for Quebec's Anglophones: while they are in part shifting towards French, this is offset by Francophone and Allophone shifts towards English (Sabourin and Bélanger, 2015). Consequently, minority Francophone communities outside of Quebec fear for the transmission and retention of French. Quebec Anglophones are mainly concerned about the protection of English-language institutions, such as schools and hospitals, and the communities they serve; the continued existence of their language as such is not at stake (SCOL, 2012, pp. 4–5). However, the respective concerns are not always clearly perceived by Quebec's linguistic groups. As Bourhis (2012, p. 313) succinctly puts it, "the 'two solitudes' often speak at cross purposes when it comes time to consider their respective fate in Quebec". In broad terms, Francophones conceive of their community as a minority on the national and continental level, facing an existential threat; Anglophones view themselves as a minority on the provincial level, preoccupied with the vitality of its local linguistic communities.

But all is not bleak in the relations between the two groups. Some 20 years after the adoption of Bill 101, Radice (2000) reported great attachment of English-speaking Montrealers to their city. They recognized the city's linguistic diversity – including specifically its Francophone character – as one of its key positive features, and a defining element of their own identity. More recently, it has been shown that Montrealers largely continue to view the city's inhabitants in binary Anglophone–Francophone terms, which is further reflected by their conception of the city along the corresponding west–east dimension. However, they also recognize a buffer zone – the downtown core – where bilingual interactions are readily expected, with bilingualism also emerging as a defining feature applied to Anglophones (Leimgruber and Fernández-Mallat, 2021). And, as evident as it may seem, it is important to note that the two linguistic communities do not live in isolation from one another. For instance, the Office of the Commissioner of Official Languages reports that 65% of Quebecers at least occasionally interact with people who mostly speak the other official language. The same proportion reports using media sources in the other official language (OCOL, 2022, p. 30). These interactions are

[7] https://censusmapper.ca/maps/838?index=3#10/45.6107/-73.7677

often sociolinguistically far-reaching. This is suggested by the fact that the use of languages in Montreal, particularly by multilingual Allophone speakers, is complex and variable to the point of questioning clear-cut boundaries between linguistic groups (Lamarre, 2013).

The state of affairs outlined so far indicates that Quebec in general, and Montreal in particular, is fertile ground for language contact. Concerning specifically the use of English, we have seen that the native Anglophone community constitutes a demographic minority, a fact remarkable in its own right. But the pool of potential English speakers is much wider; strikingly, more than half of Montreal's inhabitants exhibit official language bilingualism. Although most of them are native Francophones, there is arguably no inherent impediment to contact-related innovations circulating between different mother-tongue groups of bilingual speakers. This is further supported by the intermediary role played by Allophone communities, which link the city's neighborhoods and linguistic communities. Having placed the use of English in this context of complex interactions between different linguistic groups, I now turn to the main features of Quebec's languages.

2.2 Quebec French

This dissertation investigates the use of English in Quebec, but it does so by accounting for the cross-linguistic influence that may be exerted on it by French, which by all demographic metrics represents the majority language in the province. In this section, I will briefly present some of the main phonological, morphosyntactic, and lexical features of Quebec French. Since a comprehensive description is beyond the scope of this dissertation, I will aim to highlight the linguistic features whose traces might be found in contact-induced phenomena in Quebec English, as well as those that more generally illustrate cross-linguistic processes which operate in the province. But first, a word is due on the definition of Quebec French adopted in this work.

2.2.1 Defining Quebec French

As discussed in the previous section, French is one of Canada's two federal official languages. On the provincial level, it is the only official language of Quebec; it is also one of the two official languages, with English, in New Brunswick. The vast majority of native French speakers in Canada is concentrated in Quebec (85.6%); other sizeable French-speaking communities are mainly found in the bordering provinces of Ontario and New Brunswick (Lepage, 2017a, pp. 4–5). With discussions of Canadian French generally referring to the whole range of French varieties spoken across the country (e.g. Papen, 1998), I will stick to the term Quebec French (QF). This is motivated by the particular role played by the Quebec variety within the Canadian linguistic landscape, and by the regional focus of this dissertation.

As suggested in the discussion of the historical context, the place of French in Quebec became an issue of central importance during the Quiet Revolution, and particularly following the passage of language laws. While at the time the sociolect known as *joual* was closely associated with the use of French in the province, it gave way to more homogeneous, less

marked, usage as the Quebec norm (Barbaud, 1998, p. 182). An important development in this regard was the shift, over the second half of the 20th century, from a close alignment with Parisian usage to an endogenous norm, reflecting the way French is spoken in the province (Cajolet-Laganière, 2021). The resulting standard variety is often termed "standard French used in Quebec", defined as the socially valued usage in the province (Martel, 2006, p. 848).

Given the demographic and cultural weight of Quebec in French-speaking Canada, the French usage typical of the province tends to constitute the variety of reference for other parts of the country as well (Bigot and Papen, 2013, p. 116). It should however be noted that considerable regional variation has been described both across Quebec (e.g. Remysen, 2016; Remysen et al., 2020) and other parts of Canada (cf. Papen, 1998). In this context, the term Laurentian French is sometimes used to encompass Quebec French and the varieties deriving from it in other parts of the country, in particular west of Quebec, whereas Acadian French refers to the historically distinct varieties spoken in the Atlantic provinces (Remysen, 2019, p. 33).

In the remainder of this dissertation, Quebec French should be taken to apply broadly to the French spoken in the province. My view is not limited to standard Quebec French, as defined above; rather, it includes the whole range of registers used in the province (cf. Martel, 2006, p. 849). Moreover, given my interest in the consequences of language contact – which, as we have seen in Chapter 1, can originate from a single bilingual individual – I do not constrain this definition based on the speaker's mother tongue. Rather, following a view outlined in an admittedly different context, I extend it "to [...] Francophones, Anglophones, immigrants, in short all those who write, read, or hear [Quebec French words] on a daily basis" (Martel, 2006, p. 847; my translation). Let us now turn to some of the main characteristics of this variety.

2.2.2 Phonetics and phonology

While it largely shares its phonemic inventory with other varieties, Quebec French displays a series of characteristics that distinguish its pronunciation. As Bigot and Papen (2013, pp. 119–120) note, some of these features are also present elsewhere, but are more frequent in Quebec; others are entirely specific to the province. On the consonantal level, the authors identify the following characteristics:

- assibilation of /t/ and /d/ before high front vowels and the corresponding glides, as in *tu dis* [tˢydᶻi];
- word-final consonant cluster reduction, as in *juste* [ʒys];
- deletion of /l/ in personal pronouns *il(s)* and *elle(s)* before consonants, as well as in determinants and object pronouns *la* and *les* in an intervocalic position, as in *Roger la voit* [ʁɔʒeavwa].

As for Quebec French vowels, the most distinctive characteristics outlined by Bigot and Papen (2013, pp. 119–120) include:

- four nasal vowels, with a preserved /ɛ̃/ ∼ /œ̃/ distinction, as in *brin* vs. *brun*;
- preserved /a/ ∼ /ɑ/ distinction, as in *patte* vs. *pâte*;
- preserved /ɛ/ ∼ /ɛː/ distinction, as in *renne* vs. *reine*;

- diphthongization of long vowels, as in *rêve* [ʁaⁱv];
- devoicing or deletion of high vowels in atonic syllables, in a fricative context, as in *université* [ynivɛʁs(i̥)te];
- conditioned variability in the phonetic quality of other vowels, particularly /ɛ/, /ɑ/, and /wa/.

There are different ways in which these characteristics can be implicated in a study of contact-related lexical semantic phenomena. In general terms, phonological similarity is one of the aspects assumed to drive cross-linguistic semantic influence; this issue is explored in more detail in Chapter 3. In the specific context of Quebec English, it has been noted that French-origin lexical items vary in terms of phonological integration into English. This issue is examined by Rouaud (2019b), who specifically looks into the realization of /y/, /ʁ/, nasal vowels, and /t/ and /d/ assibilation in French-origin lexical items. Broadly speaking, French realizations (i.e. lack of adaptation to English) are associated with a higher degree of bilingualism (pp. 250–256). This is in line with the trends reported for L2 French production of English-speaking Montrealers (Blondeau et al., 2002, paras 32–33). More generally, as discussed in Chapter 1, phonological adaptation to the recipient language is seen as reflecting the degree to which a borrowing is established (cf. Poplack et al., 1988).

2.2.3 Morphosyntax

Quebec French is characterized by a series morphosyntactic features which distinguish it from other varieties. A detailed discussion of these issues is beyond the scope of this dissertation; for a more comprehensive overview, see e.g. Meney (2017, pp. 45–152). I will however briefly touch upon several characteristics that have been examined from the standpoint of L2 French use by English-speaking Montrealers. Although they are not representative of the full range of morphosyntactic phenomena characterizing the use of Quebec French, they crucially illustrate potential dynamics between native and non-native speakers of official languages in the city. In an overview of a series of studies conducted on this topic, Blondeau et al. (2002) discuss the following characteristics:

- verb negation, with an opposition between the standard *ne ... pas* pattern and the use of *pas* on its own, typical of spoken French;
- variable use of *on* and *nous* as the first-person plural subject pronoun, the former used near-categorically in Quebec French;
- stressed plural subject pronouns with *autres* (*nous autres*, *vous autres*, *eux autres*), which are highly frequent in Quebec French, with the choice depending on factors such as topic and formality;
- double subject marking (noun phrase + pronoun), similarly highly frequent;
- variable use of *on* / *tu* / *vous* as a generic pronoun, with a shift in Quebec French from *on* to *tu*.

The precise patterns reported by the authors vary from near-identical replication of native French behavior (as in the case of verb negation) to significantly lower rates of use (as in the

case of pronouns with *autres*). In most cases, the intensity of contact with French has a significant effect on the rates of use of these variables, and hence replication of native Quebec French behaviors. A particularly interesting case is that of the generic pronoun, with English-speaking Montrealers using the emerging French form *tu* at a higher rate than native French speakers; a tentative explanation is the parallel use of *you* in English, which might facilitate convergence with an ongoing change in French. More generally, these observations provide further support for the claim that non-native speakers of official languages can be closely involved in patterns of language variation and change in Montreal, an issue of central importance in defining the language communities under study.

2.2.4 Lexicon

Compared to other French varieties, the lexicon of Quebec French is distinctively characterized by a range of lexical items arising from different sources, as shown in an overview by Mercier et al. (2017, pp. 292–295). Some are related to the divergent development of Quebec French with respect to the varieties spoken in France, starting with the permanent presence of settlers in the 17th century. This has resulted in the continued use of lexical items such as *jaser* 'chat', *tantôt* 'earlier; in a while', or *présentement* 'at the moment'. They represent archaisms in most French regions, where they have been replaced by *bavarder*, *tout à l'heure*, and *actuellement*, respectively. In much the same way, some lexical items originating from the dialects spoken by the early settlers have remained limited to those dialects in France, but are widely used in Quebec French. One such examples is the verb *achaler* 'bother'.

Other lexical items specific to Quebec French result from its contact with the languages spoken in North America. This includes loanwords from Indigenous languages, most of which denote the local flora and fauna. As an example, *achigan* refers to a species of fish known as *bass* in English. It is of Algonquian origin, specifically present in the Algonquin and Ojibwe languages, and attested in French since the mid-17th century. The noun *caribou* similarly comes from an Algonquian language, likely Mi'kmaq; it was attested in French as early as 1606.[8]

Through the same process, Quebec French has also acquired a large number of anglicisms starting in the second half of the 18th century. They are more numerous, more varied, but also more negatively perceived. The affected semantic fields include work (e.g. *job*), food (e.g. *toast*), and home appliances (e.g. *blender*). Particularly remarkable is the tendency for long-established anglicisms to be strongly adapted, as in the case of *pinotte* 'peanut'. Anglicisms are negatively perceived from a prescriptive standpoint, but their use remains particularly strong in informal communication, where lexical items such as *gang* 'group of friends' or *chum* 'boyfriend' are often seen as more expressive.

Finally, some lexical items typical of Quebec French are local creations. This is often the case with terminological alternatives to English loanwords, such as *courriel* 'email' or *baladodiffusion* 'podcast', often abbreviated to *balado*. A related type of lexical specificity has to do with local referents. For instance, *cégep*, an acronym for *collège d'enseignement*

[8] All etymological information based on the *Base de données lexicographique du Québec*, available at https://www.bdlp.org/base/Québec.

général et professionnel 'general and vocational college', is a type of junior college unique to Quebec's education system. (This is in turn distinct from the use of *collège* in the French education system, which roughly corresponds to *middle school* or *junior high* in some Canadian provinces.)

In terms of ongoing contact-related sociolinguistic dynamics, the role of anglicisms is particularly important. Within the ideologically tense context of Quebec described at the beginning of this chapter, they have received much attention, predominantly negative, in public debates on the use of French; that remains the case to this day (Elchacar and Salita, 2019). Moreover, both the strong incentive to use French alternatives for many anglicisms, noted above, as well as the way in which anglicisms are phonologically integrated (cf. Côté and Remysen, 2019) differentiate Quebec French from other varieties, including those spoken in France. Beyond this distinctive role, anglicisms more generally illustrate the potential for French and English spoken in Quebec to engage in circular cross-linguistic influence. For instance, Vincent (2019) describes the use of the English borrowing *all dressed* in Quebec French to denote foods such as pizza 'with all the toppings'. But as we will see in the next section, this is one of the most regionally emblematic lexical items in Quebec English, produced by a calque of the Quebec French term *tout garni*. Cases such as this provide further evidence of cross-linguistic permeability in the use of the two languages in Quebec.

As we have seen, the position of Quebec French has shifted from that of a variety mainly defined in relation to those spoken in France, to one which has developed an endogenous norm and plays a central role in the wider context of French use in Canada. It is distinguished from other varieties of French on all levels of linguistic structures, including, as shown for the lexicon, by influence arising from its contact with English. We have also seen that native English-speaking Quebecers exhibit many of these features to varying degrees in their L2 French, meaning that they can serve as an estimate of their bilingual ability. More generally, the patterns of cross-linguistic influence noted in this section illustrate the potential for the two official language communities to influence one another. We now turn to a description of Quebec English.

2.3 Quebec English

This sections presents the main characteristics of Quebec English. As before, it first more closely defines Quebec English, and then addresses its main levels of linguistic structure – phonetics and phonology, morphosyntax, and lexicon. This is followed by a section dedicated to contact-induced semantic shifts. Note that the aim of this description is to provide a global overview of the variety under study, rather than discuss all patterns of variation occurring within it; some of these are presented in more detail in Chapter 6.

2.3.1 Defining Quebec English

Quebec English (QE) is a regional variety of Canadian English (CE), which can in turn be defined as a postcolonial English variety, and specifically an Inner Circle variety (e.g. Dollinger,

2020, p. 53). In the Concentric Circles Model of World Englishes (Kachru, 1985), this refers to the regions where English is the primary language of the population; it is opposed to the outer circle, where it is used as one of two or more languages, and has acquired an official status (e.g. India); and the expanding circle, where it is used as a language of international communication. This typology is paralleled by the often drawn distinction between English as a native language (ENL), English as a second language (ESL), and English as a foreign language (EFL) (Quirk, 1985, p. 2).

This characterization clearly indicates the current status of Canadian English, and points to its historical emergence from a well-established and progressively ever more independent group of English speakers. However, specific stages in its development are described more precisely by the Dynamic Model (Schneider, 2007). Applying his general approach to Canadian English (pp. 240–250), Schneider identifies the following stages:

(1) Foundation (1713–1812): English is brought to Canada in the aftermath of the British conquest and the arrival of Loyalists. The settlers come into initial contact with French and Indigenous speakers, as reflected by toponymic borrowings;
(2) Exonormative Stabilization (1812–1867): the settler presence is stabilized, including through immigration from the British Isles, with the London norm prevailing. Borrowing continues, principally in relation to Indigenous names for local referents such as *ouananiche* 'freshwater salmon' or *caribou*, with the process often mediated by French;
(3) Nativization (1867–c. 1910s): with a more clearly defined political status following Confederation, a locally specific usage emerges through a combination of typically British and American characteristics, as well as further integration of features from Indigenous languages and French;
(4) Endonormative Stabilization (c. 1920–c. 1970): the emergence of a more clearly defined national identity is linguistically paralleled by the codification of Canadian English in its own right, rather than in comparison with other varieties;
(5) Diversification (c. 1970): the period is marked by full sovereignty from the mother country, more diverse immigration, and growing regional and social stratification of English.

These trends are paralleled by research on Canadian English. In particular, linguistic descriptions produced over much of the 20th century mainly analyze Canadian English as a result of variable British and American influences, with a subsequent focus on identifying the features that clearly distinguish it from both. An important indicator of codification (and therefore endonormative stabilization) is lexicographic work addressing Canadian English in its own right, starting with the *Dictionary of Canadianisms on Historical Principles* (Avis et al., 1967), further discussed below. A broadly consensual view has emerged according to which Canadian English has become autonomous with regard to other World Englishes. It is moreover largely homogeneous, despite the vast distances that it covers, but this does not preclude longstanding patterns of heterogeneity (Dollinger and Clarke, 2012).

These characteristics are reflected by Boberg's (2005b) oft-cited summary of regional patterns of lexical variation across North America. His dialect survey, further discussed below, has established that

> Canadian dialect regions have more in common with one another than any of them has with the United States and that no region of Canada could be characterized as consistently more or less American in its lexicon than any other. (p. 53)

But the same analysis also identifies well-defined dialect regions within Canada, the foremost among them being that of Montreal, characterized by significant effects of contact with French.

The distinct status of Quebec among Canadian English-speaking regions is discussed in more detail by Boberg (2010, pp. 24–29). As already suggested, it is uniquely marked by a situation of language contact: the strongest influence is exerted by French, but other immigrant languages are also widely used in the province. This context is reflective of settlement patterns that differentiate Quebec from other parts of the country, making it one of several linguistic enclaves within the wider Canadian English landscape. The salience of the way in which English is spoken in the province is further confirmed by its negative perception not only by other Canadian English speakers, but also by many Quebecers.

In more formal terms, Rouaud (2019b, pp. 107–108) suggests that recent developments in the use of English in Quebec can be analyzed using Schneider's dynamic model. She posits (i) a foundation phase, triggered by the passage of Bill 101 in 1977 and the subsequent transformation of Quebec's language communities; (ii) an exonormative stabilization phase (1980s–1990s), with an emerging local identity and awareness of English use specific to the province; (iii) a nativization phase (1990s–today), with increased codeswitching, lexical borrowing, and syntagmatic and semantic innovations, mainly driven by a rise in bilingualism. While it remains to be seen if this analysis is empirically confirmed in the long term – as would be suggested, for example, by future trends towards codification – existing descriptions provide abundant evidence of regional specificity in the way English is used in Quebec.

But who is it that speaks English in the province? McArthur (1989, pp. 12–13) identifies five distinct categories of English-speaking Quebecers:

- members of the historical English-speaking community, who identify more with the rest of Canada than with Quebec, and who have limited contact with French;
- younger English speakers, often the second generation of the previous category, who use French to a larger degree, including with English insertions, and who may similarly introduce French elements into English discourse;
- Francophones who also speak English, including with numerous French-origin items resulting from processes of interference;
- Allophones variously proficient in English and French;
- a minority of speakers who are highly proficient in both English and French, and can use them without interference, as well as by voluntarily introducing elements from one into the other. (This view is reminiscent of the notion of balanced bilingualism discussed in Chapter 1.)

This analysis is broadly reflective of the different categories of English speakers presented in the discussion of Quebec's demolinguistic profile (Section 2.1.2). And while the precise composition of the language communities in the province may well have evolved in the 30 years since McArthur's analysis, his summary of the situation likely still holds:

> If this is an accurate picture of the situation, then there is no sense in which we can consider QE homogeneous entity; we cannot expect neat-and-tidy usage [...] The totality of QE includes everybody described above, in a vastly complex social interaction *as much marked by doubt and ignorance as by certainty* about what 'proper English' is. (McArthur, 1989, p. 13)

As we will see in more detail in Chapter 6, sociolinguistic studies tend to adopt restrictive definitions of speech communities, often limited to native speakers of the language under study. Although this is supported by valid practical concerns in collecting and analyzing data, I will adopt a broader view of Quebec English, perhaps best summarized as follows:

> it exists as a continuum, from long-established unilingual anglophones broadly similar to anglophones in Ontario through bilinguals of various kinds to francophones using English as a second language. (McArthur and Fee, 1992, p. 832)

In other words, I will apply the term Quebec English to any use of English by people living in the province, under the assumption that they are all at least passively exposed to French. I moreover assume that linguistic innovations arising in any of the subgroups comprising this English-speaking community (e.g. monolingual English speakers, native French speakers, and so on) are not inherently limited to that subgroup, but may circulate among speakers of all linguistic profiles.[9] This view is underpinned by a range of considerations:

- In demographic terms, we have seen that the number of native English speakers in Quebec is exceeded by the number of Quebecers who speak English at home, as well as by those who are able to speak English in general. Constraining Quebec English to native speakers would amount to delegitimizing a whole section of the population that also speaks the language.
- As discussed in Chapter 1, cross-linguistic interference leading to lexical semantic effects is not limited to specific ages or degrees of bilingualism; rather, it is facilitated by the general mechanisms underlying bilingual lexical knowledge. Excluding some types of bilinguals from the scope of the study could deprive us of observing relevant sociolingusitic behaviors, especially when we know that speakers of different linguistic profiles form an integral part of Quebec's linguistic communities (Chambers and Heisler, 1999, p. 41; Fee, 2008, p. 183).
- In terms of general sociolinguistic dynamics, the discussion of Quebec French has shown that non-native speakers often align with ongoing trends of language change (cf. Blondeau et al., 2002). Likewise, socially stratified patterns of variation in other regions, presented in Chapter 6, highlight the fact that non-native speakers can help preserve uses typical of Canadian English (cf. Dollinger, 2012). It is reasonable to expect that comparable patterns could be reproduced for Quebec English, with a potential contribution of non-native speakers to regionally-specific uses.

[9] I am grateful to Wim Remysen for pointing out this potential trend.

I now turn to a brief description of the linguistic characteristics Quebec English. I will draw on several sources of information (for a more extensive discussion of the underlying data collection methods, see Chapter 4):

- Dialect surveys, based on the use of written questionnaires, are usually distributed to tens or hundreds of informants. Their focus is principally on lexical variation, but phonological and morphosyntactic phenomena are also investigated to a lesser extent (Boberg, 2004a, 2005b; Boberg and Hotton, 2015; Chambers and Heisler, 1999; Hamilton, 1958; McArthur, 1989). A related source of information is the Atlas of North American English (Labov et al., 2006), which used phone interviews to collect data on pronunciation across North America.
- Sociolinguistic interviews targeting Quebec English have been conducted by Poplack et al. (2006) and Rouaud (2019b). This approach favors the production of spontaneous speech and a more extensive description of the informants' background, and has been used to examine the phonological, morphosyntactic, and lexical features of Quebec English. However, the practical constraints on data collection entail a lower number of participants compared to dialect surveys, leading to quantitatively limited analyses of lexical phenomena in particular. A related type of information is provided by studies on the phonetic characteristics of Montreal English (e.g. Boberg, 2004b, 2005a, 2014).
- Corpus-based analyses, mainly conducted on newspaper articles from Quebec, have been used to investigate the lexical influence of French on Quebec English, principally from a qualitative perspective (e.g. Fee, 1991, 2008; Grant-Russell, 1999; Grant-Russell and Beaudet, 1999; Russell, 1996).
- These descriptions are complemented by anecdotal reports on the use of English in Quebec (e.g. Boberg, 2012; Grant, 2010).

Drawing on these studies, my aim will be to illustrate the principal Canadian English features that constitute the core of Quebec English, as well as to underscore those that distinguish it from other regional varieties, many of which are related to contact with French.

2.3.2 Phonetics and phonology

This section discusses some of the main characteristics of Quebec English pronunciation. In order to do so, it will present key distinctive features of Canadian English, which taken together serve to distinguish it from other major national varieties of English. Broadly speaking, these features also characterize English pronunciation in Quebec; however, regional specifics exist, mainly on the phonetic level, and they will be discussed as needed. Following Boberg (2010), the terms Standard Canadian, Standard British, and Standard American English (SCE, SBE, and SAE, respectively) will be used in this discussion.[10]

[10]SCE covers the relatively homogeneous speech spanning from British Columbia in the west to Nova Scotia in the east, as observed in the usage of the social majority comprised between the working class and the upper middle class (Boberg, 2010, p. 107). SBE corresponds to the variety known as Received Pronunciation, historically based on the speech of southeastern England. SAE is taken to comprise the varieties that are not usually

While phonological and phonetic patterns do not constitute an independent object of study in this dissertation, they are nevertheless highly relevant. As noted for Quebec French, a general understanding of phonological features is necessary in order to estimate the formal similarity of lexical items affected by cross-linguistic influence, as well as their integration into the recipient language. Additionally, these features can be used as a criterion to evaluate a speaker's participation in the community under study. For the sociolinguistic interviews conducted as part of this dissertation, this analysis is implemented in Chapter 14. Note that the present section will focus on a segmental account of Quebec English; suprasegmental features are traditionally less described and are of limited relevance in the context of this dissertation.

2.3.2.1 Consonantal features

In terms of its consonantal features, SCE is in many ways a typical North American variety. It is a rhotic variety, i.e. it systematically preserves the non-prevocalic /ɹ/. This realization historically precedes the development of non-rhoticity in Southern British varieties. As such, it was transported into British North American colonies and subsequently into Canada with the arrival of Loyalists (Boberg, 2010, pp. 131–132). Moreover, it has lost the opposition between the plain voiced /w/ and the preaspirated /ʍ/, which serve to distinguish pairs like *weather* and *whether*; this contrast is now a minority conservative feature (Boberg, 2010, pp. 124–125). While its evolution presents some regional differences, the youngest groups of speakers near-categorically use the voiced /w/ variant (Chambers, 2002, pp. 362–364).

A further feature shared with SAE is *t*-flapping or tapping. Here, the post-tonic, intervocalic and postrhotic /t/ is pronounced as the flap [ɾ], resulting in homophonous realizations of pairs such as *shutter* and *shudder*. The same process occurs after /l/ and /n/, but is less generalized (Boberg, 2010, pp. 135–136). Consistent with its variable status elsewhere in Canada, *t*-flapping is reported to be progressing in Quebec both in apparent and in real time.[11] It is more frequent in intervocalic contexts than after /n/ (Boberg and Hotton, 2015, p. 289), and it presents intra-speaker variability related to the degree of formality (Rouaud, 2019b, p. 220).

SCE is further characterized by the conditioned merger of /juː/ and /uː/, whereby the palatal glide is suppressed in specific contexts; this is also known as *yod*-dropping. Distinctively, SCE has *yod*-dropping after the coronal consonants /t/, /d/, and /n/, like SAE and unlike SBE. As a result, pairs such as *new* and *noon* have the same vowel. This usage is not categorical, but the trend is clearly towards the loss of the glide in these contexts (Boberg, 2010, pp. 134–134). However, there is ample evidence indicating that Montreal, as well as Quebec more generally, constitutes a more conservative area in this regard. It exhibits noticeably higher rates of *yod*-retention, even though they are coupled with variability across speakers and lexical items (Boberg, 2004a, p. 180; Boberg, 2004c, p. 264; Boberg and Hotton, 2015, p. 292; Hamilton, 1958, p. 75; Rouaud, 2019b, p. 219).

associated with specific regions, often in the Midland and the West of the United States (Boberg, 2010, p. 124).

[11] In variationist sociolinguistics, apparent time evidence is constituted by differences in language use that are stratified by age, based on the assumption that younger speakers reflect a more recent stage in the development of the linguistic system. Real time evidence is obtained through direct comparisons with data from an earlier point in time. See Chapter 6 for a more comprehensive discussion.

Some lexically specific consonantal features have been tentatively attributed to contact with French. For example, Boberg (2004a) reports a higher rate of *yod*-dropping in *coupon* for Montreal than for Southern Ontario (34% vs. 7%), noting that the glide-less pronunciation /ˈkuːpɒn/ is closer to the French equivalent (p. 184). In Quebec City, Chambers and Heisler (1999) examine the variable pronunciation of *asphalt*, with /s/ vs. /ʃ/. They report that the /s/ variant is associated with a higher degree of use of French, which has the same pronunciation in the corresponding item *asphalte* (pp. 29–31). A reflection of this trend is found on the regional level, with a higher rate of the /s/ variant in Montreal than in Southern Ontario, but here contact has not been explicitly invoked as an explanatory factor (Boberg, 2004a, p. 188).

2.3.2.2 Vocalic features

The vowel inventory of SCE is distinctively characterized by several mergers. It has lost the distinction between the TRAP and BATH lexical sets, with the vowel in both realized as /æ/ (Boberg, 2010, p. 126), including in Quebec (Hamilton, 1958, p. 75; Rouaud, 2019b, p. 218). It is also characterized by a series of *r*-conditioned mergers, which are explained by the fact that the realization of /ɹ/ mechanically limits the feasible vocalic oppositions in the preceding position. This has led to the merger of the vowels in *spirit* and *spear*; *sorry* and *sore*; *hurry* and *her*. In addition, most of Canada merges the vowels in *marry*, *merry*, and *Mary*. Montreal is a notable exception, as it only displays the merger between the *merry* and *Mary* vowels, resulting in /ɛɹ/, but it maintains the distinctiveness of /æɹ/ (Boberg, 2004a, p. 187; Boberg, 2008, p. 142; Labov et al., 2006, p. 219).[12] However, this distinction appears to be losing ground to the general Canadian trend in other parts of the province, including the Gaspé region (Boberg and Hotton, 2015, p. 302), the Eastern Townships, and to a lesser extent Quebec City (Chambers, 2007a, p. 33).

In the low-back quadrant, SCE displays a double merger of /ɒ/, /ɔː/, and /ɑː/, meaning that the items in the LOT, CLOTH, THOUGHT, and PALM lexical sets are homophones. This trend holds for nearly all of Canada (Boberg, 2010, pp. 126–131), and is well established in Quebec (Boberg and Hotton, 2015, pp. 287–288; Rouaud, 2019b, p. 218). The phonetic quality of the resulting vowel is regionally variable (Boberg, 2010, p. 128); in Quebec, a low-back, unrounded realization represented as /ɑ/ has been noted (Rouaud, 2019b, p. 218).

This process is additionally important because on the phonetic level it gives rise to the Canadian Shift, whereby the front short vowels lower and retract (Figure 2.6). More specifically, it is posited that the low-back merger frees up space into which /æ/ can move, becoming more central and lower. This similarly makes room for /ɛ/ to move, which finally triggers the lowering of /ɪ/. First investigated in detail by Clarke et al. (1995), the Canadian Shift is considered to be a defining characteristic of the

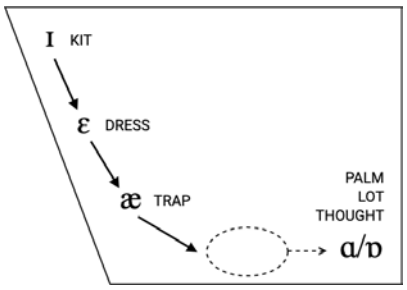


FIGURE 2.6: Schematic representation of the Canadian Shift, adapted from Clarke et al. (1995).

[12] The results in Boberg (2008) are discussed as representing Quebec, but it is noted that the participants' origin is "mostly greater Montreal" (p. 133).

English spoken in Canada: it unites the area comprised
between Vancouver and Montreal, and further distinguishes it from the bordering varieties spoken in the United States, which are undergoing a shift operating in the opposite direction (Boberg, 2010, pp. 146–147; Labov et al., 2006, pp. 219–221). In Montreal, it has been suggested that the shift more precisely involves parallel retractions of /ɪ/ and /ɛ/ rather than a chain movement (Boberg, 2005a).

Another phonetic characteristic is Canadian Raising, in which the nucleus of /aʊ/ and /aɪ/ is raised before voiceless consonants, resulting in realizations such as [əʊ] and [əɪ] for *lout* (but not *loud*) and *tight* (but not *tide*). Canadian Raising is neither uniformly distributed across the country nor limited to it – by these standards, the Canadian Shift is a better national indicator of language use – but it is stereotypically associated with Canadian English and as such occupies an important role (Boberg, 2010, p. 149–151). While Labov et al. (2006, pp. 220–221) suggest that it is limited to the inland area extending from Alberta to Ontario, Boberg (2008, pp. 138–141) shows on a larger sample that the regions absent from that initial isogloss – such as Quebec – display the same behavior, albeit with a degree of inter-speaker variability. This is further confirmed by the fact that Canadian Raising in Montreal is influenced by factors including gender (Rouaud, 2019b, p. 225) and ethnicity (Boberg, 2014, p. 68).

We should also note several lexically specific phonological variables, some of which have been extensively described in Quebec English. For instance, Montreal speakers tend to have /iː/ in *leisure*, *either*, and *lever*; they mostly use /ɒ/ in *progress* and *shone*. These observations are a matter of relative preference rather than categorical choice, with the majority rates ranging from 56% to 79% for the cited examples. However, they are relevant to note because they indicate variable preference for British and American phonological variants, going against a traditional view of Montreal as being more strongly influenced by American speech patterns (e.g. Hamilton, 1958). These choices moreover closely follow those observed in Southern Ontario, suggesting a fundamental similarity between the two regional varieties (Boberg, 2004a, p. 178). Some of the variables have also been examined elsewhere in Quebec. For example, the pronunciation of *lever* with /iː/ is similarly prevalent in the Gaspé region; however, real-time trends there indicate a shift to the American /ɛ/ variant among younger speakers (Boberg and Hotton, 2015, p. 289).

Finally, foreign *a* nativization – the adaptation of *a* in lexical items of foreign origin – tends to resolve in a general preference for /æ/ in cases such as *pasta* and *lava*. The main alternative, /ɑː/, usually only appears in cases such as *spa*, where phonological constraints preclude the use of /æ/. More recently, extraphonemic productions, occupying an intermediate position between the phonemes that exist in the Canadian English inventory, have also been observed. They affect instances such as the French loanword *façade*, with a lower, more centralized realization as [a], close to the corresponding French vowel (Boberg, 2010, pp. 138–139). This may be relevant for interpreting the integration of French-origin items in Quebec English (Rouaud, 2019b, p. 87).

2.3.2.3 Summary of phonological and phonetic features

The previously discussed characteristics of Canadian English pronunciation are summarized in Table 2.1. The features presenting a categorical (or near-categorical) realization situate Canadian English as a North American variety, distinguishing it from Standard British English (with a single exception). Within this broad trend, the features that are unique to Canada further distinguish it from Standard American English. Finally, the variable features may provide, through that very variability, evidence of the regionally specific status of Quebec in the general Canadian context, including based on its tendency to lag behind some of the trends observed in other regions.

Feature	**Realization**	**Other varieties**	**Quebec**
rhoticity	categorical	SAE	—
loss of /ʍ/	categorical	SAE, SBE	—
t-flapping	variable	SAE	—
yod-dropping	variable	SAE	conservative
TRAP-BATH merger	categorical	SAE	—
Mary-merry-(*marry*) merger	variable	SAE	conservative
other *r*-conditioned mergers	categorical	SAE	—
double low-back merger	categorical	SAE[a]	—
Canadian Shift	variable	—	—
Canadian Raising	variable	—[b]	conservative
foreign *a* nativization	variable	SAE, SBE	innovative[c]

TABLE 2.1: Main features characterizing the pronunciation of Canadian English, with the prevailing nature of their realization, the national varieties with which they are shared, and regionally specific trends in Quebec. (a) The low-back merger is less advanced in the United States, with some regions distinguishing LOT/PALM from THOUGHT. (b) Canadian Raising is also found in some areas of the United States, including those bordering Ontario. (c) This category involves multiple phonological variants, with preference variably aligning with the other national variants. The innovative role of Quebec is based on the assumption that French knowledge might facilitate the use of the emerging [a] variant.

2.3.3 Morphosyntax

In morphosyntactic terms, most Canadian English forms are largely shared with other national varieties of English (e.g. Chambers, 2010, p. 23). In this section, I will briefly discuss how Quebec English aligns with some of the broad Canadian patterns, pointing to its regional specificity. I will also address potential French influence on the morphosyntax of Quebec English. These patterns will not be directly investigated in this dissertation, but they illustrate the extent of contact-related phenomena in Quebec English and provide a basis for a discussion of their status in the next section.

A widely studied inflectional phenomenon is the emergence of irregular (strong) verbal forms in the past tense. Over the course of the 20th century, the preferred past form of *sneak* changed from *sneaked* to *snuck* across Canada, including Quebec. Within the province, it is interesting to note that Quebec City initially lagged behind by around two decades before catching up with other regions, including Montreal and the Eastern Townships (Chambers,

2007a, pp. 30–31). Similarly, *dive* exhibits near-categorical replacement of *dived* with *dove* in Southern Ontario (Chambers, 1998, p. 21) as well as Montreal (Boberg, 2004a, pp. 195, 198). In the case of this morphosyntactic change, Quebec seems to toe the national line, at most exhibiting a somewhat conservative character.

Another well-described trend, this time involving contact influence, is variable preposition use in Quebec English. Take for example the adjective *different*, which can be followed by the prepositions *from*, *to*, and *than*. In Quebec City, a relatively stronger preference for *from* has been observed among speakers more exposed to French. This has been tentatively interpreted in terms of contact influence, because the corresponding French adjective *différent* only allows the preposition *de*, a literal equivalent of *from* (Chambers and Heisler, 1999, pp. 31–32). Several similar patterns are anecdotally reported in Montreal English by Boberg (2012, pp. 497–498):

- expressing a value on a scale: "get 7 *on* 10" rather than "get 7 *out of* 10" on a test, cf. Fr. "avoir 7 *sur* 10";
- indicating street intersections: located "on St. Catherine, corner Peel" rather than "at the corner of St. Catherine and Peel", cf. QF "sur Sainte-Catherine, coin Peel";
- differences impacting specific verbs: "How much did you pay the car?" rather than "pay *for* the car" (in this case potentially explained by the Italian background of the speaker rather than French influence, although the two would be parallel).

But the influence of contact has not been confirmed in all investigated instances of morphosyntactic change. Reviewing a series of variationist sociolinguistic studies on Quebec English, Poplack (2008) excludes the possibility of convergence with French for both future temporal reference (En. *going to* vs. *will*; Fr. *aller* vs. synthetic future) and variable expression of relative pronouns (En. *that* vs. Ø; Fr. *que* vs. Ø). In both cases, language-internal constraints are reported to be different for English and French. Poplack also discusses three potential cases of divergent change: due to its minority status and the resulting isolation from mainstream Canadian varieties, Quebec English could be expected to lack innovations observed in other regions. However, all three cases – variable expression of deontic modality (and particularly the emergence of *need*), plural existentials with singular concord ("there's things that I have to do", p. 195), and the use of *be like* as a quotative verb – are reported to display the same conditioning factors as in other varieties. This is taken to confirm parallel patterns of change in Quebec English; however, innovative variants also tend to be used at a lower rate, suggesting that Quebec might lag behind in the adoption of changes originating elsewhere. Further evidence of conservative trends within the province is reported by Kastronic (2011), who finds that *be like* is adopted at a lower rate in Quebec City than in Montreal. While a direct effect of contact is excluded, this pattern is consonant with the idea of a more pronounced minority status, like in the case of Quebec City, reinforcing isolation from mainstream trends.

To summarize, in morphosyntactic change unrelated to language contact, Quebec English follows the nationwide trends, perhaps with a slight delay. In other cases, it presents clearly distinct usage which is likely explained by the influence of French. When direct effects of contact are refuted, Quebec English seems to revert to its conservative nature and lag behind mainstream dynamics; however, this also suggests that its minority status contributes to its

isolation. On the whole, then, the morphosyntax of Quebec English is not dramatically altered through language contact, but it does present observable influence of French. Even if this influence is largely limited to its conservative character, this is arguably not negligible. And yet, Poplack (2008, p. 197) is adamant that the research she reviews "offers no support for claims that QcE [Quebec English] differs from other varieties of Canadian English as a result of its minority status and sustained contact with French". Discarding the reported differences in variant use as minor, she argues that the perceived distinctiveness of Quebec English is likely related to borrowed lexical items. This is addressed in the next section.

2.3.4 Lexicon

Like the descriptions of Canadian English in general, early work on its lexicon mainly focused on differences with respect to British and American English. A turning point came from lexicography, with the work on the *Dictionary of Canadianisms on Historical Principles* (Avis et al., 1967; hereafter DCHP) explicitly investigating Canadian English for its own sake (Dollinger and Brinton, 2008, pp. 43–44).

One type of descriptions arising from this work has focused on the semantic domains contributing to the Canadian English lexicon. Historically, the importance of the local economy (e.g. fishing, fur trade) and the environment (flora and fauna) has been noted. In a reflection of the historical evolution of Canadian society, more recent lexical innovations have been reported in domains such as sports, government, and French-English relations (Dollinger and Brinton, 2008, p. 46). As for the linguistic origin of the lexical items specific to Canadian English, the role of borrowing has been noted. In the first edition of DCHP, 1,016 headwords (just over 11% of the total) were found to be borrowings. Of the 685 direct borrowings, 57% were of French origin, followed by 25% from First Nation languages, and 10% from Inuktitut (Harris, 1975, p. 36, cited by Dollinger and Brinton, 2008, p. 47). Lexical items such as these fall under the notion of Canadianism, defined as

> a word, expression, or meaning which is native to Canada or which is distinctively characteristic of Canadian usage though not necessarily exclusive to Canada. (Avis, 1967, p. xiii)

A more precise operationalization of this term was put in place for the second edition of DCHP. A clear distinction has now been drawn between six main types of Canadianisms, all of which apply to both lexical items and individual senses (except for type 3, which by definition only concerns senses):

(1) Origin: emergence in present-day Canada (e.g. *parkade* 'parking garage');
(2) Preservation: continued use in Canadian English despite obsolescence in other varieties (e.g. *pencil crayon* 'colored pencil');
(3) Semantic Change: acquisition of a new sense in Canada (e.g. *toque* 'beanie, woolen hat', which denotes other types of hats in British English);
(4) Culturally Significant: positive association with national identity or history (e.g. *Native Canadian*);

(5) Frequency: continued use in different national varieties, but with the highest frequency in Canada (e.g. *advanced green* 'traffic signal allowing a left turn');
(6) Memorial: negative association with national history (e.g. *residential school* 'school for cultural assimilation of Indigenous peoples') (Dollinger, 2022).

Note that both Avis' initial definition and Dollinger's typology highlight the fact that a usage can be considered as specific to a variety if it is relatively more frequent, but not exclusive to it.

This broad Canadian basis also constitutes the core of Quebec English vocabulary. It is moreover characterized by regionally specific lexical choices, resulting both from general, pan-Canadian regional variation, and from its local contact with French. As previously discussed, lexical descriptions of this variety comes from three main types of sources: dialect surveys; variationist sociolinguistic interviews; and corpus-based observations, mostly on newspaper texts and qualitative in nature, which are further complemented by anecdotal reports.

The remainder of this section mainly addresses the variable choice of semantically equivalent lexical items (onomasiological variation). Differences in the meaning of individual lexical items (semasiological variation) are discussed in Section 2.3.5 in terms of existing descriptions, as well as in in Chapter 3 from a theoretical standpoint. Moreover, the focus for now remains on providing an overview of general characteristics of Quebec English; the precise factors conditioning their use are explored in Chapter 6.

2.3.4.1 Regional distinctiveness of Quebec

As noted above, descriptions of Canadian English have often relied on the use of dialect surveys, which provide extensive evidence of patterns of regional variation. An important source of information on Quebec is the Dialect Topography Project, which has investigated a set of phonological, morphosyntactic, and lexical variables across Canada, starting in the early 1990s (Chambers, 1994). In Quebec City, Chambers and Heisler (1999) draw particular attention to several lexical variables indicating locally specific trends. Like elsewhere in the country, the term *couch* is preferred over *sofa* and the Canadianism *chesterfield*, but the adoption of this item is slower and follows a different path than in Ontario, with a relatively higher frequency of *sofa*. A potential effect of contact is posited for two other variants, which exhibit surface similarity with a corresponding Quebec French expression: *soft drink* (rather than *soda* or *pop*; cf. QF *liqueur douce*) and *bureau* (rather than *dresser* or *chest of drawers*; cf. QF *bureau*). Though not strongly preferred overall, these variants are particularly frequent for speakers with the strongest personal links with the city, suggesting greater exposition to the effects of contact.

In Montreal, Boberg (2004a, pp. 183–186) similarly reports a locally higher frequency for *sofa*, *soft drink*, and *bureau*, further noting that the use of *sofa* may be related to the formally identical French equivalent. He also observes the same type of preference for *supper* (rather than *dinner*; cf. QF *souper*), as well as *balcony* and *gallery* (rather than *porch* or *veranda*; cf. Fr. *balcon*, QF *galerie*). Substantially similar patterns are outlined by Boberg and Hotton (2015) in the Gaspé region.[13] Taken together, the results discussed so far indicate that different regions of Quebec tend to use lexical variants that also exist elsewhere in Canada – as is the case

[13] Boberg and Hotton (2015) report on the following cases: *sofa*, p. 291; *supper*, p. 297; *soft drink*, p. 303.

with the ones highlighted here – but with a slight divergence from national trends reflected by differences in the distribution of variants, which is potentially related to the influence of French. However, as noted by Boberg (2004a, pp. 184–185), the attribution of contact influence is not straightforward, and the use of supposedly contact-induced variants in other varieties should be taken into account.

A broader view of variation is provided by the North American Regional Vocabulary Survey (NARVS) (Boberg, 2005b). Partly inspired by the Dialect Topography Project, it systematically analyzes a larger set of lexical variables – including some typical of Quebec – in terms of regional distribution across North America, investigated on a sample of 1,800 Canadians and 360 Americans. Based on the extent of variation observed across the 44 variables retained for the analysis, it concludes that "in Canada, the strongest lexical boundaries were found to divide the English-speaking community of Montreal from neighboring regions to the east and west" (p. 53); to put it more clearly, "Montreal appears to be the most lexically distinct region in Canada" (p. 36). Most of the regionally distinctive lexical items in the city are related to French influence (Table 2.2), with highly similar patterns reported in other parts of Quebec (Boberg, 2010, pp. 170–188; Boberg and Hotton, 2015).[14] Moreover, although some variants specific to Montreal are not majority responses, their local importance is confirmed by their limited use in other provinces (Boberg, 2005b, p. 36). In Quebec, they tend to be more widely used outside Montreal, likely due to greater exposure to French (Boberg, 2012, p. 501).

Mechanism	**Quebec variant**	**Other variant**	**French source**
lexical transfer	*dépanneur*, *dep*	*convenience store*	*dépanneur*
	guichet	*ATM*	*guichet*
	stage /stɑːʒ/	*internship*	*stage*
loan translation	*all-dressed* (pizza, burger)	*everything-on-it*	*toute garnie*
	one-and-a-half, *two-and-a-half*	*studio apartment*	*un et demi*, *deux et demi*
semantic shift	*cash*	*check-out*	*caisse*
	chalet 'lakeside house'	*cottage*	*chalet*
	trio 'sandwich, fries, drink'	*combo*	*trio*
other	*copybook*	*notebook*	—
	running shoes	*sneakers*	—
	see-saw	*teeter-totter*	—
	soft drink	*pop*	(*liqueur douce*)

TABLE 2.2: The most distinctive Montreal items, as reported by Boberg (2005b, p. 36). The potential French sources for *cash* and *soft drink* are taken from the subsequent discussion in Boberg (2010, p. 173). The proposed underlying mechanisms have been added to the initial analysis. Note that the survey includes additional alternative variants beyond the ones provided here.

A particularly interesting case is the preference for *chalet* over the alternatives *cottage* and *cabin* to refer to a house in the countryside, often by a river or a lake, where people go on

[14] Boberg (2010) extends the analysis of the NARVS data on 22 participants from Quebec outside Montreal, in addition to 394 participants from Greater Montreal (mean number of participants per question; no further information on geographic origin provided). Boberg and Hotton (2015) report on 124 participants from the Gaspé region. They use a different questionnaire, but it includes many NARVS variables, including those in Table 2.2.

summer weekends. Its use is likely related to the formally identical French equivalent, whose use in Quebec might have led to a shift away from the otherwise widespread sense of *chalet* 'ski lodge'. In parallel, this shift resolves an overlap with the use of *cottage* to refer to a two-story city house, also attested in Montreal (Boberg, 2005b, p. 42). This example neatly illustrates the potential interaction between onomasiological and semasiological variation, which we will explore in more detail in Chapter 5.

2.3.4.2 Types of contact-induced lexical influence

Important information on the Quebec English lexicon, particularly as concerns additional types of contact-induced influence, is provided by corpus-based research conducted on newspaper articles from Quebec (e.g. Fee, 1991, 2008; Grant-Russell, 1999; Grant-Russell and Beaudet, 1999; Russell, 1996). Most of these studies are qualitative in nature and hence do not establish clear estimates of the diffusion of the described phenomena, but they provide dozens of attested examples of contact-related usage. Lexical influence of French is reported in fields including linguistic policy (*francize* 'teach French to immigrants', cf. Fr. *franciser*), provincial institutions and services (*cégep* 'junior college in Quebec's educational system'), culture (*vernissage* 'exhibition opening'), food (*poutine* 'French fries with cheese curds and gravy'), and transportation (*metro* 'subway', cf. Fr. *métro*, the official name of the Montreal underground railway system) (Grant-Russell, 1999, p. 477). These reports, coupled with dialect surveys and anecdotal observation, form the basis of a typology of contact-related influence in Quebec English proposed by Boberg (2012, p. 501). He distinguishes:

(1) elective direct lexical transfer: the use of a French lexical item for which a functionally equivalent English alternative exists (e.g. *dépanneur* rather than *corner store*);
(2) imposed direct lexical transfer: the use of a French lexical item related to Quebec institutions, often lacking an English alternative due to the regional specificity of the referent (e.g. *cégep* 'junior college in Quebec's education system');
(3) loan translations or calques: literal translation of the subparts of a French lexical item (e.g. *all-dressed* 'with all the toppings', cf. *toute garnie*);
(4) semantic shifts: changes in the meaning of existing English words, including older loans (e.g. *chalet* 'summer cottage').

More detailed analyses of the underlying linguistic processes have been proposed in lexicographically oriented research on French-origin items in English (Josselin, 2001; Yuen, 1994). In addition to identifying examples and the contact mechanisms behind them, corpus-based research has also provided indications regarding the constraints on the use of these items. For instance, borrowed items exhibit different degrees of integration into English, as reflected by orthographic features such as accents; flagging, i.e. setting apart the borrowed item using typographic or metalinguistic mechanisms; and morphosyntactic integration (Grant-Russell, 1999, pp. 478–481). Similarly, differences in the use of French-origin lexical items have been reported between different newspapers (Fee, 1991, p. 13) and text genres (Grant-Russell, 1999, pp. 483–484). This is reminiscent of earlier questionnaire-based observations by McArthur

(1989) showing that the acceptability of French-origin items is widely variable. Related observations in spoken language come from the sociolinguistic interviews conducted in Montreal by Rouaud (2019b). For instance, three quarters of French borrowings in her corpus are phonologically adapted to English (p. 261), with the outcome influenced by the segment in question and the speaker's degree of bilingualism (pp. 250–256). Taken together, this suggests that their use in Quebec English constitutes a prime site for variationist investigation, given variable patterns of use, likely social conditioning, and the potential for these items to convey social meanings.

2.3.4.3 Importance of language contact

The studies discussed so far reach a broad consensus in acknowledging the lexical influence of French in Quebec English. The opposite view is defended in the influential work conducted on the Quebec English Corpus (Poplack et al., 2006), which vigorously challenges the importance of lexical phenomena in this situation of language contact. By extension, it also calls into question the relevance of the object of study pursued in this dissertation. It is therefore important to take a closer look at the arguments advanced in this line of research.

The Quebec English Corpus is composed of sociolinguistic interviews with 183 speakers from Montreal, Quebec City, and the control monolingual region of Oshawa-Whitby in Ontario. Poplack et al. (2006) use it to investigated the effects of language contact, focusing in particular on borrowings and codeswitching, which are respectively defined here as single-item or multiple-item fragments of French discourse (p. 207). The authors highlight the overall rarity of both types of insertions: borrowings are found to constitute on average 0.07% of the total lexicon of a speaker (p. 210);[15] similarly, a third of the speakers never switch, and a further quarter only do so once or twice during the interview (p. 209). It is additionally underscored that both borrowings and codeswitches are principally used in metalinguistic or other "special discourse" purposes, further calling into question their effective integration (pp. 208–209). The central conclusion drawn from these observations is that "this is hardly the kind of bilingual language use that can be expected to lead to contact-induced change" (p. 210). In short, this analysis refutes any consequential impact of French on Quebec English, including its lexicon, contrary to much of the earlier research.

An important issue to be noted is the definition of contact-induced language change adopted in this analysis: it sets out to investigate an "assumption" of previous research "that lexical manifestations of contact function as agents of structural change, an idea with no basis in scientific fact" (Poplack et al., 2006, p. 186). The implication here is that lexical phenomena are inherently less important than other types of contact-induced language change, a view which is not universally supported in the language contact literature (cf. Fee, 2008, p. 184). I would further argue that the earlier studies of Quebec English reviewed in this section do not routinely invoke the potential for borrowing to lead to system-wide structural change; rather, lexical influence is investigated as a standalone effect of language contact. It has also been noted that the reported rarity of borrowings may be related to methodological choices: for instance, the

[15] It is unclear if this observation refers to types or tokens. An earlier analysis focusing on a single speaker (p. 207) reports the number of tokens.

authors exclude expressions or compounds, proper names, and loanwords whose dictionary attestation precedes the birth of the informant (Grant, 2010, p. 183). The last criterion noted by Grant is particularly important because it entails overlooking subtler differences in usage, such as a higher frequency of an existing word or the development of a new sense under the influence of French (as also noted by Fee, 2008, p. 179; Rouaud, 2019b, pp. 159-160). This consideration is especially relevant in a study of semantic shifts:

> Excluding a cognate because it appears in a Canadian English dictionary does not take into account that the meaning of such cognates may differ in QE from Canadian English or World English. [...] this approach leaves out evidence of contact, that is, integrated loanwords that are more common in QE than in [Canadian English] because of pressure from a high frequency French cognate or whose meanings have shifted, or both. (Fee, 2008, p. 179)

Moreover, the proportion of other-language items in the whole vocabulary is arguably not the best way to estimate the impact of language contact. This metric disregards the typical distribution of word frequency: a small fraction of all lexical items are highly frequent, whereas the vast majority are comparatively very infrequent (cf. Zipf, 1932). As such, borrowed forms would not be expected to constitute vast proportions of the recorded speech production to begin with; however, this does not preclude them from having high symbolic value, pointed out by Boberg (2012, p. 495). More generally, the use of French insertions may be influenced by the informant's familiarity with the interviewer (Fee, 2008, p. 183), the topic of the conversation (Rouaud, 2019b, p. 160), or other social factors. For example, the raw numbers reported by Poplack and colleagues show that both borrowings and codeswitches are used more frequently in Quebec City than in Montreal, especially by speakers born after the passage of Bill 101 (Poplack et al., 2006, pp. 208, 211), but discussion of this trend is regrettably absent.

Contrary to the position defended by the authors, I would argue that this study fails to conclusively demonstrate the limited importance of contact-induced lexical phenomena in Quebec English. In addition to potential methodological issues in quantifying the phenomena under study, it crucially disregards some instances of cross-linguistic influence whose importance is supported by other types of qualitative and quantitative evidence.

The discussion of Quebec English lexicon began by highlighting its shared Canadian basis, after which I turned to its regional specifics. First, I discussed several dialect surveys conducted across Canadian regions, which have highlighted the regional specificity of Quebec English and have attributed this feature to its ongoing contact with French. I then briefly reviewed corpus-based research that has brought to light a wider range of contact-induced lexical phenomena, and has contributed to explaining the mechanisms conditioning their use. Finally, I discussed an influential variationist sociolinguistic study which has called into question the relevance of lexical influence of French in Quebec English. I have suggested that this view is skewed by methodological choices, as well as broader issues such as word frequency distribution. Overall, the results reported in the literature outline of the potential for Quebec English speakers to actively introduce lexical material from French in contexts ranging from informal conversations to newspaper articles.

2.3.5 Previous work on contact-induced semantic shifts

To conclude the discussion of the main features of Quebec English, let us take a look at the phenomenon that is at the center of this dissertation: contact-induced semantic shifts. In keeping with the preceding sections, I will provide a brief overview of existing descriptions in order to illustrate the range of described phenomena that result from contact-related influence on the lexical semantic level. In Chapter 3, I will draw on these examples to provide a precise definition of contact-induced semantic shifts and outline the theoretical view that I adopt in investigating them. The resulting methodological implications are discussed in Chapter 5. An example of a lexicographic and corpus-based analysis of a range of specific examples is introduced in Chapter 11.

2.3.5.1 Existing descriptions

As we have just seen, the influence of French on the lexicon of Quebec English has been described in a variety of studies. Some of them describe the effects of language contact on the lexical semantic level, i.e. modifications of the meaning of existing English words. However, these accounts are often anecdotal, and lacking in terminological and theoretical clarity. For instance, Boberg (2012) uses the term *semantic shifts* to describe cases such such as *animator* 'group leader', explaining that "older French borrowings, which today have different meanings in English and French, can revert to their French meanings in QE" (p. 497). Another example is the study by Poplack et al. (2006), which briefly discusses the use of *install* 'settle' (cf. Fr. *s'installer*) (p. 195). The authors describe it as a calque, even though it does not correspond to typical examples of calques, where subparts of a linguistic expression are translated literally; in fact, this example behaves just like Boberg's semantic shifts.

Even where care is taken to properly define the phenomenon under study, another issue arises: that of analyzing its quantitative diffusion. For example, Rouaud (2019b, p. 245) uses the term *semantic loan* to describe the use of *campaign* with the sense of the French lexical item *campagne* 'countryside'; however, it only occurs once throughout her interviews, thereby precluding any generalizations. Similarly, in an earlier study, I examine a set of previously described Quebec English lexical items exhibiting the semantic influence of French (Miletic, 2018). This is based on a manual analysis of large newspaper corpora, which proves doubly challenging: in addition to the quantitative dispersion highlighted by Rouaud, it is often difficult to reliably determine which specific sense is used.

Arguably the most comprehensive analysis of contact-induced lexical semantic phenomena in Quebec English is the study conducted by McArthur (1989). From a methodological standpoint, he circumvents the issues related to quantitative evaluation by using a written questionnaire, ensuring comparable results provided by 200 respondents. And while he investigates a range of contact-related lexical effects, most of the items involve a semantic modification of a preexisting English word. For each of the 25 items, the respondents are provided with a definition and an example; they are asked to rate it on a 4-point scale of acceptability, ranging from "universally accepted" to "locally ambiguous" expressions. The results indicate wide variability both across speakers and individual items: cases such as *animator* 'group leader' and

collectivity ‘people as a whole, community’ enjoy broader support than *library* ‘bookstore’ and *demand* ‘to ask for something’ (p. 42). (These examples are related to the French lexical items *animateur*, *collectivité*, *librairie*, and *demander*, respectively.) Major dialect surveys have also investigated some instances of contact-related semantic influence, but they do so from an onomasiological perspective, i.e. as one of the lexical items that can be used to express a given meaning. This is the case of the previously discussed preference for *chalet* rather than *cottage* or *cabin* in Montreal (Boberg, 2005b, p. 36). By contrast, McArthur examines contact-related senses in their own right, as reflected by the acceptability ratings.

These phenomena have also been described in the previously discussed corpus-based research on newspaper texts. As before, the contribution of this line of research lies in identifying new examples and identifying the linguistic mechanisms which may underpin them. For instance, a two-pole distinction based on the degree of linguistic integration is proposed by Fee (2008, pp. 177, 180–181). She distinguishes:

- non-integrated borrowings, which are equated with gallicisms, *faux amis*, and false friends, without providing further distinctions between these terms (e.g. *deceive* ‘disappoint’);
- cognates that undergo semantic shifts, where a new meaning is acquired from a word with which another meanings is already shared (e.g. *population* ‘the people’).

Similarly, Grant (2010, pp. 186–187) draws an equivalence between the notions of semantic shift, semantic extension, false cognate, false friend, and *faux ami*. They are defined as English words used with the meaning of a formally similar French word (e.g. *manifestation* ‘demonstration’). But she also notes that some semantic extensions may enter accepted usage (e.g. *animator* ‘group leader’).

Moreover, both Fee and Grant discuss another type of influence, where low frequency English words are used more often because of the existence of higher frequency cognates in French (e.g. *furnish* ‘provide’; Fee 1991, p. 14). This category may additionally involve “a slight semantic shift, with the English usage reflecting nuances of the French cognate” (Grant, 2010, p. 186). In other cases, semantic shifts may affect the word’s connotation (e.g. *functionary* shifting from negative to neutral connotation) or degree of formality (e.g. *ameliorate* shifting from formal to neutral register) (Fee, 2008, p. 181).

As mentioned in the previous section, another important source of information is lexicographic analysis of these types of contact-induced phenomena (Josselin, 2001; Yuen, 1994). Unlike the other studies reviewed here, this line of research does not focus on the constraints on the use of these lexical items in the speech community, but rather analyzes them in an applied perspective. Nevertheless, they provide additional examples of contact-related semantic phenomena, as well as outline very detailed typologies of this type of behavior. They will be presented in more detail in the theoretical discussion in Chapter 3. But first, let us take a step back to discuss the main takeaways from the existing studies, as well as the questions that remain open.

2.3.5.2 Rationale for further work

The studies discussed so far demonstrate the potential for Quebec English to exhibit contact-related influence on the lexical semantic level. While various specific types of influence have been noted, the prototypical example is arguably the situation where an existing English lexical item and a phonologically similar French lexical item partly overlap in meaning. This is schematically represented by the case of En. *animator* 'creator of animated films' and Fr. *animateur* 'creator of animated films; group leader'. As a result, the English lexical item acquires the sense typical of the French one, as in *animator* 'group leader'.

The existing literature is valuable in that it provides dozens of such examples from diverse data sources, as well as tentative explanations of the underlying linguistic mechanisms. But this does not amount to a comprehensive sociolinguistic description. In the variationist framework, we would expect variable language use to be investigated in terms of the linguistic and social factors that constrain it, as well the social meaning that it conveys. When it comes to social constraints, the role of the degree of bilingualism is underscored by McArthur (1989) based on his written survey. Research on newspaper corpora has suggested regional differences between Montreal and Quebec City (Fee, 1991) and, beyond Quebec, between Montreal and Toronto (Miletic, 2018). In terms of linguistic constraints, most reviewed studies point to factors such as the degree of semantic overlap or of phonological similarity between English and French lexical items; however, they are not systematically examined. Similarly, stylistic variation is observed indirectly, based on differences between print publications (Fee, 1991; Grant-Russell, 1999). It can also be safely assumed that this type of usage is prescriptively stigmatized, as indicated by the development of software aiming to avoid it (Yuen, 1994). But we do not have reliable quantitative estimates of the extent to which these or other factors condition lexical semantic effects of language contact, or how widespread they are in the speech community.

This is not without reason. While variationist sociolinguistic research is well established, analyses of lexical semantic phenomena have remained limited (cf. Part II). As a result, a theoretical framework that can account for empirically observed differences in meaning by correlating them with linguistic and social factors is not as readily available. This issue is addressed in Chapter 3, which outlines the main principles guiding the semantic descriptions conducted in this work, and in Chapter 5, which presents the approach I adopt in analyzing lexical semantic phenomena within the variationist sociolinguistic framework. From the methodological standpoint, data collection and analysis present additional problems, especially when working with spontaneous speech. The solution proposed in this dissertation consists in relying on an interdisciplinary approach, drawing on methods from both variationist sociolinguistics and natural language processing; this is presented in detail throughout Part II.

That said, the existing descriptions may be imperfect, but they clearly illustrate that the use of contact-induced semantic shifts is variable, and likely conditioned by different types of factors. As such, it constitutes a prime site for variationist sociolinguistic exploration, which in turn motivates further work on computational models of lexical semantics.

2.4 Summary

This chapter has taken a look at at the general context in which mechanisms of language contact operate in Quebec, starting with the sociohistorical and demographic position of the province's official language communities. As we have seen, the French-speaking population was the first to arrive, with permanent settlements of New France established in the early 17th century. A turning point was marked by the British conquest in 1763, which led to the arrival of a substantial English-speaking population; despite its minority status, it occupied a position of political, social, and economic dominance, well into the 20th century. The tide began to turn in the 1960s, with the increasing affirmation of Francophone Quebecers in political life. This trend was reinforced starting from 1977, with the passage of Bill 101 leading to transformative change in the balance of power between the two linguistic communities. The minority status of the Anglophone community, its intense exposure to French, as well as a high rate of bilingualism among Quebecers in general, constitute factors facilitating the emergence of contact-induced linguistic features.

This chapter has also reviewed some of the main features of Canada's two official languages, as they are spoken in the province. Quebec French is characterized by an affirmation of an endogenous norm, i.e. it is clearly defined in its own right. Its distinctive characteristics on the phonological, morphosyntactic, and lexical levels are frequently indicative of the degree of bilingualism among English speakers; they also point to intense interaction between the two linguistic communities, including in terms of participation in ongoing language change.

As for Quebec English, it has been principally described in terms of its regionally distinctive status within the more general context of Canadian English. In terms of pronunciation, it is characterized by a typically North American phonemic inventory (e.g. low-back merger), with additional characteristics typical of Canadian English in general (e.g. Canadian Raising), as well as some that distinguish it from other Canadian varieties (e.g. *merry-marry* distinction). Its morphosyntax and to an even greater extent its lexicon present clear effects of language contact with French. On the lexical level in particular, this influence constitutes a key regional differentiator of Quebec, whether it operates through direct lexical transfer (*dépanneur*), loan translation (*all-dressed*), semantic shift (*chalet*), or other processes. More generally, for both Quebec French and English, I have argued for a broad view of linguistic communities. This specifically extends to all individuals who are able to speak the languages in the province.

Finally, the existing descriptions of the main object of study investigated in this dissertation – contact-induced semantic shifts – have also been presented. As we have seen, these accounts provide compelling evidence of lexical semantic influence of French on Quebec English. However, a more comprehensive analysis of the diffusion of these phenomena, of the social and linguistic constraints on their use, and of the social meaning that they convey is sorely needed. This, I have argued, is related to theoretical and methodological challenges in addressing language variation from a lexical semantic standpoint. I turn to the first of these two issues in the next chapter.

Chapter 3

Contact-induced semantic shifts

In the previous two chapters, we have seen that the ability to speak multiple languages can lead to different types of cross-linguistic influence at the level of individual speakers, and that these individual behaviors can give rise to community-level patterns of variation and change. One such type of contact-related influence, also described in the context of Quebec English, is the presence of semantic shifts. They will constitute the main object of study of this dissertation. However, their definition and analysis pose numerous theoretical challenges; this chapter presents the position that I adopt with regard to these issues.

Section 3.1 draws on the previously presented studies of contact-induced semantic shifts to provide a consolidated definition of this phenomenon. Section 3.2 identifies a series of issues which may impact the subsequent linguistic analyses and the way in which I will address them. Section 3.3 provides a brief summary of this discussion. Note that I will use the term *lexical item* to broadly refer to basic units of the lexicon, which associate a form with a meaning (e.g. Ilson, 1992). I will use the term *sense* to refer to one of several meanings that can be conveyed by a polysemous word (e.g. McArthur, 1992b).

3.1 Defining contact-induced semantic shifts

Lexical semantic effects of language contact are studied from different discipline-specific standpoints, both in a synchronic and in a diachronic perspective. This section provides a brief overview of the central notion of semantic shift, and more generally presents some of the views on the broadly similar issues of diachronic semantic change, synchronic semantic variation, and contact-related semantic influence. Drawing on a summary of these approaches, I then outline the definition of the object of study pursued in this dissertation.

3.1.1 A general view of semantic shifts

The term *semantic shift* is used with varying degrees of specificity and from different temporal perspectives. In the literature on Quebec English, it tends to describe the general fact that some lexical items can acquire senses related to the influence of French. In other language contact studies, the focus is similarly on the specific type of cross-linguistic influence: Haugen

(1950) applies the term *shifts* to "changes in the usage of native words [...] that are not strictly phonological or grammatical" (p. 219). His term *loanshift* is similarly used by Mott and Laso (2019) to refer to a type of borrowing consisting in the extension of a word's meaning (p. 158).

To the extent that most existing Quebec English studies are based on a comparison of regions rather than an analysis of change over time, they adopt a synchronic perspective. In lexical semantics, however, the term *semantic shift* tends to be used interchangeably with *semantic change* to refer to a diachronic change in meaning. That is the case, among others, in Geeraerts (1997, 2010) and Traugott (2017); explicit equivalence between the two terms is drawn by McArthur (1992a). In other cases, the term *semantic shift* is more specific. One such example is Koch (2016), for whom it corresponds to innovative meaning change leading to polysemy (p. 27).

A different view is presented by Koptjevskaja-Tamm (2016), who uses the term *semantic shift* as a cover notion which

> refers to a pair of meanings A and B which are linked by some genetic relation, either diachronically (cf. Latin *caput* 'head' and French *chef* 'chief') or synchronically, e.g. as two meanings of a polysemous lexeme (cf. English *head*, as in *I've hit my head*, i.e. 'top part of body', and as in *I've met my department head*, i.e. 'leader of others') (p. 1).

Although an implict diachronic dimension associated with the term is recognized, the author overtly applies it to both synchrony and diachrony. The same position is taken by other authors, such as Zalizniak et al. (2012). Similarly, Newman (2016) sees the term *semantic shift* as applying to changes in meaning over long stretches of time as well as contextual, idiosyncratic meaning modifications. But drawing on cognitive linguistic research, he argues that meaning should be viewed dynamically, as being constructed and potentially evolving in context, rather than being subject to a strong synchronic/diachronic dichotomy (p. 269).

In this dissertation, *semantic shift* will be used in the manner outlined by Koptjevskaja-Tamm, i.e. as a general notion referring to the presence of a link between two senses. Its application to contact-induced linguistic practices in Quebec English will be more clearly defined in Section 3.1.4.4. But first, let us turn to other principal approaches to differences in word meaning, starting with diachronic semantic change.

3.1.2 Diachronic semantic change

It is generally accepted that meaning change in diachrony proceeds gradually, through incremental changes to the sense inventory of the lexical item. The original meaning may only be lost over time, as a result of subsequent steps in the process of change (Traugott, 2017). By this definition, diachronic semantic change is inextricably linked to polysemy. More precisely still, "polysemy is, roughly, the synchronic reflection of diachronic semantic change" (Geeraerts, 1997, p. 6).

Accounting for diachronic patterns of semantic change can be vital in a sociolinguistic analysis of lexical semantic phenomena, particularly when it comes to determining the status

of the observed usages. Take for example the Invited Inferencing Theory of Semantic Change (IITSC) developed by Traugott and Dasher (2002), which outlines the following steps in the process of semantic change:

- shift from a conventional coded meaning to an utterance-token meaning:
  a lexical item can be interpreted as having the new meaning in a specific context;
- shift from the utterance-token meaning to an utterance-type meaning:
  the new meaning is the default interpretation, but it can be canceled out by contextual specifications;
- shift from the utterance-type meaning to a coded meaning:
  the new meaning is encoded alongside the old one and may replace it over time.

The described mechanisms are reminiscent of longstanding views on semantic change, going back at least to Paul's (1891) observation that an occasional meaning, i.e. a specific usage event, may through repetition give rise to a usual meaning, i.e. one that is encoded in the lexical item. This approach is also reflective of recent usage-based accounts of contact-related lexical influence, including in the context of Quebec English (e.g. Rouaud, 2019a). Moreover, the pragmatic dimension on which the IITSC is based is particularly relevant for contact-related change: language choice, as well as specific behaviors such as codeswitching and semantic interference, are closely related to the communicative context and issues such as the interlocutors' degree of bilingualism.

In addition to diachronic processes of change, semantic effects of language contact can also be observed in synchrony. This is the focus of the next section.

3.1.3 Synchronic semantic variation

When it comes to synchronic analyses of semantic variation, a particularly relevant framework is that of cognitive sociolinguistics. Its general aim is to bring together the preoccupations of cognitive linguistics – a usage-based view of language in which meaning occupies a central role – with those of sociolinguistics, broadly understood as focusing on the social context in which language is used (Geeraerts et al., 2010; Geeraerts and Kristiansen, 2014).

One of the issues addressed in cognitive sociolinguistics is lexical variation from a semasiological perspective, i.e. variation in the senses with which a lexical item is used at a given point in time (for a further discussion of this notion, see Section 3.2.5). As Pütz et al. (2014) point out, this complements traditional sociolinguistic approaches, where lexical variation in general, and lexical semantic variation in particular, is rarely addressed. At least part of the problem lies in the traditional definition of the sociolinguistic variable as "two alternative ways of saying the same thing" (Labov, 2004, p. 7), which is difficult to reconcile with the study of variation in the meaning expressed by a single form. Additional issues such as fuzzy boundaries between meanings further complicate the picture. A potential solution is to adopt a cognitive semantic view of meaning as a non-discrete but structured category, and analyze the way in which it varies in terms of traditional sociolinguistic variables (Pütz et al., 2014, pp. 8–9).

Importantly, cognitive sociolinguistics has been used as a theoretical background for analy-

ses of semasiological variation from a variationist sociolinguistic standpoint. Although limited in number, several studies have been published in this vein. Investigations have been conducted on the meaning of *awesome* (Robinson, 2010), *gay* (Robinson, 2012a), *skinny* (Robinson, 2012b), and *cheeky* (Bailey and Durham, 2020). These studies provide fine-grained semantic analysis, combining etymological information, occurrences in reference corpora, and data elicited through interviews or questionnaires. The analyses rely on traditional sociolingustic variables, including geographic origin, age, gender, and socioeconomic status. Another relevant study is Budinich's (2016) work on semasiological variation in Italian (e.g. *disinteresse* 'indifference' vs. 'unselfishness') related to communicative context as reflected by different topical subsections of a large corpus. Although he does not focus on speaker characteristics, this is another example of a convincing corpus-based analysis of semasiological variation.

The specific methods deployed in this line of work, as well as the more general issues related to the study of lexical semantics in variationist sociolinguistics, will be addressed more extensively in Chapter 5. For now, let us shift the focus from general processes of semantic variation and change to more specific cross-linguistic mechanisms.

3.1.4 Semantic shifts in a contact situation

Lexical semantic effects of language contact have been examined in different strands of research. I will address three of them, focusing on lexical semantic change, language contact in general, and language contact in Quebec English in particular. I will then provide an overview of the main theoretical choices across these studies.

3.1.4.1 Research on lexical semantic change

From the point of view of lexical semantics, language contact is not an issue of primary importance. It is, however, occasionally addressed in research on diachronic semantic change, specifically in the work on the mechanisms that drive it (Traugott, 2017; for a historical overview, see Geeraerts, 2010, pp. 25–44).

For example, in his classification of semantic change mechanisms, Geeraerts (1997) introduces the notion of analogical change, corresponding to the case when the new meaning of an expression "cop[ies] the semantics of another, related expression" (p. 94). This may be motivated by a syntagmatic relationship (i.e. co-occurrence), a phonetic similarity, or a semantic similarity between the two expressions. The three motivations are not mutually exclusive; moreover, traditional non-analogical mechanisms of change (metaphor, metonymy, generalization, and specialization) may also be involved. Although contact-related examples are provided, analogical change is presented as a general mechanism of semantic change which may also operate within a single language.

3.1.4.2 Research on language contact

In the context of language contact studies, lexical semantic phenomena are usually addressed within the wider focus on the lexical influence that a source language (SL) exerts on a recipient

language (RL). This goes back at least to Paul (1891), whose monograph on historical language change includes a brief discussion of calques and of semantic interference by analogy. He specifically refers to bilingual speakers using a RL word with a meaning typical of a SL word, with the two sharing another related meaning (pp. 471–472).

A similar account is found in other classical works on language contact. Haugen (1950) introduces the notion of *loanshift* as a cover term for *semantic loans* (e.g. American Portuguese *humoroso* 'capricious' acquiring the meaning 'humorous', cf. American English *humorous*) (pp. 214–215, 219). He also discusses calques or *loan translations* (e.g. Fr. *gratte-ciel* modeled on En. *skyscraper*) (p. 220). In terms of the motivation for the borrowing process, he differentiates between *analogues*, or lexical items that are both semantically and phonetically similar; *homophones*, when the similarity is only phonetic; and *homologues*, when the similarity is only semantic. He suggests that analogues are the most likely to give rise to borrowing. As for the outcome of the borrowing process, Haugen distinguishes between *loan homonyms*, when the borrowed meaning is unrelated to the conventional meaning, and *loan synonyms*, when the two meanings partly overlap.

In his seminal monograph on language contact, Weinreich (1953) further elaborates on these processes. He posits that a RL lexical item may undergo semantic extension following the model of a SL lexical item with which there is semantic or phonetic similarity. From an onomasiological perspective, this process may lead to the disappearance of a concurrent lexical item which was originally used in the RL to express the contact-related meaning. From a semasiological perspective, the affected lexical item in the RL acquires a new meaning, which in some cases entirely supplants the original one. Building on Haugen's analysis, but modifying his terminology, Weinreich distinguishes a situation where the new and the original meaning are logically linked (polysemy) and one where they are unrelated (homonymy) (pp. 48–49).

3.1.4.3 Research on Quebec English

Let us now turn to descriptive accounts of contact-related phenomena in the specific context of Quebec English. In discussing the existing sociolinguistic studies of Quebec English in Chapter 2, I reviewed varied sources of information on contact-induced semantic shifts produced in this strand of research (Boberg, 2005b, 2012; Fee, 1991, 2008; Grant, 2010; McArthur, 1989; Poplack et al., 2006; Rouaud, 2019b). Although they constitute the starting point of my work, lexical semantics rarely constitutes the primary focus of these studies. This likely explains the frequent lack of clarity regarding their theoretical and terminological positions, with limited explanations of the underlying mechanisms of semantic influence or categorizations of the resulting lexical items (for a more extensive overview of these issues, see Section 2.3.5).

The view emerging from these studies can be clarified by existing lexicographically-oriented research. Although its focus is not on the sociolinguistic characteristics of contact-induced semantic shifts, it provides fine-grained analyses of the underlying linguistic mechanisms. One such categorization is outlined by Yuen (1994). Her analysis focuses on all types of contact-related lexical influence in Quebec English, and it is in this context that she addresses semantic influence. She specifically discusses:

- borrowing of meaning, when an English word acquires a new meaning associated with a French word, with the following subcategories:
  - *faux-amis*, when there is a formal similarity between the English and French words. Partial *faux-amis* share some but not all preexisting senses, whereas full *faux-amis* share no preexisting senses;
  - transfer of primary meaning, when there is no formal similarity between the English and French words;
- gallicisms of frequency, involving an English word which is formally similar to a French word. The two partly or entirely overlap in meaning, and the English word is then used more frequently with the shared meaning. There are no borrowed elements, only frequency is impacted;
- gallicisms of usage, involving an English word which is formally similar to a French word. While they partly overlap in meaning, their senses are not identical. Typically, one word is general and the other specific, or one is abstract and the other concrete. This leads to the English word being used in different contexts, which are in fact typical of the French word;
- syntagmatic gallicisms, which are defined as calques of English word combinations (compounds and collocations). Although this is presented as a distinct phenomenon, Yuen acknowledges that around half of the cases involve a borrowing of meaning in a verbal collocate.

A similar analysis is put forward by Josselin's (2001) research on French-English and English-French borrowing in Canada and France. The following descriptive categories involve a semantic dimension:

- semantic borrowing: a lexical item in the RL acquires a meaning associated with a lexical item in the SL under the influence of their formal similarity (En. *conference* 'lecture', cf. Fr. *conférence*);
- semantic calque of a simple word: a simple word in the RL acquires a meaning associated with its translation equivalent in the SL, without the influence of formal similarity (Fr. *bienvenue* 'de rien', cf. En. *welcome*);
- calque of a complex word, and specifically one of its subtypes, which consists in using an existing RL word with a meaning typical of a SL word obtained by translating the individual morphemes (En. *attendance* 'audience', cf. Fr. *assistance*);
- calque of an expression: literal translation of a SL collocation or idiom (En. *abandon a course*, cf. Fr. *abandonner un cours*);
- borrowing of usage:
  - frequency: choosing a RL lexical item over a more widespread RL equivalent under the influence of a formally similar SL lexical item (En. *manifestation* 'demonstration');

- meaning in context: using a RL lexical item, whose decontextualized meaning corresponds to that of a formally similar SL lexical item, in a context in which that item would not otherwise be used in the RL (En. *permit* 'driver's licence', cf. Fr. *permis*);
- preservation: continued use of an otherwise obsolete lexical item in the RL under the influence of a formally similar lexical item in the SL (En. *ignored* 'not known', cf. Fr. *ignoré*).

Overall, these studies contribute valuable detail to the analysis of contact-related semantic influence. However, like in the previously discussed sociolinguistic research, other important aspects could be described in more detail. For instance, the type of cross-linguistic analogy at play (formal, semantic, or both) is generally not analyzed in a systematic manner. In most existing studies, the way in which the new meaning is integrated into the polysemic structure of the affected lexical item is not addressed; neither is the key distinction between polysemy and homonymy. Moreover, the distinctions between the different categories are not always clear: for example, Yuen (1994) and Josselin (2001) both discuss collocations affected by language contact, but their examples can also be analyzed as semantic borrowings or borrowings of usage if the focus is shifted on the impacted lexical item within the collocation. Finally, specific analyses may differ depending on the author: for instance, *manifestation* 'demonstration' is analyzed as a borrowing of usage by Josselin (2001), and as a *faux-ami* by Grant (2010).

3.1.4.4 Overview of analyses on contact-related semantic influence

While the mechanisms of contact-related semantic influence discussed by different authors vary, several dimensions involved in this process can be identified. They are summarized below.

Underlying similarity. Contact-related semantic influence is driven by some type of similarity between RL and SL lexical items, whether it be semantic, formal, or both. Most authors focus on one of the two types of similarity; if both are mentioned, they are usually addressed non-systematically. Out of the works reviewed here, only Haugen (1950) explicitly analyzes all possibilities. A summary from key sources is presented in Table 3.1.

Object of influence. Different specific realizations of contact-related semantic influence have been described. First, an individual RL lexical item can acquire a new denotational meaning. This is the most emblematic case, which in practical terms would be reflected by an additional meaning appearing in the dictionary entry for the lexical item in question. Second, an individual RL lexical item may be used in a different linguistic context, with a different connotation or degree of formality, or with an increased frequency. Third, frequent word combinations in the RL (collocations or idioms) may be modified. However, as discussed above, this may in fact involve contact-related modifications of individual items within the word combination.

Outcome. Influence driven by semantic similarity (including in combination with formal similarity) results in polysemy. Influence driven by formal similarity in isolation results in

Source	Type of cross-linguistic similarity		
	semantic and formal	**semantic**	**formal**
Paul (1891)		(semantic interference)	
Haugen (1950)	loanshifts / analogues	loanshifts / homologues	loanshifts / homophones
Weinreich (1953)	semantic extension		
Geeraerts (1997)	analogical change		
Yuen (1994)	partial faux-amis		full faux-amis
Josselin (2001)		semantic calque of a simple word	semantic borrowing; calque of a complex word

TABLE 3.1: Types of semantic influence in language contact settings. Empty cells do not indicate that the corresponding category is excluded by the author, but rather that it is not explicitly addressed.

homonymy. This distinction is principally discussed by Haugen (1950) and Weinreich (1953).

3.1.4.5 A definition of contact-induced semantic shifts

Drawing on the preceding discussion, I will adopt a broad view of contact-induced semantic shifts in Quebec English, understood as the presence of a specific sense in a preexisting English word that is explained by the presence of the equivalent sense in a formally and/or semantically similar French word. The earlier discussion of different objects of semantic influence in the context of language contact is further echoed by the traditionally established distinction between denotational meaning, corresponding to "the basic referring function of language", and non-denotational meaning, related to "emotive or stylistic overtones" that lexical items may carry (Geeraerts, 1997, p. 18). Building on these considerations, I propose an analysis of contact-related semantic influence on three levels of meaning. They are presented below, together with typical examples from Quebec English.

Denotational meaning. In this case, the affected Quebec English lexical item presents a sense associated with a formally and/or semantically similar French lexical item. This may involve the general process of innovative meaning change leading to polysemy (Koch, 2016), which is reflected by the process of semantic borrowing in the context of language contact. It may also involve an increase in the frequency of a preexisting sense, including obsolete senses. Contact-related innovations range from a clear change in referent (*animator* 'group leader' in addition to 'creator of animated films') to more nuanced cases involving phenomena such as generalization or narrowing of meaning (*entourage* 'circle of friends' in addition to 'group of people attending an important person').

Connotational meaning. The Quebec English lexical item has an unchanged inventory of senses, but it presents an emotive or stylistic value that is associated with a formally and/or

semantically similar French lexical item. For instance, *souvenir* 'memory' is described as literary by the OED, whereas in Quebec English it is routinely used in neutral contexts.

Collocational meaning. The Quebec English lexical item, used with a preexisting sense, frequently cooccurs with another lexical item, due to comparable collocational properties of a formally and/or semantically similar French lexical item. For instance, *abandon* may appear in the collocation *abandon a course* instead of *drop a course* (cf. Fr. *abandonner un cours*). This does not seem to involve a change in denotational meaning, as one of the conventional senses of *abandon* is 'give up completely (a practice or a course of action)'; rather, the modification appears to be limited to a more frequent use in this specific context.

It should be acknowledged that the distinction between the three levels is not always clear-cut. For instance, Yuen (1994) suggests that in some cases slight shifts in denotational meaning may accompany collocational differences. However, this overview clearly indicates the types of cross-linguistic influence that are considered relevant in this dissertation, and to that extent provides important guiding principles for the analyses presented in the following chapters. It should also be complemented with more specific views on the analysis of the semantic structure of lexical items; this issue is addressed in the next section.

3.2 Describing contact-induced semantic shifts

The examples of semantic shifts as well as their categorizations presented in the previous section point to several issues which may arise while analyzing them. Since I consider that semantic shifts affect preexisting lexical items, it is important to be able to distinguish occurrences which correspond to one lexical item from those that correspond to another, particularly in the case of phonological identity. And since one of the main ways in which semantic shifts manifest themselves involves the acquisition of a new sense, it is equally important to be able to distinguish between different senses of a lexical item. More generally, this process can be addressed both from the standpoint of different lexical items conveying the same meaning, as well as that of the sense inventory of a single lexical item.

The remainder of this section outlines the position I take with regard these issues. I will first discuss the adopted view of word meaning, before focusing on how it is reflected by the distribution of lexical items in linguistic contexts. I will then address several types of indeterminacy: on the level of senses, polysemy and vagueness; on the level of lexical items, homonymy and heterosemy. The section will conclude with a discussion of two possible perspectives in the study of word meaning: onomasiology, which starts from a sense and looks at the lexical items which can be used to express it; and semasiology, which starts from a lexical item and analyzes the senses associated with it.

3.2.1 A general view of meaning

The descriptive analyses conducted in this dissertation will rely on empirically occurring linguistic data. In this context, bottom-up approaches to the analysis of word meaning play an important role. One such view is formulated by Taylor (1992): his representation of semantic structure starts with individual occurrences of a lexical item on the lowest level, which are then progressively linked together into ever more abstract senses. He posits that the most salient senses – those that we routinely access in language production and interpretation – are situated at an intermediate level of representation, corresponding to that of basic level concepts (cf. e.g. Geeraerts, 2010, pp. 199–203). A more recent but substantially similar view is put forth by Gries (2015). It is based on the idea of individual usages represented as points in multidimensional space. In this conception, the senses of a lexical item correspond to groupings of usages that speakers identify in the semantic space. The senses are not static; depending on the context, speakers may approach the space from different angles, they may identify similarities between different points, or they may condense parts of the space (pp. 482–483).

This cognitive linguistic view is reminiscent of other usage-based operationalizations of word senses. For example, Kilgarriff (1997) argues that senses correspond to the groupings of similar occurrences of a lexical item produced by lexicographers. This approach crucially highlights the mutable status of senses: those that are posited for a single lexical item may vary depending on the purpose for which they are defined and the corpus that is used in that process. A key takeaway remains that senses roughly correspond to similar uses that speakers make of a lexical item, and that precise boundaries between senses are difficult to establish.

This dissertation will also rely on top-down information on the sense inventory of lexical items. This will be provided by a range of lexicographic sources, used in guiding initial explorations of the data and in validating more complex analyses. Further details on this approach are presented in Chapter 5. For now, let us turn to another usage-based approach to word meaning: distributional semantics. It similarly draws on individual occurrences of a lexical item to form a representation of its meaning, and it will play a central role in the corpus-based analyses conducted in this dissertation.

3.2.2 Distributional patterns

As Lenci (2008) notes, distributional semantics can be seen as a group of related approaches sharing the basic assumption that the semantic behavior of a lexical item can be characterized, at least to some extent, by its statistical distribution in linguistic contexts. A central notion in this framework is that of similarity. That is also the case for many other theories of lexical semantics; the distinguishing feature here is the assumption that the similarity between two lexical items, and all derived observations, can be defined in terms of linguistic distributions (pp. 1–2).

The general view has come to be known as the Distributional Hypothesis, which can be formulated as follows:

> The degree of semantic similarity between two linguistic expressions *A* and *B* is a

> function of the similarity of the linguistic contexts in which *A* and *B* can appear. (Lenci, 2008, p. 3)

The origins of distributional semantics can be traced back to structuralism, and particularly Harris (1954). In his work, the general distributional approach first applied to the study of other levels of linguistic structure was extended to lexical semantics, providing an empirically grounded way of studying some aspects of word meaning. But distributionalism was soon left behind by theoretical linguistics; in the following decades, it was dominated by emerging currents such as generativism and, in the case of lexical semantics, cognitive linguistics. Distributional semantics nevertheless survived, and even thrived, in the field of corpus linguistics. Firth's (1957) oft-cited dictum, "You shall know a word by the company it keeps" (p. 11), outlines the basic idea underpinning systematic computational modeling of word meaning. The diffusion of distributional semantics has paralleled the rise in importance of corpus-based approaches over the last four decades (Lenci, 2008, pp. 4–6).

But before we get to computational implementations, it should be noted that adopting a distributional semantic approach is not theoretically inconsequential. In the present dissertation, a weak Distributional Hypothesis is adopted: I do not argue that linguistic distributions represent, say, the mental organization of semantic meaning; rather, I assume that the meaning of a lexical item determines its distributional behavior, and that an analysis of distributional contexts can uncover some relevant semantic characteristics (cf. Lenci, 2008, p. 14). Put otherwise, I argue that distributional representations capture word meaning not such as it is represented in our minds, but such as it is attested in the texts from which the representations are constructed (Sahlgren, 2008, p. 49).

While this implies that there are aspects of word meaning that are not captured by distributional representations, they still have important advantages, as Boleda (2020) points out. First, distributional semantic models are based on attested linguistic data, so they are "radically empirical" (p. 215), unlike many other currents of lexical semantics. Second, the distributional representations used in current computational implementations are highly multidimensional; that is to say, they encode many different types of linguistic information, whereas traditional descriptive methods tend to focus on a very limited number of features due to practical constraints. Finally, distributional representations are graded: for instance, similarity between these representations is measured using continuous values. This is reflective of many possible degrees of similarity between two lexical items, of the ability to observe similarity along some – but not all – dimensions of lexical meaning, and so forth (pp. 214–216).

I will come back to the principles of distributional semantics when I present the related methodological issues in Chapter 5. But first, let us turn to several more fine-grained theoretical distinctions, starting with different types of indeterminacy.

3.2.3 Delimiting senses: vagueness, polysemy, semantic relations

As implicitly suggested by the discussion of bottom-up approaches to meaning, I take it that most lexical items in a language can be used with different senses. This issue is closely related to different types of indeterminacy and semantic relations, which are discussed in this section.

Geeraerts (1993) addresses the distinction between polysemy and vagueness, which is related to determining "whether a particular piece of semantic information is part of the underlying semantic structure of the item, or is the result of a contextual (and hence pragmatic) specification" (p. 228). For instance, *neighbor* does not specify whether the denoted person is male or female; this information is not contained in its semantic structure and is inferred from context. Contrast that with *plain*, which may carry the meaning 'simple' or 'ugly': while context may help to disambiguate it, this semantic information is encoded in the lexical item itself. Although it is clearly important to take this distinction into account, Geeraerts also shows that the tests that are traditionally used to differentiate vagueness from polysemy produce unreliable results. He therefore contends that the distinction between vagueness and polysemy is in fact unstable.

In a similar vein, Tuggy (1993) outlines a model based on cognitive linguistic theory which posits a continuum between ambiguity and vagueness, with polysemy occupying a central position. Prototypical examples of the three cases are *bank* ('river bank', 'financial insitution') for ambiguity; *paint* ('color a wall', 'create an art piece' etc.) for polysemy; and *aunt* ('mother's sister', 'father's sister') for vagueness.

A further descriptive implication advanced by these studies is the empirically observed continuity between phenomena such as homonymy and polysemy: for instance, there are borderline cases with an existent but tenuous semantic link between word senses. More broadly, it is essential to keep in mind that not all indeterminacy is the same, and that more robust analyses can be provided by determining if specific usages are related to contextual specifications or to the polysemic structure of the item at hand.

Finally, the description of the sense inventory of a lexical item affected by language contact can benefit from a specification of the semantic relations at play. This is particularly relevant because different types of semantic change can give rise to different semantic relations (e.g. Koch, 2016, p. 31). As suggested by the previously reviewed examples of semantic shifts, contact-induced senses are often linked to the preexisting senses with the following relations:

- cohyponymy, in the case of a clear change in referent:
  animator 'group leader' and 'creator of animated films', where both senses can be analyzed as hyponyms of 'skilled worker';
- hypernymy, in the case of semantic generalization:
  entourage 'circle of friends' relative to 'group of people attending an important person';
- hyponymy, in the case of semantic narrowing:
  permit 'driver's license' relative to 'legal document granting permission'.

We now move from the level of senses to the level of lexical items to discuss additional types of indeterminacy.

3.2.4 Delimiting lexical items: homonymy, heterosemy

The importance of some types of indeterminacy emerges from the exploratory analyses conducted as part of this dissertation. Take for example the English noun *dodo*: it is attested in

the Montreal data with the meaning of 'sleep', which is typical of its French homograph and entirely unrelated to the conventional English meaning referring to the extinct flightless bird. Given the lack of etymological or semantic relationship between the two, *dodo* 'sleep' represents an instance of homonymy. As such, it could be described as a borrowing (of a lexical item and its associated meaning) rather than a semantic shift (as it is understood in this dissertation; see above). This issue is explicitly discussed by Weinreich (1953, p. 49): he underscores the difficulty of determining which of the two processes is at play, but does not provide a direct answer on how it should be addressed.

Another related issue is the change of grammatical category. A particularly recurrent scenario in the Montreal data is the occasional use of common nouns as proper nouns: for instance, *plateau* is frequently used in Montreal to refer to the neighborhood of Plateau-Mont-Royal. Different positions are adopted in the literature with respect to the change of grammatical category. For instance, in cognitive linguistics, Gries's (2006) corpus-based analysis of the verb *to run* includes a discussion of nominal uses. By contrast, Budinich (2016) explicitly limits his analysis of semasiological variation in Italian to a single part of speech. As for the existing descriptive studies addressing semantic shifts in Quebec English, they mostly circumvent this issue. For example, Yuen (1994) describes the nominal use of *polyvalent*, referring to a type of secondary school in Quebec, as borrowing of meaning coupled with change of category, but she does not provide further detail.

From a theoretical standpoint, Lichtenberk (1991) points out that polysemy involves a single lexical item without syntactic differences (i.e. differences in grammatical category). He introduces another notion, that of heterosemy, to analyze semantic similarities reflected in lexical items of different grammatical categories originating from the same etymon. He underscores that, despite their historical links, the semantic features of heterosemous lexical items ultimately might not be shared or even similar (p. 480). This can be interpreted as an argument either for the exclusion of semantic shifts involving a change of grammatical category or for a specific treatment of these cases.

3.2.5 Distinguishing perspectives: semasiology, onomasiology

As suggested above, different perspectives can be adopted in analyzing lexical semantic phenomena. A central terminological and conceptual distinction in this respect is that between onomasiology, which analyzes the lexical items used to express a given meaning, and semasiology, which analyzes the meanings associated with a given lexical item. This distinction is routinely made in semantic change research (e.g. Traugott, 2017) and lexical semantics at large. It often has to do with the granularity of research: for example, Koch (2016) defines lexical change as any change affecting the lexicon, and it is within this notion that he introduces the distinction between onomasiological and semasiological perspectives. He argues that a complete analysis can only be provided by combining the two, by analyzing meaning change from a semasiological point while setting it against the backdrop of onomasiological phenomena (p. 23).

A comparable position is present in the classical literature dealing with contact-related se-

mantic change. Paul (1891) addresses semantic change from a semasiological perspective, but he also alludes to onomasiological consequences in terms of the adaptation of the rest of vocabulary. A similar idea – that the semantic extension of a word may lead to the disappearance of a concurrent form that was previously used with the same meaning – is advanced by Haugen (1950) and Weinreich (1953, pp. 53–56) in their discussions of borrowing. But while this distinction can be teased out from the text, it is never made explicit by the authors.

In more recent work, Geeraerts (1997) splits the mechanisms of semantic change into onomasiological and semasiological, with analogical change – the mechanism allowing for contact-related influence – in the latter category. But he also underscores that a strict distinction between the two categories comes down to the perspective that is adopted. Moreover, onomasiological mechanisms include semasiological ones because semasiological extension is one mechanism of onomasiological change (pp. 94–95). From another perspective, analogical change necessarily involves an onomasiological perspective given that it analyzes how different lexical items influence one another (Geeraerts, 2010, p. 54).

Given my focus on the semantic influence of formally similar lexical items across different languages, I will take these lexical items as the starting point and then investigate whether their senses are affected by language contact. The implications of this perspective for a variationist sociolinguistic study conducted in synchrony are further discussed in Chapter 5. Moreover, although this semasiological perspective constitutes the main point of view adopted in this dissertation, it will be complemented by more focused onomasiological analyses to clarify the patterns of use of the lexical items of interest.

3.3 Summary

This chapter has reviewed key theoretical issues related to the analysis of contact-induced semantic shifts, aiming to provide a clearer understanding of the view that I adopt in this dissertation. I first focused on a definition of this object of study, drawing on existing research on diachronic semantic change, synchronic semantic variation, and lexical semantic effects of language contact. I proposed a broad view of contact-induced semantic shifts in Quebec English, corresponding to the presence of a specific sense in a preexisting English word that is explained by the presence of the equivalent sense in a formally and/or semantically similar French word. I further suggested that this phenomenon may involve effects on the denotational, connotational, and collocational levels of meaning.

Based on this definition and the discussed examples, I then briefly presented several issues which have implications for the resulting description of semantic shifts. Specifically, I outlined an empirically grounded view of word meaning, which assumes that most lexical items are polysemous, that their individual occurrences provide a starting point in identifying their senses, and that the result of this process is not immutable but rather depends on the adopted perspective and the data used for the analysis. I then more extensively presented the notion of distributional semantics, which formalizes some of these general principles and will be used as the basis of the corpus-based analyses conducted in this dissertation. I further underscored the

importance of distinguishing between different types of indeterminacy affecting the meaning of a given lexical item, as well as between formally identical but semantically or grammatically different lexical items. Finally, I drew a distinction between onomasiological and semasiological perspectives on the study of lexical semantic phenomena.

In the remainder of this dissertation, I will mainly address the general phenomenon of contact-induced semantic shifts through an analysis of synchronic semasiological variation, contrasting the meanings typical of Quebec English with those used in other regional varieties or by specific subgroups of speakers. The interpretation of these patterns will be complemented as needed with a diachronic perspective (examining the emergence of specific meanings in Quebec English over time), as well as with onomasiological considerations (analyzing the use of other lexical items that are similar in meaning). The methodological background for this approach is presented in Part II.

Part II

An interdisciplinary approach

Part I has provided a general background on the object of study pursued in this dissertation: it investigates contact-induced semantic shifts, seen here as an effect of bilingualism and specifically studied in the context of Quebec English. In order to examine this behavior empirically and systematically on the scale of a speech community, I draw on methodologies developed in two disciplines with very different traditions: sociolinguistics and natural language processing.

Sociolinguistics can be defined as "that part of linguistics which is concerned with language as a social and cultural phenomenon" (Trudgill, 2000, p. 21). In practice, this is an umbrella term referring to many different strands which utilize both qualitative and quantitative methods to investigate the relationship between language and society from different standpoints. All of these positions share a strong empirical orientation and base their descriptions on analyses of documented language use; however, this work specifically adopts a variationist sociolinguistic perspective. This approach is grounded in and motivated by the notion of "orderly heterogeneity" in language (Weinreich et al., 1968, p. 100), i.e. the fact that all linguistic systems display variability which is nevertheless structured in nature. With this in mind, the aim of variationist studies is to observe language variation and uncover patterning which may explain it. In doing so, reference is made to both internal (linguistic) and external (social) constraints on language use. One motivation behind studying synchronic language variation is to use it as a reflection of diachronic processes helping to uncover the complex principles behind language change (Labov, 1994, 2001, 2010). Another is to understand the social meaning that speakers seek to convey when they make different linguistic choices (Eckert, 2000).

As we will see in the pages that follow, despite the wealth of studies conducted over the last six decades or so (i.e. since Bright, 1966), variationist sociolinguistics has remained firmly focused on phonological and, to a lesser extent, morphosyntactic phenomena. While the description pursued in this work is strongly informed by variationist principles, the traditional methodologies developed within the discipline must be complemented to ensure an exhaustive account. That is why I also turn to natural language processing (NLP), which provides methods to efficiently collect and process vast amounts of data, as well as model lexical semantic phenomena at scale so as to identify linguistic patterns of descriptive interest. This is not the first work to apply NLP methods to language variation, as demonstrated by the emerging field of computational sociolinguistics (Nguyen and Cornips, 2016) and numerous studies on computational models of diachronic semantic change (Tahmasebi et al., 2021). But although these computational approaches are promising, their descriptive contribution is yet to be established (Boleda, 2020); implementing them in such a way that they provide informative and reliable linguistic descriptions remains a challenge in its own right.

The next chapters will discuss in more detail how and why variationist sociolinguistic and NLP methods can be brought together to address the descriptive issue defined at the outset. Chapter 4 presents the criteria and practices to construct different types of corpora capturing language variation. Chapter 5 overviews the strategies to isolate patterns of semasiological variation in the collected data. Chapter 6 presents different ways of accounting for the observed linguistic variation, both in terms of the factors that motivate it and of the social significance it achieves. Drawing on this overview, Chapter 7 more precisely defines the research objectives and outlines the main elements of the approach implemented in the remainder of the study.

Chapter 4

Data for language variation

Although they rely on very different methods, both variationist sociolinguistics and natural language processing are empirical scientific disciplines. This chapter reviews some of the main ways they provide of collecting naturally occurring linguistic data to study language variation within and across communities of speakers. Section 4.1 addresses this issue from the standpoint of variationist sociolinguistics. It particularly focuses on the structure of the classic sociolinguistic interview, but it also explores other data collection methods commonly used in the discipline. Section 4.2 discusses corpus construction relying on publicly available social media data, focusing in particular on Twitter. This type of communication has attracted some sociolinguistic interest as a variant in its own right. However, I will mainly view it as an alternative data source, likely similar in nature to face-to-face communication, which crucially facilitates the construction of very large linguistic corpora. It will become clear in the next chapter that this is a key practical requirement for the NLP methods implemented in this study. Finally, Section 4.3 provides a summary of the main points.

It should be noted that Sections 4.1 and 4.2 place emphasis on the methodological issues considered as central in the respective disciplines, but they overall address the same set of problems: the characteristics of the targeted communities of speakers and their linguistic behaviors; the practical process of collecting and filtering data; and the limitations of each of the approaches. Note also that the scope of this chapter is limited to an overview of common approaches to collecting data which capture language variation. The ways in which this variation can be modeled and explained are respectively presented in Chapters 5 and 6. Data collection carried out as part of the present study is discussed in Chapter 8 (for Twitter data) and Chapter 12 (for sociolinguistic interviews).

4.1 Sociolinguistic corpora

Research conducted within the variationist sociolinguistic framework is dependent on reliable linguistic data, as was made clear by William Labov's founding studies: "our initial approach to the speech community is governed by the need to obtain large volumes of well-recorded natural speech" (Labov, 1972, p. 208). While variationists record the language production of carefully chosen speakers, in much of this tradition individual ways of speaking are merely

seen as a means to understanding the language use of the community at large. Or, to put it in Labov's words, "the community is prior to the individual" (Labov, 2006, p. 5).

Building on this view, this section will first provide a more precise definition of speech communities. It will then present the criteria used to select a sample of speakers, collect data, and prepare it for analysis. Finally, the limitations of this approach will be discussed.

4.1.1 Defining speech communities

In presenting general characteristics of bilingual language use, Chapter 1 defined language communities based on the language that their members use. While this is useful in understanding broad social dynamics between speakers of different languages, the analysis of variation in a given language relies on a different perspective. In variationist sociolinguistics, it is traditionally considered that

> the speech community is not defined by any marked agreement in the use of language elements, so much as by participation in a set of shared norms; these norms may be observed in overt types of evaluative behavior, and by the uniformity of abstract patterns of variation which are invariant in respect to particular levels of usage (Labov, 1972, pp. 120–121).

This view initially arose from Labov's work on New York City, which found that despite considerable variation between different speakers in using the examined linguistic variables, their subjective evaluations of language use were uniform and distinct from speakers from other regions (Labov, 2006, p. 6). Crucially, as Patrick (2002, p. 586) points out, "Labov's conception requires *reference* to a set of shared norms – not deference or uniform adherence". In other words, even though departures from general trends may be observed, it is important to identify these trends.

Another prism by which communities of speakers are analyzed is that of social networks. A social network can be defined as "the aggregate of relationships contracted with others" (Milroy, 2002, p. 549); these relationships are then analyzed in terms of their structures and properties. For practical reasons, these analyses focus on personal social networks, which are still assumed to exist within a wider social framework. A crucial observation arising from sociolinguistic work focusing on this issue is that

> networks constituted chiefly of strong (dense and multiplex) ties support localized linguistic norms, resisting pressures to adopt competing external norms. By the same token, if these ties weaken conditions favorable to language change are produced. (Milroy, 2002, p. 550)

A related approach relies on the notion of community of practice, defined as "a collection of people who engage on an ongoing basis in some common endeavor" (Eckert, 2006, p. 683). This concept was initially developed in a social theory of learning (Lave and Wenger, 1991; Wenger, 2000), and was first used in sociolinguistics to study language and gender (Eckert and

McConnell-Ginet, 1992a,b). It moves beyond an analysis of communities based on geographic location or social characteristics and focuses instead on shared practice; it is understood that this practice may involve the development of a specific linguistic style. In comparison with the speech community view, this approach crucially allows to associate broad patterns of variation with the meanings that speakers construct in specific communicative situations (Eckert, 2006).

A final point to be noted is related not to the definition of speech communities but to their comparison. This idea is at the basis of the comparative method in sociolinguistics, which consists in

> comparing the patterning of variability in each possible source. If the conditioning effects on the variable linguistic features show patterns approximating those found in a putative source, we can conclude that they represent structures drawn from that source. [...] On the other hand, where there are dissimilarities, we have grounds for concluding that the phenomena in question belong to different linguistic systems. (Tagliamonte, 2002, p. 732)

This view underpins the idea of comparing language uses originating from different speech communities, and will become particularly relevant in later stages of this work. For now, I turn to the general principles guiding data collection in variationist sociolinguistics.

4.1.2 Creating a sociolinguistic corpus

Having established that sociolinguistics is based on the study of language data produced by speakers, let us now define which specific type of speech production is targeted. It is understood that all speakers use a range of styles; in other words, they adapt some of their linguistic choices depending on factors such as context and topic. These styles can be ordered based on the amount of attention paid to speech. The focus of sociolinguistic studies is the vernacular, "the style in which the minimum attention is given to the monitoring of speech" (Labov, 1972, p. 208). The search for vernacular data is closely related to another central methodological issue known as the observer's paradox: "the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation" (Labov, 1972, p. 209).

In order to obtain speech production approaching the vernacular, the impact of the observer's paradox must be reduced as much as possible. This problem guides much of data collection in variationist sociolinguistics. Different solutions to it have been proposed; I will discuss them in relation to the sociolinguistic interview, the standard method of data collection, as well as some of the alternative approaches. But first, let us take a look at another practical issue: how to choose the members of the speech community to include in a study.

4.1.2.1 Sampling speakers from a community

The criteria and the specific way in which members of a speech community are chosen to participate in a sociolinguistic study depend on the aims of the study and the deployed data collection technique. Construction pipelines for other types of corpora, including those based on

social media data, more heavily rely on post-collection data filtering to identify the speakers of interest (see Section 4.2.2). By contrast, in variationist sociolinguistics all sampling decisions are taken before any data collection occurs.

The way in which a sample is constructed strongly influences how representative the data are of the wider speech community and hence how strong the conclusions drawn from the data are. In *random sampling*, anyone from the sample frame – some list of the population, such as a phone directory – has an equal chance of being drawn to participate in the study. A subtype of this approach is *stratified random sampling*, where relevant social categories (e.g. age, gender etc.) are first identified, and then each of the cells corresponding to the final category is filled with a random sample of the population corresponding to that category. In *quota* or *judgment sampling*, target categories are similarly defined at the outset, but the choice of informants is led by the researcher's judgment rather than any random method. In practical terms, the quotas are often filled using the "snowball" technique, i.e. through the social network of initial participants. Also known as the "friend of a friend" technique, this method crucially helps establish a rapport between the researcher and the participants. Although it is not representative of the wider population, a judgment sample is easier to implement than random sampling and it can provide important evidence of linguistic variation. However, given its subjective nature, it is important for inclusion decisions to be theoretically grounded (Milroy and Gordon, 2003, 24–33). The same range of approaches, including random sampling and social network methodology, are used in variationist studies of language contact (Poplack, 1993, pp. 262–265).

As for the categories used to stratify the sample, they depend on the specific aim of the study, but they routinely include age, gender, and some indicator of socioeconomic status (Tagliamonte, 2006, p. 23). In language contact studies, the degree of bilingual ability is another crucial explanatory variable, even when it does not constitute an inclusion criterion (Poplack, 1993, p. 255). We will see in detail how these and other factors may explain language variation in Chapter 6.

The number of categories used in constructing the sample directly influences another important decision: how many speakers are to be included in the study. In principle, the larger the sample, the more reliable the analysis. In practice, however, sample size must take into account the time and resources available to process the recorded data. Studies where an unrealistically large number of speakers is interviewed run the risk of never exploiting some of the recordings (Tagliamonte, 2006, pp. 32–33). In filling the cells which constitute the structure of the sample, five persons per cell is often considered as an adequate lower limit (Feagin, 2002, p. 29), with many influential studies using samples of well under 100 participants (Milroy and Gordon, 2003, p. 29). The sampling and recruitment procedure used for the sociolinguistic interviews conducted in this dissertation are presented in Section 12.2.2.

Once the sample is designed and recruited, data collection begins. The most commonly used method in the discipline is the sociolinguistic interview; it is presented in the next section. This is followed by an overview of other frequently used data collection approaches.

4.1.2.2 The sociolinguistic interview

Pioneered by William Labov, the sociolinguistic interview is a carefully structured exchange, usually conducted one-on-one, in person. Other settings have also been used, including group conversations and phone surveys. The questions asked of the participants aim both at engaging them in spontaneous interaction, which provides linguistic data, as well as obtaining a detailed overview of their sociodemographic background. In addition to general characteristics such as age and gender, extensive information is usually collected on the participants' residential history, the languages that they speak, their socioeconomic status, and so forth; this background is central in explaining the observed patterns of variation. The length of the interview varies; traditionally it is considered optimal to aim for 1 to 3 hours of speech, although shorter periods may be useful for phonological data. The interview should ideally last long enough for the speaker to be comfortable enough to speak in their most informal style. However, speakers do not lower their degree of formality in a linear manner, but may move back and forth depending on a variety of factors (Milroy and Gordon, 2003, pp. 57–61).

It is around this behavior – style shifting – that the traditional interview structure is centered. Labov (1994, p. 157) distinguishes between

- casual speech, which implies the least attention to the way of speaking, is the closest to the vernacular, and in practice corresponds to emotionally involved speech;
- careful speech, which corresponds to the majority of the interview;
- controlled styles, used in reading tasks and opposed to spontaneous speech, covering casual and careful speech.

Within the structure of the interview, different devices are used to elicit a range of behaviors corresponding to different degrees of formality and attention to speech. Labov (2006, pp. 59–63) defines the following interview contexts, starting with the least formal:

- Context A, corresponding to casual speech;
- Context B, corresponding to the exchanges that the participant perceives as taking place within the formal structure of the interview;
- Context C, when the participant reads a text which elicits the pronunciation of targeted linguistic features;
- Context D, when the participant reads a word list, which may be extended to Context D', corresponding to a list of minimal pairs for a feature of interest.

In terms of execution by the interviewer, the most challenging of these is Context A. It is by definition contradictory, as the aim is for the subject to behave during an interview as if they were in a natural communicative situation. Labov distinguishes five specific contextual situations corresponding to this style:

- speech outside of the formal interview, e.g. surrounding an interruption;
- speech with a third person, such as a family member;
- speech not in direct response to a question, as when a person provides a long and potentially irrelevant response. It is argued that these types of responses should not be interrupted in order to ask other questions, but rather be seen as an opportunity to obtain

more natural speech production;

- childhood rhymes and customs, e.g. when participants are asked to recite something they memorized as children, making conscious control of speech production unlikely;
- the danger of death, i.e. a question which would lead to the respondent reliving an emotionally charged moment and thereby relaxing their conscious control of language. A traditional example is to ask about a near-death experience.

These contextual cues are complemented by channel cues indicative of casual style: a change in tempo, pitch range, volume, or rate of breathing; and laughter, which constitutes a subtype of a change in the rate of breathing (Labov, 2006, pp. 64–72). The traditional structure described here constitutes the basis of the interview protocol used in this dissertation, which is introduced in Section 12.1.

Note however that even with all the strategies put in place, the recorded responses may be influenced by the natural barrier between the researcher and the participant induced by the formal setting of the interview. Lowering that barrier is an important aim and requires considerable effort on the part of the researcher, going beyond the design of the interview. For instance, it is important to adapt to the local context, particularly when studying an unfamiliar culture or language. This can involve choices such as clothing or seating arrangements during the interview in order to show respect or solidarity as appropriate (Feagin, 2002, pp. 24–26). In studies aiming to elicit bilingual language behavior such as codeswitching, Poplack (1993) argues that data should be collected by interviewers who are members of the speech community, are perceived as such, and themselves use the linguistic phenomena under study (p. 260).

While the sociolinguistic interview has been successfully used in numerous studies, it also has limitations which other data collection methods have sought to address. We will now take a look at some of the main alternatives and the issues at stake.

4.1.2.3 Alternative approaches to data collection

Although the structure of the sociolinguistic interview aims to reduce the effect of the observer's paradox by controlling for style shifting, it is still unlikely that even the least formal production in the interview corresponds to the way participants speak when they are not observed. A key issue at play is the position of the interviewer as external to the speech community. A method aiming to overcome this issue is participant observation.

In this approach, the investigator does not limit the interaction on a single interview session, but rather embeds themself in the group under study, becoming a member of the community and participating in its activities. By building trust over time, the investigator can gain access to speech productions of far greater descriptive accuracy. A prime example of this approach is Penelope Eckert's study of Detroit-area schools (Eckert, 2000), which she conducted by observing and interacting with students over the course of two years, with extensive recordings of 200 participants. This approach is particularly valuable in small, well-defined communities, where the role of the outside observer would pose significant challenges, as well as in studies of bilingual behavior, where outsiders have difficulty accessing crucial patterns such as language choice and codeswitching. The main drawbacks include the extraordinary requirements in

terms of time, effort, and emotional implication needed to collect the data; inefficiency, in the sense that many more hours of data are collected than transcribed and analyzed; and difficulty in placing the observed local patterns within a wider sociolinguistic context (Milroy and Gordon, 2003, pp. 68–72).

Another strategy aimed at attenuating the observer's paradox as well as the time-consuming nature of the sociolinguistic interview is the rapid and anonymous survey. It allows to investigate a very precise phenomenon, which is elicited as a likely response to a carefully designed question, asked in spontaneous interaction. This was first developed in Labov's New York department stores study (Labov, 1972), where he analyzed the use of (r) by asking department store employees for the location of an item for which the response would be *fourth floor*. The main drawbacks are the focus on a tightly defined linguistic phenomenon and the very limited data on the background of the interviewee (Milroy and Gordon, 2003, pp. 56–57). For instance, Labov approximated the age of the respondents, and he inferred social class based on that of the clientele of the store in question.

Perhaps the most extreme way of overcoming the observer's paradox is surreptitious recording, which consists in recording speakers without their knowledge. However, this practice is not considered acceptable in sociolinguistic work given the potential legal and clear ethical issues that it raises. (On standard ethics requirements, see the discussion in Section 4.1.3.) It has also been argued that covert recordings carry practical disadvantages, including compromising the researcher's relationship with the community and leading to a recording of poor quality (Milroy and Gordon, 2003, pp. 81–83). Moreover, they provide no reliable information on the informants' background, constraining the ability to explain the observed patterns of language variation.

A more general alternative to the use of audio data recorded in the context of a conversation consists in using written questionnaires, which can be more easily distributed to a larger number of speakers, and allow for an efficient collection and analysis of directly comparable data (Dollinger, 2015; Schleef, 2014). These and other closely related methods developed in the tradition of dialectology are particularly well suited to the study of the lexicon, as they allow for aggregate analyses that can more easily be extended to large numbers of variables (Nerbonne, 2018, pp. 235–237). Their importance in the description of Canadian varieties of English, including Quebec English, has been noted in Chapter 2. They are further presented in Chapter 5 with regard to the study of lexical semantic variation; a questionnaire-based task is also integrated in the face-to-face interviews conduced in this dissertation, as discussed in Section 12.1.2.

Finally, not all data collection methods are concerned with speech production; in many studies, the aim is to understand how speakers perceive a language use (Preston, 2002). A particularly well-known approach is the matched guise technique, first developed to study attitudes to different languages, starting with English and French in Canada (Lambert et al., 1960). Participants hear recordings of the same passage which is read by a single person in two different ways (using different languages or characteristics of different varieties). The attitudes associated with the tested languages or ways of speaking are then elicited; the fact that the speaker remains the same ensures that the responses are related to the difference in the perception of

languages or varieties, rather than the characteristics of the speaker's voice, for example. Although first developed in social psychology, approaches such as this have been used extensively in variationist sociolinguistics, going back to Labov's work on New York City which lead to his definition of the speech community (Labov, 2006). These approaches can be integrated in the structure of the sociolinguistic interview, and their interest for semasiological variation will be addressed in more detail in Chapter 5.

Once recordings of speakers are made, they must be processed before any analyses can be conducted. I now turn to the standard practices in this regard.

4.1.2.4 Data processing

In sociolinguistic studies relying on audio recordings, the stage of processing the collected data represents an important challenge. This is a notoriously tedious and time-consuming task, with one estimate from the literature stating that at least ten hours of work are required to fully analyze an hour of recorded speech (Milroy and Gordon, 2003, p. 72); another says that only the initial transcription of one hour of recorded data requires at least four hours of work, which may vary depending on factors including sound quality and familiarity with the recorded variety (Tagliamonte, 2006, p. 54).

The orthographic transcription of the recordings is the first step leading to the analysis of the recorded audio data. It is important for the transcription to follow a defined protocol in order to ensure systematicity in transposing speech patterns specific to oral communication into written form. Standard orthography and punctuation are commonly used, with specific conventions determining how to represent behaviors such as false starts, partially produced words, pauses, laughter, and other contextual information. The transcription of words that do not exist in dictionaries must also be defined, as in the case of variety-specific nonstandard words and morphological features (Tagliamonte, 2006, pp. 53–65).

The initial orthographic transcription results in a machine-readable corpus that can be analyzed using concordancers and other standard tools in corpus linguistics. Depending on the aim of the study, however, additional levels of transcription and annotation may be needed. In studies focusing on phonological variation, a phonological transcription follows. On the segmental level, this can include a citation-phonemic representation, which represents the phonemes directly associated with the transcribed lexical items; a broad phonetic transcription, which represents the actual pronunciation at the contrastive phonological level (including phenomena such as consonant deletion and vowel reduction); a narrow phonetic transcription, which more closely corresponds to the actual pronunciation, including at the allophonic level; or an acoustic phonetic transcription, which precisely indicates different elements and phases occurring in the production of a sound (Delais-Roussarie and Post, 2014, pp. 54–57).

Once a manual orthographic transcription is produced, a phonological transcription can be automatically generated (Strik and Cucchiarini, 2014), as well as aligned to the orthographic information and the audio recording using different available tools (Gorman et al., 2011; McAuliffe et al., 2017; Rosenfelder et al., 2011). These tools have also been used to automatically code a range of phonological variables (Bailey, 2016; Gupta and DiPadova, 2019;

Milne, 2014; Yuan and Liberman, 2011). These approaches tend to produce more errors than humans, but it is usually argued that the increased error rate may be acceptable depending on the specific aims of the study.

However, it is the orthographic transcription that is the main bottleneck in data treatment. Attempts to fully automate it use automated speech recognition, which tends to produce a considerably higher error rate compared to human transcribers. The fully automated approach is therefore still based on the assumption that errors on the lexical level are not burdensome in some contexts, for example if the erroneous items contain the same vowels as the true tokens, and the study focuses on vowel analysis (Reddy and Stanford, 2015; Coto-Solano et al., 2021). This clearly does not hold true for any study dealing with lexical phenomena, where manual orthographic transcription remains necessary. This is an important practical limitation of traditional sociolinguistic data; more general issues are explored in the next section.

4.1.3 Limitations

Sociolinguistic studies involve work with human participants. Although it is generally considered to be minimally invasive, this work by definition entails ethical requirements. The key principles in modern sociolinguistic research are informed consent, meaning that participants are asked to voluntarily partake in the study before any data collection takes place; and anonymization, which typically involves substituting or removing personally identifiable information from the collected data, and storing the data in a secure manner (Milroy and Gordon, 2003, pp. 79–81). The specific extent to which data should be anonymized depends on factors including the size of the community under study (and hence the likelihood of the participants being identified) and the access policy for the completed corpus. More extensive anonymization involves not only the substitution of informants' names with speaker codes, but also the removal of names of any other individuals, narrowly defined places, or any other information that would allow indirect identification of the speaker (Childs et al., 2011). In addition to the basic ethical requirement of not doing any harm to the participants, a more involved, advocacy position has also been put forward. A central argument from this standpoint is that researchers have an obligation to use the knowledge based on the data collected in a community for the benefit of that community (Milroy and Gordon, 2003, pp. 84–87). In line with these recommendations, ethics approval was sought – and obtained – for the sociolinguistic interviews conducted in this dissertation; this is discussed in Section 12.2.1.

On a more general note, the decades of studies conducted within the framework of variationist sociolinguistics have led to the development of reliable research practices in data collection and analysis. However, the vast majority of these studies, just like the very structure of the sociolinguistic interview, are centered around phonological and to a lesser extent morphosyntactic variation (e.g. Durand et al., 2014). In addition to the research focus being placed on these issues (Labov, 1994; Tagliamonte, 2006), it is also the case that, given the very demanding requirements at all stages of data collection and processing, the size of sociolinguistic corpora is usually insufficient to study lexical features in a reliable way. This is more specifically related to the fact that word frequency distribution follows Zipf's law (1932), i.e. a word's frequency

is inversely proportional to its rank, meaning that all but the most frequent words are overall very rare. Obtaining comparable lexical data from different speakers in spontaneous communication is therefore challenging, which is one reason behind the previously discussed use of written dialect questionnaires.

That is not to say that the interview approach is of no interest for the study of other phenomena, including semasiological variation: it enables a detailed and reliable description of the respondent's sociodemographic profile; it provides a suitable context to elicit the targeted type of information; and it provides complementary information on other levels of linguistic structure, the foremost among them being phonology. The implementation of this approach in the present study will be presented in Chapter 12.

On the whole, we have seen that data collection in variationist sociolinguistics is driven by the objective of describing the linguistic patterns typical of a speech community, with the analysis of individual ways of speaking traditionally seen as a means of producing a generalizable description. The choice of participants therefore has profound implication in terms of representativity. Traditional data collection methods are structured around the need to access the vernacular, often resorting to different elicitation devices to record speech in a range of styles, as well as obtaining detailed sociodemographic information which are instrumental for subsequent analyses. Data processing involves at a minimum an orthographic transcription of audio recordings, and likely other levels of transcription and annotation; much of this still requires painstaking manual work. Overall, the care and planning involved in all stages of data collection ensure the reliability of the results, but they also limit the size of sociolinguistic corpora, making them of limited use in the study of lexical phenomena. That is why I now turn to another type of data: Twitter-based corpora.

4.2 Twitter-based corpora

The large amount of publicly available data on Twitter and the relative ease with which they can be accessed have led to widespread use of Twitter as a data source in a variety of scientific disciplines. A range of analytical methods have been applied on both structured and unstructured data, with datasets ranging in size from a few thousand to millions of tweets. The number of Twitter studies across most disciplines, including NLP, has been steadily increasing over the past decade, and this trend is expected to continue into the future (Karami et al., 2020).

In the present work, I will mainly be interested in Twitter as a source of linguistic data enabling analyses of language variation which would otherwise be difficult to conduct. This is one of the main applications of Twitter data in the field of computational linguistics, motivated by the ability it offers to conduct both large-scale and fine-grained analyses based on unobtrusive observation of language used in different social contexts (Nguyen, 2021). I will come back to the way variation is studied on Twitter in the next chapter. For now, I will explore the general characteristics of communication of Twitter, the main ways in which data is collected and filtered to create linguistic corpora, and the key limitations of this data source.

4.2.1 Characteristics of communication on Twitter

This section provides an overview of the main features that distinguish Twitter from other types of communication. It specifically addresses the formal constraints and communicative conventions; the range of users and their interactions; and the linguistic features resulting from this context.

4.2.1.1 Features and conventions of Twitter

Twitter is a microblogging service created in 2006. The primary way in which users communicate on the service is by posting tweets, or text messages up to 280 characters in length.[1] Tweets can contain different types of special tokens, such as `#hashtags`, usually indicating a topic, and `@mentions`, indicating another user. They can also include URLs, as well as multimedia objects such as images or videos. Users can choose to associate their geographic location to the tweets they post. They are required to explicitly opt-in to the service, which then allows them to indicate a point of interest to be associated with the tweet from a dropdown list. Mobile devices additionally allow for precise geolocation, meaning that the user's geographic coordinates at the time of tweeting can be associated with the tweet.[2]

Each user has a profile page, which presents all of their tweets as well as basic metadata. Additional information can be provided, including a profile photo, a profile description (up to 160 characters in length), and a free text location. Users can interact with the content posted by others by liking it or replying to it. They can reproduce it by retweeting it (forwarding it in its original form) or quote tweeting it (adding their own comment to the original post). Users can form ties with other users by following them, and hence regularly seeing their posts in their own timelines. These ties are asymmetrical, i.e. a followee has no obligation of following back their follower. By default, Twitter accounts are public, but they can be made private; in that case, followers are first approved by the private account, and are only then able to see that account's tweets.[3]

4.2.1.2 Users and interactions

As of 2019, Twitter had 330 million monthly active users globally (Twitter, 2019).[4] This, however, does not constitute a representative sample of society at large. According to a survey conducted in 2020, 42% of online adults in Canada have a Twitter account. The user base is slightly skewed towards men. Twitter is considerably more prevalent among younger users, with 65% of 18-24s having an account, compared to 27% of over-55s. It also tends to be used more by higher earners and those with a university degree. The trends are on the whole stable over time, compared to a 2017 survey (Gruzd and Mai, 2020, p. 12). Another point of note, reported in research conducted in the United States, is that tweet production is not

[1] The maximum length was 140 characters until November 2017.
[2] https://help.twitter.com/en/using-twitter/tweet-location
[3] https://help.twitter.com/en/safety-and-security/public-and-protected-tweets
[4] More recent precise data is not available. Following its Q1 2019 earnings release, Twitter switched from reporting the number of monthly active users to monetizable daily active usage. These metrics are not directly comparable.

evenly distributed across users: the top 10% of users account for 80% of all tweets (Wojcik and Hughes, 2019, p. 2).

Users engage in different types of behavior on Twitter. An influential early analysis by Java et al. (2007) posited three categories based on interaction links between users: *information sources*, who have a large number of followers and represents a hub in the user network, even though they may tweet at varying intervals; *information seekers*, who tweet rarely but follow other users; and *friends*, who are situated between the other two categories in terms of behavior and cover the majority of relationships (p. 63). Moreover, in addition to legitimate human users, Twitter is characterized by the presence of bots, i.e. accounts involving automated activity. They exhibit behavioral differences (e.g. bots retweet more often, post significantly more URLs, and form fewer reciprocal relationships with other accounts), but legitimate users may interact with automatically generated content (Gilani et al., 2019). This still represents an issue in corpus construction; we will come back to it in the next section.

Although Twitter was initially designed as a microblogging platform, where broadcasting one's opinion rather than interacting with others was the primary goal, it was observed early on in its adoption that it was actively used in conversations between users, mainly through the use of user mentions (Honeycutt and Herring, 2009). The conversational dynamics of Twitter are also characterized by retweets, or reproductions of other users' messages. The reasons behind retweeting are numerous, and include amplifying tweets, commenting someone's tweet (when additional content is added), showing public agreement, and so on (boyd et al., 2010).

Moreover, despite the asymmetric nature of the ties between users mentioned earlier, Gruzd et al. (2011) find that Twitter can be used to construct communities, in which participants interact and exhibit a sense of community. On a physical level, this is reflected by the impact of geographic distance on social ties formed on Twitter. A plurality of ties between followers and followees are formed within a single metropolitan area, with geographic distance a key predictor for the creation of the remaining ties (Takhteyev et al., 2012). In terms of tweet content, this is reflected by the ability to use linguistic features to detect communities of Twitter users (Ramponi et al., 2019).

Further, it has been argued that Twitter users who engage in interaction driven by similar interests constitute communities of practice, where they form interpersonal bonds around a common interest and take on different roles in the community depending on their own ability (Malik and Haidar, 2020). These observations are complemented by research on other types of online communities, which has found that socialization within language communities is reflected by language use, including accommodating behaviors potentially leading to language change (Nguyen and Rosé, 2011). This brings us to another important issue: the linguistic characteristics of Twitter-based communication.

4.2.1.3 Linguistic features

As already mentioned, a key formal characteristic of Twitter communication is the 280-character limit on message length. This is an obvious constraint on language use compared to other types of written (and oral) communication, but it has been argued that, in conversational terms at

least, it does not represent a limitation. Rather, it "allows [messages] to be produced, consumed, and shared without a significant amount of effort, allowing a fast-paced conversational environment to emerge" (boyd et al., 2010, p. 10). However, in terms of structural features, the impact may be felt differently: for instance, it has been found that the shift from 140 to 280 characters led to a decrease in the use of abbreviations and other space-conserving features, which are also associated with informality (Boot et al., 2019). This contrasts with earlier observations, which found that non-standard abbreviations (e.g. *ur* meaning *your* or *you're*) more frequently appeared in shorter tweets than the corresponding standard forms. If their use were associated with the character limit, they would be expected to appear in longer tweets, thereby allowing them to be published (Eisenstein, 2013, p. 361). While the normalization of nonstandard tweets has been proposed, both lexically (Baldwin et al., 2015) and syntactically (Kaufmann, 2010), nonstandard orthographic features may carry descriptive interest, as studies on other messaging platforms have found the use of nonstandard forms to be involved in language variation phenomena (Peersman et al., 2016; Squires, 2007; Tagliamonte, 2016; Tagliamonte and Denis, 2008).

When compared to a range of other written corpora, Twitter data exhibit the lowest average word and sentence length, the highest out-of-vocabulary rate, and the highest proportion of ungrammatical spans of text. They are the most similar to a corpus of YouTube comments, and the most different from Wikipedia (Baldwin et al., 2013). Similarly, in a distributional semantic comparison of lexical usage on Twitter and Wikipedia, Tan et al. (2015) found considerable differences in nearest neighbors, i.e. the words sharing the same cooccurrence patterns as the target word. Many cases were reflective of informal language on Twitter (e.g. *ill* used as the contraction *I'll* rather than as a synonym of *sick*; p. 660). However, language on Twitter is not universally informal or reflective of conversational style. For instance, Paris et al. (2012) compared two different communities of Twitter users, finding statistically significant differences in the frequency of use of informal lexical items (e.g. contractions, abbreviations) as well as emotive and personal language (e.g. repeated exclamations, first-person pronouns reflecting personal opinions). This points to overarching differences in style between these subsets of users.

Twitter users are free to tweet in any language they like, and so they do. Much like in face-to-face communication, language choice reflects factors including interlocutors and communities. Focusing specifically on minority languages in the Netherlands, Nguyen et al. (2015) found that language choice was influenced by the interlocutor's dominant language and the language of the tweet to which a reply was given. The use of the majority language was likely when trying to reach a wider audience. Moreover, multilingual users participate in distinct types of communities in terms of linguistic links. This ranges from the "gatekeeper" network, where members of the two linguistic communities with whom a given user interacts are very weakly connected between themselves, to the "integration" type, where one linguistic group exists within another (Eleta and Golbeck, 2014). Twitter users also engage in linguistic behaviors typical of other multilingual speakers, such as codeswitching, which has been observed in a variety of language pairs (Lynn and Scannell, 2019; Rudra et al., 2019; Vilares et al., 2016). It can also intersect with Twitter-specific structures: for instance, a codeswitch can occur be-

tween a hashtag and the remainder of the tweet, with the hashtag nevertheless fulfilling the same functions it does in monolingual communication (Jurgens et al., 2014).

In summary, we have seen that Twitter provides a very specific means of communicating, and its characteristics must be borne in mind when analyzing the data that it provides. It enables its users to post messages and to create an online presence in the form of their profile, as well as to interact with other users and the content that they post. It is demographically biased towards the younger and the more well-to-do. It gives rise to conversational dynamics specific to the platform, while the linguistic production is constrained by the character limit and presents other medium-specific features. However, Twitter also exhibits features shared with other types of communication, including typical bilingual behaviors and strong community bonds, while the trend towards informal communication arguably constitutes an advantage for studies of language variation. This, coupled with the vast amount of available data, justifies the use of Twitter to construct linguistic corpora. I review the practical details of this process below.

4.2.2 Construction pipelines

In this section, I will first present the general way in which data can be accessed through Twitter. I will then address different filtering steps that are routinely applied in order to make the data usable in linguistic research, focusing on three common issues: language identification, removal of unwanted content, and normalization.

4.2.2.1 Data collection

Twitter provides different ways of accessing its data. The ability of researchers and other users to access the publicly available data is based on Twitter's Terms of Service, which stipulate that the content posted by users can be made available by Twitter to third parties.[5] The use of these data is regulated by Twitter's Developer Agreement and Policy, which impose restrictions on issues including sensitive information, publishing and sharing data, and matching online profiles with the individuals behind them.[6] The main point of access is Twitter's API (application programming interface), which provides a systematic way of querying the available data.[7] In practical terms, the API also limits the amount of data available to developers, for instance by imposing rate limits (a limited number of requests can be sent over a given time period) and only providing samples of the data to the public. These data contain tweet text as well as a range of metadata, including tweet and user identifiers, tweet language, geolocation, special entities included in the tweet (hashtags, user mentions, URLs etc.), account-level information (number of followers, followees, tweets etc.), and so forth. Different API versions and access levels exist; the focus throughout this dissertation will be on the free access to API version 1.1. I will now present the general principles and differences between two key ways of accessing

[5] https://twitter.com/en/tos
[6] https://developer.twitter.com/en/developer-terms/agreement-and-policy
[7] https://developer.twitter.com/en/docs/twitter-api

Twitter data, and will then discuss specific ways in which they have been implemented to create linguistic corpora.

The Search API provides the ability to query the archives of previously published tweets. The search is conducted on a non-exhaustive sample of tweets published over the preceding six to nine days; it is unclear what percentage of total tweets is included in the sample. The search involves the use of keywords looked up in the content of the tweet, as well as a number of other parameters, including tweet language and location.[8] Another option is the Streaming API, which provides access to a sample of all tweets as they are published in real time. The sample is either entirely random or filtered using a similar set of operators as for the search API. The sample output by the Streaming API, whether random or filtered, is capped at roughly 1% of all tweets. If the number of tweets corresponding to filtering parameters is lower than that threshold, then all corresponding tweets are returned.[9] Multiple random samples streamed in parallel overlap nearly entirely in content, making it difficult to circumvent the 1% cap (Joseph et al., 2014).

Both Search and Streaming APIs allow for different types of information to be looked up in the data, including keywords (contained in the text of the tweet), tweet language, and geolocation. Several differences should however be noted. Keywords are required for the Search API, and optional for the Streaming API. This means that real-time tweets can be sampled based solely on location, language or other available parameters, whereas searches through the archive must always include a linguistic expression. As for geolocation, in Search API, it is indicated as a radius around a point defined in terms of latitude and longitude; in Streaming API, it is specified as a bounding box defined by the coordinates of the southwest and the northeast corner. More importantly, the two approaches do not resolve geolocation in the same manner. Streaming API only takes into account tweet-level location data: precise geolocation, when the tweet is tagged with the geographic coordinates of the user's location at the time of tweeting; or manual geolocation, when the user chooses the place associated with the tweet from a list of proposed options or by looking up a specific place. These features are only available on mobile devices and are actively used by a fraction of all users, which limits the availability of geotagged tweets. By contrast, Search API maximizes the amount of data returned in geographic queries by interpreting non-geotagged tweets as sent from the location indicated in the user profile.

These ways of accessing data, sometimes coupled with crawling user timelines or extracting patterns of interaction, are implemented in different ways in order to build linguistic corpora. A random sample of tweets can be created using the Streaming API without any specific criteria (Petrović et al., 2010). Corpora aiming to include tweets produced in a specific language have used the Streaming API with a set of language-specific keywords (Basile and Nissim, 2013; Kreutz and Daelemans, 2020; Scheffler, 2014), in some cases supplementing this by crawling the timelines of previously identified users (Tjong Kim Sang and van den Bosch, 2013). Timeline crawls have also been used on target accounts identified manually (Bergsma et al., 2012)

[8] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/guides/standard-operators

[9] https://twittercommunity.com/t/diffence-between-sample-and-filter-streaming-api/15094/2

or by querying Twitter for language-specific keywords using the Search API (Ljubešić et al., 2014). Streaming API coupled with geographic coordinates has been used to collect tweets from a region spanning multiple countries and languages (Laitinen et al., 2018) or tweets published from a specific country in a geographically widespread language (Barbaresi, 2016), in the latter case coupled with a subsequent timeline crawl. Studies examining large-scale variation on Twitter have often been conducted on English using geotagged data from the United States obtained through the Streaming API (Bamman et al., 2014; Blodgett et al., 2016; Eisenstein, 2013), but data collection and filtering parameters are surprisingly often underspecified.

As we have seen in Section 4.1, the reliability of sociolinguistic corpora relies on a careful choice of speakers before data collection, complex elicitation devices during data collection, and (mostly manual) transcription and annotation after data collection. By contrast, Twitter does not allow for the required sampling precision, meaning that data are collected with the assumption that they are likely interspersed with noise. Reliability is then ensured by filtering the data once they have been collected; given the large corpus sizes at play, this necessarily involves automatic methods. Let us take a look at some of the main procedures, starting with language identification.

4.2.2.2 Language identification

A basic requirement in constructing a corpus is ensuring that the data are written in the target language. All tweets are associated with a language tag provided by Twitter's in-house language identifier (Twitter, 2015). It indicates a single language tag per tweet from a set of 70 language tags; additionally, some tweets are tagged as undetermined, for example if the linguistic information they contain is too limited to identify a single language.[10]

Extensive investigations into, and direct quantitative comparisons with, the performance of Twitter's tagger are limited by the fact that the Developer Agreement precludes benchmarking of Twitter's services. However, considerable research has gone into language identification of Twitter messages. As Lui and Baldwin (2014) point out, this was motivated by issues including a lack of language information in Twitter metadata until 2013, limited coverage of the world's languages, and attempts to improve performance on the in-house identifier. To address these issues, they introduced a simple majority-vote system using off-the-shelf language identifiers, showing it to outperform any individual language identifier. While they do not provide quantitative results for legal reasons, they report that the accuracy of Twitter metadata "is not substantially better than the best off-the-shelf language identifiers" (p. 24).

More recently, interest has shifted onto some of the specific challenges related to Twitter. Considerable drops in language identification performance are specifically related to (i) tweet length, with optimal performance for tweets longer than 60 characters, and dramatic drops in performance for tweets with 20 characters or fewer; (ii) the presence of multiple languages in a single tweet, particularly when one of the languages covers a limited span of text; (iii) similar languages, as in the case of poor accuracy on Galician due to its similarity with Spanish

[10] https://developer.twitter.com/en/docs/twitter-api/enterprise/powertrack-api/guides/operators

and Portuguese (Zubiaga et al., 2016). A closely related issue is that of codeswitching in tweets. Issues that have been addressed include word-level language identification (Nguyen and Doğruöz, 2013), prediction of codeswitch sites (Papalexakis et al., 2014), and detection of intra-word codeswitches (Nguyen and Cornips, 2016).

In addition to implementing various language identifiers, different types of data have also been used in order to improve language identification performance. For instance, geolocation data has been used to complement preexisting language ID information to improve results on similar languages (Williams and Dagli, 2017). Working on the distinction between English and non-English tweets, Blodgett et al. (2017) complemented a language identifier with demographic data. They specifically used external information on the linguistic structure of geographically defined communities. They found that this systematically improved performance, particularly in the case of short tweets. However, the generalizability of their approach depends on the availability of demographic data and the type of geolocation used. Moreover, it has also been underscored that the link between geolocation and language data is not direct (Graham et al., 2014). This leads us to another issue of central importance, namely the use geographic information contained in Twitter data.

4.2.2.3 Geolocation

There are two main ways in which geographic information is provided on Twitter: tweet-level location, indicated as a latitude/longitude coordinate pair associated with an individual tweet, and free text location indicated in the user profile. Different other types of information can be used to infer a user's location, such as their social networks (Jurgens et al., 2015) or places mentioned in the text of the tweet (Ajao et al., 2015). In this section, however, I will focus on the two readily available types of information which are provided in the metadata. They are routinely used in corpus construction, but their individual reliability as well as the extent to which they correspond to one another are important to understand.

In an analysis of a sample of geolocated tweets from four metropolitan areas, Graham et al. (2014) compared tweet-level geolocation and the location indicated in the user profile. Only around half of profile locations placed the user within the bounding box from which the tweet was sent; the remaining cases corresponded to genuine locations outside of the bounding box, non-geographic text, or generic locations. In addition to the pervasive discrepancies between the two ways of geotagging data, the authors found that most users tended to opt for one of the two methods, rather than using both.

In terms of consequences for corpus creation, Pavalanathan and Eisenstein (2015b) found that samples based on tweet-level and profile-level geolocation led to different linguistic generalizations. For instance, they observed a higher frequency of nonstandard, geographically specific terms in the sample containing tweet-level geolocation, further arguing that this might be related to underlying demographic differences in the use of geolocation services. This intuition is supported by Sloan and Morgan (2015), who contrasted the demographic characteristics of users who use profile-level and tweet-level geolocation, finding differences in age, gender, socioeconomic class, and language. Although they suggested that the differences might be tol-

erable when Twitter data is used in research, they underscored that using geolocation data was not representative of the general Twitter population.

A final stage of filtering that I will address is related to the removal of unwanted content. I turn to it next.

4.2.2.4 Unwanted content

The presence of unwanted content represents an important problem in corpus construction, particularly as it may bias frequency information used in subsequent analyses. Specific issues include the presence of automatically generated content, such as tweets providing links to external websites; automated accounts; as well as otherwise repetitive messages, for example driven by a user's topical interests. A more detailed overview based on the data collected in this dissertation is presented in Section 8.3.3. The discussion here is limited to two general approaches to addressing these issues in Twitter-related studies: those focusing on the presence of unwanted users and of unwanted messages.

In a study focusing on unwanted users, Yardi et al. (2010) identified account-level features which could characterize their behavior. They drew a distinction between what they termed spam users and legitimate users, showing that they did not differ significantly in terms of account age or follower-to-followee ratio. However, spam users had a slightly higher number of retweets and replies, as well as a statistically significantly higher mean number of tweets per day, number of hashtags, and total number of followers and followees. In addition to account-level trends such as these, features characterizing tweet content (e.g. number of characters per tweet, number of hashtags per tweet, etc.) have been used to implement methods allowing to automatically detect spam users. An early example is the work by Benevenuto et al. (2010), who found that the most important predictor was the fraction of tweets with URLs, followed by the age of the user account and the average number of URLs per tweet. The focus in subsequent studies has included identifying additional types of features, such as geographic usage patterns (Guo and Chen, 2014), with the central difference among the approaches remaining the choice of features used in classification (Wu et al., 2018). Similar methods have been used to differentiate specific types of accounts. For instance, a number of studies have focused on the distinction between corporate and personal accounts, the former often exhibiting spam-like behavior and hence being of limited interest for linguistic research (Ljubešić and Fišer, 2016; McCorriston et al., 2015; Wood-Doughty et al., 2018).

Another related issue is that of repetitive content. To address this, Tao et al. (2013) adopted a five-degree scale of tweet similarity, distinguishing between (i) exact copies; (ii) nearly exact copies; (iii) strong near-duplicates (same core message with additional information in one of the tweets); (iv) weak near-duplicates (same core message with personal information or differing pieces of information); (v) low-overlapping tweets (semantic similarity but realized using few common words). They devised a set of features to detect such content and eliminate it at the level of tweets, rather than excluding whole accounts. Specifically, they used surface features (e.g. Levenshtein distance), semantic features (e.g. overlap in topics), enriched semantic features (including content from linked websites), and contextual features (e.g. temporal dif-

ference between the tweets). They used a logistic regression obtaining an F-score of 0.46 with the full set of features. This suggests that near-duplicate exclusion is a difficult task; it also explains why heuristic solutions are frequently used in other studies to deal with similar issues.

This section has illustrated how Twitter data can be accessed and filtered in constructing linguistic corpora. In general terms, a choice can be made between using a sample of tweets published in real time or looking up Twitter's archives; this can be complemented by crawling user timelines. The choice of the method as well as query parameters (linguistic expressions, geographic position, etc.) depend on the specifics of the study; whatever the case, the data are likely to require filtering before they can be used. The main steps I have reviewed include verifying language identification and geolocation data, and removing unwanted content. While computational methods for many of these tasks are readily available, some open questions remain; these are particularly relevant if high precision is required. In any event, even though Twitter data are comparatively easy to collect, corpus construction requires time and effort. Twitter data also come with limitations, which are addressed by the next section.

4.2.3 Limitations

In addition to the issues addressed through filtering, the use of Twitter data entails other potential problems as well. A key practical issue is the specificity of the language used on Twitter compared to other types of corpora, with implications for the performance of standard preprocessing tools. For instance, Jørgensen et al. (2015) tested three POS taggers, including two specifically created for Twitter. They observed overall low performance, which further dropped on non-standard language. They obtained the best results with Gate, a tagger designed for Twitter (Derczynski et al., 2013), which was nevertheless judged unsatisfactory for descriptive linguistic studies: accuracy stood at 79% for data reflecting African American Vernacular English (AAVE) and at 83% for non-AAVE data.

Another problem affecting the validity of linguistic analyses has to do with the biases involved in Twitter data collection. As we have seen, in the vast majority of cases researchers access data which is sampled from Twitter's archives or real-time posts. However, the sample itself is not perfectly random (Pfeffer et al., 2018); more importantly, the demographic profile of Twitter users is not representative of the general population (Jørgensen et al., 2015). This population bias is compounded by a range of other potential issues, including behavioral biases (e.g. different interactions on Twitter compared to face-to-face communication), content production biases (e.g. language use specific to a subset of a population or a given context), and linking biases (e.g. differences in behavior correlating with follower count) (Olteanu et al., 2019, pp. 6–9).

More generally, reproducibility of Twitter-based research is considerably limited by legal constraints on data diffusion. In particular, Twitter's Developer Policy prohibits the public distribution of tweet content, and only allows the distribution of lists of tweet or user identifiers. This is in principle sufficient to reconstruct a corpus by downloading the same tweets, but this is a time-consuming process which moreover results in an imperfect replication of the original dataset. Twitter users can delete existing tweets, make their accounts private, or delete them

entirely; any one of these actions will render the tweet unavailable. As a result, Zubiaga (2018) reports, decay in reconstructed datasets increases over time, with fewer than 70% of original tweets and unique users available after four years. While original metadata, such as user profile descriptions and follower and followee counts, are also likely to change over time, the textual content of the decayed dataset is generally representative of the original corpus, constituting in effect a subsample of the initial data. That said, these constraints have led to a range of proposed solutions, including more time-efficient systems to distribute preexisting collections of tweets (McCreadie et al., 2012) and calls for social media archiving initiatives (Vlassenroot et al., 2021). Closely related to this is the fact that Twitter's Developer Policy and Agreement evolve over time, and some information central to previous studies may become unavailable. For instance, user-level interface language and timezone were removed from the metadata in the course of the data collection I conducted, described in Chapter 8.

Finally, ethical considerations must also be taken into account. As previously mentioned, the data provided by Twitter are publicly available, and this access is legally granted through its Terms of Service. In the majority of studies done on Twitter, this is taken to represent sufficient license to use the data, but calls for further reflection on ethical issues have also been made. Williams et al. (2017) surveyed a sample of UK Twitter users, finding that the majority of them (84%) were not at all or only slightly concerned by their data being used in academic research. However, 80% would expect to be asked for consent before their posts were published, and 90% would expect their content to be anonymized (p. 1156). While Twitter communication may be seen as a public place where people would not reasonably expect privacy, and as such is not subject to ethics evaluations, the situation is complicated by the fact that Twitter communication is in fact often intended for a more limited imagined audience, and not the Internet as a whole (pp. 1159–1160). Since Twitter Terms of Service preclude the anonymization of Twitter content when it is published, the authors argue that an informed consent should be sought for publication, including opt-out consent if the user is not in a vulnerable category and content is not sensitive (pp. 1161–1163). These findings are echoed by Fiesler and Proferes (2018), who additionally indicate that users' perceptions depend on the specific context in which their tweets would be used. For instance, the use of a tweet in an aggregate analysis is perceived more positively than in an analysis with a few dozen other tweets (21% vs. 47% of respondents, respectively, would feel somewhat or very uncomfortable); the same goes for the publication of the tweet with or without the username (56% vs. 26% of respondents, respectively, would feel somewhat or very uncomfortable) (p. 8). While these observations are unlikely to fundamentally alter the general practices of the scientific community, they provide concrete evidence for the importance of ethical considerations in Twitter-based research.

4.3 Summary

This chapter has presented two distinct but complementary approaches to collecting data reflective of language variation. In variationist sociolinguistics, the guiding objective is that of describing the linguistic practices of a speech community and of obtaining information on the

background of the selected speakers so that the observed patterns of variation can be explained. The process of data collection relies on careful consideration in terms of the choice of speakers, design of the data collection method, and data processing. This requires both considerable effort and skill, not least in direct interaction with participants which is supposed to put them at ease and facilitate the production of spontaneous speech.

However, the data collected in this way are quantitatively insufficient for a systematic study of lexical phenomena. The use of social media data, such as that available on Twitter, provides a potential response to this problem. Twitter is particularly suitable because it provides relatively easy access to large amounts of geolocated linguistic data, which tend to be informal in nature and can be traced back to individual users. Typical communicative practices, including those typical of bilingual communities, are widely represented on Twitter, as are complex interaction patterns. While the available demographic data are considerably more limited, Twitter corpora are usually several orders of magnitude larger than traditional sociolinguistic corpora. This is one important advantage when it comes to studying lexical variation; another is the fact that Twitter entirely avoids the observer's paradox.

While it is tempting to consider the divergences between these two data sources as the prevailing takeaway, it is also important to underscore the similarities between them and their disciplines. On the one hand, both types of corpora share some of the same problems, even if they are affected by them differently. These include representativity issues caused by demographic skews in the respective samples, as well as legal and ethical limitations in terms of data distribution and reproducibility. On the other hand, both sociolinguistics and NLP are empirical disciplines which are firmly grounded in the use of attested linguistic data. Just like in the case of data collection, they also provide different but complementary methods to analyze variation across communities of speakers. This is what I turn to in the next chapter.

Chapter 5

Modeling semasiological variation

We have seen in Chapter 4 that different potential data sources can provide information on language variation. We now turn to the issue of using these data to model language variation, focusing on different strategies to isolate attested patterns of semasiological variation. Section 5.1 outlines the potential methodological solutions that exist in variationist sociolinguistics and other related disciplines, as well as the main theoretical challenges in studying semasiological variation within this framework. Section 5.2 reviews the related research conducted in NLP on computational models of semantic variation and change. Section 5.3 summarizes the key points in this discussion.

5.1 Sociolinguistic approaches to lexical semantics

Despite the high salience of lexical phenomena, variation in word usage and, *a fortiori*, in word meaning is the area least studied by sociolinguistics, in large part due to the inherent methodological challenges in systematically observing and quantifying these phenomena (Durkin, 2012). As we will see in this section, variationist studies of semasiological phenomena are not without precedent, but they are overall very rare.

Unlike the work conducted on other levels of linguistic structure, semasiological variation has still not been subject to extensive theoretical and methodological discussion. That is why this section begins with a focus on a key notion in the variationist theory, that of linguistic variable, and the extent to which it can be applied to semasiological variation. I will then turn to existing studies which provide methodological paths forward and illustrate the descriptive relevance of this type of variation.

5.1.1 Linguistic variables

The basic unit of variationist sociolinguistic analysis is the linguistic variable,[1] traditionally defined as "two alternative ways of saying the same thing" (e.g. Labov, 2004, p. 7). In more precise terms, Wolfram (1993, p. 195) defines a linguistic variable as being "made up of a class

[1]The term *sociolinguistic variable* is also widely used in the literature, often to underscore to the link between linguistic and external (social) variables, discussed below. For clarity, I will stick to the term *linguistic variable*.

of variants – varying items that exist in a structurally-defined set of some type"; he further notes that the variants may in principle be found at any level of linguistic structure (see also Tagliamonte, 2006, p. 75). The traditional definition suggests that a variable and its variants are linked by a form–meaning relation. From this standpoint, variation corresponds to the existence of multiple forms expressing a single meaning; the reverse – multiple meanings expressed by a single form – is a case of ambiguity (Anttila, 2002, p. 210). We will come back to this issue in Section 5.1.1.2, as it is central to the applicability of the construct of linguistic variable to semasiological variation.

But first, let us take a look at some of the general features of linguistic variables. Labov (1972) identifies several characteristics of variables that are the most suitable for description: (i) high frequency, enabling observations of the item in unstructured communicative exchanges such as the sociolinguistic interview; (ii) structural nature, with a higher degree of integration into the larger system assumed to provide the item with more inherent linguistic interest; (iii) highly stratified distribution in terms of age or other social factors. Moreover, the feature should ideally be sufficiently salient to enable the study of social attitudes, but not so salient as to be subject to conscious distortion (p. 8). It should also be easy to quantify on a linear scale. Some of these criteria evolved in subsequent studies: for instance, Labov (2006) argues for the avoidance of conscious suppression, rather than distortion, of the examined features. He also suggests that these characteristics are recommendations rather than hard requirements (p. 32).

The linguistic variable should moreover be correlated with an external (i.e. social) variable (Labov, 1972, p. 237). A distinction can also be made in relation to the level of social awareness, as reflected by the following traditional typology of linguistic variables: (i) *indicators*, which exhibit social but not stylistic stratification, i.e. their use varies depending on social characteristics of the speaker but is consistent across different styles; (ii) *markers*, which exhibit both social and stylistic stratification; (iii) *stereotypes*, which are overtly commented upon and are involved in phenomena of correction and hypercorrection (Labov, 1972, p. 237; Labov, 1994, p. 78).

Finally, a defining component of variationist sociolinguistic methodology is known as the *principle of accountability*, which Labov (1972) describes as follows: "we will report values for every case where the variable element occurs in the relevant environments as we have defined them" (p. 72). In other words, analysis of language variation is not led by the raw frequency of a variant of interest, but the proportion of cases in which the variable was realized with that variant. A variable is therefore defined in terms of a closed set of variants. Moreover, this approach also entails the need to circumscribe the variable context, i.e. determine in a systematic manner if any occurrence of the variable should be disregarded. One such example is neutralizations, or contexts in which it cannot be determined with certainty if a given variant was realized, for example due to the linguistic elements that surround it (Tagliamonte, 2006, pp. 86–94).

We have seen so far the general principles on which the linguistic variable is based. We now turn to its application to lexical variation.

5.1.1.1 Linguistic variables at the lexical level

The construct of linguistic variable was largely developed in studies of phonetic and phonological phenomena. Sankoff (1980) was among the first to overtly argue "that variability occurs and can be dealt with at levels of grammar above (or beyond) the phonological" (p. 82), drawing on several cases of syntactic and semantic variation. However, this and other early studies dealing with non-phonological phenomena (e.g. Sankoff and Thibault, 1977; Weiner and Labov, 1983) received some pushback because of the attempt to extend the the notion of linguistic variable. For example, Lavandera (1978) argued that phonological variants carry no referential meaning, making it straightforward to determine the functional equivalence of those variants. This, she claimed, was fundamentally different from the form–meaning relationship on other levels of linguistic structure, with the issue of determining semantic equivalence impeding analyses. Other opinions, similarly focusing on the semantic equivalence of syntactic variants, followed (e.g. Cheshire, 1987; Romaine, 1984).

Focusing on lexical variables, Barysevich (2012) points to three issues frequently addressed in the literature: (i) the problem of semantic neutrality, raised by the interpretation of the denotational meaning with which a polysemous item is used in a given context; (ii) the problem of stylistic neutrality, related to the potential presence of different connotational meanings associated with the variants of a lexical items, even if their denotational meanings are identical; (iii) the problem of lexical quantification, principally involving the low frequency of content lexical items compared to structural linguistic elements, with potential solutions including the use of meta-variables (cf. e.g. Armstrong, 1998) or seeking to construct larger corpora (pp. 24–36). In practical terms, these issues are usually addressed by carefully circumscribing the variable context.

A particularly relevant strand of research on lexical variation is the one taking the notion of semantic field as a starting point. Introduced by Sankoff et al. (1978), this approach analyzes the variable choice of lexical items from the same semantic field. Importantly, the different senses of the lexical items are analyzed in terms of specific semantic features. In this way, the variants of a lexical variable are clearly delimited in terms of their semantic equivalence. For instance, the authors analyze the verbs meaning 'to reside' in Montreal French, including the lexical items *rester*, *vivre*, *demeurer*, and *habiter*. Since these verbs do not overlap in all of their senses, the authors define the specific semantic features which a sense should cover to be included in the analysis; only these uses constitute the variants of their lexical variable. In other words, the authors investigate an onomasiological lexical variable for which the internal (linguistic) conditioning criteria are defined on the semasiological level.

A number of subsequent studies in Canada applied a similar approach. The semantic fields addressed in the 1978 paper – 'to reside', 'work' (*travail*, *job* etc.), and 'thing' (*chose*, *affaire* etc.) – as well as that of 'car' (*char*, *voiture* etc.) have been extensively analyzed in French-speaking communities across Quebec and Ontario (Barysevich, 2012; Bigot, 2016; Nadasdi, 2005; Nadasdi and Mckinnie, 2003; Nadasdi et al., 2004, 2008; Sankoff, 1997, among others). Research directly drawing on this tradition has also been conducted on Canadian English, mainly focusing on the use of adjectives. For instance, Tagliamonte and Brooke (2014) ex-

amine the adjectives in the semantic field of strangeness (*strange*, *weird*, *odd* etc.), finding evidence of ongoing lexical change leading to the use of *weird* as the dominant form. Adopting a comparative perspective, Tagliamonte and Pabst (2020) investigate the use of adjectives of highly positive evaluation (*great*, *awesome*, *cool* etc.) in Toronto (Canada) and York (UK). They find that despite a similar inventory of forms in the two varieties, their relative frequencies are different, with the dominant forms not evolving in parallel. The semantic field approach has similarly been applied to other English varieties (e.g. Jauhiainen, 2020; Stratton, 2020) and other languages (e.g. Stratton, 2022).

As noted in Chapter 4, lexical variables are also investigated using written dialect surveys. In Canadian English, considerable work has been done on regional variation using this approach (Boberg, 2005b, 2010, 2016; Chambers, 1995, 1998, 2000; Dollinger, 2012), including with a specific focus on English spoken in Quebec (Boberg, 2004a,c, 2012; Boberg and Hotton, 2015; Chambers and Heisler, 1999). While the structure of the linguistic variable is the same as in variationist sociolinguistics – it is composed of different lexical items carrying the same (denotational) meaning – the way in which the variants are delimited is different. Instead of looking at occurrences in spontaneous speech and circumscribing the variable context, dialectological surveys elicit lexical items, often in response to a definition or a picture of the referent. It is this stimulus, rather than the delimitation of contexts of occurrence, that guarantees the semantic equivalence of the lexical variants (see also Underwood, 1968).

Another related line of work is variationist sociolinguistic research into discourse-pragmatic variation. In the Canadian context, numerous studies have investigated features such as quotative verbs (D'Arcy, 2004, 2007, 2017; Gardner et al., 2021; Tagliamonte and D'Arcy, 2004, 2009; Tagliamonte and Hudson, 1999; Tagliamonte et al., 2016, among others) and general extenders, or expressions occurring at the end of the utterance (Denis, 2015, 2017; Tagliamonte and Denis, 2010). While these variables are not purely lexical in nature, they involve a choice in terms of the lexical item used to convey a discoursive function.

To recapitulate, despite the initial reluctance to extend the notion of linguistic variable to non-phonological variation, extensive work on lexical and discourse phenomena, in variationist sociolinguistics and dialectology, has demonstrated the interest of this type of variation. Methodological issues can be suitably addressed through the choice of variants and variable contexts. Lexical variables defined in such a way exhibit the characteristics observed on other levels of linguistic structure, such as intralinguistic conditioning, and social and stylistic stratification. Moreover, they also convey social meanings and are as such involved in processes of indexical positioning and identity construction; we will see this in more detail in Chapter 6. However, the issue of whether and how to apply the notion of linguistic variable to lexical semantic variation remains open.

5.1.1.2 Linguistic variables at the semasiological level

As we have seen in the previous section, the analysis of lexical variables always involves a semantic component. In particular, the specification of variants involves circumscribing the variable context by defining the target meanings, perhaps most explicitly addressed by Sankoff

et al. (1978) and the studies that directly followed them. But how can we analyze the variations in the meaning of a single lexical item within the variationist sociolinguistic framework?

As Cerruti (2011) underscores, following Hasan (2009), the general lack of variationist sociolinguistic studies on lexical semantic variation is principally related to two issues: (i) the nature of meaning at the basis of the linguistic variable, which is purely referential, whereas a fine-grained lexical semantic analysis requires a contextual view of meaning; and (ii) technical difficulties in quantifying the occurrences of the variable (p. 218). These issues notwithstanding, a potential approach to studying semasiological variation would involve constituting a linguistic variable whose variants are different senses of a single lexical item (on the semasiological perspective and other theoretical issues related to lexical semantics, see Chapter 3). In the Quebec English context, this would involve analyzing a variable such as the use of the noun *animator* in terms of variants that correspond to its senses 'creator of animated films' and 'group leader'. However, the author also suggests that this would risk "radically destructuring the very concept of variable" (p. 221; my translation): the linguistic variable traditionally consists in a relation between multiple forms and a single function, not the reverse.

This theoretical reservation is nevertheless tempered by Labov (2004) overtly stating that the study of variation also applies to "situations where there are alternative meanings conveyed by the same form" (p. 7). In addition, semasiological analyses are foreshadowed by work such as Sankoff et al. (1978), who precisely quantify the number of occurrences of individual senses for each of the lexical items under study. While they do so in the process of formulating onomasiological lexical variables, they illustrate, perhaps inadvertently, how semasiological variability can be investigated: by analyzing the distribution of the senses associated with a single onomasiological *variant* (i.e. lexical item), whose entire sense inventory constitutes a semasiological *variable*. And although Wolfram (1993) does not explicitly entertain this type of variation, it could be argued that semasiological variables fit his broad description according to which variables are comprised of "varying items that exist in a structurally-defined set of some type" (p. 195).

If we let go of the traditional form-meaning definition of the linguistic variable, it becomes evident that semasiological variables exhibit most properties outlined at the beginning of this chapter. Given an adequate dataset, there is no inherent reason why a semasiological variable could not present a high frequency, a highly stratified distribution in terms of social constraints, or be highly salient. Their structural nature is admittedly different than that of phonological variables, but as Cerruti (2011) points out, they are integrated in the wider linguistic system by means of semantic relations (p. 225). The existing studies, which I will review in the next two sections, also demonstrate that semasiological variables correlate with social variables, are subject to different degrees of social awareness, and can be used to convey social meaning. The criterion of linear quantification of linguistic variants is not easily satisfied, but that is also the case for other non-phonological variables, and may be partly overcome by corpus-based methods presented in Section 5.2.

On the whole, if we abstract from the form-meaning issue, semasiological variables represent a type of linguistic variable that can be readily analyzed using standard variationist practice, including accountable quantitative analysis, and may provide valuable information on so-

ciolinguistic behaviors. As Dollinger (2017) puts it, in discussing regional semantic variation of the verb *take up* with a particular focus on the differences between Canadian and American usage:

> If narrow semantic variables are indeed more widespread than isolated cases [...], they deserve special attention in the Canadian context – however unimportant they may be in other parts of the English-speaking world. [...] TAKE UP #9 demonstrates that even tiny linguistic differences – differences that some speakers outright 'correct' or reject as ungrammatical – may have considerable social and diatopic significance in border contexts. (p. 100)

I now turn to existing empirical studies of semasiological variation conducted within the framework of dialectology and variationist sociolinguistics. Unlike the descriptions of Quebec English presented in Chapter 2, which are informative but to a certain extent anecdotal, this discussion will be limited to the studies addressing semasiological variation explicitly and systematically.

5.1.2 Dialectological questionnaires

Investigations of the lexicon in most recent dialect surveys conducted in Canada are limited to the onomasiological perspective, with questions eliciting the lexical item used to denote a given referent (e.g. Boberg, 2005b, 2016; Boberg and Hotton, 2015). However, two exceptions to this general trend are worth examining in detail, as they clearly outline methodological approaches to, and descriptive contributions of, semasiological variation.

Drawing on the data from the Dialect Topography project, previously discussed in Chapter 2, Chambers (2007b) analyzes the use of positive *any more* in the Golden Horseshoe region. The adverbial *anymore* is generally used to signify 'no longer', and it is used with negative polarity: it can only appear in syntactic contexts containing a negative marker. By contrast, positive *any more* conveys a meaning that can be paraphrased as 'nowadays', and is used with positive polarity, as in Chambers' example *John smokes a lot any more* 'John smokes a lot nowadays' (p. 38). Although this item is analyzed as a syntactic variable, the distinction between the two uses arguably contains a semantic element. What is more, Chambers overtly probes the different meanings associated with this item, making this analysis highly relevant from a semasiological perspective.

The item is addressed by four multiple choice questions dispersed throughout the written questionnaire, each containing an example of the target item with the investigated use. Two questions elicit the choice of a proposed rephrased sentence corresponding to the meaning conveyed by the example; one question focuses on the target lexical item, asking for a synonym; and the final question investigates the acceptability of the sentence. While roughly half of the respondents interpret the examples as having the investigated meaning, the reported rate of personal use stands at just below 10%. The responses correlated with age indicate a clear decline in the rate of use starting in the 1950s.

A more comprehensive study of semasiological variation is conducted by Dollinger (2017), who investigates the use of the phrasal verb *take up*. He specifically focuses on the meaning 'to provide and explicate a model solution', which, based on extensive lexicographic research, is argued to be specific to Canada. This variable is moreover characterized by having contextual (but not syntactic) restrictions, as well as "a semantically narrow yet clearly identifiable meaning" (p. 84).

This usage was studied using a written dialect questionnaire, administered online. The use of the variable was tested using the example "The professor took up our test in math class this morning", where the target lexical item is used with the tested meaning. The questionnaire elicited the perceived meaning of the sentence, a paraphrase, and spontaneous observations (p. 86). A total of 608 responses were collected across the Canadian provinces, as well as Pennsylvania and the UK Midlands. The semantic variable was analyzed in binary terms, with the target use classified as "recognized" or "not recognized"; positive judgment was coded in a restrictive manner, with any sign of hesitation or ambiguity leading to the interpretation as "not recognized".

The rate of recognition of the target meaning is the highest in Ontario (67%) and the lowest in the United States (10%). It stands at 43% in the Prairies (Alberta, Saskatchewan, and Manitoba), and ranges from 22% to 28% in the remaining Canadian provinces. An analysis in apparent time, also taking into account migratory trends and diachronic lexicographic evidence, suggests that the investigated sense of *take up* is a Canadian innovation originating in Ontario and spreading westward, this diffusion being associated with economic migration to Alberta.

These two examples convincingly illustrate the insights that can be obtained by studying semasiological variation in dialect surveys. Although both variables are comprised of multiple meanings associated with a single form, contrary to the variationist sociolinguistic tradition, they are associated with linguistic conditioning factors as well as correlated with social variables such as age. In both studies, the target use is analyzed as a binary variable; note however that the variants for positive *anymore* are its two potential interpretations, whereas *take up* is made into a binary variable through the coding procedure opposing the usage under study to the rest of the sense inventory.

More generally, these examples illustrate good practices in written questionnaire design described by Dollinger (2015). For instance, both studies use interrelated questionnaire items: multiple questions address the examined variable in different contexts or from different standpoints, ensuring that the respondent must provide more than one (potentially chance) correct answer (p. 241). They also illustrate different types of questions: they use both questions directly examining language variation as well as eliciting acceptability judgments (p. 12). They further combine elements of self-reporting (describing own language use) and community-reporting (describing the language use of other speakers) (p. 235).

Although these decisions improve the reliability of the results, written questionnaires by design cannot capture language variation in spontaneous speech, and they provide limited background information on the respondents compared to that obtained in a sociolinguistic interview. This is the context that I address in the next section.

5.1.3 Sociolinguistic interviews

A rare – if not sole – example of semasiological variation studied in an interview setting is the work by Robinson (2010, 2012a,b, 2014). Drawing on the principles of cognitive sociolinguistics, broadly construed as a study of meaning and variation, she investigates semasiological variation in adjective use in British English. The sample includes 72 speakers from South Yorkshire, and is balanced for age, gender and socioeconomic position.

Each speaker participated in a face-to-face interview designed to elicit the meanings of the examined lexical items. A total of 15 adjectives were investigated: eight with recent semantic shifts, and seven controlling variables (polysemous adjectives with no recent semantic shifts or monosemous adjectives). For each adjective, the following pattern of questions was used:

> Question: Who or what is *gay*?
> Answer: My school.
> Question: Why is your school – *gay*?
> Answer: Because it is boring. (Robinson, 2012a, p. 43)

In other words, this procedure elicits the referent for an adjective, and a justification for that answer. Reported uses were also recorded, corresponding to the senses that participants recognize, but state that they do not use them or only do so in a specific context. The referent elicitation task was followed by a conversation about the use of the target lexical items, aimed at recording the associated perceptions and attitudes. The first step in the subsequent analysis consisted in forming sense clusters based on the answers provided by the participants. This was then analyzed against their main sociodemographic characteristics: age, gender, and socio-economic status.

The strongest effects on the choice of sense variants were those of age, reported for the independently described variables – *awesome* (Robinson, 2010), *gay* (Robinson, 2012a), and *skinny* (Robinson, 2012b) – as well as in an aggregate analysis of all eight target adjectives (Robinson, 2014). This was interpreted as a demonstration of semantic evolution of these adjectives in apparent time. A key observation is the fact that the overall most frequent sense was often found to be the same for all generations; however, significant relative differences in the use of the remaining senses were observed based on age, thereby reflecting both newly invented and disappearing senses. As for the effect of gender and socio-economic status, they were less systematic, but nevertheless provided important insights into individual patterns of variation. For instance, the innovative use of *gay* 'unmanly' and *gay* 'lame' appeared to be led by younger males, which was tentatively interpreted as a way for them to distance themselves from the idea of homosexuality (Robinson, 2012a, p. 50).

On the whole, Robinson's observations echo the results from dialectological studies: they confirm that it is possible and relevant to analyze semasiological variables within the variationist sociolinguistic framework. However, several limitations should also be pointed out. The procedure proposed here elicits the most salient senses for the participants, but not their entire sense inventory (Robinson, 2012a, p. 53); this has potential implications in terms of the systematicity of the description. The direct applicability of the procedure additionally depends on

the type of lexical item under study: for instance, given the focus on the referent, it is difficult to imagine how the same question pattern would be used with abstract nouns. On a similar note, Robinson (2012b) herself cautions that the interpretation of the results should also take into account the referent denoted by a given sense, as in the case of *skinny* 'low fat' (p. 225). The suggestion here is that some cases of variation may be driven by external factors affecting the referent rather than language use.

The method introduced by Robinson is an important step towards integrating the study of semasiological variation into traditional sociolinguistic interviews, but, like in dialect surveys, her results remain disconnected from spontaneous speech. Moreover, although the analyses produced by both of these approaches are fine-grained, they remain focused on a limited number of linguistic variables. An attempt can be made to overcome these issues using large-scale computational semantic models, which I explore in Section 5.2. But first, a note is due on other sources of information used in analyzing contact-induced semantic shifts.

5.1.4 Further information on lexical items of interest

In addition to the data on the use of contact-induced semantic shifts provided by the informants recruited for a face-to-face interview or a written questionnaire, a comprehensive analysis of this issue also requires background information documenting the use of the target lexical items. In the course of the analyses conducted in this dissertation, I principally relied on the existing sociolinguistic descriptions (cf. Chapter 2), as well as a range of lexicographic sources:

- the Canadian Oxford Dictionary (COD; Barber, 2004), as it is widely accepted as the reference for Canadian English usage;
- the Dictionary of Canadianisms on Historical Principles (DCHP-2; Dollinger and Fee, 2017), which provides detailed information on the development and usage of lexical items specific to Canada;
- the Oxford English Dictionary (OED),[2] as it provides a broad, detailed, and historically well-documented description of the English lexicon in general;
- the Trésor de la langue française informatisé (TLFi; Dendien and Pierrel, 2003), which provides detailed descriptions of the French lexicon in general;
- Usito (Cajolet-Laganière et al., 2014), which describes French as it is used in Quebec;
- the bilingual WordReference dictionary (WR),[3] which has the advantage of complementing the standard English-French and French-English directions with the "reverse" function. For a given lexical item, it indicates all the entries in the other language where the target lexical item is provided as a translation equivalent.

Monolingual English dictionaries were principally used to establish the most prominent conventional English sense, as well as to determine if a potentially contact-related sense was already attested. Monolingual French dictionaries were used to precisely identify the sense which

[2] https://www.oed.com
[3] https://www.wordreference.com

is posited to have become associated with the contact-affected English lexical item. Usito was mainly used when no apparent explanation could be found in the TLFi. Bilingual information was used to narrow down the extent of overlapping senses between the target lexical items in English and French. While these sources were used to confirm descriptions produced throughout this dissertation, they were particularly important in validating a core set of semantic shifts used both in computational (Chapter 11) and sociolinguistic (Chapter 12) analyses.

We have so far seen a range of potential solutions and inherent challenges to the study of contact-induced semantic shifts in the variationist sociolinguistic framework. I now turn to the potentially complementary solutions based on computational analyses of large corpora.

5.2 Computational models of lexical semantics

Recent years have seen considerable interest in the use of computational meaning representations to investigate semantic variation and change. While a range of methods have been proposed (Tahmasebi et al., 2021), our focus will be on the use of vector space models, which arguably constitute the predominant approach (Kutuzov et al., 2018). This section will first review different types of models, and then discuss the specifics of their application in analyses of semantic variation and change.

5.2.1 Vector space models

Most commonly used computational representations of meaning are rooted in the principles of distributional semantics. The origin of this view of lexical semantics is usually situated in the tradition of structuralism (Harris, 1954), as well as analytically related but more pragmatically oriented work (Firth, 1957). The general approach can be summarized by the distributional hypothesis, according to which words appearing in similar linguistic contexts are expected to have a similar meaning. This points to the more general observation that a word's meaning is reflected by the linguistic contexts in which it occurs (see e.g. Sahlgren, 2008).

The computational models that rely on distributional principles represent a word's meaning as a vector, which is essentially a list of numbers reflecting the word's co-occurrence statistics in a given corpus (which are taken to be representative of the word's usage). They are therefore known as vector space models (VSMs) (Turney and Pantel, 2010). This is the term that will be used throughout this dissertation, as it is the most readily applicable to the range of approaches that will be implemented, but note that the terms distributional semantic model (DSM) (Baroni and Lenci, 2010) and word space model (WSM) (Sahlgren, 2006) refer to the same basic concept. A crucial characteristic of VSMs is the ability to quantify the distance between two vectors, which is taken to reflect the difference in meaning of the words represented by the vectors. This has important implications in terms of permitting systematic, empirical studies of lexical semantics (Boleda, 2020).

VSMs are mainly created from very large, generic corpora (composed of newspaper or Wikipedia articles, varied content obtained through web crawls, and so forth); in most ap-

proaches, their content is not extensively questioned. Starting from the chosen dataset, VSMs can be created in different ways; the most commonly used architectures are presented below. I will distinguish type-level representations, where a single vector is produced for the entire range of contexts in which a word occurs, and token-level representations, which reflect the meaning of an individual occurrence of a given word.

5.2.1.1 Type-level representations

Two main approaches to creating type-level VSMs will be presented: count-based models, which incorporate information on word co-occurrence frequencies, and neural models, which are created by training a neural network.

Count-based models. The most direct implementation of the distributional hypothesis is found in count-based models. In the most basic approach, the words in a corpus are represented based on a co-occurrence matrix, an example of which is provided in Table 5.1.

	eat	cheese	gravy	developer	engineer	system	...
poutine	52	16	5	-	-	-	
fries	24	24	10	-	-	-	
software	4	-	-	129	64	24	
design	-	-	-	6	26	97	
...							

TABLE 5.1: Sample co-occurrence matrix.

The meaning of each target word is represented by a vector, which conventionally corresponds to a row in the matrix. The columns represent the target word's linguistic contexts, i.e. the words with which it appears in a span of text of a predefined length; in technical terms, each column constitutes a vector's dimension. The values indicate how many times a target word co-occurs with each of the context words. In general, there are as many rows and as many columns as there are words in the vocabulary, but this depends on specific processing decisions. For example, context words can be specified in terms of syntactic dependencies, thereby capturing more precise distributional patterns (Padó and Lapata, 2007).

The sample matrix also illustrates other general principles of the distributional hypothesis. For instance, it is obvious at first glance that the vectors for the words *poutine* and *fries* are similar. They are very different from the vectors representing the words *software* and *design*, which in turn resemble one another. As mentioned before, this intuitively observed similarity in meaning can be systematically quantified. This is most commonly done using the cosine similarity, which reflects the angle between the vectors in multidimensional space. Its value ranges from -1, for opposite vectors, to 1, for identical vectors; it stands at 0 if the vectors are orthogonal. Based on the sample matrix used here, the cosine similarity between *poutine* and *fries* would stand at 0.87, indicating high similarity; the one between *poutine* and *software* would stand at 0.02, suggesting a lack of relatedness in meaning.

Raw co-occurrence frequencies are helpful in understanding the underlying mechanisms, but in practice they are rarely used to create VSMs. They are instead usually weighted in some

manner in order to limit the skew introduced by highly frequent words. A commonly used weighting function is the Positive Pointwise Mutual Information (PPMI) (Bullinaria and Levy, 2007). It compares the probability of two words occurring together and their probability of occurring independently in order to prioritizes more informative target–context pairs.

Another common practice consists in reducing the number of vector dimensions, for example using Singular Value Decomposition (SVD) (Deerwester et al., 1990; Landauer and Dumais, 1997). Simplifying, this approach combines vector dimensions that carry similar information (i.e. linguistic contexts associated with similar distributions of target words), allowing to address the issue of context dispersion. It is usually implemented so as to produce a significant reduction in the number of dimensions – from the tens of thousands to the hundreds – leading to more efficient models. However, this also leads to a loss of direct interpretability of the linguistic information captured by the matrix: it is no longer clear which linguistic contexts are represented by each of the dimensions.

Neural models. More recently, neural network architectures have been deployed to produce VSMs. Neural networks are comprised of layers of artificial neurons; these are computing units that take multiple values as input and produce a single value as output (Jurafsky and Martin, 2022, p. 133). Neural networks are extensively used in classification tasks, and this is central to their application to VSMs.

Following initial work on this topic (e.g. Bengio et al., 2003; Collobert and Weston, 2008), arguably the most influential method for type-level neural representations was introduced by Mikolov et al. (2013). Known as `word2vec`, it produces the same general type of word meaning representations as count-based models. However, these vector representations are low-dimensional and dense, i.e. they contain significantly fewer dimensions than count-based models and none of those are empty. These representations are known as word embeddings, and their low-dimensional dense nature makes them computationally more efficient and overall better performing. The success of `word2vec` is additionally related to the highly efficient training procedure that it implements (Jurafsky and Martin, 2022, Ch. 6).

Rather than producing meaning representations in an unsupervised manner, as in the case of count-based models, in this approach a classifier is trained based on co-occurrence patterns in the data. This is a supervised classification task: the fact that some words co-occur in the text, and others do not, implicitly provides a binary label for the training data. The training process relies on pairs of (*target*, *context*) words, extracted form a context window of a predefined size around the target word. To illustrate this more clearly, consider the following example adapted from the corpus of tweets used in this dissertation:

(2) The regular **poutine** is cheese curds, fries and gravy.

Using a window size of two words on each side of the target word *poutine*, we would extract the following pairs: (*poutine*, *the*), (*poutine*, *regular*), (*poutine*, *is*), (*poutine*, *cheese*). Each pair would be provided as a training example, but the way it would be used would depend on the chosen `word2vec` algorithm. The continuous bag of words algorithm (CBOW) aims to predict

the target word based on its linguistic contexts. Conversely, the skip-gram algorithm (SG) aims to predict the linguistic contexts based on a target word.

In more precise terms, both algorithms start learning from randomly initialized vectors, which are iteratively updated as more training examples are provided to the model. The training objective of the CBOW algorithm is to maximize the conditional probability of observing the target word vector given the context word vectors provided as input. As for the SG algorithm, it outputs multiple vectors corresponding to the context words, its training objective being to minimize the prediction error for all context word vectors, given the target word vector provided as input. The meaning representations produced by both algorithms correspond to the weights learned in the hidden layer on the classification task.

The SG algorithm incorporates several specific features. In addition to using positive training examples (those that actually occur in the data), it also includes negative sampling (hence the term skip-gram with negative sampling, or SGNS). This consists in drawing random context words from the corpus that do not co-occur with the target word so as to generate negative training examples (those that reflect patterns absent from the data, which the model should learn to reject as a potential output). The SG algorithm also performs subsampling, meaning that it eliminates a certain number of positive examples involving highly frequent words. This is conceptually similar to the weighting procedures adopted in count-based models; it has been shown to be mathematically equivalent to a method using pointwise mutual information and and SVD-based dimensionality reduction (Levy and Goldberg, 2014b).

Compared to earlier models, `word2vec` led to significant improvements in terms of computational efficiency, as well as performance on a range of evaluations (Baroni et al., 2014). Other neural models have also been released, including adaptations of `word2vec` to include syntactic contexts (Levy and Goldberg, 2014a) or character n-grams (`fasttext`; Bojanowski et al., 2017). Despite these indications of success, several shortcomings should also be noted. SGNS models suffer from instability due to inherent randomness in the algorithm (Pierrejean and Tanguy, 2018) as well as data features such as corpus size and document order (Antoniak and Mimno, 2018). More generally, just like in the case of count-based models, their performance depends on hyperparameter settings and corpora (Caselles-Dupré et al., 2018; Chiu et al., 2016); I now turn to this issue more closely.

Methodological choices. In addition to the choice of the data used to train a VSM and the general model architecture, a series of other settings more precisely determine the way in which the model is trained. For both count-based and neural models, a window size must be set, corresponding to the number of context words surrounding the target word that are taken into account. Previously reported choices in an early review range from 2 to 25 words on each side of the target word, with the suggestion that the contexts closer to the target word are more important for determining its meaning (Turney and Pantel, 2010, p. 170). The default window size for `word2vec` is 5, but this is a dynamic window, meaning that the context words are weighted: the farther they are from the target word, the less weight is attributed to them (Levy et al., 2015, p. 214).

Both count-based and neural models require that the number of vector dimensions be de-

termined. For count-based models, this corresponds to the number of context words taken into account, typically in the order of the tens of thousands. If dimensionality reduction such as SVD is applied, this number is reduced drastically, typically in the order of the hundreds. This is the same range as for `word2vec`, where the default number of dimensions stands at 100. Other hyperparameters that can be manipulated for `word2vec` include the subsampling rate and the negative sampling rate. All of these choices can potentially influence model performance (e.g. Baroni et al., 2014); different combinations will be explored in Chapter 10.

5.2.1.2 Token-level representations

The vector representations described so far have been extensively used in NLP, arguably becoming the main way in which the meaning of a lexical item is represented in downstream tasks. More recently still, another family of models has gained currency, largely superseding the earlier distributional representations in many of those. The foremost among them is BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). Like a number of other successful models (Brown et al., 2020; Conneau and Lample, 2019; Yang et al., 2019), it is based on Transformers, a type of deep neural network characterized by the use of self-attention, a mechanism that modulates the weight attributed to different input values and crucially does not process the data in a sequential manner (Vaswani et al., 2017).

The focus in this dissertation will be on BERT. It is trained on a masked language model task, where it learns to predict multiple tokens masked in an input sequence based on their surrounding context, helping to produce a bidirectional representation (i.e. taking into account both left and right contexts non-sequentially); and a next sentence prediction task, where it learns to predict if two sentences follow one another, thereby capturing patterns between sequences (Devlin et al., 2019). Compared to earlier types of meaning representations, BERT and similar models feature considerably more complex architectures. For instance, the base version of BERT contains 12 hidden layers, each comprising 768 dimensions. Training these models from scratch requires vast amounts of data[4] and is prohibitively computationally expensive, meaning that *de facto* only pre-trained models, produced by corporations or large research consortia, are actively used. They can be fine-tuned – i.e. partly retrained on new data, a new task, or both – and in that way adapted to the NLP application at hand. Their use nevertheless raises important issues, not least the fact that the end user has no control over the input data and the potential biases that it may encode (Bender et al., 2021).

That being said, these models obtain excellent results on standard evaluations for complex NLP tasks including question answering, language inference, and sentiment analysis (Devlin et al., 2019). It is therefore reasonable to assume that the vectors they compute could also be useful on a simpler level, as distributional representations of word meaning. They offer an important added advantage of providing contextual (token-level) representations: when a text sequence is fed into a pre-trained BERT model, a vector representation is computed for each token in a way that it reflects the context of that specific occurrence. Given the assumption that

[4]BERT was trained on the BookCorpus dataset (Zhu et al., 2015), containing around 11,000 unpublished books from a range of genres, and the entire English Wikipedia, for a total of 3.3 billion words.

the senses of a polysemous word can be distinguished based on its distributional patterns, this provides a promising way of analyzing polysemy (Garí Soler and Apidianaki, 2021; Wiedemann et al., 2019), although reservations persist as to the specific token-level information that BERT encodes (Haber and Poesio, 2020) and the kinds of distinctions it can draw (Haber and Poesio, 2021).

Token-level meaning representations can be obtained by extracting one or more hidden layers for the target token. If multiple representations are used, they are combined in some way, usually by summing, averaging, or concatenating the individual representations. The decision on which layers to use is crucial, but the understanding of the best choices to make is still limited. For example, there is evidence that semantic knowledge in general (Jawahar et al., 2019) and word sense information in particular (Coenen et al., 2019) is encoded in higher layers, but type-level lexical information seems to be best captured by lower layers (Vulić et al., 2020). This is reflective of the continuously evolving knowledge on the linguistic information encoded by BERT (Rogers et al., 2020). In practical terms, token-level representations reported in the literature have been generated using a variety of layers and their combinations; the implementation chosen for this dissertation will be discussed in Chapter 10.

We have seen so far that a range of computational methods can be used to produce word meaning representations based on co-occurrence patterns. However, each step in the evolution of VSMs has come with improvements as well as drawbacks: count-based models are computationally inefficient and tend to obtain comparatively poorer results in downstream tasks, but their contexts are directly interpretable; neural models such as `word2vec` improve on efficiency and performance, but they suffer from a lack of interpretability, as well as problems related to instability; deep learning models such as BERT provide contextualized representations, opening up new research avenues, but they offer no control over the training data, and they further exacerbate the "black box" nature of word embeddings. These difficulties notwithstanding, VSMs are powerful tools that have the potential to enable new types of linguistic analyses. Let us now take a closer look at their use in semantic change studies.

5.2.2 Using vector space models for semantic change detection

As mentioned at the beginning of this section, a considerable number of recent NLP studies use computational methods, and in particular different types of VSMs, to analyze semantic variation and change. The majority of the work has gone into modeling diachronic semantic change, but interest in different types of synchronic semantic variation seems to be growing. I will first present the main methods deployed in these two types of studies. I will then address common evaluation practices, descriptive contributions provided by the existing work, and some of the alternative computational approaches to language variation. For a brief overview of alternative approaches to semantic change detection and language variation in general, see Section 5.2.3.

5.2.2.1 Diachronic semantic change

The basic computational approach to semantic change consists in obtaining word representations specific to different time periods and then quantifying the differences between these representations. Echoing the Distributional Hypothesis, this is based on the assumption that a word whose meaning stays stable appears in similar linguistic contexts, and hence has similar vector representations, across time periods. A word whose meaning changes over time has the opposite characteristics.

Most existing studies aim to detect the words that, within the whole vocabulary, change the most over time. This usually relies on a semantic change score indicative of the distance between the vectors for the same word across different time periods. This goal is pursued differently depending on the vector representations that are used (type-based or token-based), and on the specific way in which they are compared. In any case, a large, generic diachronic corpus is used; it is generally split into time bins corresponding to the time periods under study (e.g. the 19th vs. the 20th century).

The standard approach based on type-level representations consists in training a separate VSM for each of the time periods using the corresponding subcorpus. If metrics such as the cosine similarity are used to quantify the distance between vectors from different time periods, the models must be aligned, i.e. they must define the same vector space. For count-based representations with directly interpretable linguistic contexts, such as PPMI models, this consists in ensuring that vector dimensions (columns) are the same, and that they are ordered in the same way, across all models (Gulordava and Baroni, 2011). This is in turn underpinned by the assumption that most linguistic contexts that are shared between two periods tend to remain stable over time. But this approach cannot be applied to count-based models with dimensionality reduction (e.g. SVD) or neural models (e.g. `word2vec`), as their dimensions are not interpretable. In this case, the models are usually aligned using a linear mapping such as the Orthogonal Procrustes analysis, which involves a series of matrix operations whose objective is to maximally reduce the distance between the vectors represented in different matrices (Hamilton et al., 2016b). Most words are expected to remain semantically stable, and therefore have representations which are close to one another in the resulting vector spaces; if their representations remain distant, it is likely that they have changed more than the others.

While alignment procedures introduce noise in the representations (Dubossarsky et al., 2017), several alternatives may avoid this issue. For instance, the Temporal Referencing method trains a single model for all time periods at once. It does so by tagging the target words for the time period in which they are attested (e.g. `gay_1920`, `gay_1970`). Their linguistic contexts stay the same across the time periods, meaning that a single vector space is produced (Dubossarsky et al., 2019). In a different approach, Kim et al. (2014) begin by training SGNS vectors for an initial time period. These are then used to initialize the vectors in the model for the subsequent time period, and are updated using the data specific to that period. Dynamic word embeddings (Rudolph and Blei, 2018; Yao et al., 2018) and deep neural architectures (Rosenfeld and Erk, 2018) have also been proposed to incorporate temporal information at the time of training.

In addition to identifying the words which change the most in meaning, other tasks have

also been formulated. For instance, Kulkarni et al. (2015) use a time series analysis to detect the point in time in which a semantic change occurs. Moreover, besides measuring the distance between a word's vectors from different time periods, other ways of quantifying semantic change have been proposed. Hamilton et al. (2016a) compute a second-order similarity vector, finding that it captures changes in word usage that are more closely related to cultural processes such as technical innovations. Gonen et al. (2020) quantify usage change based on the number of overlapping nearest neighbors for a given word in the models that are being compared, arguing that this improves on the stability and interpretability of competing methods.

More recently, pre-trained deep neural models have been applied to semantic change detection. BERT has been used in type-level analyses of semantic change, essentially as a potential replacement for traditional VSMs. One approach consists in modeling individual occurrences of a target word, clustering them, and quantifying the differences in clustering patterns across time to estimate semantic change (Giulianelli et al., 2020; Martinc et al., 2020b). Another solution averages over token-level representations for a given time period to produce time-specific type-level representations, roughly replicating the standard approach (Martinc et al., 2020a).

Several studies have more directly leveraged BERT's ability to produce sense-level information. Hu et al. (2019) use BERT to model dictionary definitions of potential senses as well as target word occurrences. This allows them to tag the sense with which each occurrence is used and to observe how sense distributions evolve over time, capturing distinct quantitative patterns. In a similar perspective, Montariol et al. (2021) cluster token-level representations for all time periods, implicitly splitting the occurrences into senses, and observe how their distribution across clusters evolves over time. They generate keywords to characterize the clusters in order to improve interpretability.

I now turn to studies of semantic variation in synchrony. Although they are more limited in number than diachronic studies, they have been conducted from a variety of perspectives and provide promising paths forward.

5.2.2.2 Synchronic semantic variation

Methods very similar to those described in the previous section have been applied to meaning variation across dimensions other than time. The key difference is the type of data that is used: these studies rely on comparing corpora representing different communities, text types, regions, and so on.

For instance, Del Tredici and Fernández (2017) use an extension of the SGNS algorithm introduced by Bamman et al. (2014) to analyze meaning variation across communities on Reddit, an online messaging platform. Fišer and Ljubešić (2018) contrast a Twitter corpus and a reference corpus for Slovene, under the assumption that usage specific to social media may reflect innovative trends indicative of incipient diachronic language change. Despite a high rate of noise due to preprocessing errors (90 out of the 200 top-ranking words), they report promising results, but they also suggest that the general approach is likely best suited for a semi-automated setup. Similarly focusing on different text types, Schlechtweg et al. (2019) evaluate the performance of a range of methods on the detection of synchronic semantic differences between

general and domain-specific language. Similar methods have also been adopted to examine differences in viewpoint as reflected by speeches from different political parties (Azarbonyad et al., 2017) as well as general patterns of language variation across national varieties of English (Kulkarni et al., 2016).

This work is complemented by studies adopting a multilingual or a contact linguistic perspective. Uban et al. (2019) investigate the semantic divergence of cognates in six Romance languages. They train VSMs for each language independently and then align them. This allows them to quantify the distance between language-specific vectors, finding that frequency and polysemy are positively correlated with cross-lingual semantic divergence. A similar mapping procedure for language-specific VSMs is used by Takamura et al. (2017), who investigate the use of English loanwords in Japanese by contrasting their distributional properties to those observed in English. To the best of my knowledge, no computational studies have investigated the influence of language contact in Canada, or the semantic effects of language contact on inherited (i.e. non-borrowed) lexis.

I now turn to the most common ways of evaluating these methods.

5.2.2.3 Evaluation

Evaluation of semantic change detection methods is challenging and is often limited to a qualitative analysis of a restricted number of examples. For example, Hamilton et al. (2016b) examine the top 10 semantic change candidates output by each of the models they test, contrasting their use in the 1900s and in the 1990s. They consider cases such as *gay* shifting from 'happy' to 'homosexual' to be true positives. They also report several false positives, for example due to topical variation.

Systematic quantitative evaluations have only recently become available. They generally rely on manually annotated datasets containing (usually several dozen) words whose meaning is either stable or subject to change (Basile et al., 2020; Del Tredici et al., 2019; Gulordava and Baroni, 2011; Pivovarova and Kutuzov, 2021; Schlechtweg et al., 2019, 2020). The words are associated with binary labels or semantic change scores, and the models are evaluated on a ranking or binary classification task. Most recent datasets were produced using the Diachronic Usage Relatedness (DURel) framework (Schlechtweg et al., 2018). In this approach, human annotators are asked to judge the semantic similarity of contextualized examples of word usage on a scale from 0 to 4. Each annotator rates multiple example pairs per word, with a final graded semantic change score obtained by averaging over the individual ratings.

The first evaluation using this type of dataset was conducted by Schlechtweg et al. (2019) on German. They reported the most robust results for `word2vec` models trained using the SGNS algorithm and aligned using the Orthogonal Procrustes analysis. Systematic comparisons of an ever increasing range of methods have continued through a series of shared tasks on several European languages. Results on English, German, Latin, and Swedish (Schlechtweg et al., 2020) as well as Italian (Basile et al., 2020) highlighted unexpectedly strong results of type-level models, which clearly outperformed contextualized representations produced by BERT and similar models. This trend inverted for Russian (Pivovarova and Kutuzov, 2021): the best

performing system used contextual representations produced by XLM-R, another Transformer-based model, achieving double the score of the best performing type-level model. This may be explained by the more complex implementation compared to previous attempts at using pre-trained models on this task.

While this trend in evaluation has put methodological choices on a much firmer ground, it relies on a limited number of lexical items. There are still comparatively few studies examining the behavior of semantic change detection models on the whole vocabulary, even though this is important in order to understand their practical utility. One exception is the study by Basile and McGillivray (2018), who evaluate the top semantic change candidates proposed by different systems against an external lexicographic resource. The large number of identified candidates entails a relatively high recall (0.104–0.849), but all models obtain extremely low precision scores (0.003–0.005), with the maximum F1 score at 0.01. This is in stark contrast with the relatively high performance on test sets. Similarly, Shoemark et al. (2019) evaluate the ability of their models to detect the top semantic change candidates. They use synthetic corpora, where the distributional patterns of specific words are altered in a controlled way so that they reflect different types of semantic change. While this leads to informative methodological recommendations, it does not reflect important real-life issues such as the impact of noise in the data.

As Hengchen et al. (2021) observe, more work on evaluation is needed, and this remains a challenging endeavor. The results discussed here suggest that the practical applicability of the models in descriptive research may be limited, but this question is yet to be addressed head on. That being said, examples of potential descriptive applications of these methods exist, and I turn to them next.

5.2.2.4 Descriptive contributions

As illustrated by the discussion in the previous sections and underscored by Boleda (2020, p. 218), the numerous recent studies on semantic change detection mostly aim to demonstrate that NLP systems can detect semantic change, rather than provide a descriptive contribution relying on these systems. This is compounded by a broad understanding of semantic change in NLP (Tahmasebi et al., 2021, p. 15), often unconstrained by theoretical or methodological considerations of the linguistic issues at stake (p. 12). Existing work nevertheless clearly illustrates the descriptive potential of these approaches.

A series of studies have used VSMs to empirically investigate the validity of theoretical linguistic hypotheses. For example, Xu and Kemp (2015) evaluate two competing laws of semantic change: the law of differentiation, which posits that near-synonyms diverge in meaning over time, and the law of parallel change, according to which words that are similar in meaning follow similar semantic change trajectories. They report overwhelming evidence for the latter.

In a similar vein, Dubossarsky et al. (2015) draw on regularities in their models to formulate the law of prototypicality, according to which the degree of prototypicality is negatively correlated with the rate of semantic change. Hamilton et al. (2016b) propose the law of conformity, corresponding to a negative correlation of word frequency and semantic change rate;

and the law of innovation, corresponding a positive correlation between polysemy and the rate of semantic change. However, Dubossarsky et al. (2017) have subsequently reported that these observations were largely biased by the inherent negative influence of low frequency on the quality of vector representations. This underscores the potential for teething problems with semantic change detection methods to impact the linguistic descriptions they produce.

Another type of descriptive application consists in using the models to facilitate exploratory analyses by domain experts. For instance, Rodda et al. (2017) use VSMs for Ancient Greek to detect semantic shift candidates and then analyzed their nearest neighbors to establish meaning change. Similarly, Peirsman et al. (2010) examine the distributional profiles of religion names before and after 9/11, finding an increase in negative associations with Islam in the later period. De Pascale (2019) studies regional lexical variation in Dutch using token-level models to identify cases where competing words were used with equivalent meanings. This enables a more precise definition of linguistic variables in the study of onomasiological variation. These and other studies in the tradition of dialectometry, which model aggregate patterns in regional language variation, present an obvious interest for this dissertation.

More generally, computational analyses of language variation often correspond to comparisons of linguistic patterns in different corpora, and numerous methods beyond VSMs can be used to address them. I briefly review some of them below.

5.2.3 Other computational approaches to language variation

Computational detection of semantic change can also rely on approaches involving different types of information in addition to vector-based meaning representations. This includes methods using co-occcurrence statistics of some kind (Basile et al., 2016; Tang et al., 2013), topic models (Lau et al., 2012; Cook et al., 2013; Frermann and Lapata, 2016), and architectures based on word sense induction (Mitra et al., 2014, 2015; Tahmasebi, 2013; Tahmasebi and Risse, 2017). However, given the general focus on vector space models and their strong performance on a range of evaluations, these methods will not be discussed in further detail; a comprehensive review can be found in Tahmasebi et al. (2021).

Moreover, beyond analyses focusing on semantics, a range of computational approaches allow for large-scale investigations into language variation across different corpora. One widely used approach is the Sparse Additive Generative Model (SAGE) (Eisenstein et al., 2011), which estimates the deviation in log-frequencies of terms in a corpus of interest relative to their log-frequencies in a background corpus using the maximum-likelihood criterion, with a regularization parameter ensuring that rare terms are not overemphasized. It has been used to identify lexical variation driven related to geography (Eisenstein, 2018), age and gender (Pavalanathan and Eisenstein, 2015b), national identity (Shoemark et al., 2017b), and the use of hate speech (Chandrasekharan et al., 2017). More specific work on lexical variation includes the induction of lexical variables, understood as comprising functionally equivalent lexical variants (Shoemark et al., 2018). Approaches such as these do not constitute the core of the work presented in this thesis, but I have used them in specific steps (cf. Section 9.1) and will reference them as needed in the coming chapters.

The sheer number and variety of studies discussed in this section illustrate the vitality of the research on computational modeling of semantic variation and change. At this point, fairly clear methodological recommendations and evaluation practices have been established. However, much of the work continues to be devoted to introducing new methods and evaluating them on standard datasets, which do not reflect all of the aspects central to the utility of these approaches in linguistic work. While their potential is clear in principle, and is supported by the existing descriptive studies, it is yet to be demonstrated at scale.

5.3 Summary

This chapter reviewed wide-ranging methodological, and some theoretical, considerations related to extracting and analyzing patterns of semasiological variation from linguistic data. From the standpoint of variationist sociolinguistics, this type of variation is rarely studied, in large part due to a lack of adequate methods. This is compounded by a theoretically unwieldy nature of semasiological variables within the variationist framework. Although existing studies in dialectology and variationist sociolinguistics are limited in number, they provide possible methodological solutions, as well as demonstrate the descriptive interest of semasiological variation.

However, as relevant as these studies are, it is clear that they cannot account for large-scale quantitative patterns. That is why I also turn to computational studies of semantic variation and change. These methods are numerous, promising, and should allow a systematic, bottom-up study of semasiological variation. But this is also a relatively recent field, with many open questions relating to the optimal implementation and evaluation of the models, as well as their utility in descriptive research.

To sum up, the possible methodological choices resemble a balancing act. The sociolinguistic approaches are likely only applicable to a limited number of variables, but can provide detailed sociodemographic information on the speaker, which can be interpreted against the backdrop of well-established variationist theory. The computational approaches allow for systematic, vocabulary-level analyses, but provide little to no information on the speakers, and come with uncertainties as to their descriptive validity. Bearing this in mind, the choice made in this dissertation is to implement both types of approaches in a complementary manner; this will be outlined in detail in Chapter 7. But first, we turn to the issue of accounting for the variation patterns extracted from the data.

Chapter 6

Accounting for language variation

In the previous chapter, we saw that a range of approaches, operating at very different scales, can be used to identify patterns of semasiological variation in linguistic data. Some of the analyses proceed in a top-down fashion, examining in detail a predefined set of semasiological variables; others adopt a bottom-up approach, aiming to spontaneously uncover traces of semasiological variation. Whatever the case, the fact that patterns of potential interest are turned up by these methods does not in itself constitute a comprehensive sociolinguistic description. These linguistic patterns must also be explained; in other words, the constraints that influence their use must be accounted for.

This is the focus of the present chapter. Section 6.1 discusses the criteria that are used to establish if an observed case of language variation is in fact related to contact. Section 6.2 outlines the internal (linguistic) and external (social) factors that may condition language variation, focusing specifically on the use of semantic shifts in Quebec English. Section 6.3 addresses the potential for contact-induced semantic shifts to convey social meaning. Shifting the focus from variationist to computational analyses, Section 6.4 presents the methods that can be used to investigate standard sociolinguistic factors in large datasets. Finally, Section 6.5 provides the main takeaways from this discussion.

6.1 Establishing the effect of language contact

By definition, a study of contact-induced semantic shifts entails the need to establish if the observed patterns of language variation can be reliably ascribed to language contact. In defining the criteria which will guide this decision, I turn to existing studies of language contact phenomena conducted in variationist sociolinguistics.

Lexical effects of language contact have often been examined within the broader focus on the use of other-language items. This line of research has directed considerable effort into distinguishing between different types of other-language insertions, in particular codeswitching and borrowing. This work has resulted in comprehensive proposals on how these phenomena can be differentiated (e.g. Poplack and Meechan, 1998; see Chapter 1 for a summary of this debate). It has also fostered a reflection on adapting some of the cornerstones of the variationist theory to the study of language contact.

A case in point is the principle of accountability, according to which both realizations and non-realizations of all variants of a linguistic variable should be quantitatively reported whenever possible (see Chapter 5). Poplack (1993) argues that it cannot be applied to borrowings in a straightforward manner because of the difficulty in determining which recipient language words might be replaced by borrowings, and in defining the full set of synonyms for a borrowed lexical item. She therefore points to the solution adopted by Poplack et al. (1988): instead of comparing individual other-language items to their recipient language alternatives, the analysis consists in identifying differences within the entire set of other-language items. The types of other-language items that are identified in that way are then correlated with sociolinguistic factors, much as if they were the variants of a traditional linguistic variable (Poplack, 1993, pp. 277–278). But that is not the only way to frame this issue. For instance, dialect surveys have shown that it is in fact possible to define a comprehensive set of variants for a lexical variable which also includes borrowed forms, and that this can provide valuable information on the effects of language contact (e.g. Boberg, 2012; Boberg and Hotton, 2015; Chambers and Heisler, 1999).

However, the issues at stake are not identical when it comes to contact-induced semantic shifts. If a lexical item is fully borrowed, the influence of language contact is self-evident in that the surface form is present in the donor language and absent from the recipient language (except for ambiguous cases, which nevertheless likely remain marginal). By contrast, semantic shifts involve a modification in the meaning of a lexical item that already exists in the recipient language, so the influence of the donor language is less directly observable.

This is reminiscent of structural (i.e. morphosyntactic) effects of language contact. Whereas the studies of contact-induced lexical phenomena mainly aim to differentiate distinct types of other-language items, a key issue in this line of research is that of determining whether an observed morphosyntactic pattern is in fact a product of cross-linguistic influence. For instance, Poplack et al. (2012a, p. 204) outline a series of precise criteria in this respect:

> A conclusion in favor of contact-induced change should rest on the demonstrations that the candidate feature
>
> (i) is in fact a change,
> (ii) was not present in the pre-contact variety,
> (iii) is not present in a contemporaneous non-contact variety,
> (iv) behaves in the same way as its putatively borrowed counterpart in the source variety, and
> (v) differs in non-trivial ways from superficially similar constructions in the host language, if any.

It is certainly useful to define strict criteria for establishing the influence of contact. However, the position outlined here implies a binary view, in which the development of a linguistic feature must exclusively depend on the donor language for it to be considered as contact-induced. Otheguy (2012) points out that this precludes the possibility for contact to act as a contributing factor, whereby the presence of a linguistic feature in the donor language accelerates the development of a corresponding feature in the recipient language, even if the same

feature also exists in non-contact situations (p. 227). But Poplack et al. (2012b) retort that this requires determining the rate of change in the absence of contact, which they suggest cannot be done. They instead underscore the importance of external speaker-level variables, such as the degree of contact or of bilingualism (pp. 250–251).

This view fails to answer the underlying question, even though the authors acknowledge the need for further debate. I would also argue that it underestimates the methodological implications arising from the five criteria presented above. Note in particular that points (ii) and (iii) suggest that the variety under study should be compared diachronically to a pre-contact variety, and synchronically to a non-contact variety. But another comparison can also be made: that between the pre-contact variety and the contemporaneous non-contact variety. This of course makes little sense if we require that the contact-induced feature be absent from both those varieties. But if we instead focus on subtler differences in the rate of use of a linguistic feature that is variably present across the varieties, this third comparison provides the missing information – albeit imperfect – on how that feature might evolve in the absence of contact.

Another note is due on the notion of language change. It is widely accepted in variationist sociolinguistics, as well as reiterated by Poplack et al. (2012a) among many others, that synchronic variation necessarily precedes diachronic change, but that not all cases of variation ultimately lead to change. Bearing this in mind, it is worth reiterating that this dissertation views contact-induced semantic shifts as a general phenomenon of cross-linguistic influence on the lexical semantic level. It studies them mainly from the standpoint of synchronic semasiological variation, but it also validates these observations against diachronic data. The theoretical basis for this view was discussed in Chapter 3, the methodological implications will be fully presented in Chapter 7, and the link between variation and change will be further clarified in relation to age as an external sociolinguistic factor (Section 6.2.2.3). Note moreover that this dissertation assumes that synchronic variation is important in its own right, whether or not it leads to stable change, as illustrated by the social meanings that it can convey (Section 6.3).

Finally, as mentioned in Chapter 2, much has been said in the literature on Quebec English regarding the relative importance of lexical and structural contact-induced change, and whether the former can lead to the latter (e.g. Boberg, 2012; Boberg and Hotton, 2015; Fee, 2008; Poplack et al., 2006). This specific debate is beyond the scope of this dissertation, which simply assumes that lexical semantic effects of language contact can be independently observed and described. In order to pursue this description, however, we must first define the sociolinguistic factors which might explain the observed cases of variation. This issue is addressed by the next section.

6.2 Sociolinguistic factors

In the previous chapter, I discussed the notion of linguistic variable: it enables us to correlate the choices of different linguistic variants with a range of explanatory sociolinguistic factors. The focus of this section is precisely on the factors – both internal (linguistic) and external (social) – that may explain the use of contact-induced semantic shifts.

6.2.1 Internal factors

Like many other aspects of variationist sociolinguistic analysis, the use of internal factors, i.e. quantifiable linguistic features which might influence the use of a linguistic variable, was pioneered in studies of sound change (cf. Labov, 1994). Some of the existing studies of semasiological variation (see Chapter 5) address linguistic properties indirectly, in circumscribing the variable context in which a lexical item occurs, but none of them explore their explanatory potential. In order to do so, I will draw on other related lines of research, particularly on lexical borrowing and on diachronic semantic change, as some of the features they examine may also be expected to apply to contact-induced semantic shifts.

Part of speech. Research on lexical borrowing has established that "major-class content words such as nouns, verbs, and adjectives are the most likely to be borrowed" (Poplack and Meechan, 1998, p. 127). This reflects the findings reported by Van Hout and Muysken (1994), who additionally provide an empirically observed, five-level hierarchy of borrowability: the first level comprises common and proper nouns; the second includes adverbs; and the third includes adjectives and verbs (p. 60). The impact of part of speech has also been examined in computational studies of semantic change. Contrary to the hierarchy of borrowability, Dubossarsky et al. (2016) have suggested that verbs change more than nouns, and nouns more than adjectives. However, methodological choices, and specifically the measures that are used to quantify semantic change, may prioritize verbs and nouns differently. This is what Hamilton et al. (2016a) report; their study is in turn based on the more general claim that nouns are more likely to be affected by irregular patterns of semantic change arising from cultural influence (Traugott and Dasher, 2002, pp. 3–4).

Frequency. In lexical borrowing, frequency can be addressed from the perspective of the donor language as well as of the recipient language. Starting with the former, Van Hout and Muysken (1994) report that a high donor language frequency may facilitate borrowing, but they also caution that it may interact with other factors such as part of speech (pp. 52–54, 59–60). Working on different language pairs, Zenner et al. (2017) similarly report that loanwords that are more frequent in the donor language are also more likely to be borrowed (pp. 124–126). But these analyses are not without methodological challenges: for instance, Zenner et al. (2012) decide against examining donor language frequencies absent a comparable corpus in the donor language which would ensure reliable frequency measures (p. 768). As for recipient language frequency, it is often used in variationist studies of borrowing, but generally not as an explanatory variable directly accounting for patterns of language variation. Rather, it serves as a basis for differentiating related outcomes of contact, such as nonce borrowings and established loanwords (e.g. Poplack, 2012, p. 645).

The impact of frequency has also been examined in computational studies of semantic change. Working on monolingual English data, Hamilton et al. (2016b) find that frequency is negatively correlated with semantic change; that is to say, the more frequent a word is, the less it changes over time. However, as previously discussed, this observation may be affected

by the inherent influence of frequency on corpus-based analyses; this extends to cosine distance, which is used to quantify semantic change across VSMs (Dubossarsky et al., 2017). This is further supported by the fact that Uban et al. (2019) report the opposite trend from a multilingual standpoint. They study the evolution of cognates in several European languages, finding that semantic divergence between pairs of cognates over time is positively correlated with frequency. These results overall indicate the likely relevance of frequency, but they also underscore the potential impact of methodological choices: Hamilton et al. (2016b) limit the analysis on the top 10,000 most frequent lexical items and measure frequency in the initial time period (before any potential change), whereas Uban et al. (2019) focus on a more limited and carefully constructed set of cognates, and measure frequency in the final time period (after any potential change).

Semantic properties. A range of lexical semantic properties have been investigated in relation to semantic change, including in some of the same computational studies that have examined the effects of frequency. Specifically, the degree of polysemy was found to be positively correlated with the rate of semantic change both for monolingual English data (Hamilton et al., 2016b) and for cognates observed across multiple languages over time (Uban et al., 2019). Another semantic characteristic related to diachronic semantic change, that of prototypicality (Geeraerts, 1997), was computationally examined by Dubossarsky et al. (2015). They found an effect operating in the opposite direction, i.e. the degree of prototypicality is negatively correlated with the rate of semantic change. However, in addition to the same methodological caveats as for the studies investigating frequency (Dubossarsky et al., 2017), these analyses involve the additional difficulty of automatically quantifying the degree of polysemy and prototypicality, which may introduce a bias of its own (see e.g. Hamilton et al., 2016b, p. 1496).

Dialectometric studies such as Franco et al. (2019) have investigated regional differences in lexical diversity, defined as "the amount of lexical variation a particular concept shows" (p. 206), and how it relates to concept characteristics such as vagueness and salience. They report that the characteristics of the semantic field to which a concept belongs mediate the effect of these concept characteristics on lexical diversity (pp. 235–236). Likewise, Zenner et al. (2012) report that loanwords are more readily adopted if they belong to specific lexical fields, in particular those that are culturally associated with the donor language (p. 781). As in the case of polysemy and protoypicality, a key methodological challenge in these analyses consists in operationalizing the semantic features under study. While some solutions are provided in existing research, it is unclear how easily they can be applied to a very large corpus, or whether these operationalizations remain reliable even as the theoretical complexity of the underlying features increases.

Phonological integration. The degree of phonological integration of borrowed lexical items is a central issue in sociolinguistic research. An influential empirical claim, put forward by Poplack et al. (1988), states that the degree of phonological integration of a loanword is correlated (i) positively with its diffusion in the speech community, as reflected by the number of speakers who use it; (ii) positively with its age of attestation; and (iii) negatively with the

bilingual ability of individual speakers, specifically as regards their proficiency in the donor language (pp. 70–75). These observations are supported by subsequent studies, including in the context of Quebec English. For instance, Rouaud (2019b) reports a statistically significant effect of the degree of bilingualism on the phonological integration of French loanwords, with original pronunciation generally produced by speakers who are also proficient in French (pp. 250–256). On the language-internal level, Poplack et al. (2020) claim that the degree of integration is influenced by the other-language segment that needs to be adapted. The reported effect is so strong that the link with bilingualism is no longer found to have an impact; note however that monolingual speakers were excluded from this analysis.

As for contact-induced semantic shifts in Quebec English, the effect of phonological integration can only be observed on a subset of these items, those that are sufficiently phonologically similar to their French equivalents for them to be subject to French pronunciation. However, phonological integration remains a worthwhile dimension to pursue given the previously reported links with constraints on both the internal and external level. It is moreover indicative of cross-linguistic similarity, to which I now turn.

Cross-linguistic similarity. Recall that the definition of semantic shifts adopted in this dissertation entails a degree of semantic and/or phonological similarity between corresponding lexical items in the donor and recipient language. The extent of this similarity may affect the characteristics of their use, which is also reflected by the distinctions drawn in existing classifications of semantic shifts (e.g. Haugen, 1950, pp. 219–220; see Chapter 3 for an overview).

A systematic account of cross-linguistic similarity would involve, on the semantic level, an estimate of the difference between the meaning of the recipient and donor language words. For example, Uban et al. (2019) obtain this by computing the distance between vectors for the languages that they examine in a multilingual vector space model. As for phonological similarity, it is investigated by Zenner et al. (2017) under the notion of foreignness. In a study of English–Dutch lexical borrowing, they estimate foreignness by computing the difference between native English pronunciation and the corresponding Dutch pronunciation if it were adapted in the simplest possible way (pp. 123–124). While they find that its effect on loanword adoption is not significant (p. 127), it is unclear if the same pattern should be expected for contact-induced semantic shifts, which are not underpinned by the same mechanisms.

Attestation history. Loanword research has shown the explanatory importance of attestation history regarding the diffusion of borrowed items. As an example, Poplack et al. (1988) report a correlation between, on the one hand, the number of speakers who use a borrowed item, and, on the other, the probability of that item being attested as well as its age of attestation. Put otherwise, borrowed lexical items that are more widespread in the community also tend to be better established as indicated by lexicographic evidence (pp. 58–59). The fact that a lexical item or a specific sense is attested in a dictionary provides evidence of its status within the speech community and is central to existing descriptions of lexical variation (e.g. Dollinger, 2017, pp. 84–85; Rouaud, 2019b, pp. 245–246). This type of information is particularly relevant as an indication of historical trends and can therefore help establish the influence of language

contact on observed patterns of variation (see Section 6.1).

In summary, a range of factors related to the linguistic structure can be expected to affect the use of contact-induced semantic shifts, based on reports in previous studies of semantic change, lexical borrowing, and lexical variation. But these linguistic behaviors can also be influenced by external factors, related to speakers' sociodemographic characteristics; I turn to them next.

6.2.2 External factors

The potential impact of a series of external factors on language variation is well established in variationist theory as well as in existing sociolinguistic studies of Canadian and Quebec English (many of them previously introduced in Chapter 2 and, with specific reference to lexical variation, Chapter 5). Drawing on this background, this section will first address two factors related to the focus on regionally specific consequences of language contact, namely geographic origin and language use. It will then discuss the core set of factors used in variationist analysis: age, gender, socioeconomic status, and ethnicity.

6.2.2.1 Geographic origin

We begin this discussion by exploring the influence of the speaker's place of origin and residential history. The explanatory power of regional origin in studies of linguistic variation decreased over the course of the 20th century, in parallel with increased social mobility and the resulting trend towards linguistic homogeneization (Chambers, 2000, pp. 173–176, discussing Johnson, 1996). This reflects the fact that geographic linguistic patterns interact with other social factors (Trudgill, 2000, ch. 8), as we will also see. In this dissertation, however, they are of central importance in their own right: describing the way English is spoken in Quebec by definition includes a geographic dimension. Moreover, they are particularly relevant for language contact studies, especially when a contrastive perspective is adopted in order to establish if the presence or absence of another language may affect sociolinguistic behaviors in different communities (e.g. Poplack et al., 2006).

Dialectological and variationist sociolinguistic studies investigating the use of English in Quebec often restrict their scope to Montreal, more precisely defined as the Greater Montreal (Boberg, 2005b) or the West Island of Montreal (Rouaud, 2019b). Other studies have focused on specific areas of Quebec outside of Montreal such as Quebec City (Chambers and Heisler, 1999) and the Gaspé region (Boberg and Hotton, 2015). A comparative focus has also been adopted, contrasting Montreal with the remainder of Quebec in general (Boberg, 2010; McArthur, 1989), with specific areas of Quebec in particular (Chambers, 2007a), or with monolingual areas in the rest of Canada (Poplack et al., 2006).

These comparisons are usually motivated by the claim that the rate of exposure to French is higher outside of Montreal. In general terms, McArthur (1989) reports a difference between the speakers raised in Quebec and those raised outside of the province, which, coupled with their linguistic profile, influences their perception of gallicisms in Quebec English. When it

comes to individual linguistic variables, Boberg (2010) notes that, despite generally parallel trends, some cases pattern differently in Montreal compared to the rest of Quebec. For instance, the term *dépanneur* or its abbreviation *dep* is more strongly preferred to alternatives such as *convenience store* in Montreal than it is in the rest of Quebec, making Montreal more distant from the national mean (p. 173). But the opposite trend is also attested, as in the case of the term for the evening meal, with *supper* less frequent in Montreal than elsewhere in Quebec (p. 181). This second tendency is echoed by Boberg and Hotton (2015), who report that lexical gallicisms are overall more frequent in the Gaspé region than in Montreal, further confirming the effect exerted by more intense contact with French as well as by a higher degree of isolation from other English-speaking areas (p. 307). However, variables unrelated to language contact also present complex patterns. Chambers (2007a) reports that, compared to other regions of Quebec, Montreal is more advanced for one language change (the irregularization of the past tense of *sneak* to *snuck*), but lags behind for another (the *Mary–marry–merry* merger). This suggests that Montreal is a conservative dialect region in some respects, "a population centre that maintains its own integrity and forms a pocket of resistance to the norms surrounding it" (p. 34).

In addition to defining a region to be investigated, each study decides what kind of residential history is deemed acceptable. Most examples cited in this section implement traditionally strict criteria. Poplack et al. (2006) include "only anglophones born, raised, and currently residing" in Montreal, Quebec City, and the control region of Oshawa-Whitby in Ontario (p. 187). The North American Regional Vocabulary Survey is agnostic as to the birthplace, but it only retains the respondents who "grew up entirely in one region and still live in that region today". This is justified by the claim that relaxing this criterion could obscure otherwise clear regional patterns (Boberg, 2005b, p. 29). Similarly, Boberg and Hotton (2015, p. 285) recruit participants who grew up in the Gaspé region and focus their analysis on those still living there. Rouaud (2019b) likewise includes speakers who were born and/or raised from an early age, and spent most of their life, in the Greater Montreal area, with at least one parent presenting the same residential history (pp. 183–184). In a word, the standard minimum criterion is birth or early arrival as well as near-continuous residence in the target geographic area.

One exception to this trend is the Dialect Topography Project, which includes around 300 respondents from the Quebec City region. However, they are not limited to people born in the region:

> In the Dialect Topography sample, we aim for a demographic cross-section of the survey area rather than a population of indigènes as in traditional dialect surveys. Our reasoning is straightforward: some proportion of urban speech communities is made up of people who were born outside the community, and the variants they use in their speech are heard in that speech community and have some status in it. We want to know what those variants are and the extent of their use. (Chambers and Heisler, 1999, p. 41)

The dialect questionnaire includes a series of questions on the respondents' and their parents' residential history. This information is then used to compute a Regionality Index. It

ranges from 1, for those born, raised, and living in the Quebec City region, and whose parents were born there; to 7, for those living in the Quebec City region, but born and raised outside of the province, with parents likewise born outside of the province (p. 41). Its utility is confirmed by the fact that it is the most robust correlate for two lexical variables in Quebec City. For example, the limited but non-negligible use of *bureau* (rather than *chest of drawers* or *dresser*) is strongly associated with a low Regionality Index, which is indicative of strong local ties. This is in turn explained by a potential link with the corresponding Quebec French term *bureau* (pp. 43–46).

The same approach has also been successfully applied outside of Quebec. In a comparative study of English in Vancouver and Washington State, Dollinger (2012) uses Regionality Index to split his participants into two groups: locals, i.e. those who grew up in the target area, and non-locals. This study is notable because the latter group is large (45% of all participants) and consists mostly of L2 speakers, mirroring the situation in Vancouver. Crucially, many of the observed patterns are explained in terms of language use; I discuss it below.

6.2.2.2 Language use

Investigations of language contact describe patterns of language use exhibited by respondents of different linguistic profiles. Reflecting the criteria on geographic origin, most studies on Quebec English only recruit native English speakers. But even so, their knowledge of other languages is evaluated in order to determine if it has an effect on their English use; this is closely related to the issue of quantifying bilingual proficiency, discussed in Chapter 1.

A typical approach is illustrated by Rouaud (2019b), who recruits native English speakers, and then scores their French skills using the information obtained during the sociolinguistic interviews. She computes their rate of bilingualism based on their (i) self-reported level of proficiency and frequency of use; (ii) passive exposure to French; (iii) age and mode of acquisition of French; and (iv) domains of use of French. The information provided during the interviews is also used to measure attitudes towards French language policies on a three-point scale: positive, neutral, or negative (pp. 204-209). The practical applicability of the approach is demonstrated by the fact that the rate of bilingualism, operationally split into three categories, exhibits a significant effect on the phonological integration of French borrowings (pp. 251–253). This is similar to the approach adopted by Poplack et al. (2006), who also recruit participants identifying as anglophones, and measure their French proficiency using two numerical indices. They are comparable in nature to Rouaud's proficiency and attitude scores, but insufficient details are provided to reproduce the exact scoring procedure (pp. 191–192). Moreover, once established, the indices are not used in the ensuing analysis.

More diverse linguistic backgrounds are present in other studies. McArthur (1989) combines the information on language use and geographic origin in order to operationally categorize his participants. He reports that Francophone Quebecers are the most likely to perceive contact-related use as acceptable, followed by Quebec-born Anglophones, and finally English speakers raised elsewhere. Importantly, English-speaking residents of Quebec raised outside of the province exhibit some patterns that do not align with the varieties of English spoken outside

of Quebec. They instead tend towards French usage, indicating the potential for contact with French to actively influence English when living in the province (pp. 72–76).

Similarly to its approach to geographic origin, the Dialect Topography Project also allows for a variety of linguistic profiles. It analyzes them using the Language Use Index, based on the reported frequency of the use of English in four contexts (at home, at work, with friends, and with relatives). The index theoretically ranges from 0, for respondents who always use English in all contexts, to 12, for those who never use English in any context (Chambers and Heisler, 1999, p. 28). Its explanatory value is illustrated by several patterns reported in Quebec City, such as the preposition used with the adjective *different* – *from*, *to*, or *than*. The use of *from* is dominant overall, but increases with the Language Use Index and becomes categorical among the speakers scoring 8 on the scale. This is interpreted as potentially reflecting the directly equivalent French usage of *différent de* (Chambers and Heisler, 1999, pp. 31–32).

The descriptive relevance of this approach is further confirmed by Dollinger's (2012) previously discussed study of Vancouver English, in which the Language Use Index is a leading predictor for most examined lexical variables. This is indicative of the central role played by non-native speakers in Vancouver, who reinforce some ongoing changes and resist others. Crucially, they are the ones driving the preservation of some distinctive features of Canadian English varieties (e.g. preservation of the glide in *news* /nju:z/) and thereby helping to distinguish the use of English in Vancouver from that in Washington State (pp. 529–531).

6.2.2.3 Age

Analyzing language variation in relation to age is a central component of sociolinguistic research. It is grounded in the apparent time hypothesis, which posits that synchronic stratification of linguistic variants by age may be indicative of diachronic language change. This is based on the assumption that speakers are likely to learn a linguistic feature and retain it throughout their life, subsequently reflecting the state of the linguistic system at the time of acquisition. This pattern must be distinguished from age grading, where stratification by age reflects changes to the individual's way of speaking over the course of their lifetime. Distinguishing the two often involves the introduction of a real time component, i.e. historical evidence against which the evolution suggested by generational patterns can be validated (see e.g. Sankoff, 2006).

In addition to this general caveat, Boberg (2010, pp. 188–189) calls for caution in applying the apparent time hypothesis to lexical variables, echoing similar concerns voiced by Labov (2001, p. 123). The apparent time analysis hinges upon the stability of linguistic features once they have been acquired, whereas evidence suggests that some speakers may adopt innovative lexical forms later in life (cf. Boberg, 2004c). In a concrete illustration of this concern in Canadian English, Chambers (1998) reports an aggregate analysis of several phonological, morphosyntactic, and lexical variables, characterized by a high degree of variability among the older respondents. He argues that one potential explanation is the subsequent adoption of innovative forms by older people; another is the presence of broader patterns of variability in the older generation reflective of more pronounced heterogeneity of the Canadian society in the

first half of the 20th century (pp. 28–30). Uncertain interpretations of this kind illustrate the importance of real time validation in some contexts.

That being said, numerous other studies of lexical variation in Canadian and Quebec English demonstrate the interest of the apparent time hypothesis. A well-known example is Chambers's (1995) report of a strong decline in the use of the Canadianism *chesterfield* and its near total replacement by the synonym *couch*. Based on the age of the respondents coupled with lexicographic evidence, he argues that *chesterfield* became the standard Canadian variant in the 1920s, it started losing ground in the 1950s, and entered a definitive decline in the 1970s. Interestingly, the Quebec City data show that *couch* is the dominant choice in that region as well, but they also paint a more complex picture in apparent time. Unlike elsewhere in Canada, *chesterfield* was briefly the majority choice, but never a strongly dominant variant, in Quebec City. It was first replaced by *sofa*, and then by *couch*, with *sofa* still reported by around 20% of respondents in the youngest cohort at the time of the survey (Chambers and Heisler, 1999, pp. 33–35).

The influence of contact as reflected by age is more closely examined by Boberg and Hotton (2015). They rely on apparent time patterns to show that Gaspé English exhibits trends of convergence with Montreal English, such as increased use over time of contact-related lexical variants including *chalet* and *trio*, also preferred in Montreal (p. 300). But the same apparent time evidence also points to some divergent patterns, including a decrease in the use of lexical variants such as *soft drink*, replaced by *pop* (p. 303). On a more general level, a potential interaction between age and language use in Quebec is pointed out by McArthur (1989), who posits two separate categories of predominantly English-speaking Quebecers: those who are members of longstanding Anglophone communities, who identify more with the rest of Canada than with Quebec, and whose exposure to French is limited; and younger Anglophones, often the second generation of the first category, who are increasingly in contact with French, speak it with English elements, and may introduce French elements into their English (p. 12). He does not include age in his analysis, so it is impossible to ascertain if this hypothesis is borne out by the data. However, it is interesting to note this potential trend given that the "younger Anglophone" category corresponds to people born in the years surrounding the adoption of Bill 101 and hence growing up with its consequences.

This brings us to the practical issue of splitting speakers into age groups. This is often driven by the availability of the data and the characteristics of the sample, with choices including three (Boberg and Hotton, 2015) or four large age categories (Boberg, 2005b), or finer-grained, 10-year age brackets (Chambers and Heisler, 1999). In studies on Quebec English involving a smaller number of categories, the adoption of Bill 101 in 1977 is often used as a cut-off point, given its strong effects on the use of languages in Quebec (Poplack et al., 2006; Rouaud, 2019b). In this setup, the influence of French is a strong candidate in explaining any differences observed between the two age groups.

6.2.2.4 Gender

Another extensively used and theorized explanatory variable is that of gender. Summarizing the results of a range of studies, Labov (2001, ch. 8) argues that, in stable variation, women use fewer stigmatized and more prestige variants than men. Parallel to this trend, women lead change from above, i.e. the conscious adoption of prestige forms. However, they are also found to lead change from below, i.e. the adoption of innovative forms without conscious awareness. Taken together, this points to what Labov terms the gender paradox: "Women conform more closely than men to sociolinguistic norms that are overtly prescribed, but conform less than men when they are not" (p. 293). While these observations are reflective of consistently reported general trends, they also interact with other variables such as age and socioeconomic status (Labov, 2001, ch. 9). Moreover, the correlational approach to gender differences obscures the role of gender in context-specific language use aimed at constructing identities (cf. Bucholtz, 2002); we will come back to this issue in Section 6.3.

But let us first take a look at the place of gender in some of the existing research. Analysis of this variable is strikingly absent from most studies on Quebec English, frequently due to a strong overrepresentation of women in the samples, which is known to affect sociolinguistic studies (e.g. Boberg and Hotton, 2015, pp. 285–286). Where it *is* used, gender is usually applied to phonological variation, including the sound changes overviewed in Chapter 2. For instance, Rouaud (2019b) examines Canadian Raising in Montreal, reporting that women lead /aɪ/-raising, whereas no effect of gender is observed on /aʊ/-raising (p. 225). This constitutes a supporting factor for a separate treatment of the two diphthongs (p. 231). Focusing on the Canadian Shift, Boberg (2005a) finds that women lead the retraction of /æ/ and /ɒ/, as would be expected, but not /ɪ/, pointing to a potentially complex social embedding of sound change in Montreal (pp. 147–148).

Trends contrary to the traditionally expected role of women in language change have also been observed on the lexical level elsewhere in Canada. For example, Franco and Tagliamonte (2021) investigate the nouns used to refer to an adult man in Ontario, such as *man*, *guy*, and *dude*. They report a very pronounced diffusion of *guy* at the expense of all other lexical variants, which appears to be led by men. However, interactions with other variables provide the tentative explanation that women of higher socioeconomic status may in fact be distancing themselves from this male-led change. Other interactions of gender with variables including socioeconomic status and ethnicity have been reported in the Canadian context; they will be discussed in the corresponding sections below.

6.2.2.5 Socioeconomic status

As Dollinger (2020) notes, Canadian English "has not yet developed widely perceived social distinctions in speech of the kind that are found in areas that have longer settlement histories or less pervasive forces of homogeneity" (p. 61). On a general level, this is consonant with Chambers's (1998) report of higher linguistic variability among older speakers as a potential reflection of social heterogeneity (see Section 6.2.2.3).

While this is an important trend to note, previous descriptions of Quebec English have

also examined the effect of socioeconomic status, operationalizing it in different ways. For instance, McArthur (1989) reports the occupation of his respondents. A strong skew towards the well-educated can be noted, with nearly three-quarters of the respondents working as language teachers. However, no effect of occupation is reported on language use (p. 29).

A more complex description is pursued by Rouaud (2019b). She uses a scoring system based on five criteria: occupation, education, filial breadwinner's occupation, housing type, and residential area as a proxy for income. Different score ranges correspond to five social classes, with Rouaud's sample split between the middle class and the upper middle class (pp. 186–189). Similarly, Poplack et al. (2006) infer socioeconomic status based on the respondents' occupation, education, and linguistic market ranking. The last point is of particular note for the Quebec context. Introduced by Sankoff and Laberge (1978), it estimates the importance of the language under study for the respondent's socioeconomic life based on ratings provided by a panel of judges (Poplack et al., 2006, pp. 188–190).

While socioeconomic status is only used in these two studies to describe the sample, it features as an explanatory variable in work on lexical variation conducted elsewhere in Canada. In a survey of Ottawa English, Woods (1999) reports that the rate of retention of Canadian lexical markers (e.g. *chesterfield* rather than *couch*) is the highest for the upper middle and lower upper class, followed by the middle class. However, this pattern interacts with other factors, specifically age and gender, with older speakers and female speakers also retaining more Canadian markers (pp. 261–262). A comparison of the Ottawa data with a survey conducted in Vancouver similarly finds that conservative lexical trends are associated with female speakers, older speakers, and higher socioeconomic status (De Wolf, 1996, p. 145). This again underscores the potential importance of socioeconomic background in explaining lexical phenomena in Canadian English.

6.2.2.6 Ethnicity

A final sociolinguistic factor of note is ethnicity.[1] It has been investigated in several studies of Montreal English, where an interaction can be observed between the minority status of English and the preservation of speech patterns specific to different ethnic groups. Boberg (2004b) investigates phonetic variation in the vowel production of Montrealers of Irish, Jewish, and Italian origin, finding clear distinctions between the three groups. Two complementary explanations are put forward: the minority status of English in Montreal, which entails fewer opportunities for immigrants to assimilate to the established local variety; and a high degree of ethnic homogeneity in the neighborhoods in which these groups tend to live.

Building on these initial findings, Boberg (2014) notes that the features specific to the ethnic groups are retained as far down the transmission line as the third generation of speakers, even in the absence of active knowledge of the group's heritage language (p. 71). However, this influence is not straightforward; rather, ethnicity interacts with age and gender. For example, Italians overall diverge from the general Canadian trend to front /uː/ and /aɪ/, the latter outside

[1] Note that in sociolinguistic studies this factor is analyzed based on self-reported information provided by participants rather than an external assessment.

of Canadian Raising contexts. But this behavior is stratified by gender, with female speakers exhibiting significantly more fronted realizations than male speakers. This is consistent with previous reports of /uː/-fronting being female-led, but it also suggests that the more backed realizations may be associated with a stereotypically masculine Italian-American pronunciation (Boberg, 2014, pp. 76–77).

Interestingly, the persistence of ethnolectal features across multiple generations distinguishes Montreal from Toronto, another multicultural Canadian city with ethnically homogeneous neighborhoods. In a study of Torontonians of Chinese and Italian descent, Hoffman and Walker (2010) find that group-specific patterns, reflecting substrate transfer, do not persist beyond the first generation (p. 21). Boberg (2014) interprets this finding in terms of the different status of English in the two cities. In Toronto, it dominates communication and is accessible to all speakers. In Montreal, it has no official status, and there are fewer native English speakers than there are those whose mother tongue is a language other than English or French (p. 75). This suggestion is further supported by results observed in Montreal French, whose role is arguably closer to that of Toronto English. Blondeau (2020) reports that only the first generation of Montreal French speakers exhibit substrate influence, precisely like in Toronto English. Second generation speakers, as well as first generation speakers arriving at an early age, tend to fully acquire the local sociophonetic patterns (p. 168).

Summing up, this section has discussed the use of a range of external factors in explaining observed patterns of language variation. The reviewed literature on Canadian and Quebec English indicates distinct trends that can be expected to affect the use of contact-induced semantic shifts. But while this analysis is likely to reveal social stratification whose significance can be interpreted thanks to well-established theoretical background, it fails to fully account for the social meaning that speakers convey by choosing one linguistic feature over another. I address this issue in the next section.

6.3 Social meaning of variation

The social meaning conveyed by variable choices that speakers make is frequently addressed within the framework of indexicality. I will first discuss the main principles underlying this notion, and will then take a closer look at how it can be applied to semasiological variation, bearing in mind the resulting interaction between lexical and social meaning.

6.3.1 Indexicality and representations

As Eckert (2008, p. 455) has noted, the study of sociolinguistic variation in the Labovian tradition, where explanations for observed language variation are sought in correlations with social factors, can indicate empirically valid aggregate patterns, but it says little about the motivations underlying the choices made by individual speakers. To put it more clearly:

> Because these macrosocial categories [such as class, age, and gender] are fundamental to the social order, they correlate regularly with linguistic variation. This

> is not because the categories themselves engage directly with linguistic practice, but because their intersections structure the conditions and everyday experiences of life on the ground, and variation takes on meaning in the local social practice that unfolds in response to these conditions. (Eckert, 2019, p. 751)

Drawing on Silverstein's (2003) notion of indexical order, Eckert (2008) argues that a linguistic variable can index different kinds of characteristics; in other words, it can convey different social meanings. All of the potential meanings of a variable constitute its indexical field, "a constellation of meanings that are ideologically linked" (p. 464).

Take the example of the southern accent in the United States, discussed by Eckert (2019). In a straightforward manner, this linguistic feature indexes the speaker's geographic origin. However, this first-order indexicality may in turn evoke other characteristics associated with southerners, such as the "redneck" stereotype. It is precisely this accumulation of associations that constitutes an indexical field (p. 754). Importantly, an indexical field represents a communicative resource on which speakers actively draw: in choosing a linguistic variant, they aim to index a specific value, whether it is preexisting or created on the spot (Eckert, 2008, p. 464).

Contact-related linguistic features can also carry social meaning, in a process that is likely to implicate the speaker's relationship with the respective language groups (e.g. Rodríguez-Ordóñez, 2021). In this context, an efficient approach to analyzing social meaning consists in eliciting representations, i.e. qualitative, verbally expressed information reflective of linguistic attitudes (Gueunier, 2003). For instance, the sociolinguistic interviews conducted by Rouaud (2019b) contain a series of questions related to identity and language use, enabling a detailed qualitative description of her informants (pp. 213–217). While in this case the focus was on speaker-level trends, a similar approach can be adopted to elicit representations associated with individual linguistic variables; this will be discussed in Chapter 12.

A challenge specific to the study of the social meaning of semasiological variation is the fact that, by definition, it involves multiple types of meaning. Let us take a look at how this issue can be addressed.

6.3.2 Lexical and social meaning in semasiological variation

Just like in the case of phonological features discussed above, the use of a single lexical item may convey a range of social meanings. One line of supporting evidence is provided by research on terms of address. Take for example Kiesling's (2004) study of *dude*, which is used mainly, but not exclusively, by young men to address other young men. He argues that the term indexes the stances of solidarity and nonintimacy, which can be deployed together to index a stance of cool solidarity. This value explains not only the importance of *dude* in discourses of young masculinity, but also its diffusion among women. In the latter case, the stance of cool solidarity is indexed separately and distinctly from the value of masculinity. From a multilingual standpoint, Alimoradian (2014) reports similarly complex patterns in the use of the vocative *mate* by Australians of a non-English-speaking background. In addition to the well-established association with masculinity, its use here is partly explained by ethnic identity: the

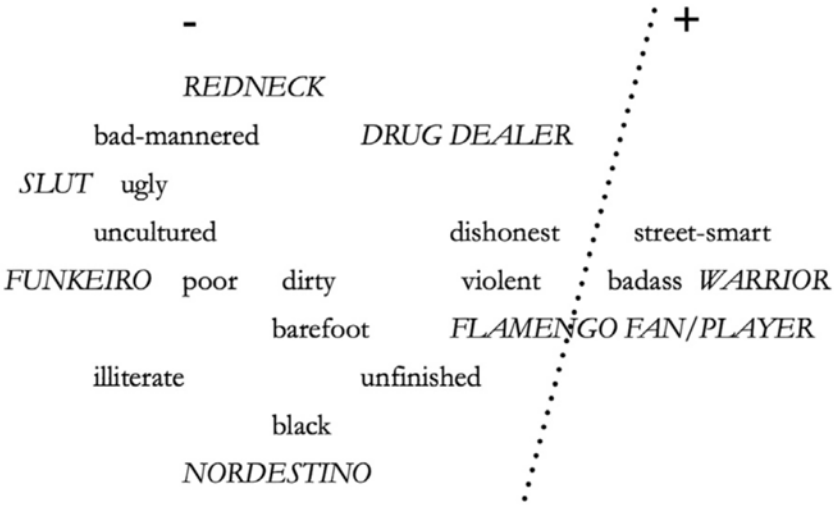


FIGURE 6.1: Indexical field for *favelado*, reproduced from Beaton and Washington (2015). Lower case terms indicate permanent qualities, upper case terms indicate social types, and the dotted line separates positive and negative qualities and social types.

respondents who indicate weaker ties with their ethnic background are more likely to both use *mate* and to be called *mate*, including in a conscious expression of Australian identity.

The interplay between the use of multiple languages and lexical choice is explored by Hultgren (2014). Focusing on the context of Danish universities with a strong international presence, she investigates the variable use of equivalent English and Danish scientific terms. This choice is found to be partly associated with indexicality, as Danish variants convey values such as national pride and intellectualism. This, however, does not explain all variable choices; they are also context-dependent (e.g. limited to specific workplaces) and partly driven by communicative efficiency, which is frequently associated with English terms.

Another important line of research, which involves an overt analysis of semasiological patterns with regard to indexical positioning, addresses the reappropriation of slurs. This is the mechanism by which insulting terms are reclaimed by the group that they originally target, and are then used to express in-group solidarity. An illuminating example is Beaton and Washington's (2015) analysis of the Brazilian Portuguese term *favelado*: in addition to denoting a person living in a *favela*, a slum, it exhibits a range of indexical values. While they principally evoke negative qualities associated with slum-dwellers, reappropriated use points to positive features. The entire indexical field is presented in Figure 6.1.

In applying indexicality to an analysis of lexical meaning, the authors note a crucial difference with respect to other levels of linguistic structure. Phonological variants, for example, do not denote anything in the external world, hence it is understood that their indexical meaning points to the speaker. By contrast, lexical items have a denotational meaning, and it likely does not coincide with the speaker. Here, *favelado* denotes a referent indicated by the utterance in which it is used; it is with this referent that it indexically associates a characteristic of the *favela*. The potential indexical meanings correspond to the permanent qualities and social types in Figure 6.1. Note however that stance – the expression of the speaker's relationship to their

talk and to their interlocutor (Kiesling, 2009) – can nevertheless contextually emerge from this use and point back to the speaker.

How can an analysis of indexicality be applied to contact-induced semantic shifts in Quebec English? Much like in the preceding examples, the use of an English lexical item with a sense typical of French is likely to convey social as well as lexical meaning. Specifically, we can expect it to evoke an association with the use of French or with Quebec. This first-order indexicality may in turn trigger other social meanings associated with the linguistic or regional background. It remains to be seen whether the indexical values point back to the speaker or are instead projected onto an unrelated referent: the latter was argued for *favelado*, but bilingual behaviors are more likely to operate below the level of consciousness, additionally complicating the picture. I would also suggest that the indexical value that is the most immediately associated with a semantic shift – that of French or that of Quebec – might vary depending on the sociolinguistic status of the lexical item; this will be further addressed in Chapter 14.

This approach, coupled with a quantitative analysis of sociolinguistic factors, provides a comprehensive basis for a variationist sociolinguistic account of contact-induced semantic shifts in Quebec English. However, since the analysis will in part depend on Twitter data, we must also understand how these characteristics can be modeled at scale; this is the topic of the next section.

6.4 Corpus-based patterns

Whereas traditional sociolinguistic analyses can draw on decades of theoretical and methodological background, Twitter-based research is significantly more recent. Although it provides promising opportunities for studying previously inaccessible types of language variation and change, the methodology used to model social patterns in Twitter corpora is less consolidated. This section explores the practical implications of this situation, specifically addressing large-scale analyses of regional variation, indirect estimates of sociodemographic background, and studies focusing on patterns of interaction.

6.4.1 Regional variation

Perhaps the most intuitive way of putting Twitter corpora to use consists in analyzing regional patterns reflected by geotagged posts. The objective here is similar to that of dialectology, with the large amount of available data routinely presented as an advantage compared to traditional surveys. While this is counterbalanced by the lack of demographic information, Twitter corpora do introduce an important innovation in enabling bottom-up analyses which aim to uncover patterns of regional variation in the data without predefined constraints.

As an example, Grieve et al. (2018) examine significant rises in frequency over time to semi-automatically identify lexical innovations in American Twitter (e.g. *amirite* 'am I right?'). They then map the spread of these items across the United States, observing five distinct geographic patterns underpinned by different hubs of innovation and pathways of diffusion. Eisenstein

(2018) applies a similar approach based on spatial trends. He quantifies differences in frequency across metropolitan statistical areas (MSAs) to detect lexical items specific to different regions of the United States. Other related studies have focused on identifying topical variation (Eisenstein et al., 2010), training computational models that can detect linguistic features specific to a dialect (Demszky et al., 2021), and learning computational representations of the geographic areas from which linguistic data originate (Purschke and Hovy, 2019).

In addition to this bottom-up trend, predefined sets of lexical items have also been used to investigate linguistic diffusion. Eisenstein et al. (2014) analyze around 2,600 English lexical items obtained through a preliminary semi-automatic analysis. They use census information for 200 largest American MSAs to explore how these items spread through social media. The diffusion of a lexical item between two MSAs is facilitated by demographic similarity, the main predictor being the proportion of African American population. Similar analyses have also been conducted on the global level. Gonçalves and Sánchez (2014) examine a predefined set of variables in geotagged Spanish tweets collected throughout the world. They find that the main distinction in the data is the one between urban and rural areas, with Donoso and Sánchez (2017) confirming the same trend on the country level for Spain.

While results such as these highlight broad interactions of regional and demographic factors, their interpretation is often detached from established sociolinguistic theory. But other studies have drawn on typical sociolinguistic practice, such as defining linguistic variables as sets of corresponding variants and then circumscribing their variable contexts. For instance, Russ (2013) investigates a previously described lexical variable, *soda* / *pop* / *soft drink*, finding that its geographic distribution in American Twitter data coincides with traditional dialect surveys. Building on this validation, he then independently investigates a new variable, the intensifier *hella* / *very*. Similarly, in an earlier study I examine previously described lexical variables known to exhibit regional variation in Canadian English (Miletic, 2016). The observed patterns are similarly reflective of previous reports, with some indicating potentially ongoing change.

The issue underlying these studies – the extent to which the regional patterns observed in Twitter corpora correspond to those established in traditional studies – has also been addressed more explicitly. Doyle (2014) focuses on potential skews in the distribution of Twitter data, introduced for example by the fact that it only provides positive data points (i.e. attested uses of a linguistic item without non-occurrences). He proposes a statistical method resulting in a description which improves the correlation with previous dialect surveys. Shifting the focus to British English, Grieve et al. (2019) analyze 139 lexical variables, comparing the dialect maps produced using Twitter data with those obtained in a traditional study. The spatial patterns described by the two methods coincide, but to varying degrees: Twitter maps are more accurate for lexical variables that present clear regional distinctions, limited polysemy, and low frequency. The findings are nevertheless interpreted as validating the use of Twitter corpora for broad regional analyses.

As we have seen earlier in this chapter, regional trends tend to interact with other external factors, which ought to be included in a sociolinguistic analysis. I now turn to potential solutions to this requirement when working with Twitter data.

6.4.2 Sociodemographic information

This section addresses some of the same sociolinguistic factors already presented in Section 6.2: age and gender, ethnicity, and socioeconomic status. However, the focus here shifts from theoretical grounds for analyzing this information to practical ways in which it can be extracted from Twitter data. This still represents a major obstacle to comprehensive Twitter-based descriptions of language variation.

As shown in the preceding discussion, the standard approach in variationist sociolinguistics consists in identifying patterns of language variation and then correlating them with factors such as age and gender. Given its central explanatory role, sociodemographic information is collected uniformly for all speakers, in a process that is independent from any targeted patterns of variation. By contrast, most Twitter-based studies reverse the perspective, aiming to infer sociodemographic characteristics from the linguistic patterns exhibited by the speakers; direct use of information generated in this way would pose a risk of circularity in sociolinguistic descriptions. However, the strategies used to collect ground truth data for these approaches indicate ways to obtain demographic information independently of linguistic production, and will be explored in this section. This will be complemented by two other approaches: analyses based on the projection of external demographic information, and the occasional descriptively oriented study.

6.4.2.1 Age and gender

Approaches to establishing the age and gender of Twitter users have relied on mining textual information from tweet content and metadata, manually annotating textual and visual information, and adapting data collection strategies. Specific solutions to determine age include extracting recurring textual patterns used to indicate age in profile descriptions (Gauthier et al., 2015; Sloan et al., 2015) and collecting "happy birthday" tweets, in which the target user, or another person wishing them a happy birthday, explicitly states how old they are (Al Zamal et al., 2012; Morgan-Lopez et al., 2017). Manual annotation has been used to determine age based on profile information and tweet content (Nguyen et al., 2013) or visual inspection of profile photos (Jung et al., 2018). Procedures to collect tweets have also been adapted so as to target different generations of users, for example by searching for profiles containing the term *junior* or *freshman* in the description field (Rao et al., 2010).

Similar strategies have been applied to gender identification. This includes running manual annotations based on the profile photo (Ciot et al., 2013) or a wider range of metadata (Nguyen et al., 2013), extracting gender from an external profile indicated as a URL in the description field (Burger et al., 2011), and adapting data collection by initiating Twitter crawls with a gender-specific set of seed users (Rao et al., 2010). Another widespread solution is the use of census data and other similar sources to determine the user's gender based on their first name (Al Zamal et al., 2012; Bamman et al., 2014; Fink et al., 2012; Gauthier et al., 2015; Mislove et al., 2011). Off-the-shelf tools have been designed to produce a gender label based on the username (Knowles et al., 2016), additional information such as profile description (Vicente et al., 2019), visual information from profile photos (Jung et al., 2018) or posted images

(Alvarez-Carmona et al., 2018), or multimodal data combining most of the above (Tellez et al., 2018).

As suggested at the beginning of the section, most of these studies aim to develop computational systems capable of predicting age and gender based on linguistic patterns. A notable descriptive exception is the study by Bamman et al. (2014). They implement a standard computational analysis, assigning gender to Twitter users based on their first name and then predicting it based on textual features. But the rest of their analysis is fully descriptive: first, they obtain the linguistic features associated with each of the genders; then, they explore the users' social networks, finding that the use of gendered linguistic features is associated with the extent to which the same gender is represented in their social networks. This is interpreted as further empirical evidence that gender is not a binary category directly conditioning the choice of linguistic resources; rather, large-scale quantitative associations reflect the active use of gender-indexing linguistic features in a process of stylistic positioning influenced by the audience.

6.4.2.2 Ethnicity

Similarly to other personal attributes, existing studies have often analyzed ethnicity at the level of individual users based on estimates comparing the users' last names to census data for racial distribution (e.g. Mislove et al., 2011). Another approach avoids profile-level analysis altogether, instead using census information in conjunction with geotagged tweets to estimate the influence of ethnicity at a larger scale (Blodgett et al., 2016; Mohammady and Culotta, 2014). An important caveat regarding these approaches is the frequent lack of clarity as to the definition of ethnicity, compounded by the general potential for external evaluations to differ from self-identification (Cesare et al., 2017).

Ethnicity has also been explored in Twitter corpora from a descriptive sociolinguistic standpoint. Jones (2015) collects tweets containing terms associated with African American Vernacular English (AAVE), retaining only those that contain precise geolocation. Validating the overall approach, he observes a general correspondence between the geographical patterns established by Twitter data and African American population density across the United States. He then shows that this usage forms dialect regions that are distinct from those observed for white English, and align instead with historical migration patterns of African Americans.

Another illustrative example is Ilbury's (2020) micro-level manual analysis exploring the interplay of ethnicity with other factors. He analyzes ten Twitter users living in the south of England, and presenting as white gay males, aged between 18 and 25. The analysis indicates a systematic presence of linguistic features typical of AAVE in their tweets. Ilbury argues that they are used to project a specific persona – a social type linked with ways of being and speaking (D'Onofrio, 2020) – termed here the "Sassy Queen". Specifically, the AAVE linguistic features are used in this context to index the view of Black women as sassy, which is in turn associated with the identity of a flamboyant gay man, known as a queen. These analyses overall illustrate the potential for Twitter data to reflect complex interactions between different sociodemographic characteristics.

6.4.2.3 Socioeconomic status

Research aiming to infer socioeconomic status is comparatively more limited, presumably due to the greater complexity of the concept and even fewer explicit indications on Twitter; however, the general approach remains comparable to other characteristics. For instance, Sloan et al. (2015) look up occupation names, drawn from a standard classification, in the profile description field. The occupation determined in this way is used as a proxy for social class. Preoţiuc-Pietro et al. (2015) similarly rely on a list of occupation names, but they integrate them in corpus construction, using them as criteria for crawling user profiles. They map each occupation to its average salary, using that information as an estimate of the user's income. The dataset produced in this way has subsequently been used to explore how socioeocnomic status is linked to stylistic variation (Flekova et al., 2016) and network structure (Aletras and Chamberlain, 2018).

While it may be possible to obtain more precise estimates by combining detailed geolocation data with income patterns (Abitbol et al., 2019), it is unclear whether this effectively improves the reliability of the information – or if it is ethically warranted. It should be noted that Twitter's Developer Agreement considers information such as ethnic origin, sexual orientation, and negative financial status as sensitive; as such, they cannot be used as a basis to profile individuals. While it is open to interpretation whether the notion of profiling applies to linguistic research, Twitter's Developer Policy additionally prohibits standalone use of geographic information originally associated with Twitter context, and limits the conditions in which Twitter profiles can be associated with off-Twitter information. It is uncertain if the most complex approaches used to infer sociodemographic information meet these criteria.

Another way of examining sociolinguistic behaviors is to look at the factors influencing the patterns of interactions exhibited by speakers. This research direction is discussed below.

6.4.3 Interactions and identity

As we have seen earlier on, the social meaning conveyed by variable linguistic choices is a central aspect of variationist theory, but it has only recently started garnering attention in computational studies of language variation (Nguyen et al., 2021). In addition to Twitter-based analyses of the interplay between sociodemographic characteristics and identity construction, briefly presented in the preceding discussion, a related issue that has been explored more extensively is the way in which language variation in Twitter is connected to topic and audience.

For instance, Pavalanathan and Eisenstein (2015a) investigate the use of over 200 lexical items that are either regionally-specific for areas of the United States, or are non-standard items typical of online communication. They examine how the use of these variables varies depending on the audience of the tweets in which they appear. A tweet is said to target a wider audience if it contains a hashtag, a limited audience if it contains a user mention, and a local audience if the mentioned user is from the same metropolitan area as the author of the tweet. The authors report that the use of local and non-standard lexical items increases as the audience becomes more local and smaller.

Adopting an approach closer to standard sociolinguistic practice, Shoemark et al. (2017b)

examine a set of lexical variables, each comprising one or more Scottish English lexical items and one or more Standard English equivalents. They analyze differences in their use for Scottish Twitter users, grouped into two categories depending on whether they express support for or against Scottish independence. The independence-supporting users are found to use more Scottish variants overall; however, this appears to be mediated by topic, as the rate decreases in overt political discourse.

Drawing on these observations, Shoemark et al. (2017a) directly investigate the influence of both audience and topic. Again focusing on Scottish usage, they compare the behavior of users geotagged as Scottish, and of those expressing support for Scottish independence. In the first case, they find independent effects of audience and topic on variable use, which moreover follow the previously reported patterns relative to audience size. However, this pattern is not confirmed in the user group created based on topic (Scottish independence), highlighting the importance of validating descriptive claims across different datasets. Overall, though, these studies confirm that Twitter users exhibit a range of behaviors observed in face-to-face communication, and that these can be successfully modeled at scale.

Taking a step back, this section has shown that most analyses that are central to variationist sociolinguistics are still subject to teething methodological challenges when dealing with Twitter data. Nevertheless, existing studies demonstrate that it is possible to study regional patterns of variation, to indirectly estimate speaker-level as well as large-scale sociodemographic characteristics, and to analyze the factors guiding the interactions between different users. While the issue of reliability inevitably permeates this discussion, it is counterbalanced by the vast amount of data offered by Twitter, as well as complementary sources of information in an interdisciplinary setup that will be outlined in Chapter 7.

6.5 Summary

Building on the preceding discussion on collecting data and isolating patterns of semasiological variation, this chapter explored possible approaches to accounting for these patterns. This overview mainly drew on existing Canadian and Quebec English studies grounded in the variationist sociolinguistic framework, as well as existing Twitter-based research on language variation.

In particular, I first proposed that the effects of language contact can be established based on a strict set of criteria while also accounting for relative, rather than categorical, differences in usage across speech communities and time periods. I then reviewed a series of internal factors – such as frequency, semantic, and phonological properties – as well as external factors – including age, gender, and language use – which could provide systematic quantitative explanations for the use of contact-induced semantic shifts. Complementing this bird's-eye view of variation, I argued for the use of the construct of indexicality to explain the interactions between lexical and social meaning in the use of semantic shifts, with potential further implications regarding the communicative mechanisms at play and the status of variation observed in this manner. Finally, I turned to more practical issues of estimating the external factors in

Twitter corpora, outlining a series of challenges as well as practical ways to address them.

On the whole, this discussion is reflective of the conclusions drawn regarding data sources and linguistic patterns of variation: given the specificity of contact-induced semantic shifts, a comprehensive and accountable description of this phenomenon can only be produced by bringing together different methods. Each of them comes with its own opportunities as well as limitations; I submit that it is possible to benefit form the former while minimizing the latter. The specific way in which that will be done in this dissertation is presented in the next chapter.

Chapter 7

Overview of the method

The preceding chapters have presented the central issues comprising the theoretical, methodological, and descriptive background of this dissertation. A range of approaches have been discussed; their specific place in the overall method I propose, as well as the links between them, will now be explicitly addressed. This chapter will briefly summarize the adopted theoretical and methodological position (Section 7.1), present the high-level aims and hypotheses that are pursued (Section 7.2), and outline the key stages of the computational (Section 7.3) and sociolinguistic (Section 7.4) analyses.

7.1 Research background: a summary

As discussed in Part I, bilingual speakers of different profiles, in terms of their manner of acquisition as well as bilingual proficiency, exhibit cross-linguistic influence in their speech, including on the lexical semantic level. This may result in the use of contact-induced semantic shifts; in the specific context of Quebec English, I defined them as preexisting English lexical items used with a meaning typical of a phonologically and/or semantically similar French lexical item. In doing so, I drew on existing studies, which have described dozens of such lexical items. However, the methods deployed so far have failed to provide a comprehensive description: we still have limited understanding of the diffusion of contact-induced semantic shifts, of the factors that condition their use, and of the social meanings that they convey.

But why is that? I would argue that multiple issues are at play. On the theoretical level, the current treatment of semantic and contact-related phenomena in variationist sociolinguistics is not readily applicable to fine-grained patterns of variation such as contact-induced semantic shifts. On the methodological level, traditional approaches to data collection and analysis are at odds with systematic study of lexical phenomena in spontaneous speech. The discussion in Part II has addressed these issue in more detail and provided potential solutions.

From the theoretical standpoint, I argued, first, that the traditional construct of linguistic variable can be effectively used in a study of contact-induced semantic shifts. A lexical item subject to the potential influence of French can be analyzed as a semasiological variable comprised of different senses with which it is associated. Secondly, I drew on variationists research on language contact to propose adopting a set of established criteria to determine if an observed

linguistic feature does reflect the influence of contact. I suggested that a three-way comparison – between the contact variety, a contemporaneous non-contact variety, and a historical pre-contact variety – can provide sufficient evidence of language contact. This specifically extends to establishing the influence of another language based on relative differences in usage patterns rather than the traditionally prioritized categorical differences. And complementing systematic investigations of sociolinguistic factors, I finally argued for the use of the framework of indexicality to analyze the social meanings conveyed by contact-induced semantic shifts. Importantly, an interaction is expected between the lexical and the social meaning conveyed by a lexical item.

Moreover, we have seen that different types of data sources come with different challenges. Traditional sociolinguistic interviews collect detailed sociodemographic information as well as spontaneous speech production, all the while controlling for phenomena such as style shifting. This however presents limits in terms of studying lexical phenomena, because the production of target lexical items in spontaneous speech is difficult to control, and the resulting amount of occurrences is often insufficient for quantitative analyses. Alternative approaches include using dialectological questionnaires or eliciting meanings within an interview context, but they remain focused on a limited number of linguistic variables. This points to another key issue: in order to systematically study a large number of semantic shifts, and thereby improve upon previous analyses, vast amounts of linguistic data are necessary. A pragmatic solution to this challenge is to construct a corpus of social media posts, like those published on Twitter. While this comes with its own problems, such as limited control over demographic profiles and behaviors such as style shifting, Twitter corpora provide access to many more speakers than can be recruited in a traditional survey, and circumvent longstanding methodological challenges such as the observer's paradox.

Once adequate data are collected, the use of semantic shifts must be quantified. Analyzing them as cases of semasiological variation specifically consists in determining the rate at which a target lexical item is used with a contact-induced sense versus its conventional senses. And once these patterns are extracted from the data, they must be accounted for: the rates of use must be correlated with sociolinguistic factors, such as the speakers' geographic origin and bilingual profile. I have discussed two general ways to go about these two issues. On the one hand, variationist sociolinguistic and dialectological methods provide a means of describing a predefined set of variables while also obtaining reliable information on a comparatively limited number of speakers. On the other hand, computational models of lexical semantics provide ways of systematically identifying differences in word usage within the entire vocabulary of a large number of speakers, and associating them with general demographic trends. The first approach is a top-down analysis which is limited in scope and challenging to apply to spontaneous speech; however, it provides detailed descriptions of individual speakers based on longstanding methodological principles. The second approach is a bottom-up analysis with the potential to provide a system-wide account based on spontaneous speech; but this comes at the expense of speaker-level information and established descriptive validity. In a word, the scale at which these analyses operate is inversely proportional to the degree of descriptive detail they provide.

In order to produce as comprehensive a description as possible, I will rely on both these ap-

proaches, benefiting from their complementary nature. The aims and hypotheses underpinning their use, and the specific way in which they are implemented, are described below.

7.2 Aims and hypotheses

This dissertation adopts a two-pronged objective: descriptive and methodological. The following specific aims are pursued on the descriptive level.

(1) Determine the diffusion and status of contact-induced semantic shifts in Quebec English.

Diffusion can be addressed in terms of language-internal phenomena, i.e. the proportion of the vocabulary affected by this issue, as well as community-level patterns, i.e. the subset of Quebec English speakers who exhibit of this behavior.

Relatedly, status refers here to the extent to which contact-related use is established within the speech community. This is expected to range from a strong association with an imperfect command of English by French-dominant speakers, to widespread regional use typical of Quebec in general. From another perspective, status can be analyzed in terms of the diachronic stability of a pattern of variation observed in synchrony.

(2) Establish the sociolinguistic factors influencing the use of contact-induced semantic shifts.

This includes both internal factors – related to inherent characteristics of a lexical item, such as its frequency or cross-linguistic similarity – as well as external factors, including standard variables such as age, gender, and bilingual language use.

(3) Identify the social meanings conveyed through the use of contact-induced semantic shifts.

This objective is grounded in an analysis of indexical positioning, under the assumption that this process is interactive in nature: some social meanings may be consciously conveyed by the speakers, but others may arise from the perception of the speaker's behavior on the part of their interlocutor.

These aims are associated with a set of corresponding high-level hypotheses.

(1) The diffusion of semantic shifts within the vocabulary is wider than previously indicated. This assumption is based on the discussed lack of systematicity in previous sociolinguistic descriptions, and the pervasiveness of the underlying psycholinguistic mechanism of semantic interference.

Individual speakers are likely to exhibit different rates of use of semantic shifts, primarily related to their linguistic profile. Individual semantic shifts are likely to present different status and diffusion within the community, corresponding to the two poles indicated for aim (1) above. Both claims are supported by existing evidence (McArthur, 1989).

(2) The use of contact-induced semantic shifts is expected to be facilitated by sociolinguistic factors associated with bilingualism, both internally (e.g. strong cross-linguistic similarity) and externally (e.g. speakers who actively use French, speakers who are younger and

hence more exposed to French in Quebec etc.). This is grounded in the same evidence as the previous hypothesis.

(3) The use of contact-induced semantic shifts is likely to index French knowledge or Quebec origin, which may in turn trigger other related associations. This expectation is supported by the high symbolic value reported for contact-related lexical variants in Quebec English (Boberg, 2012; Boberg and Hotton, 2015).

More generally, first-order indexicality is expected to reflect the status of the semantic shift: a primary association with French would indicate a variation in usage related to bilingualism; an association with Quebec would point to established regional variation.

These hypotheses will be restated more precisely and operationalized when they are examined in the coming chapters.

As for the methodological aims, they are subsumed under one general objective: the implementation of an approach that can provide a systematic description of contact-induced semantic shifts in Quebec English, as described above. This, of course, involves multiple components, grounded in the idea that a combination of computational and sociolinguistic methods can provide the most comprehensive outcome. Specifically, the computational methods should:

- identify, within the entire vocabulary, the lexical items that are the most likely to be influenced by the use of French;
- uncover broad factors underlying this variation, as reflected by corpus data;
- for the lexical items that are the most affected by contact, isolate the individual occurrences that directly reflect the influence of French.

These methods entail additional aims related to the data that are necessary for them to be implemented:

- create a corpus that is (i) sufficiently large to ensure the reliability of the computational methods; (ii) regionally diverse, to enable a comparative approach; and (iii) contains sufficient background information on the speakers who created it so as to allow for a broad sociolinguistic description;
- create a benchmark to systematically validate the computational methods.

In short, the use of computational methods constitutes a macro-level, bottom-up analysis which should produce a set of lexical items affected by language contact, and specific occurrences in which that contact-related usage can be observed. This output represents the starting point for the sociolinguistic analysis, which examines the same items more closely in an interview setting. The specific aims include:

- developing an interview task targeting contact-induced semantic shifts, which will provide both comparable quantitative estimates of their use and elicit representations that are associated with them;
- analyzing the data to identify the sociolinguistic factors conditioning the use of semantic shifts, as well as the social meanings that they convey;

- using the obtained results to further investigate the descriptive validity of the computational methods.

A note is also due on the statistical methods that will be used to analyze the results. A key consideration in this respect is my focus on a large number of lexical items, including to uncover previously undescribed linguistic phenomena, coupled with working across large and disparate corpora, derived representations, and complementary empirical information. In this context, I will largely rely on exploratory multivariate methods such as clustering and principal component analysis. They are particularly useful in highlighting meaningful patterns in complex datasets, including to facilitate large-scale manual explorations by the linguist. They will also be used to guide the description of key trends observed in the sociolinguistic interviews; given strong practical constraints on participant recruitment (cf. Chapter 12), the resulting sample is not deemed robust enough to implement confirmatory analyses, such as the logistic regression, which are otherwise routinely used in sociolinguistics. More generally, the implemented computational methods as well as the semantic information collected in face-to-face interviews are compatible with a graded view of (differences in) meaning; where possible, the observed patterns will be analyzed in terms of a continuum rather than being split up based on arbitrary thresholds.

I now turn to an overview of the specific solutions that were chosen to fulfill the aims outlined above. Following the order of implementation, I will first discuss computational models and then sociolinguistic interviews.

7.3 Computational models

The first step in implementing the computational analyses consists in creating the data necessary to implement the lexical semantic models. As previously noted, the analyses rely on a corpus of tweets constructed specifically for this study (Chapter 8). Its construction is grounded in comparative sociolinguistic principles: it comprises data from three Canadian cities – Montreal, Toronto, and Vancouver. The first constitutes the target region for this study, as the one city where French is the majority language; the remaining two cities are control regions, in which the use of French is comparatively limited. As a result, language use that is specific to Montreal, and absent from Toronto and Vancouver, is expected to be related to the influence of French. Moreover, this stage of the study also involves the creation of a test set enabling a systematic evaluation of the computational models. It contains validated cases of semantic shifts as well as stable words, based on external sources (previous studies and lexicographic evidence) and subsequent validation in the Twitter corpus.

The lexical semantic analysis is implemented using different methods. Type-level models are used to automatically detect the lexical items whose distributional semantic profiles are the most different in the Montreal subcorpus and which to that extent constitute semantic shifts candidates. This represents a synchronic implementation of the standard NLP task of unsupervised semantic change detection, with the added constraint of contact influence. The results of type-level analyses are moreover used in a a multidimensional analysis, which is implemented

to facilitate manual data exploration as well as better understand the mechanisms behind the observed trends. A token-level analysis is also conducted. Taking as its starting point a set of already validated semantic shift candidates, it analyzes the individual occurrences of each of those items. Contextualized vector representations are used to automatically group the occurrences into clusters in which the target item is used in a similar manner. This further facilitates manual corpus inspection, enables more extensive quantification of the observed patterns, and leads to a better understanding of the errors affecting type-level approaches. The methods described here are first implemented in an exploratory analysis (Chapter 9), followed by a more thorough investigation of the patterns captured by the models and the data (Chapter 10), and finally a systematic evaluation of type-level models and a more comprehensive deployment of token-level analyses (Chapter 11).

On the whole, these analyses provide recommendations for the implementation of different semantic models, clear estimates of their descriptive utility, and precautions to be taken in standard evaluation practices. Descriptively, these results confirm the presence of previously described semantic shifts in the data, detect new cases, and identify factors potentially explaining their use. The detected lexical items and their most distinctive occurrences constitute the basis of the sociolinguistic survey presented in the next section.

7.4 Sociolinguistic survey

The preparation of the sociolinguistic survey begins by defining an interview protocol. To do so, I draw on a well-established method which includes standard tasks on a range of linguistic structures, especially focusing on pronunciation, and produces a detailed sociodemographic description. It has previously been adapted to contact situations, including in Quebec, which provides a point of comparison with earlier data. Adding to the standard protocol structure, I develop an acceptability rating task to investigate the use of the contact-induced semantic shifts output by the computational methods. It provides standard quantitative information on the perception of these items, spontaneous comments on their use, and comparable phonological evidence reflecting internal sociolinguistic factors. The protocol and the recruitment procedure are described in detail in Chapter 12; an overview of the sample, pointing to a range of bilingual profiles, is outlined in Chapter 13.

The quantitative analysis of the collected data is principally based on the acceptability ratings and the properties of the lexical items, analyzed against the backdrop of the speakers' sociodemographic characteristics, including reported and production-based estimates of bilingual ability. It provides clear indications as to the differences in status between different contact-induced semantic shifts, as they span the whole range of acceptability ratings and are associated with distinct factors. On the qualitative level, the use of semantic shifts is analyzed by reconstructing indexical fields based on the representations expressed during the interview. This analysis is discussed in Chapter 14.

Finally, these results – both quantitative and qualitative – are used as a new basis to evaluate the computational methods. This specifically concerns the performance of vector space mod-

els, with differences between variation metrics suggesting the need for qualitative profiling of lexical semantic change in addition to its detection. The validity of Twitter corpora as a sociolinguistic data source is also addressed, additionally drawing on the comments provided by the participants regarding their use of social media and its perceived relationship with face-to-face communication. These results are presented in Chapter 15.

A final note is due on the interdisciplinary ties holding this approach together, over and above the sociolinguistic phenomenon being described. It is in fact the case that, at numerous stages of this work, the imperatives related to one discipline have informed the methodological choices specific to the other. For instance, the construction of the Twitter corpus follows standard computational practice; however, (i) its regional structure directly reflects a comparative sociolinguistic view; (ii) specific data processing steps are included to improve descriptive reliability; and (iii) estimates of key speaker characteristics are introduced with the sociolinguistic objectives in mind. Similarly, as already mentioned, the results of the computational analyses are used to both formulate general hypotheses regarding the factors behind semantic shifts, as well as to identify individual cases that can be examined through the sociolinguistic interview. A final case in point is the interview structure itself, as the semantic shift task is specifically designed so as to provide readily usable data for further evaluation of computational methods. Taken together, these examples illustrate the underlying ambition of the approach outlined in this chapter: its aim is not to simply test an existing computational tool on a new task, but rather to address a real methodological need and, in doing so, provide mutually informed contributions to both disciplines.

Part III

Corpus-based analyses

The chapters in this part of the dissertation present the corpus-based analyses used to analyze contact-induced semantic shifts. Chapter 8 describes the creation of a large corpus of tweets allowing for a regional comparison of language use, and more precisely the identification of characteristics that are specific to Montreal and to that extent potentially related to language contact. Chapter 9 discusses two exploratory analyses of the collected data, respectively focusing on the detection of regionally specific lexical items and meanings. It provides initial evidence of the regional nature and comparability of the data. It moreover highlights methodological issues affecting the use of type-level vector space models in this context. Building on these observations, Chapter 10 aims to provide a more thorough understanding of the patterns captured by the data and the implemented models. In a series of experiments, it more clearly outlines the shortcomings related to type-level models, implements a multidimensional analysis to facilitate further exploration of the data, and introduces token-level vector space models as a potential way of accelerating this analysis. Finally, Chapter 11 more formally addresses some of the observed challenges. It introduces a test set for the detection of contact-induced semantic shifts, and uses it to systematically evaluate type-level models. The target set of lexical items is further analyzed using token-level models, providing an initial characterization of their use. This in turn constitutes the basis for the sociolinguistic inquiry presented in Part IV.

Chapter 8

Collecting tweets to investigate regional variation

As suggested by the methodological overview in Part II, the computational analyses conducted in this dissertation require a highly specific corpus. Some of the criteria are determined by the descriptive aim of investigating contact-induced semantic shifts, specifically by observing patterns of regional semasiological variation across Canada; other criteria are related to the methods used to identify these patterns. I argue that one possible solution to satisfying both types of requirements – the only solution that was readily available to me – consists in constructing a carefully designed and filtered corpus of tweets. This is the issue addressed by the present chapter.

The motivation behind this approach is clarified in Section 8.1, which contrasts the design criteria for this study with existing corpora. The adopted method of data collection is outlined in Section 8.2, and the filtering pipeline is discussed in Section 8.3. The structure of the resulting corpus is presented in Section 8.4. A brief summary concludes the discussion in Section 8.5. Note that this chapter is limited to the implementation of data collection and filtering adopted in this work. It builds upon the broader discussion of Twitter data in Chapter 4, which provides general background for the specific methodological decisions presented here.

8.1 Motivation for using Twitter data

Drawing on the main descriptive goals and methodological requirements, this section defines the design criteria for the corpus used in the computational analyses described in the following chapters. It then shows that these requirements are not fulfilled by any existing corpus containing English data from Canada, providing a strong case for the use of Twitter data.

8.1.1 Corpus design criteria

As stated above, the computational analyses conducted in this dissertation aim to identify patterns of semasiological variation across Canadian regions. Adopting a comparative sociolinguistic view, this approach specifically focuses on the patterns that differentiate Quebec from

Canadian provinces where French is not widely spoken, so as to capture behaviors that are potentially related to language contact. A range of methods are deployed in analyzing the data, but all of them involve some type of vector space representations (cf. Chapter 5). This methodological framework translates to the following corpus design requirements:

(1) the corpus should reflect the specifics of the English used in Canada, as opposed to corpora of other national varieties of English or more generic datasets;

(2) additional geographic metadata is necessary to compare different regional varieties of Canadian English: the province of origin of individual utterances in the corpus is required as a minimum;

(3) each regional subcorpus must meet a minimum size threshold of $\approx$ 100 million tokens in order for the proposed data processing methods to produce reliable results;

(4) the reliance of these methods on features such as co-occurrence frequencies entails the need to limit sources of bias such as an irregular distribution of content across authors or a pervasive presence of spam or other types of noise;

(5) sociolinguistic analysis of ongoing synchronic language variation requires data that are recent, largely contemporaneous, and produced in a reasonably spontaneous communicative context by individually traceable speakers;

(6) the identification of individual speakers should allow us to examine inter-speaker variation within the local community: a description of the languages the individuals speak is necessary at a minimum, given Canada's multilingual environment and my focus on language contact.

Computational studies of diachronic semantic change, which constitute the basis of the methodology adopted here, usually rely on large generic diachronic corpora of English. While these are readily available, that is not the case when it comes to fulfilling the criteria outlined above. Let us see more specifically how some of the available corpora fit this picture.

8.1.2 Existing corpora

Existing publicly available corpora of Canadian English are presented in Table 8.1.

Corpus	**Tokens**	**Geographic information**	
Strathy	50m	country	text metadata
GloWbE	134m	country	text metadata
iWeb	308m	country	website domain
NOW	898m	country	text metadata
ENCOW16	222m	city	website IP address
JSI	1.3b	city	place of publication

TABLE 8.1: Existing corpora containing Canadian English data, with the size of the Canadian section and the granularity and origin of geographic information. Corpus size corresponds to the best estimate at the time of completion of the Twitter corpus, i.e. November 2019.

The existing corpora include the Strathy Corpus of Canadian English (Strathy Language Unit, 2011), comprised of written and oral texts covering a variety of genres and historical periods, as well as the Canadian sections of multinational corpora such as Global Web-based English (GloWbE) (Davies, 2013a), News on the Web (NOW) (Davies, 2013b), and iWeb (Davies, 2018). However, these are all of limited utility in studies of regional variation, as the only provided geographic information is the country from which individual texts originate.

City-level geolocation is available in two large web-based corpora with Canadian content, but it is of questionable reliability. ENCOW16 (Schäfer and Bildhauer, 2012; Schäfer, 2015) derives geographic information from website IP addresses, meaning that it locates the servers hosting the websites rather than their users. In contrast, the JSI Newsfeed Corpus (Bušta et al., 2017) geotags online journalistic content based on its place of publication, but the solidity of this information is counterbalanced by considerable divergences in the amount of data originating from different Canadian regions.

Other key design criteria, such as the ability to identify all linguistic content produced by the same speaker, are not met by any of the 6 cited corpora. This, of course, is not the case in corpora of sociolinguistic interviews, which provide detailed descriptions of each individual speaker. However, they are generally not publicly available, in addition to being far too limited in size for large-scale computational analyses. To the best of my knowledge, the largest sociolinguistic corpus of Quebec English is the one introduced by Poplack et al. (2006); with a total of 2.8 million tokens across three regions, it remains well below the 100-million-token threshold suggested above.

Having established the lack of an adequate existing corpus, we now turn to the construction of a new dataset.

8.2 Data collection

As we have seen in Chapter 4, Twitter-based corpora are increasingly used to study differences in language use across large geographic areas, as well as other dimensions of variation. However, Twitter data can be accessed and filtered in different ways; this entails a range of methodological decisions with repercussions in terms of both the efficiency of the corpus construction process and the content retained in the resulting dataset. The choices made in this dissertation are aimed at finding a reasonable balance between efficiency (completing the corpus in months rather than years) and reliability (regionally representative and comparable data).

Similarly to some of the previous work on collecting geotagged Twitter data (Barbaresi, 2016; Ljubešić et al., 2014), I propose a data collection pipeline comprising two main steps: (i) an initial data collection which principally aims to identify Twitter users in geographic areas of interest; and (ii) a subsequent crawl of the indexed users' tweets. This is shown in Figure 8.1.

The first step was implemented by repeatedly querying Twitter's Search API in conjunction with geographic and linguistic filters (on the technical characteristics of the Search API, see Section 4.2.2.1). I used as search terms the 20,000 most frequent word bigrams in the 1-billion-word Corpus of Contemporary American English (COCA) (Davies, 2011). COCA is

composed of texts that are roughly equally distributed over 30 years (1990-2019) and 8 genres, ranging from academic to spoken language. While the most frequent bigrams in the list are sequences of function words (e.g. *of the*), the majority include content words in commonly occurring patterns (e.g. *they work*, *my car*, *interest in*). This approach is similar to the use of mid-frequency words to crawl web corpora (e.g. Baroni and Bernardini, 2004; Schäfer and Bildhauer, 2012), but like Scheffler (2014) I found that high-frequency search terms were more efficient on Twitter. As further discussed below, this stage allowed me to identify English-speaking users living in Toronto, Montreal, and Vancouver, and more generally to gain an initial insight into the gathered data.

The second step consisted in collecting all available tweets published by the initially indexed users. The aim was to increase the amount of available data while balancing the size of the regional subcorpora, as well as to obtain enough tweets published by individual users to analyze speaker-specific linguistic patterns. Tweets written in all languages were initially retained to enable a description of the overall linguistic profile of the corpus.

The collected data were then filtered by (i) verifying user profile locations to confirm that they reference one of the targeted cities; (ii) excluding tweets in languages other than English; and (iii) excluding near-duplicate tweets to limit the impact of repetitive or automatically generated messages.

The remainder of this section presents the main data collection steps in more detail. The implemented data filtering approaches are discussed in Section 8.3.

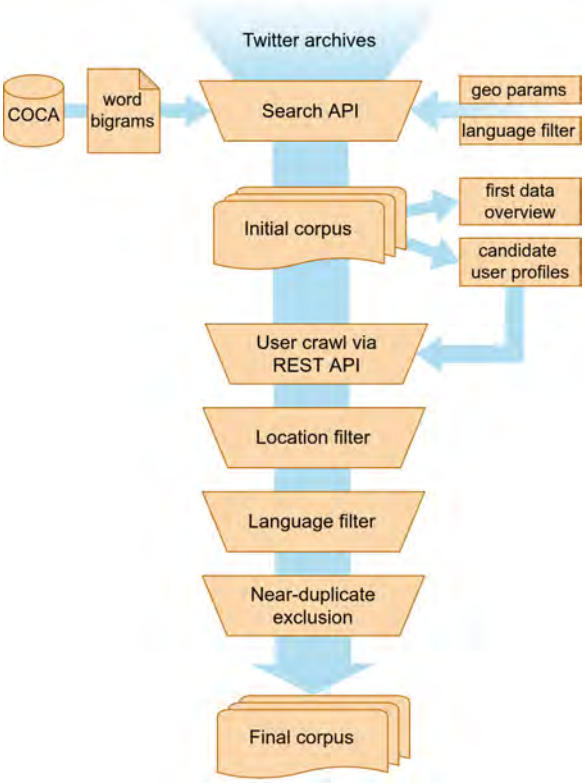


FIGURE 8.1: Data collection and filtering pipeline

8.2.1 Choice of geographic areas

Tweet collection was geographically limited to Canada's three most populous cities: Toronto (Ontario), Montreal (Quebec), and Vancouver (British Columbia). From a practical point of view, the choice of these cities was motivated by the need for a sufficiently large local user base allowing for the collection of enough data over a reasonably short period of time. Moreover, and more importantly, this geographic focus enables a study of Quebec English in a comparative perspective: the three cities belong to distinct lexical dialect regions, as is also evidenced by their demographic profile. Both issues are extensively discussed in Chapter 2; here, it will suffice to recall several key characteristics.

Montreal is home to 74.2% – or around 534,000 – of Quebec's mother-tongue English

speakers, but they represent only 13.2% of the city's population. A total of 91.4% of Montrealers report knowledge of French; the rate is lower but still high in the native Anglophone population (71%). Conversely, in Toronto and Vancouver the dominant language is English, with 8% and 7.1% of the population, respectively, reporting knowledge of French. Native Francophones constitute less than 2% of the population in both cases (Statistics Canada, 2017c).[1] In comparing these three cities, the aim is to detect contact-related phenomena as well as limit the impact of those deriving from unrelated regional variation. This is done by identifying the linguistic properties that are specific to Montreal and distinguish it from both Toronto and Vancouver.

Data collection was limited to tweets sent from the metropolitan areas of the three cities, all of which are highly multicultural. This means that the corpus may contain messages posted by non-native speakers of English. I experimented with creating corpora of smaller, more homogeneous communities within the three cities or the surrounding area (West Island of Montreal; Oshawa–Whitby, ON; Victoria, BC), but this led to a multifold decrease in collected data and was deemed too inefficient. Moreover, the wider geographic scope is coherent with the broad definition of linguistic communities adopted in this work, which extends to non-native speakers (see Chapter 2).

8.2.2 Initial tweet collection

An initial corpus was created using Twitter's Search API, which looks up queries in a sample of recently published tweets. The queries were filtered geographically by indicating the targeted areas as a radius around a point of interest, defined using geographic coordinates. Since this stage only aimed to identify English speakers, data collection was restricted to tweets tagged by Twitter as written in English. Moreover, search parameters were used to exclude retweets, i.e. reproductions of other users' posts, from the results. The diffusion of content posted by others may be indicative of the popularity of different subjects across regions, but my focus is on individual users' linguistic production rather than their topical interests.

As mentioned above, I queried the Search API using the 20,000 most frequent word bigrams from COCA. For each bigram in the list, all available tweets in the targeted geographic areas were collected. As a single iteration over the entire list took an average of 5 days due to the rate limits imposed by Twitter, iterations were repeated so as to move chronologically through Twitter's archives. By the time an iteration was completed, the temporal window of available tweets (6–9 days before the query) would also have shifted. The next iteration would then mostly return previously unavailable data.

A total of 50 iterations (i.e. a total of 100,000 queries) were completed between mid-January and mid-November 2019. The resulting corpus contains 58,451,998 distinct tweets published by 679,785 different users. As shown in Figure 8.2, 50.6% of users were identified in the first 5 iterations, but subsequent queries still provided a constant and non-negligible flow of new data. However, the number of collected tweets per user varies considerably (the top 1% of users

[1]The reported statistics refer to the mother tongue and knowledge of official languages, as per the 2016 Census for the corresponding Census Metropolitan Areas. These and other language variables used by Statistics Canada are discussed in Chapter 2. The counts reported here include bilingual speakers.

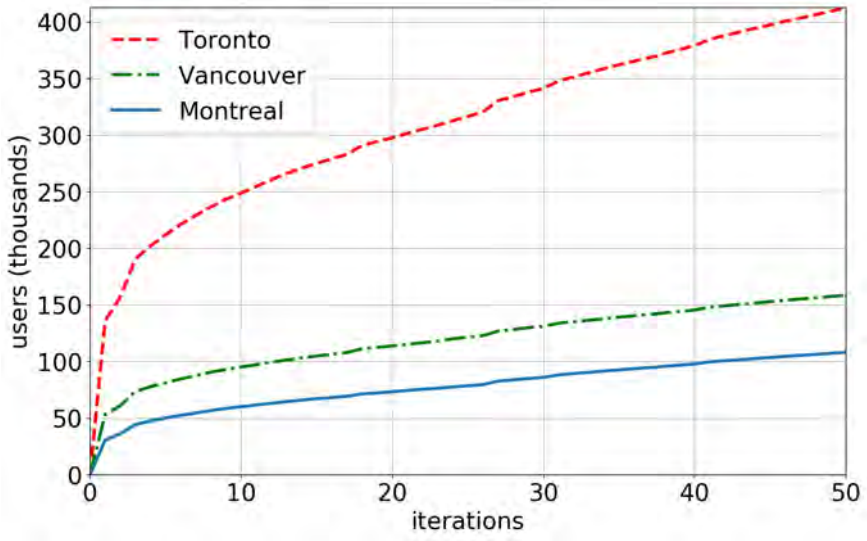


FIGURE 8.2: Cumulative number of identified users per subcorpus

account for 36.6% of tweets), as does the number of identified users across regions (108,383 in Montreal, 158,762 in Vancouver and 412,640 in Toronto). That said, this initial dataset is a valuable starting point for a more controlled user-level tweet collection.

The search method was chosen over the Streaming API (cf. Section 4.2.2.1), which returns a real-time sample of tweets, as it yielded considerably more data. For comparison, I ran the Streaming API for 30 days in October 2019 with comparable geographic parameters, obtaining 925,668 tweets published by 57,218 individual users. Over the same period of time, 6 iterations of the Search method were completed, yielding 8,332,629 tweets published by 303,538 users. In other words, the use of the Streaming API led to a roughly ninefold decrease in collected data and a fivefold decrease in identified users compared to the other approach.

As we have seen in Chapter 4, this is largely due to the fact that the Streaming API only takes into account tweet-level location data. They are only available on some devices and are actively used by a fraction of all users, which limits the availability of geotagged tweets. In this specific case, the problem is compounded by tight geographic constraints and a comparatively small number of targeted users (especially English-speaking Montrealers). An alternative solution was proposed for the German Twitter Snapshot (Scheffler, 2014), which collected tweets published in Germany by tracking words specific to German rather than applying geographic filtering. I could not implement this method, as only a fraction of all English-language tweets are posted in Canada.

As for the Search API, it maximizes the amount of data returned by geographic queries by interpreting non-geotagged tweets (93% of my initial dataset) as sent from the location indicated in the user profile. In previous work, corpora were created for three closely related languages with limited coverage in the Streaming API – Croatian, Serbian and Slovene – by querying the Search API using words specific to the targeted languages, without geographic parameters (Ljubešić et al., 2014). While this approach avoids issues related to the reliability of geolocation, it is not directly applicable to the present case. Lexical variants distinguishing dialect regions are both less numerous and less frequent than words differentiating distinct

languages, which would impact the efficiency of this method and would potentially introduce a bias towards speakers who use regionalisms more frequently.

Although both tweet-based and profile-based geolocation may introduce a demographic bias (Pavalanathan and Eisenstein, 2015b), the reliance on manually indicated user profile location is justified by a considerable increase in collected data as well as by its sociolinguistic significance: this location corresponds to the place users consciously associate with their online presence. Moreover, the types of tweets for which precise geolocation is enabled have repeatedly evolved, including in the course of the data collection described here.[2] This has the potential to affect data collection pipelines relying on this type of geographic information.

8.2.3 User profile crawling

Once individual users were identified, their entire timelines were crawled, subject to the Twitter-imposed limit of 3,200 most recent tweets per user (including retweets). In order for the final regional subcorpora to be comparable in size, the crawl included the 108,383 users indexed in the initial Montreal subcorpus, as well as the same number of randomly sampled users in each of the larger Toronto and Vancouver subcorpora. The crawl was performed for two batches of users, in April and November 2019, respectively.

In addition to excluding retweets based on Twitter metadata, common practice was followed in eliminating the messages that contain the "RT @" string in their text. This case, affecting 0.7% of collected tweets, corresponds to comments of other users' messages embedded into tweet text; as such, it has the potential to distort user-level word frequencies. Moreover, I only retained the tweets with at least 2 words in addition to any hashtags, user handles and URLs. While this led to the exclusion of 8.7% of collected tweets, it ensured that each retained tweet contained at least some linguistic content as opposed to being a list of Twitter-related entities. Unlike in the initial data collection, no language restrictions were sent to the Twitter API in order to allow for a subsequent analysis of the languages that are actively used by individual users. This stage of data collection was followed by a series of filtering steps; they are the focus of the next section.

8.3 Data filtering

In order to improve the reliability of the corpus, filtering was implemented in order to verify the location and language of the tweet, as well as to limit near-duplicate content.

8.3.1 Location filtering

Since the corpus should reflect the linguistic communities of Montreal, Toronto, and Vancouver, it was important to restrict data collection to the users who explicitly declare that they live in these cities. While the geographic parameters used with the Search API correspond to these

[2] https://twitter.com/TwitterSupport/status/1141039841993355264

areas, some users in the corpus may have been identified independently of their profile locations, based solely on individually geotagged tweets. Others still may have been retained even though multiple cities are indicated in their profile.

I therefore used a heuristic to additionally filter the places indicated in the `location` field in the user profile. In order for a user to be retained, the field was required to include the name of the examined city (e.g. *Montreal*). Accepted additional information included the name of the corresponding province (e.g. *Quebec*), the name of the country (e.g. *Canada*), and generic geographic descriptors (e.g. *north, greater, metro* etc.). No other elements were accepted.

In the Montreal subcorpus, profile locations were indicated in 7,719 distinct ways (after being lowercased and stripped of punctuation and diacritics). Of these, 46 met the above criteria and were used by 69% of the identified users. The individual realizations differed in terms of the order and precision of included information (*Montreal* vs. *Montreal West, Quebec*), orthographic choices (*Montreal* vs. *Montréal*), use of abbreviations (*Quebec* vs. *QC*) and punctuation. Out of the 7,673 rejected locations, 6,872 (used by 22% of users) indicated multiple targeted cities (*Montreal & Toronto*), places outside of the search area (*Ottawa*) or insufficient geographic information (*Canada*). The remaining 801 locations (used by 9% of users) referred to neighborhoods (*Plateau Mont-Royal*) or points of interest (*McGill University*) in the search area, but were excluded due to the presence of lexical items which were too specific to incorporate in the filtering heuristic. Based on the number of classified users, the Montreal subcorpus heuristic obtained an F-score of 0.94. Comparable patterns were also observed in the Toronto and Vancouver subcorpora.

8.3.2 Language identification

As previously mentioned, the populations of Montreal, Toronto, and Vancouver are all highly multilingual. While the initial data collection parameters ensure that the identified users have sent at least one tweet tagged as English, crawling their entire timelines provides a clearer picture of the languages they actually use. The distribution of language tags outlined in Table 8.2 shows that English is by far the most frequent language in the corpus, but, in addition to the expected use of French in Montreal, immigrant languages are also present. Since I only aim to investigate regional differences affecting English, tweets tagged as written in other languages (15.5% overall) were excluded.

Montreal		**Toronto**		**Vancouver**	
en	69.7%	en	93.4%	en	92.4%
fr	22.6%	es	1.2%	es	1.6%
es	2.3%	tl	.8%	pt	1.1%
pt	.7%	pt	.7%	tl	.9%
ar	.6%	fr	.6%	fr	.6%
other	4.1%	other	3.3%	other	3.5%
total	100.0%	total	100.0%	total	100.0%

TABLE 8.2: Distribution of tweets across the top language tags (components may not sum to totals due to rounding)

The decision to use Twitter-provided language tags was preceded by an evaluation of third-party systems on a manually annotated sample of 494 monolingual English tweets and 420 monolingual French tweets, grouped into balanced categories with 2, 5, 10, 15 or 20 words per tweet. The focus on English in French is related to the fact that, in addition to being Canada's two official languages and the center of my research objectives, they correspond to the most frequent language tags in the corpus. I tested three widely used off-the-shelf language identification systems – `langid.py` (Lui and Baldwin, 2012), `cld2` (McCandless, 2014) and `langdetect` (Nakatani, 2010) – as well as a majority-vote system combining the three methods, proposed in an earlier evaluation (Lui and Baldwin, 2014). The results in Table 8.3 show that all systems are consistently reliable except on very short tweets. As expected, the vote-based system performs on par with or improves on the best individual F-scores.

System	Words per tweet					
	2	**5**	**10**	**15**	**20**	**all**
langid	.822	.964	.989	**.994**	**1.000**	.963
langdetect	.896	.917	.989	.989	**1.000**	.963
cld2	.793	.898	.971	.967	**1.000**	.935
vote	**.902**	**.976**	**.994**	**.994**	1.000	**.979**

TABLE 8.3: Macro-averaged F-score on manually annotated English and French tweets of different lengths

The performance of the evaluated systems was further compared to the language tags indicated in tweet metadata. While the quantitative results cannot be reported because Twitter's developer policy[3] prohibits the benchmarking of their services, they suggest that it is not necessary to implement a third-party language identification system in the filtering pipeline. The systems I evaluated on English and French occasionally provide marginal improvements compared to Twitter's tags, but their performance is overall less consistent.

But the use of Twitter's language tags raises another potential issue. Practices such as borrowing and codeswitching are frequent among bilingual speakers, meaning that multiple languages may be used in a tweet, whereas only one language tag is indicated in the metadata. This problem was evaluated on a balanced sample of 1,000 tweets tagged by Twitter as English or French. I manually identified other-language content in 65 tweets: 60 written in these two languages, and 5 written in English or French and another language. Note that most identified tweets (56 out of 65) were tagged as French.

An attempt was made to automatically identify the languages in the 65 multilingual tweets using the top two predictions produced by each of the tested language identification methods. A majority vote system was also implemented based on the two most frequent language tags from the individual predictions. The best accuracy was obtained by `langdetect`, which correctly analyzed 25% of tweets.

Given the relative rarity of other-language items and the poor performance of the tested language identification systems, multilingual content filtering has not been implemented. Word-level language identification may provide more precise results and is a possible direction of

[3] https://developer.twitter.com/en/developer-terms/agreement-and-policy

future work.

8.3.3 Near-duplicate exclusion

A frequent issue in Twitter-based corpora is the presence of near-duplicate messages generated by both automated spam accounts and prolific human users. Attempts are usually made to filter out this content as it can bias word frequencies. A common approach, presented more extensively in Chapter 4, consists in excluding accounts that exceed defined cut-off points in terms of the number of tweets, followers, followees etc., or in excluding all tweets containing URLs or other specific strings. These methodological decisions are based on the potential link between these user account features and spam production (Yardi et al., 2010).

But solutions such as these do not take into account the fact that user behavior on Twitter is often heterogeneous. To explore this trend, I manually analyzed the 20 users in the corpus with the highest number of posted tweets in their profiles. Out of these, seven accounts did exclusively publish near-duplicate content such as song titles played by radio stations, while another two posted a mix of similarly generated tweets and spontaneous messages. However, the remaining 11 accounts were all consistent with genuine human communication.

As two of these were corporate Twitter profiles where different social media managers interact with the public, I focused on the nine accounts used by individual speakers. To varying extents, they all produced genuine tweets as well as ones that were automatically generated by, for example, posting content on other social media sites. In some cases, the high number of published tweets was actually driven by retweets, while the content of original posts was similar to that of average accounts. Moreover, while some tweets containing URLs simply referenced external content (e.g. titles of linked videos), others included fully acceptable messages.

Taking into account this variety of behaviors, I implemented a system whose aim is not to exclude all tweets posted by the users most likely to produce spam, but rather to distinguish, within the production of each individual user, the tweets that are of genuine interest from near-duplicate content. For each user, a distance matrix was calculated for all their tweets. I used Levenshtein's distance, which quantifies the difference between two strings of characters as the number of edit operations (character insertions, deletions or substitutions) necessary to modify one string of characters into the other.

As my aim was to exclude messages with similar linguistic content independently of Twitter-specific entities, I removed hashtags, user handles and URLs from tweet text. In calculating the absolute Levenshtein's distance, replacement operations were assigned a weight of 2 in order for the distance between entirely different strings of characters to be equal to the sum of their lengths. This distance was then normalized by dividing it with the total number of characters in a pair of tweets. A normalized score of 0 corresponds to identical strings, and a score of 1 to strings with no overlapping characters.

After calculating the distance matrix, near-duplicate tweets were identified using hierarchical clustering. I excluded all clusters where the distance between individual tweets did not exceed 0.45. This cut-off point was determined empirically; an important assumption was that any accidental loss of non-repetitive data would be outweighed by the benefits of cleaner, less

repetitive content. Moreover, the exclusions produced by this method are related to structural similarity rather than, say, specific topics, so they are not expected to negatively affect co-occurrence statistics. While the identification of near-duplicates published by different users may further improve the quality of the data, it is computationally prohibitively expensive with the current method.

8.4 Corpus description

This section presents the corpus produced using the data collection and filtering pipeline described so far. It specifically discusses the structure of the corpus, an estimate of user-level linguistic characteristics, and access to the collected data.

8.4.1 Corpus content

The corpus obtained after crawling individual user profiles, performing language and location filtering and excluding near-duplicate content contains 78.8 million tweets posted by 196,431 individual users. Following the initial step of data collection, 325,000 Twitter profiles were crawled across the three cities. Of these, nearly 11,000 were inaccessible at the time of the crawl because they had been deleted or had become private a short time after their initial identification. While this is a tolerable loss of data (3.2% of accounts), the efficiency in other similar pipelines could be improved by crawling individual user profiles as soon as they are identified by the Search API. More significantly, 118,000 accounts (36.2%) were excluded based on their profile location. Out of the 132 million tweets retained after the user-level geographic filtering, 15.5% were rejected because they were not written in English and a further 24.7% were excluded as near-duplicates. The implemented filters led to a considerable reduction in corpus size, but they ensure the reliability of collected data.

The corpus was tokenized and POS tagged using `twokenize` (Gimpel et al., 2011; Owoputi et al., 2013), which was specifically developed to take into account the specifics of Twitter-based communication, including emojis, URLs, ambiguous tokens, and so forth. The data were then lemmatized using the NLTK WordNet lemmatizer (Bird et al., 2009). After these steps, it contains 1.3 billion tokens. On average, 401 tweets were collected per user; the top 1% of users account for only 6.2% of tweets, in a considerable improvement compared the initial stage of data collection. The data are roughly equally distributed across the three regional subcorpora.

The structure of the final corpus is presented in Table 8.4. Token counts were limited to the metadata-indicated display text range, i.e. tweet text stripped of tweet-initial user handles referring to conversation chains and of tweet-final URLs mostly used to embed media. This represents 97.7% of analyzed text content. No further removal of Twitter-related entities was performed, as they are often syntactically integrated in the tweet text and can also provide insights into bilingual communication (e.g. hashtags used in a language different from the rest of the tweet).

As shown in Table 8.4, an additional, smaller version of the corpus was also created. With the amount of collected data well above the minimum threshold defined at the outset, this step

	Base corpus			Subsampled corpus		
Subcorpus	**Users**	**Tweets**	**Tokens**	**Users**	**Tweets**	**Tokens**
Montreal	72,305	23,469,526	384,740,451	54,726	11,318,184	193,228,246
Toronto	64,164	28,442,928	481,126,844	51,245	12,465,659	222,508,471
Vancouver	59,962	26,924,158	473,322,674	47,697	11,381,080	213,200,523
Total	196,431	78,836,612	1,339,189,969	153,668	35,164,923	628,937,240

TABLE 8.4: Corpus structure

aimed to further limit potential sources of bias in word frequency. Specifically, I removed the content posted before 2016 in order to reduce any short-term diachronic effects. I then excluded the users with fewer than 10 tweets in the corpus. A maximum of 1,000 tweets per user were retained, with random subsampling performed where this was exceeded. These decisions were respectively aimed at reducing the impact of potentially aberrant tweeting behaviors, as well as that of a limited number of highly active individuals. An average of 229 tweets were retained per user, with the top 1% of users accounting for 4% of tweets.

While the base corpus remains an important, larger source of information on the use of Canadian English on Twitter, the additional precautions taken in creating the subsampled version arguably make it better suited for comparative analyses of regional variation. Consequently, it is the subsampled version that is used in the computational experiments presented in the following chapters.

8.4.2 User-level linguistic profiles

Twitter-based corpora provide limited user-level information, especially compared to the kind of detailed descriptions that are obtained through sociolinguistic interviews. However, it remains possible to infer basic speaker characteristics based on the metadata provided by Twitter. Given the focus on language contact, a key descriptor here is the linguistic profile exhibited by individual users: it is essential to understand whether a user is monolingual or bilingual, as well as which languages they speak.

A simple metric was implemented in order to approximate this information. Before the exclusion of non-English-language content from the corpus, the users (N = 196,431) were analyzed according to the languages they use on Twitter. For each user, I computed the proportion of English language tweets (out of all English and French tweets) and the proportion of tweets in English and French (out of all tweets). The distribution of users according to these scores is shown in Figure 8.3.

This distribution suggests that the data collection I implemented identified predominantly English-speaking individuals, as well as some demonstrably bilingual speakers. As expected, the use of French appears to be more frequent in the data collected in Montreal compared to the other two cities, whereas the use of non-official languages (i.e. languages other than English and French) is roughly comparable across the subcorpora.

While these observations serve to validate the basic assumptions behind the corpus con-

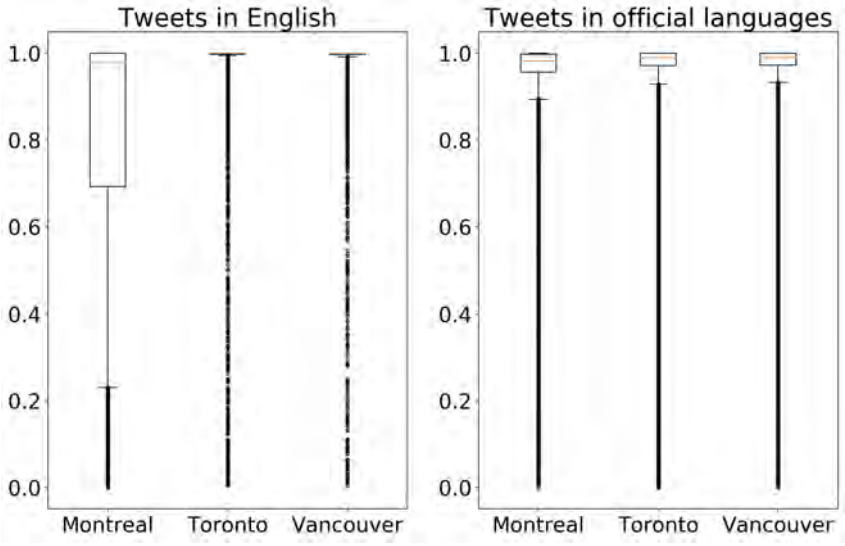


FIGURE 8.3: Left: proportion of tweets in English per user (out of tweets in English and/or French). Right: proportion of tweets in English and French per user (out of tweets in all languages). Results based on language tags produced by Twitter prior to the exclusion of non-English content. N = 72,305 (Montreal), 64,164 (Toronto), 59,962 (Vancouver).

struction, it must be emphasized that the scores only provide a very general idea of language choice on Twitter. They are far simpler than the measures of bilingualism that I implemented in sociolinguistic interviews, presented in Chapter 12. It is also unclear to what extent real-life language choices are reflected specifically by these measures, and more generally by behaviors observed on Twitter; this issue is explored in Chapter 15. That said, these scores are transparent, simple to compute, and empirically grounded, and as such are suitable for large-scale analyses conducted in the coming chapters.

8.4.3 Data distribution

The base corpus has been released as a list of tweet IDs together with instructions on how to collect the complete data using off-the-shelf software.[4] Publishing tweet content is not allowed due to restrictions under Twitter's Developer Policy. Their stated aim is to allow users to retain control over their data, for example by deleting tweets or restricting access to their profile. As discussed in Chapter 4, this limitation affects the reproductibility of Twitter-based research, as published datasets decay relatively quickly due to changes in the availability of original tweets. However, I am unaware of any clear solutions to this issue.

8.5 Summary

This chapter has presented the construction of a large Twitter-based corpus that will be used as the basis for computational analyses of regional semasiological variation aiming to identify patterns specific to Quebec English. I began by outlining a series of precise corpus design criteria, arising from the intersection of descriptive objectives and methodological requirements. Given

[4] http://redac.univ-tlse2.fr/corpora/canen.html

the lack of an existing corpus which could fulfill these criteria, I introduced a series of steps to collect and filter Twitter data, aiming to strike a balance between efficiency and reliability.

Specifically, an initial data collection identified a set of users from Montreal, Toronto, and Vancouver who had previously tweeted in English. In order to address the skew in the amount of data across regions and users, as well as extend the amount of user-level information, this was followed by a crawl of individual Twitter profiles. The collected data were then filtered for location and language, and near-duplicates were automatically excluded. This led to a non-negligible decrease in the amount of available data, but, as I have claimed, this is justified by the improved descriptive utility of the dataset.

The 11-month research effort described in this chapter resulted in a corpus containing 1.3 billion tokens, corresponding to 78.8 million tweets posted by 196,000 users. It meets the initially defined design criteria: it is sufficiently large for data-intensive modeling methods as well as fine-grained user-level analysis, and it maintains a reasonably balanced distribution of reliable content across regions and users. Moreover, it mirrors both national and regional specificities of Canadian English, as I will show in the next chapter.

Chapter 9

An exploratory overview of regional variation

In the previous chapter, I introduced the corpus of tweets that was used in the computational analyses conducted in this dissertation. As we have seen, the data collection procedure and the resulting structure of the corpus – with a geographic distinction between Montreal, Toronto, and Vancouver – are based on the assumption that linguistic behaviors which distinguish Montreal from the two other cities may be related to the influence of contact with French. We now turn to two experiments which constitute the first step towards verifying this claim, with a particular focus on regional specificity and comparability. Their aim is not to address the central research questions in a definitive way, but rather to help understand the global patterns in the data and in that way provide a sound basis for more systematic analyses.

Section 9.1 presents an experiment on the unsupervised detection of lexical items that are the most specific to Montreal, providing insights into different types of variation. The focus is restricted to the lexical semantic level in Section 9.2, which explores the applicability of a standard embedding-based method to regional semasiological variation. Practical considerations in dealing with Twitter data are outlined in Section 9.3, and a brief summary is presented in Section 9.4.

9.1 Unsupervised detection of regionally specific lexical items

The first analysis conducted on the corpus aims to identify the lexical items which are the most specific to Montreal, i.e. which are overrepresented in that subcorpus compared to the Toronto and Vancouver subcorpora. This will provide a basic understanding of the types of variation that can be captured through the regional structure of the corpus. By looking at the corresponding most underrepresented items, i.e. those that are comparatively more frequent in Toronto and Vancouver, we can further understand whether the three regional subcorpora are comparable in terms of topic and other general characteristics.

9.1.1 Experimental setup

The most distinctive regional lexical variants in the corpus were identified using the publicly available Python implementation[1] of the Sparse Additive Generative model (SAGE) (Eisenstein et al., 2011). SAGE estimates the deviation in log-frequencies of terms in a corpus of interest relative to their log-frequencies in a background corpus using the maximum-likelihood criterion, with a regularization parameter ensuring that rare terms are not overemphasized. A high value of the deviation estimate indicates that a term is overrepresented in a given corpus, and a low value that it is underrepresented. As noted in Chapter 5, this approach has been extensively used to identify lexical items related to variation in terms of a range of factors reflected by the structure of a given corpus; here, the focus is on regional variation.

SAGE was used to compare the frequency of lowercased POS-tagged lemmas in each regional subcorpus against that observed in the entire corpus of tweets. The deviation estimates were used to extract the 400 most overrepresented items per subcorpus and the 400 most underrepresented items per subcorpus. The items tagged as proper nouns, foreign words, numerals, punctuation, and Twitter entities (hashtags, username handles, URLs) were not included in the results due to their limited interest for a study of regional lexical variation (as opposed to topical effects, for instance). I experimented with the number of included lexical items: the analysis was run two times, limited first to the 10,000 and then to the 20,000 most frequent items from each subcorpus. The outcome partly overlapped; in order to retain a more varied and comprehensive list of candidate items, the two individual lists were collapsed.

The remainder of the analysis focused on the most overrepresented items, with those that are underrepresented mainly providing interpretative context. The list of lexical items was manually inspected. I excluded those whose use appeared to be driven by a limited number of highly prolific users[2] as well as those without clear links with regional variation (i.e. items for which the inspection of co-occurrence contexts did not indicate evident differences across the regional subcorpora). The remaining 264 cases were then more closely examined; based on qualitative observation, they were grouped into different categories of lexical phenomena in order to illustrate the key patterns of variation present in the corpus. On the range of information taken into account in analyzing corpus occurrences, see Section 9.3.

9.1.2 Analyzing the captured types of variation

An overview of the proposed categories of lexical phenomena and representative examples is provided in Table 9.1. They are discussed in more detail below.

Borrowings. The analysis highlights a number of previously described lexical items of French origin, confirming the presence of contact-related language use in the corpus. In an echo of the

[1] https://github.com/jacobeisenstein/SAGE

[2] An item was excluded if its most prolific user (i.e. the one with the highest number of occurrences) accounted for more than 10% of its occurrences in the subcorpus, or if the coefficient of variation (ratio of the standard deviation to the mean) for the number of occurrences per user exceeded 200%. Although arbitrary, this heuristic simplified the subsequent qualitative analysis by limiting the number of cases which were perceptibly affected by individual users.

Category	Examples
borrowings	*cegep*, *chalet*, *dep*, *metro*, *poutine*
semantic shifts	*café*, *encore*, *entourage*, *supper*, *specially*
French items	*bah*, *bonjour*, *du*, *le*, *merci*
spelling variants	*center*, *color*, *im*, *theater*, *week-end*
chatspeak features	*fkn*, *lolll*, *ouf*, *oups*, *tmr*
local referents	*blizzard*, *drouin*, *habs*, *kk*, *snowstorm*

TABLE 9.1: Categories of lexical items specific to the Montreal subcorpus

earlier discussion of the Quebec English lexicon (see Chapter 2), these items fit into different categories of lexical influence outlined by Boberg (2012). The list includes instances of imposed direct lexical transfer, corresponding to the choice of a French term influenced by official use, such as the previously discussed cases of *cégep* 'junior college' and *metro* 'subway'. The list also points to examples of elective direct lexical transfer, where a French borrowing is used despite the existence of a readily available English-language alternative, as in the case of *dep* 'convenience store' and *chalet* '(summer) cottage'. Importantly, alternative variants for items from both categories – e.g. *subway* and *cottage* – feature among the top underrepresented items in the Montreal subcorpus, providing support for the regional representativeness and comparability of the data. This is illustrated in the examples[3] below, the first posted in Montreal and the second in Toronto:

(3) Oh it's one of those days where I forget to get off the **metro** at my stop. I see.

(4) Made it to work for 8am with coffee in hand despite the unappealing conditions and a **subway** delay. A small victory.

Semantic shifts. The analysis also picks up examples which fall under the general notion of semantic shift, as defined in Chapter 3. For example, the adverb *specially* usually expresses the meaning 'for a special purpose', whereas *especially* is used to signify 'above all' or 'to a great extent', particularly in formal contexts (Lindberg, 2012, p. 295). In French, all of these senses are expressed by the adverb *spécialement*, which may explain the higher relative frequency of *specially* in Montreal. The traditionally proscribed usage is attested in tweets like this one:

(5) It's hard to move on, **specially** when you don't want to...

Another interesting case is that of *supper* 'evening meal'. As noted in Chapter 2, this variant has been described in studies of lexical variation from an onomasiological perspective, which have shown that it is usually preferred to its equivalent *dinner* in rural Canada, but is also more frequent in Montreal than in other cities (Boberg, 2010, p. 181). More recently, it has increased in frequency among younger speakers elsewhere in Quebec; this suggests that it is undergoing

[3]The examples from the Twitter corpus will be limited to their textual content. In order to limit the impact on the privacy of the tweet authors, no metadata or personally identifiable information will be included. User names will be replaced by the "<username>" token (except for public figures, where relevant for context). The lexical item under discussion will be displayed in bold. If the tweet contains content in French, the translation will be provided in italics.

diffusion, possibly due to the similarity of the QF equivalent *souper* (Boberg and Hotton, 2015, p. 297). The fact that it exhibits a higher frequency in the Montreal subcorpus is coherent with this previously reported trend. From a semasiological standpoint, however, this example represents a more marginal case of contact-related influence, which appears to be limited to a higher rate of use of an already existing sense.

French function words. The presence of French function words in the Montreal subcorpus is indicative of codeswitching, a typical feature of language contact. Recall that the corpus only contains tweets that were originally tagged by Twitter as written in English, with the evaluation discussed in Chapter 8 confirming that this system is largely reliable. Bearing this in mind, occasional slip-ups in language identification might occur, but other-language content should be overall limited, and used together with English elements.

This is confirmed by the manually inspected examples. Items such as the French definite article *le* and the preposition *de* 'of' are systematically attested within larger spans of French text, usually in tweets that also contain some English. However, the precise codeswitching patterns vary, as shown by the following examples:

(6) **On devrait juste interdire les commentaires.** That's it. Then again, no more FB or Twitter...
They should just forbid comments. *That's it. Then again, no more FB or Twitter...*

(7) My calendar here in Montreal reads **"premier jour du printemps"** which I've come to learn is properly translated from the Quebec dialect of French as "still winter".
My calendar here in Montreal reads ***"first day of spring"*** *which I've come to learn is properly translated from the Quebec dialect of French as "still winter".*

(8) Hi there, guys! We always appreciate the support. You're the best! **Merci!** #GoHabsGo
Hi there, guys! We always appreciate the support. You're the best! ***Thanks!*** *#GoHabsGo*

In example (6), the user produced a complete sentence in French and then switched to English for the remainder of the tweet; the switch was possibly triggered by the use of the fixed expression *that's it*. In example (7), the author reported a single French phrase, originally seen in that language, in a tweet otherwise written in English. As for example (8), it was sent in response to a message from the United States, so the isolated use of *merci* can be seen as an expression of local identity. These and other observed patterns show that codeswitching is deployed on Twitter in a variety of ways which are reminiscent of its use in face-to-face communication (cf. Chapter 1). This provides further support for the presence of bilingual speakers on Twitter and the descriptive interest of the data they produce.

Spelling variants. The current spelling conventions in Canadian English represent a compromise: British spelling is used in words such as *colour* (vs. *color*) and *centre* (vs. *center*), and American spelling in cases like *program* (vs. *programme*) and *optimize* (vs. *optimise*) (Boberg,

2010, p. 40). Despite general trends, these patterns continue to be variably realized. This is also reflected by the analysis conducted here, with Montreal exhibiting a preference for the typically American forms such as *color* and *theater*. However, the results also include the form *realise* as well as *optimize*, suggesting a less clear-cut opposition in this case.

A related issue is the higher prevalence of apostrophe dropping in Montreal (e.g. *im*, *wouldnt* rather than *I'm*, *wouldn't*). Similarly, the Montreal subcorpus presents a more variable orthography of compound nouns, reflected by nonstandard forms such as *week-end* (vs. *weekend*) and *bestfriend* (vs. *best friend*). While *week-end* is a clear transposition of French spelling in English-language tweets, the link with contact is less evident in other cases.

Chatspeak features. The most regionally specific items include nonstandard abbreviations and other informal orthographic variants typical of online communication. It is not immediately clear why some of the identified examples are more frequent in Montreal (e.g. *fkn* 'fucking', *tmr* 'tomorrow'), but a relatively straightforward link with language contact can be postulated in other cases. These include the exclamation *ouf*, directly borrowed from French, as well as the adapted variant *oof*, which is also overrepresented in Montreal. Both forms roughly correspond to *phew* or *ugh*, depending on context. Similarly, *oups* is the French orthographic variant of the existing English exclamation *oops*.

Expressions of laughter, frequently examined in studies of computer-mediated communication (Tagliamonte and Denis, 2008; Tagliamonte, 2016), also present regional usage patterns in this analysis. The variant *lol* 'laughing out loud' is typically realized in Toronto with an orthographic lengthening of the vowel (e.g. *lool*, *loool*), whereas the forms salient for Montreal emphasize the final consonant (e.g. *loll*, *lolll*). This is consistent with the pattern typical of the corresponding French initialism *mdr* (e.g. *mdrr*, *mdrrr*), based on the expression *mort de rire*, literally 'dead of laughter'. Like in the case of spelling variants, these examples illustrate an added value of Twitter data in enabling an analysis of medium-specific variation phenomena.

Local referents. Some lexical items are more frequent in Montreal than elsewhere because of the importance that their referents have in the city. This is the case of locally familiar proper nouns, such as those that refer to sports teams (e.g. *habs* 'Habs', the nickname of the Montreal Canadiens hockey team) or their players (e.g. *drouin* 'Jonathan Drouin'; *kk* 'Jesperi Kotkaniemi'). This category also includes terms linked to isolated events, such as *blizzard* and *snowstorm*: their frequency in the Montreal subcorpus spiked on 20 January and 13 February 2019, at the outset of two severe winter storms in Quebec. These items are reflective of topical differences rather than underlying lexical variation, but their presence is significant because it confirms the regional nature of the data.

These initial observations show that the Twitter corpus captures key types of contact-related influence in Quebec English. They also point to other types of regional regularities, including those related to orthographic variation and nonstandard forms typical of online communication. Let us now take a closer look at these medium-specific phenomena: this will allow us to better understand how generalizable the initial observations are, as well as provide an opportunity to

further explore this type of variation.

9.1.3 Extending the analysis: Twitter-specific usage

This step focuses on a subset of the initially observed phenomena: spelling variation, apostrophe dropping, and abbreviations. The SAGE analysis only provided individual lexical items specific to a given subcorpus; in contrast, sociolinguistic analyses rely on the notion of linguistic variable (cf. Chapter 9) to the relative preference for sets of functionally equivalent variants, such as *endeavor* and *endeavour*. In order to obtain this descriptively vital information, I drew on the general overview of patterns provided by the SAGE analysis to construct a series of linguistic variables. For each variable, I extracted the frequency of POS-tagged lemmas for all the variants in the three subcorpora, retaining the variants with a minimum of 5 occurrences in at least one subcorpus. Moreover, I only analyzed the variables where all subcorpora presented a minimum of 5 occurrences of at least one individual variant.

The chi-square test was used to examine the relation between the variants and the subcorpora; for statistically significant results ($p < .05$), Pearson's residuals were computed to evaluate individual contributions to the obtained chi-square value. Note that this analysis is not taken to represent a definitive confirmation of regional trends; rather, it is used as an aid in exploring the tendencies attested in the corpus. Corresponding hypotheses require further validation in another dataset.

9.1.3.1 Spelling variation

The VarCon database[4] of English spelling variants was used to define 220 linguistic variables reflecting unstable spelling in Canadian English. The breakdown of the results in Table 9.2 suggests that realizations vary both within and across the three targeted spelling patterns. This heterogeneity is consistent with the fact that, despite standardized recommendations, spelling choices in CanE depend on individual words and vary across regions (Pratt, 1993).

Pattern		Example		N	*
our	or	*colour*	*color*	72	43
tre	ter	*theatre*	*theater*	15	4
ise	ize	*realise*	*realize*	133	59
total				220	106

TABLE 9.2: Spelling patterns, with the number of analyzed and statistically significant variables

The distinction between *our* and *or* is characterized by an overarching preference for American spelling in Montreal (62 out of 72 variables). For the 43 variables yielding a significant chi-square test, Pearson's residuals further confirm that the American variants, in *or*, are most strongly associated with the Montreal subcorpus in all but three cases. For instance, the spelling *neighborhood* is used in 52% of occurrences in Montreal vs. 24% and 25%, respectively, in

[4] http://wordlist.aspell.net/varcon/

Toronto and Vancouver. This trend may reflect a diachronic lag with regard to other Canadian English regions, where spelling in *or* largely fell out of use in the 1980s (Dollinger, 2010).

A similar tendency characterizes the *tre* vs. *ter* category, where the traditionally American form in *ter* has the highest relative frequency in Montreal for all but three variables. Like before, Pearson's residuals indicate that the differences between the regions are consistently driven by usage in Montreal: for instance, *center* is used in 50% of cases in Montreal, compared to 24% in both Toronto and Vancouver.

The choice between *ise* and *ize* follows a different trend, with the mean preference for the American variant at or above 90% in all subcorpora. However, for the 59 statistically significant variables it is the British spelling, in *ise*, that exceeds the expected frequency. It is typically associated with the Vancouver subcorpus (51 variables, e.g. *recognise* in 9% of occurrences in Vancouver vs. 4% and 3%, respectively, in Montreal and Toronto). Association with Montreal is considerably less frequent (8 variables, e.g. *utilise* in 12% of cases in Montreal vs. 2% and 5%, respectively, in Toronto and Vancouver).

9.1.3.2 Apostrophe dropping

Another set of variables was generated in order to investigate the use of verbal contractions without apostrophe (e.g. *im* instead of *I'm*). This included the combinations of subject pronouns and the clitic forms of the auxiliary verbs *be*, *will*, and *have*, as well as the appending of the negative particle *n't* to auxiliary and modal verbs. Ten variables were excluded due to homography occurring once the apostrophe is removed (e.g. *he'll* > *hell*), and another six due to low frequency (e.g. *oughtn't*). For the remaining 30 variables, apostrophe dropping is a minority behavior: it occurs on average in 6.7% of cases in Montreal, 4.8% in Toronto and 4.3% in Vancouver. However, the higher propensity for apostrophe dropping in the Montreal subcorpus is systematic; this finding is illustrated for several items in Figure 9.1.

With no evident crosslinguistic explanation for this regional distinction, an extralinguistic factor may have come into play. Gouws et al. (2011) have shown that graphemic variation, including apostrophe dropping, varies according to the device used to access Twitter. In the Montreal subcorpus, some tweets may have been posted by bilingual speakers using a French smartphone keyboard; this would have disabled English autocorrect settings, leaving irregular contractions apostrophe-less. But the impact of social factors should not be discounted either. For instance, Squires (2007) found that men were significantly more likely to drop apostrophes than women, echoing the generally established tendency for men to use more nonstandard linguistic forms. The differences between the three regional subcorpora are not expected to reflect gender (although the metadata included in the corpus do not provide a way of verifying this claim). They may however be indicative of the use and perception of nonstandard features in the linguistic communities to which they correspond.

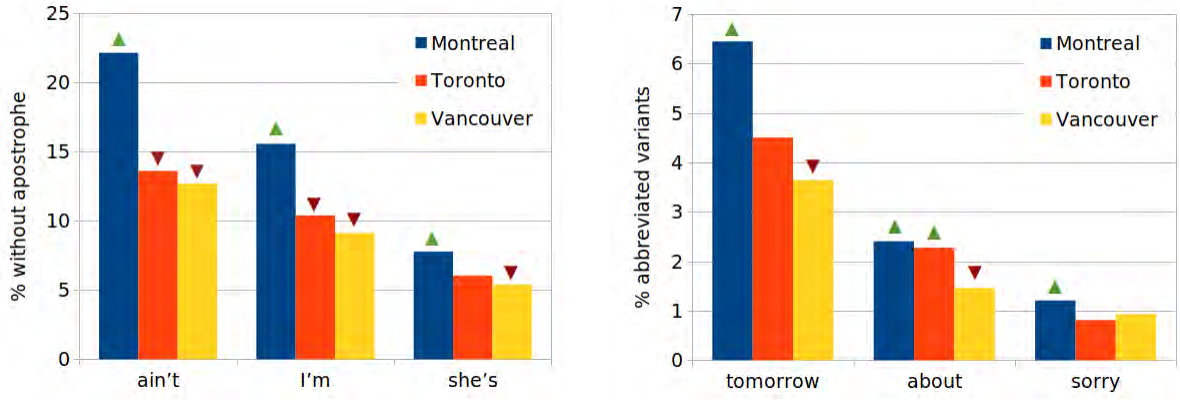


FIGURE 9.1: Proportion of realizations without apostrophe (left) or as an abbreviation (right), for selected variables. Positive and negative Pearson's residuals (absolute value > 2) are respectively indicated by ▲ and ▼.

9.1.3.3 Abbreviated variants

The final area of focus concerns abbreviated variants typical of online communication. This analysis was based on the UTDallas lexical normalization dictionary,[5] which contains 3,800 nonstandard tokens from tweets, normalized by human annotators (Liu et al., 2011, 2012). I extracted the normalized items presenting any of the following patterns in at least one nonstandard variant: deletion of three or more characters (e.g. *birthday* > *bday*); absence of vowels (e.g. *please* > *pls*); use of numerals (e.g. *before* > *b4*); word-final substitution of *er* with *a* or *ah* (e.g. *brother* > *brotha*). For each item identified in this way, all corresponding nonstandard variants were retained. The initial list was manually narrowed down to 35 variables, comprised of at least one standard form and one informal abbreviation (e.g. *sorry* vs. *sry* / *srry* / *soz*).

The propensity to use abbreviated forms, plotted for representative examples in Figure 9.1, exhibits considerable variability. It is the strongest for the term *introduction*, with the variant *intro* used in 64% of occurrences in Montreal, 56% in Toronto and 58% in Vancouver. At the other extreme, the forms *nvr* and *nevr* occur as an alternative of *never* in 0.06% of cases in Montreal, 0.04% in Toronto and 0.03% in Vancouver. This contrast may however be related to the fact that, in addition to the propensity to abbreviate, the regional subcorpora also differ in the choice of some of the abbreviated variants. Take for example the abbreviations *abt* and *bout*, referring to the preposition *about*: the form *abt* presents a lower relative frequency in Toronto (used in 30% of abbreviations) than in Montreal and Vancouver (36% and 35%, respectively).

Whether language contact can be linked to these regional differences remains unclear, particularly given the potential effect of patterns of interaction. Previous studies on computer-mediated communication have associated the use of regional lexical variants, including nonstandard realizations, with interlocutors from the same area (Pavalanathan and Eisenstein, 2015a) as well as a higher degree of local dialect vitality (Peersman et al., 2016). A similar effect may be at play in Montreal, as it is the most lexically distinct Canadian region (cf. Chapter 2).

The additional exploration of linguistic features typical of Twitter has confirmed that the

[5] http://www.hlt.utdallas.edu/~yangl/data/Text_Norm_Data_Release_Fei_Liu/

regional specificity of the lexical items output by SAGE is underpinned by regular patterns of variation. Statistical testing was used to guide the qualitative analysis rather than provide a definitive explanation for regional variation on Twitter. I have suggested several hypotheses for the observed patterns, including the conservative nature of Montreal as a dialect region, extralinguistic and social factors, and patterns of interaction. These claims should however be validated on a different dataset. Although the potential influence of language contact is not immediately clear for most of the examined variables, this analysis has further confirmed the comparability of the subcorpora, the presence of clear regional trends, and the relevance of well-established types of variation for Twitter data. We now turn to a more direct exploration of semantic shifts.

9.2 Regional variation in vector space representations

The first computational experiment on contact-induced semantic shifts aims to provide an initial assessment of the feasibility of embedding-based analyses of contact-induced semantic shifts. I used a previously developed method to conduct a bottom-up, vocabulary-level analysis in order to identify the lexical items which present the largest regional differences in meaning as reflected by vector representations. The aim was not to determine the optimal implementation or provide a detailed description, but to explore the usefulness and shortcomings of vector space models when applied to the corpus that I constructed and the research objectives that I pursue.

9.2.1 Experimental setup

The setup adopted in this experiment drew heavily on approaches used in computational studies of diachronic semantic change, discussed more extensively in Chapter 5. In the diachronic version, a typical method involves training a vector space model for each section of the corpus corresponding to a different time period, and then comparing vector representations over time. In keeping with this logic, I trained a model for each regional subcorpus, under the previously stated assumption that regionally specific vector representations may be used to identify semantic patterns that distinguish Montreal from both Toronto and Vancouver, and in that way point to potential contact-induced semantic shifts.

Specifically, I used the `gensim` implementation (Řehůřek and Sojka, 2010) of `word2vec` (Mikolov et al., 2013) to train a type-level vector space model for each regional subcorpus. The skip-gram architecture was used with default parameters: the window size was set to 5, the vector dimensions to 100, the negative sampling rate to 5, the subsampling rate to 10^{-3}, and the number of iterations to 5. The model was trained on POS-tagged lemmas, with minimum word frequency set to 100.

Recall that `word2vec` produces a low-dimensional vector space, in which vector dimensions are not directly interpretable. This complicates comparisons across models, as their vector dimensions are not naturally aligned to the same coordinate axis and hence are not immediately comparable. Following existing work on diachronic word embeddings (Hamilton et al., 2016b), I addressed this issue by aligning the models using Orhtogonal Procrustes, available in a Python

implementation.[6] As previously noted, this approach aims to reduce the distance between the corresponding vectors of all word in a pair of models. The distance that nevertheless persists between pairs of vectors is then taken to reflect actual differences in their representations rather than the effect of misaligned vector spaces.

Following this step, I quantitatively compared each word's vectors in all pairs of models in order to detect the most prominent divergences in Montreal. Given a word w, I measured the cosine distances (CD) between the word's vectors in the Montreal ($\vec{w}_m$), Toronto ($\vec{w}_t$), and Vancouver ($\vec{w}_v$) subcorpora. While these distances can be used to estimate variation in different ways, my aim was to prioritize the words whose meaning is different in Montreal, but varies as little as possible between Toronto and Vancouver. I therefore computed a variation score as follows:

$$var(w) = \frac{CD(\vec{w}_m, \vec{w}_t) + CD(\vec{w}_m, \vec{w}_v)}{CD(\vec{w}_m, \vec{w}_t) + CD(\vec{w}_m, \vec{w}_v) + CD(\vec{w}_t, \vec{w}_v)}$$

Cosine distance is defined as $(1 - cosine\ similarity)$, so it theoretically ranges from 0 (identical vectors) to 2 (diametrically opposed vectors). The maximum theoretical value of the variation score is 1, if the Toronto and Vancouver vectors are identical, but the Montreal vector is different. The larger the distance between the Toronto and Vancouver vectors, the smaller the resulting variation score (given the same distance from the Montreal vector).

9.2.2 True positives: a qualitative analysis

An inspection of the list of top semantic shift candidates, i.e. lexical items with the highest variation score, confirmed that this approach could successfully identify the presence of contact-induced meanings. Findings reported in previous studies are reflected by examples such as *exposition* 'exhibition' (Fee, 1991, p. 14) and *terrace* 'restaurant patio' (Fee, 2008, pp. 179–180). Newly identified cases such as *definitively* 'definitely' present comparable contact-related influence: the unconventional meanings are all likely related to phonologically and semantically similar French lexical items (*exposition*, *terrasse* and *définitivement*, respectively).

In order to more precisely understand how these differences in meaning are captured by vector space models, let us take a closer look at the case of *exposition*. One way to examine the meaning represented by a vector space model is to inspect a word's nearest neighbors, i.e. the words whose vectors are the closest to that of the target word. By definition, nearest neighbors have similar cooccurrence patterns. Broadly speaking, they are also semantically related; specific relations at play include synonymy and hyponymy. For ease of interpretation, the nearest neighbors for *exposition* are plotted in a two-dimensional space in Figure 9.2. The method used is the t-distributed stochastic neighbor embedding (t-SNE) in the `scikit-learn` (Pedregosa et al., 2011) implementation. The vectors are represented so that those that are similar in the original high-dimensional space are shown close together in the two-dimensional space. While the distances between the vectors should not be interpreted literally, the global patterns are expected to be reliable.

[6] https://github.com/williamleif/histwords

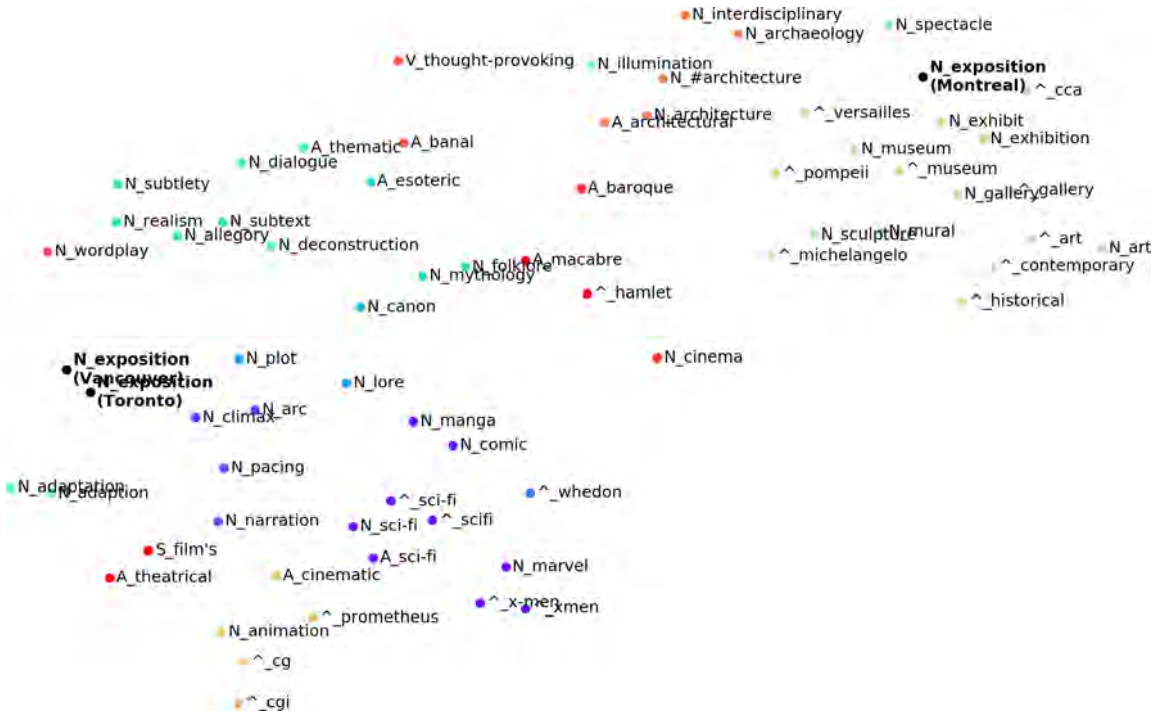


FIGURE 9.2: Two-dimensional (t-SNE) projection of the vectors for *exposition* and their nearest neighbors. The target word vectors, represented in bold, were extracted from the Montreal, Toronto and Vancouver vector spaces. A union of the top 20 nearest neighbors in the three models was then defined; vectors for these words were extracted from the Montreal vector space. This projection can be taken to illustrate the target word's meaning in the three regions from the viewpoint of the Montreal model. POS tags: `N`: common noun; `V`: verb; `A`: adjective; `S`: possessive form; `^`: proper noun.

The plot shows the Montreal vector for *exposition* as being distant from the Toronto and Vancouver vectors, which are close to one another. This is indicative of a word whose meaning is similar in Toronto and Vancouver, but different in Montreal, and therefore potentially influenced by French. Indeed, the Montreal vector's neighbors are mostly related to art (*gallery*, *sculpture*), suggesting a meaning similar to that of *exhibition*. In Toronto and Vancouver, the word seems to refer to the opening section in a work of fiction (cf. *narration*, *plot*).

This global overview provides the neat idea of one meaning being used in Toronto and Vancouver, and another in Montreal, the latter influenced by French; in a word, it corresponds to a prototypical contact-induced semantic shift. But if we go back to the data in the corpus, the situation appears to be much more nuanced. Consider the following examples, all taken from the Montreal subcorpus:

(9) I really want to go to an art museum or an art **exposition** :(

(10) On parle de notre **exposition** Brown's Hill! // An article about our **exhibition** Brown's Hill

(11) Canada's centennial year saw Montreal host the 1967 International and Universal **Exposition** (or Expo 67)

(12) Three straight scenes of clunky dialogue filling in for **exposition**. Yup, it's a Schwarzenegger film!

(13) A brilliant **exposition** of dietary fiber & the wonders it can perform for human health.

In tweet (9), *exposition* is used to refer to a public display of artwork. Onomasiologically speaking, the English word *exhibition* would be expected in this context in English. This case can be interpreted as influenced by the French word *exposition* 'art exhibition'. However, in example (10), *exposition* is attested with precisely the same meaning, but in French, in a codeswitched tweet that also includes the English term *exhibition*. The occurrence in tweet (11) denotes a large public exhibition of trade goods; the distinction from the previous sense is admittedly subtle, but it is noted in lexicographic sources. Example (12) refers to an opening section of a film, and example (13) to a comprehensive explanation of an idea.

Overall, the situation appears to be far more complex than suggested by the nearest neighbor analysis. The meaning initially proposed for Montreal is only one of the attested senses; those from examples 11–13 are also present in the Toronto and Vancouver subcorpora. The vector representations are not entirely invalidated, since a regional difference does exist, with the contact-related sense limited to Montreal. However, this suggests that type-level vectors such as those produced by `word2vec` do not equally capture the different senses of a word, but strongly emphasize only one of them, likely the most frequent. Limiting the analysis to nearest neighbors is therefore insufficient. It is vital to closely examine the data on which models are trained; otherwise, crucial phenomena – codeswitching, homography, polysemy – can be misinterpreted.

In addition to cases such as *exposition*, which despite their complexity present a descriptive interest for this dissertation, other types of lexical items were also picked up by the vector space models. They are addressed in the next section.

9.2.3 False positives: distinguishing types of noise

Many of the top semantic shift candidates do not present an apparent link with language contact. They are nevertheless subject to regional differences in use, which explains the fact that they were detected by the models. However, they do not fall under the definition of contact-induced semantic shifts adopted in this dissertation; to that extent, they constitute false positives. They are described in more detail below.

Local referents. Regional prevalence of some uses is related to locally specific referents. For example, *plateau* frequently denotes the neighborhood of Plateau-Mont-Royal in Montreal, as opposed to the conventional sense of 'an area of fairly level high ground' or 'a state of little or no change following a period of activity or progress'. The first sense is not related to language contact, but to the fact that the neighborhood is located – and therefore more frequently discussed – in Montreal. (Note that this is a proper noun, but it is tagged as a common noun in the Twitter corpus presumably due to the use of determinants, as in *the Plateau*.)

Topical variation. Regional differences in the use of some senses are related to topical variation, which is in turn explained by cultural and social factors. That is the case of *unsupervised*, which principally refers to childcare in Toronto and Vancouver, and to machine learning in

Montreal. The latter meaning has little to do with the influence of French; it is instead related to Montreal's thriving IT industry. Interestingly, the same reason could explain why some previously described semantic shifts go undetected, such as *animator* 'group leader' (cf. Fr. *animateur*). Described since at least McArthur (1989, p. 53), the contact-related sense is occasionally attested, but most occurrences in Montreal refer to animated films or video games. This is likely due to the fact that the city is a global center for the animation industry.

Prolific users. A single, highly productive Twitter user can strongly influence the entire semantic representation of a word. For instance, the verb *waffle* typically means 'speak or write at length' or 'be undecided', but in the Montreal data it is overwhelmingly used to signify 'make waffles'. This is due to a single user account that is entirely dedicated to waffle recipes. Given the controlled number of tweets per user and the relatively high minimum word frequency of 100, this issue likely affects the words that are infrequent and are repeatedly used by an account focusing on a single topic.

French homographs. In some cases, high variation scores are driven by the presence of French homographs of English words. For example, the form corresponding to the English verb *pour* is widely attested in the Montreal corpus as the French preposition 'for', usually in codeswitched tweets. As a result, its cooccurrents are very different from those in the Toronto and Vancouver subcorpora, leading to strong divergences in vector representations. And while the presence of codeswitching in the final corpus could be seen as a shortcoming of the filtering pipeline presented in Chapter 8, it only affects a minority of tweets, as I have already noted. This is supported by the fact that the influence of homography tends to be limited to the forms that correspond to function words – and hence have a high frequency – in French.

Misspellings. A related issue is the presence of misspellings, such as *trough* 'through', which are reflective of an imperfect command of English. Their prevalence in the Montreal data is likely explained by a higher degree of bilingual and non-native speakers of English. These cases arguably represent typos rather than the introduction of a new sense due to the influence of language contact.

Given the number and variety of false positives, the use of vector space representations to detect *only* contact-induced semantic shifts – as defined in this dissertation – appears to be a highly complex task. The problem is further exacerbated by the fact that type-level vector representations obscure some of the complexity in the data, as shown in the discussion of true positives. A more precise understanding of the mechanisms operating both in the models and in the data is therefore necessary; this is the focus of Chapter 10. The exploration of the initial results also brought to light practical considerations in dealing with Twitter data; they are briefly addressed below.

9.3 On the linguistic analysis of Twitter data

In the previous section, I suggested that the tendencies highlighted by large-scale quantitative analyses were often underpinned by more complex patterns in the data. As intuitive as this claim may seem, it is worth discussing it in more detail from a practical standpoint.

The importance of this trend became fully apparent when I began exploring the output of the methods presented in this chapter. Both SAGE and embedding-based analyses produce somewhat cryptic results, which consist in a list of lexical items sorted based on a relatively abstract score. While explanations for some results – such as *metro* or *exposition* – might be immediately intuitive, for others that was not the case. This led to a detailed manual inspection of the underlying occurrences in the corpus, which turned out to be invaluable: it was only at this point that the full sense inventory of a lexical item, its prevalent regional use, and its connotational features became clear.

This process in turn raised issues specific to the use of Twitter-based corpora. First, tweets are limited in length to 280 characters, meaning that they often provide insufficient linguistic context to fully understand the use of a given lexical item. As a result, it was often necessary to seek additional information, including by contextualizing a tweet with respect to the original interaction (if it was part of a thread), taking into account the users' linguistic or geographic profile, and contextualizing an individual occurrence with respect to the user's remaining tweets. Although this analysis was strictly limited to publicly available content, the additional information obtained in this way was not stored so as to limit the impact on the users' privacy. It was however crucial in refining the linguistic interpretation of the observed patterns.

Furthermore, since I am not a member of the linguistic community under study, some lexical items were challenging to interpret even in the context of the original tweets. This led to time-consuming research into locally specific abbreviations, obliquely referenced events, and a wide variety of cultural referents. These include local bands, sports teams and their players, cafés and restaurants, neighborhoods, as well as the social issues affecting the three cities included in the corpus. These steps might appear to be beyond the scope of a linguistic description, but I would argue that they are in fact necessary to ensure the reliability of the reported observations. The issues described in this section more generally show that, even with the help of computational tools, the linguist's analytical skills remain vital for a comprehensive interpretation of the observed results.

9.4 Summary

This chapter has presented two exploratory corpus-based experiments, whose aim was to assess the presence of regional trends in the corpus and the feasibility of vector-based analyses of contact-induced semantic shifts. The first experiment confirmed that the Montreal subcorpus is characterized by contact-induced features, such as borrowings and semantic shifts, as well as other typically bilingual behaviors such as codeswitching. It also brought to light extensive variation in the use of medium-specific features such as informal abbreviations, whose use appears to follow regional regularities and involve a range of explanatory factors. These

observations confirm the regional specificity and comparability of the collected data, as well as the relevance of Twitter-based corpora in studying language variation.

In the second analysis, I implemented a method developed in diachronic semantic change detection, showing that it can be used to capture regional semasiological patterns in synchrony. It highlighted both previously described and newly identified contact-induced semantic shifts, confirming the interest of the approach for this dissertation. But a range of potential issues also emerged. On the one hand, even when Montreal-specific usage presents clear links with language contact, the underlying patterns are often far more complex. They often involve the presence of conventional senses in Montreal, as well as that of contact-related senses in the other two cities, but to different extents. On the other hand, not all instances of regional semasiological variation are related to language contact, with different types of noise obscuring the output of the analysis.

Taken together, these results point to a sociolinguistic phenomenon which cannot be neatly captured using readily available methods, at least not without introducing numerous methodological adjustments with often unclear indications as to the best practices. As previously stated (Chapter 5), computational studies on diachronic data – on which my implementation of vector space models is based – often adopt a very broad view of semantic change; by contrast, a sociolinguistic description of synchronic semantic variation requires a more precise analysis of coexisting linguistic variants (in this case, different senses), with the variant of interest potentially appearing in a very limited number of occurrences. The choice of vector representations, and particularly their ability to capture different senses, is therefore crucial. It also seems that the hypothesis underpinning the proposed methodological design – according to which the linguistic patterns distinguishing Montreal from both Toronto and Vancouver are likely related to language contact – does not hold in a strong version, since many other sources of variation are also captured by the corpus. However, it might be possible to characterize these types of noise and to reduce their impact on the overall analysis. I have also discussed manual corpus exploration; it will continue to play an important role, particularly because there are no existing benchmarks for contact-induced semantic shifts in Quebec English. That said, the existing sociolinguistic descriptions can help assess semantic shift candidates as they are attested in the corpus, but this means that facilitating manual data exploration should also be considered. I turn to these issues in the next chapter.

Chapter 10

Towards a better understanding of variation in the models and the data

The initial experiment on the use of vector space models for semantic shift detection has shown that this method is promising, as suggested by its ability to detect relevant examples, but it also poses methodological challenges, reflected by various types of noise in the results. In order to implement this method more efficiently with a descriptive objective, it is important to understand the precise mechanisms that are at play, particularly when comparing representations across different models. It is also important to assess whether this type of semantic information interacts with other data-driven measures, as well as whether any alternative approaches might help address the shortcomings of the initial method. This is the focus of the present chapter.

The performance of type-level models is addressed in Section 10.1, which compares a range of model configurations, complementing the regionally-specific experimental setup with a control condition. A multidimensional analysis, presented in Section 10.2, is then used to explore the contribution of different types of linguistic information, and further circumscribe the characteristics of the lexical items that are likely to present a descriptive interest. Section 10.3 introduces token-level vector representations, used to produce a finer-grained analysis of previously identified items of interest. Section 10.4 provides a brief summary.

10.1 Variation and instability in type-level models

The first experiment in this chapter examines type-level representations, focusing on the specific mechanisms that underpin – and may potentially skew – comparisons of vector representations obtained from different models. In particular, it investigates the impact of model architectures and hyperparameters, which can affect the resulting vector representations and, consequently, the methods that incorporate them.

At the time when these experiments were conducted, the one available systematic evaluation of semantic change detection methods (Schlechtweg et al., 2019) indicated that some methodological choices systematically yielded better results, but it also highlighted differences depending on the task at hand (synchronic semantic variation across text types vs. diachronic semantic change). Given the specifics of the work conducted in this dissertation – the focus on

synchronic regional variation, the indirect modeling of cross-linguistic influence – it appears important to better understand model performance in this particular setup.

10.1.1 Experimental setup

This section presents the corpora, model configurations, and semantic variation measures used in the experiment.

Corpora. The vector space models were trained using the corpus presented in Chapter 8. In addition, a shuffled corpus was also created for this experiment. It simulates a control condition, in which we would not expect to observe any semantic differences related to regional variation or idiolectal preferences, and which should help detect the noise affecting the semantic change detection methods.

The shuffled corpus was created by splitting up the regionally-specific tweets between three new subcorpora. Specifically, for each user, I iterated over their chronologically ordered tweets. The iterations were done in batches of three tweets; for each batch, each of the shuffled subcorpora was randomly assigned one of the tweets. This ensured that all users were represented across the shuffled subcorpora to a near-identical extent. The structure of the experimental and control condition corpora is presented in Table 10.1.

Subcorpus	Users	Tweets	Tokens
Montreal	54,726	11,318,184	193,228,246
Toronto	51,245	12,465,659	222,508,471
Vancouver	47,697	11,381,080	213,200,523
Total	153,668	35,164,923	628,937,240

Subcorpus	Users	Tweets	Tokens
A	153,668	11,721,641	209,665,214
B	153,668	11,721,641	209,612,901
C	153,668	11,721,641	209,659,125
Total	153,668	35,164,923	628,937,240

TABLE 10.1: Structure of the experimental corpus (left) and the control condition corpus (right)

The shared vocabulary (minimum frequency = 100 occurrences per subcorpus) contains 35,814 POS-tagged lemmas for the experimental condition, and 44,373 for the control condition. This difference points to the presence of regionally-specific lexical items which do not exceed the minimum frequency threshold in all three regional subcorpora.

Model configurations. A total of 18 model configurations were examined by varying the methodological choices summarized in Table 10.2. A more detailed overview of model architectures and the associated parameters is presented in Chapter 5; the discussion below will focus on the specifics of this experiment.

Category	Values
Method	SGNS, PPMI
Alignment	AL, SR
Vector dimensions	100, 300 (SGNS only)
Window size	2, 5, 10

TABLE 10.2: Tested model configurations

In terms of methods, SGNS refers to `word2vec` models trained using the skip-gram algorithm with negative sampling; for the hyperparameters with which I did not experiment, default values were used (the negative sampling rate was set to 5, the subsampling rate to 10^{-3}, and the number of iterations to 5). PPMI refers to count-based models weighted using positive pointwise mutual information. I experimented with different window sizes (the number of words occurring around the target word that is taken into account by the model). For SGNS models, I also experimented with different vector dimensions; this is not applicable to PPMI models, where the number of dimensions corresponds to the size of the vocabulary.

In terms of model alignment, two main approaches were tested. AL refers to training three separate models, one per subcorpus, and then aligning them in a subsequent step. For SGNS models, this was done using the previously described Orthogonal Procrustes approach (see Section 5.2.2.1 for a general discussion and Section 9.2.1 for the implementation used in this dissertation). For PPMI models, the adopted solution, known as column intersection, consists in retaining only those dimensions (i.e. context words) which are shared across the models. They are then ordered in the same way for each model, thereby ensuring the direct comparability of the vectors. This solution draws on the interpretability of vector dimensions in count-based models and as such cannot be applied to SGNS models.

In the second alignment approach, a single model was trained using the entire corpus. Target words were tagged so as to be specific to the subcorpus in which they appeared. For instance, the lemma *N_exposition*, already POS-tagged as a noun, was respectively modified as *m_N_exposition*, *t_N_exposition*, and *v_N_exposition* in the Montreal, Toronto, and Vancouver subcorpora. Context words were the same across the subcorpora (in this case, *N_exposition*, independently of the subcorpus). As a result, the meaning representations are specific to each subcorpus, but they share the same vector space and are directly comparable. This corresponds to the Temporal Referencing method, which was introduced in diachronic studies with the aim of limiting noise in model alignment (Dubossarsky et al., 2019). For clarity, I will refer to this method as Spatial Referencing (SR) given the focus of this dissertation.

For SGNS models, three runs of each configuration were performed in order to investigate the instability specific to this method (cf. Pierrejean and Tanguy, 2018). AL SGNS models were trained using `gensim` (Řehůřek and Sojka, 2010), SR SGNS models were trained using `word2vecf` (Levy and Goldberg, 2014a), and PPMI models were trained using DISSECT (Dinu et al., 2013). The Orthogonal Procrustes alignment was based on a Python implementation of the method introduced by Hamilton et al. (2016b).[1]

Measuring semantic variation. Following common practice, the difference between vector representations was measured using the cosine distance (CD). Specifically, pairwise cosine distances were computed for each word in the shared vocabulary between all pairs of subcorpora (Montreal-Toronto, Montreal-Vancouver, Toronto-Vancouver, for the experimental condition; A-B, A-C, B-C, for the control condition).

In addition, the experimental condition requires a derived measure to identify the meanings specific to Montreal. Three such measures were tested, starting with the following:

[1] `https://github.com/williamleif/histwords`

$$avg(w) = \frac{CD(\vec{w}_m, \vec{w}_t) + CD(\vec{w}_m, \vec{w}_v)}{2}$$

where the word w is represented by its vectors corresponding to the Montreal ($\vec{w}_m$), Toronto ($\vec{w}_t$), and Vancouver ($\vec{w}_v$) subcorpora. In simple terms, this measure corresponds to the mean of the Montreal-Toronto and Montreal-Vancouver cosine distances. It was further used to compute two other measures:

$$diff(w) = avg(w) - CD(\vec{w}_t, \vec{w}_v) \qquad ratio(w) = \frac{avg(w)}{CD(\vec{w}_t, \vec{w}_v)}$$

These measures correspond to the difference and the ratio, respectively, between the mean Montreal cosine distance and the Toronto–Vancouver cosine distance.

The first measure is the simplest, as it does not account for the variation between the two control areas. To this extent, it is also the closest to the standard approach in diachronic studies, which consists in measuring the cosine distance between vector representations corresponding to different time periods. The remaining two measures attempt to incorporate the differences relative to the control areas similarly to the initially devised variation score from Chapter 9, but they are more easily interpretable.

The analysis of semantic variation was limited to the words tagged as belonging to the open classes (common nouns, verbs, adjectives, and adverbs).

10.1.2 Variation in the control condition

We begin by examining the control condition – models trained on shuffled corpora – in order to assess the stability of vector representations in this setup, with their use expected to be largely the same across the subcorpora. In addition, this step establishes word-level instability scores, which will be used in subsequent analyses of regional variation.

For each word in the vocabulary (N = 44,373; minimum frequency = 100 occurrences per subcorpus), I computed the cosine distances for all pairs of vectors corresponding to the three shuffled corpora (AB, AC, and BC). I then computed the mean of the three pairwise cosine distances for each word. The distribution of these values is plotted in Figure 10.1 for all model configurations.

The plot indicates that the vector representations for a given word often differ somewhat even across shuffled corpora. For SGNS models, the median pairwise cosine distance stands at around 0.2, with outliers reaching up to 0.7. The values for PPMI models are considerably higher, but this may be due to inherent differences between the two methods (in particular, significantly more dimensions for PPMI vectors) rather than poorer overall performance.

The mean pairwise cosine distances are generally strongly correlated *across* different configurations, with Spearman's rho ranging from 0.684 to 0.972. This suggests that some semantic representations exhibit a degree of instability independently of the method used to create the model. In addition, it is also interesting to look at the three individual pairwise cosine distances (corresponding to subcorpus pairs AB, AC, and BC). They are very strongly correlated *within*

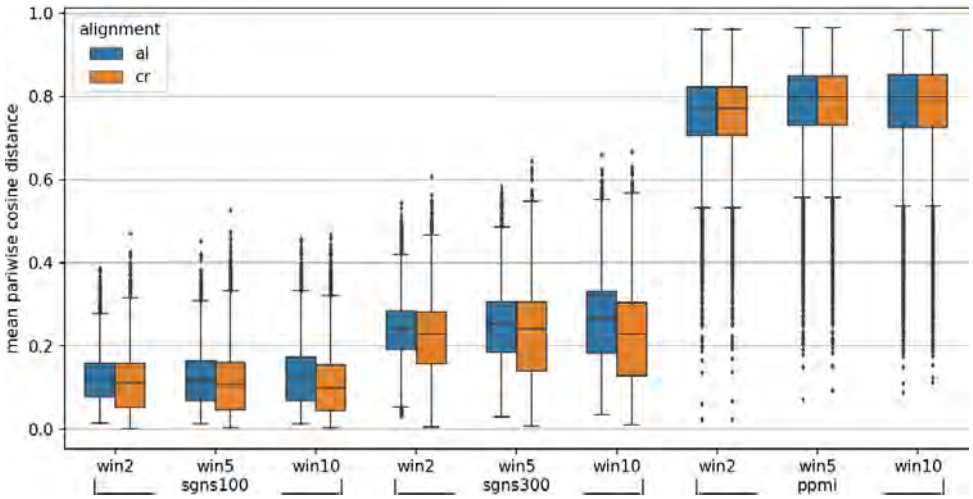


FIGURE 10.1: Distribution of mean pairwise cosine distances for models trained on shuffled corpora

each of the model configurations, with Spearman's rho ranging from 0.904 to 0.975.[2] This suggests that, for a given model configuration, the stability or instability of words remains globally similar across the three pairs of subcorpora.

In connection with this issue, the impact of word frequency on cosine distance has been noted in previous studies of diachronic semantic change (Dubossarsky et al., 2017). The correlation between frequency and the mean pairwise cosine distances is plotted in Figure 10.2. As noted in previous reports, there is a strong negative correlation between the two (i.e. less frequent words tend to be more unstable).

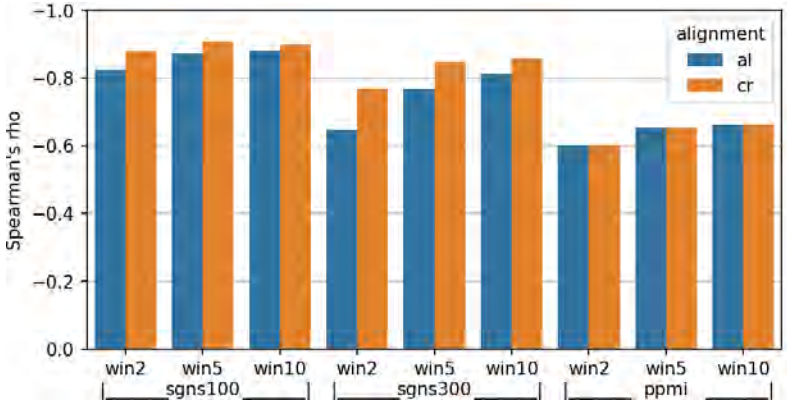


FIGURE 10.2: Correlation between mean pairwise cosine distances and word frequency

The impact of frequency may also be related to another tendency. In contrast to vocabulary-level trends, the subset of words that are the most stable or unstable varies considerably depending on the configuration. I extracted the 100 words with the lowest and the highest mean pairwise cosine distances for each configuration. On average, around half of the items in any given two lists overlap. Taking into account all pairwise comparisons of model configurations, the number of different items out of 100 ranges from 0 to 95 (median = 57) for the most stable words, and from 0 to 90 (median = 65) for the most unstable words. This suggests that the most

[2]The reported values correspond to the mean of the three correlation coefficients (AB-AC, AB-BC, and AC-BC) computed for each model configuration.

extreme values of common variation measures are not entirely reliable, which has important implications for bottom-up detection of instances of variation. (Large overlaps mainly occur when comparing PPMI models, with few differences between AL and SR implementations.)

In summary, this experiment has shown that (i) a word's semantic representations often differ even across shuffled corpora, with a clear tendency for an increase in vector dimensions to be paralleled by larger differences; (ii) pairwise cosine distances – reflecting a word's instability – are strongly correlated both within and across model configurations, suggesting largely consistent vocabulary-level trends; (iii) instability scores are strongly negatively correlated with frequency; (iv) the words that are the most stable or unstable vary considerably depending on the configuration, pointing to variability at the extremes of the vocabulary. All of these issues may have implications for semantic comparisons across regional subcorpora, which are addressed in the next section.

10.1.3 Regional variation

I now turn to variation in regionally specific models, and focus on the three previously introduced measures – `avg`, `diff`, and `ratio` – aimed at identifying the words whose meaning is the most different in Montreal. Like in the previous section, I will address vocabulary-level trends, the impact of frequency, and the words with the highest variation scores.

To begin, let us take a look at the patterns exhibited by the three variation measures *across* different model configurations (i.e. models which differ in terms of alignment, window size, and – for SGNS models – vector dimensions). For each of the measures, I calculated the correlation for all pairs of model configurations based on the variation scores produced for the entire vocabulary. The summary of these results is shown in Table 10.3. They indicate that `avg` scores are strongly correlated across different model configurations; for `diff` and `ratio` scores, correlation tends to be weak to moderate. This suggests that `avg` scores are less sensitive to model configurations. Moreover, correlation is particularly weak when comparing an SGNS and a PPMI model, and particularly strong between pairs of PPMI models; this suggests, not unexpectedly, that model architecture has the strongest impact on the resulting variation scores.

	All models			**SGNS models**			**PPMI models**		
	`avg`	`diff`	`ratio`	`avg`	`diff`	`ratio`	`avg`	`diff`	`ratio`
mean	0.826	0.367	0.382	0.894	0.419	0.417	0.942	0.737	0.746
min	0.603	0.165	0.177	0.720	0.268	0.273	0.877	0.539	0.553
max	1.000	0.993	0.992	0.972	0.690	0.686	1.000	0.993	0.992

TABLE 10.3: Spearman's rho for pairwise correlations between different configurations, based on each of the three variation measures. There are overall 18 model configurations (153 pairwise comparisons); these include 12 SGNS model configurations (66 pairwise comparisons) and 6 PPMI model configurations (15 pairwise comparisons).

The trends for the three variation scores are corroborated by correlations between the scores *within* a model configuration (Table 10.4). In line with the previous observations, `avg` is weakly correlated with both `ratio` and `diff`, which are in turn strongly correlated with one another.

	avg-diff	avg-ratio	diff-ratio
mean	0.088	-0.058	0.963
min	-0.160	-0.215	0.929
max	0.284	0.111	0.995

TABLE 10.4: Spearman's rho for different variation scores in a given model configuration

The properties of the three measures are further clarified by their correlation with frequency, plotted in Figure 10.3. It shows that `avg` is strongly negatively correlated with frequency. This additionally translates to a strong positive correlation with the instability scores computed in the control condition models. These results overall indicate that a high cosine distance may simply reflect poor vector representations related to a low frequency. The pattern is much less pronounced for `diff` and `ratio`, suggesting that the way they are calculated implicitly neutralizes some of the noise captured by the models.

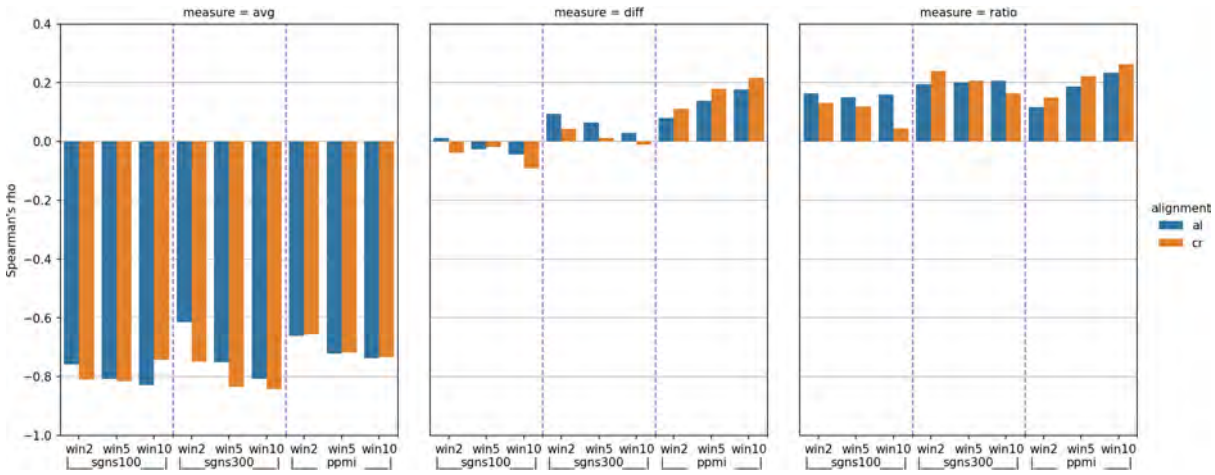


FIGURE 10.3: Correlation between different semantic variation measures and frequency

Next, I used each of the three measures to extract 100 words with the highest variation scores for each of the 18 model configurations. The lists produced by different model configurations and variation measures differ significantly. When two lists are compared, the median number of different words stands at 72;[3] the values range from 1 to 95 different words. I additionally looked at the differences between different SGNS runs for the same model configuration. On average, when comparing two lists based on models trained in subsequent runs with identical parameters, there is a difference of 22 (`avg`), 26 (`diff`), or 43 (`ratio`) words between the two lists. Moreover, most words with the highest variation scores appear in very few different lists. The median number of model configurations that include a given word in the top 100 list stands at 3 for `avg`, 2 for `diff`, and 1 for `ratio` (out of a total of 18 configurations).

Table 10.5 presents the top 30 words per measure, ranked by the number of models which include them in the top 100 list. Beyond the striking difference of the three lists, the qualitative patterns are reminiscent of the exploratory analyses presented in Chapter 7. At first glance, some of the words are intriguing. Take for example the adverb *definitively*: its use in English is

[3]Calculated as $100 - n$, where n is the number of overlapping words. The order of the words is not taken into account.

more restricted than that of *definitely*, but its form is closer to the more general French adverb *définitivement*, which might be conducive to a shift in usage. Similarly, both *saison* and *encore* are used in highly specialized contexts in English (the first denotes a type of beer; the second refers to a repeated performance), whereas their orthographically identical French equivalents have very broad meanings. Examples such as these warrant further investigation.

But for most items in the list, the underlying explanations are much simpler. Many are French grammatical words that have English homographs and appear in codeswitched tweets (*ont*, *ce*, *pour*). Others are related to topical variation (*mortgage*, *housing*, *detached*, etc. are linked to Vancouver's saturated property market) or to typing practices and tokenization errors (*bec*, as in *Québec*, is occasionally attested detached from the rest of the word). Note moreover that most of these items are infrequent. This is particularly true for the lists based on the `avg` and `diff` measures, where the median raw frequency of the retained items in most model configurations is well below 1,000 for the whole corpus. The lists based on `ratio` emphasize more frequent words, with most medians in the range between 1,000 and 10,000 occurrences.[4]

avg				**diff**				**ratio**			
V_ca	18	V_ont	12	A_sous	18	V_ont	12	N_overdose	18	N_housing	11
R_bec	18	**N_pour**	12	N_saison	18	**N_trough**	12	N_saison	18	N_cpp	11
N_nt	16	N_svp	12	N_trustee	18	N_mb	12	A_sous	17	N_mortgage	11
N_vag	16	N_gorge	12	A_detached	18	N_svp	12	N_trustee	17	A_plus	10
R_definitively	16	A_od	12	N_parfait	18	N_cpa	12	N_buyer	17	N_ton	10
R_afterall	15	N_vers	12	N_loin	18	V_tout	12	N_price	17	**N_nt**	10
R_now-	15	N_matt	12	N_overdose	17	N_dispatch	12	N_parfait	16	N_affordability	10
N_bm	14	N_cp	12	R_den	16	R_definitively	12	N_sq	16	N_aux	10
V_;s	14	N_bd	12	**N_pour**	14	N_vers	12	R_den	15	N_renovation	10
R_alway	14	R_obvs	12	V_pour	14	V_zone	12	V_pour	14	V_tout	10
N_cf	13	V_is/was	12	N_rapport	14	N_prix	12	**N_pour**	14	N_bedroom	10
R_defiantly	13	N_plateau	11	A_immaculate	14	N_encore	11	N_loin	14	A_spacious	10
R_lastly	13	**N_trough**	11	A_desirable	14	N_aux	11	A_detached	13	A_desirable	10
R_consequently	13	N_ln	11	**N_nt**	12	N_still	11	N_rapport	12	N_antioxidant	10
N_chum	12	N_vu	11	N_plateau	12	N_le	11	N_le	11	**N_trough**	9

TABLE 10.5: Top 30 words with highest scores for different variation measures based on the number of model configurations that detect them (out of 18; number indicated next to the words). The words in bold appear in the top 30 words for all three measures.

To recapitulate, this experiment has shown that (i) the three semantic variation measures are positively correlated across model configurations, but this is more pronounced for `avg` than for `diff` and `ratio`; (ii) `avg` is weakly correlated with both `diff` and `ratio`, which are in turn strongly correlated with one another, suggesting that they capture different information; (iii) raw cosine distances and hence `avg` scores are strongly correlated with frequency and with instability scores computed in the control condition, suggesting that they are of limited reliability in detecting semantic shifts; (iv) lists of words with highest variation scores differ significantly between model configurations, as well as between different SGNS runs of the same configuration; (v) most of the words with highest variation scores are picked up by very few different models, at least based on the simple method examined here.

[4]For context, a raw frequency of 1,000 occurrences in this corpus corresponds to a normalized frequency of 1.6 per million words.

Taking a further step back, the results discussed so far highlight the impact of model configurations and variation measures on the obtained results. They also suggest that some vector representations are of questionable quality, given that they vary somewhat even when they would be expected to remain stable. This issue affects different variation measures to different extents, and it further highlights the central role of word frequency. These observations are hardly surprising; for example, discussion is already ongoing on whether frequency primarily facilitates semantic change (e.g. Hamilton et al., 2016b) or introduces a bias into the models (e.g. Dubossarsky et al., 2017). However, given the trends observed in the exploratory experiment, a more precise understanding of the mechanisms which might affect descriptive results was a necessary precaution. The underscored trends should be accounted for when comparing vector representations across different models; one way to do so is discussed in the next section.

10.2 Exploring the dimensions of variation

We have seen that vector space models are capable of capturing some trends that are useful for the detection of contact-induced semantic shifts in Quebec English, but the targeted results are partly obscured by other phenomena. In a step towards addressing this issue, the next experiment draws on multiple types of information to better circumscribe the area of the vocabulary exhibiting cross-linguistic semantic influence.

I am unaware of any existing test sets which could be readily applied to the detection of contact-induced semantic shifts in Quebec English. As a result, the bottom-up, exploratory approach is maintained in this experiment, but the observation of trends in the vocabulary is guided by the existing sociolinguistic descriptions of the phenomenon under study. They are used both directly, in examining the way in which previously described lexical items are represented in the models and in the corpus, as well as indirectly, in determining the likely characteristics of potential semantic shifts.

10.2.1 Experimental setup

This experiment uses principal component analysis (PCA) to examine the interaction between regional semantic variation, as captured by vector space models, and additional linguistic information. The analyses discussed so far have highlighted the potential role of a wide range of variables:

- the measures directly output by vector space models (cosine distances) and the SAGE-based specificity score (cf. Chapter 9), which are expected to capture general regional trends;
- the derived variation scores introduced above (`avg`, `diff`, `ratio`), which are expected to emphasize regional variation in meaning driven by the use in Montreal;
- frequency in the Twitter corpus, as it has been shown to affect the performance of a range of measures and methods;

- variation across the shuffled subcorpora, which serves as a control condition indicative of the noise captured by the methods;
- information reflecting the presence of French content, as it is an important source of noise;
- variability of the context in which the word appears, which may reflect uses that are too specific (e.g. those related to local referents).

Starting from a set of input variables, PCA identifies principal components, i.e. new variables (linear combinations of the initial ones) which are uncorrelated between them, and are obtained in decreasing order of explained variance. It provides a better understanding of the correlations between input variables, and therefore of the information that they capture with respect to one another. Principal components summarize these correlations and can be used to create a low-dimensional space, facilitating data exploration. The contact-induced semantic shifts described in the literature, together with the examples retained from the initial analyses, constitute a starting point for this exploration: it can be expected that the full range of information outlined above will facilitate the identification of other words exhibiting similar characteristics.

Computing principal components. PCA was computed using the `scikit-learn` implementation (Pedregosa et al., 2011) with default parameters. Input variables (see below) were mean-centered and scaled to unit variance. Note that separate input matrices were created for each of the 18 vector space model configurations. Although some differences in the output exist, the qualitative trends are substantially similar across the configurations. For clarity, I will report the results for a single configuration: SGNS architecture, 300-dimensional vectors, window size of 5, alignment using Orthogonal Procrustes.

Input variables. The following information was used as input:

- pairwise cosine distances (Montreal-Toronto, Montreal-Vancouver, Toronto-Vancouver);
- measures of regional semantic variation (`avg`, `diff`, `ratio`);
- word frequency for the three subcorpora;
- specificity scores for the three subcorpora (output by SAGE, presented in Chapter 9);
- shuffle-based instability score, calculated as the mean of a word's cosine distances for all pairs of models trained on the shuffled subcorpora (AB, AC, and BC) for a given model configuration;
- frequency of the target word in a large French corpus, FrWaC (Baroni et al., 2009) (aiming to identify French homographs);
- FrWaC frequency of the target word's context, calculated as a weighted mean of the frequencies of the word's cooccurrents in a symmetrical 10-word window (aiming to identify phenomena such as English borrowings used in French-language tweets);
- context variability score, calculated for a given target word as the mean cosine distance between its context words, the assumption being that restricted referential use will be reflected by a more limited variability of contexts, as suggested by Del Tredici et al.

(2019).

All frequencies were log-transformed. The context variability score was calculated on a maximum of 1,000 contexts, which were randomly sampled where that number was exceeded. I also experimented with additional information, including French context frequencies in a different window, differences in frequency across subcorpora, as well as a measure of SGNS instability. However, each of these measures was strongly correlated with some of the other input variables, so they were not included in the final analysis.

Previously described semantic shifts. In order to facilitate parts of the analysis, a list of 52 previously described semantic shifts was used. It was based on descriptions in a range of sociolinguistic studies (Boberg, 2012; Fee, 1991, 2008; Grant, 2010; McArthur, 1989; Rouaud, 2019b). The posited contact-related senses were clarified using an analysis based on lexicographic sources (cf. Section 5.1.4). In an echo of the theoretical discussion in Chapter 3, some instances include an English word which is used with a clearly distinguishable new sense (e.g. *circulation* 'traffic' rather than 'circular movement [of air, blood in the body, money, etc.]', cf. Fr. *circulation*). In other cases, the posited distinction between the senses is subtler, often more general (e.g. *militant* 'activist, campaigner' rather than 'combative, aggressive activist', cf. Fr. *militant*). Finally, the list also includes items that are mainly described in the literature as being used more frequently or in a wider range of registers under the potential influence of French (e.g. *furnish*, cf. Fr. *fournir*).

At this stage, the items were not used to directly evaluate the ability of the underlying information to capture their characteristics. Rather, they represented tentative starting points enabling further data exploration. A more comprehensive analysis of a subset of these items is presented in Chapter 11.

10.2.2 Components of interest

We begin by inspecting the obtained principal components. Table 10.6 presents their component scores (i.e. coefficients used to compute the new variables), which indicate their association with the original variables. The input variables are of equal statistical importance given the fact that they were mean-centered and scaled to unit variance before the analysis.

The first component (48% of explained variance) is associated with pairwise cosine distances, `avg`, frequency, and shuffle instability score. All of these variables were previously found to be correlated to one another; I have moreover suggested that they point to unreliable vector representations due to the link with variation across control condition corpora. The second component (14% of explained variance) is orthogonal to the first one (i.e. independent from it), meaning that it should capture trends unrelated to frequency. It is associated with the `diff` and `ratio` scores, which are less affected by frequency than `avg` and therefore potentially more informative for semantic shift detection. The third component (11% of explained variance) is associated with French word frequency and with high specificity to Montreal, suggesting that it more directly captures the presence of French elements. The next two components are strongly associated with Vancouver and Toronto specificity scores, respectively, indicating re-

	cos_mt	cos_mv	cos_tv	avg	diff	ratio	fr	fr_win	freq_m	freq_t	freq_v	sage_m	sage_t	sage_v	context	shuff
1	-0.3342	-0.3323	-0.3398	-0.3376	0.0106	0.0923	0.1313	0.0419	0.3396	0.3420	0.3409	-0.0073	0.0302	0.0013	0.2283	-0.3452
2	0.1590	0.1555	-0.0211	0.1593	0.5840	0.5655	0.1722	0.2169	-0.0013	0.0313	0.0652	-0.2779	-0.0385	0.2780	0.1590	0.0254
3	0.1229	0.1306	0.1465	0.1284	-0.0599	-0.0963	0.4183	0.4600	0.1457	0.0554	0.0542	0.4824	-0.3029	-0.2680	0.3133	0.0941
4	-0.0137	0.0044	0.0953	-0.0047	-0.3244	-0.2947	0.1469	0.1118	-0.0148	-0.0132	0.0976	-0.3145	-0.4386	0.6759	0.0806	-0.0066
5	0.1042	0.1057	0.1747	0.1063	-0.2233	-0.2051	0.2159	0.2090	0.0397	0.1522	0.0854	-0.4288	0.6812	-0.0950	0.2447	0.0718
6	0.1628	0.1820	0.1685	0.1746	0.0181	0.0022	0.1389	-0.7718	0.2441	0.2287	0.2384	0.0815	-0.0703	0.0259	0.2501	0.1505
7	-0.0645	-0.0657	-0.0705	-0.0660	0.0154	0.0219	0.8113	-0.1867	-0.0997	-0.0923	-0.1007	-0.0276	0.0652	-0.0208	-0.5005	-0.0617
8	0.1178	0.1602	0.1491	0.1408	-0.0282	0.0029	-0.1706	0.1999	0.3433	0.3508	0.3440	-0.1714	-0.1643	-0.1730	-0.6348	0.0004
9	-0.0365	0.0496	0.0120	0.0067	-0.0172	-0.0144	0.0464	-0.1086	-0.1291	-0.0872	-0.1167	-0.5652	-0.4383	-0.5713	0.2044	-0.2528
10	-0.0645	-0.0355	0.1499	-0.0506	-0.6503	0.7274	0.0046	-0.0011	-0.0127	-0.0218	-0.0234	0.0062	-0.0190	-0.0253	0.0124	0.1236
11	-0.4429	0.6618	0.1245	0.1119	-0.0420	0.0313	-0.0221	-0.0146	-0.0233	-0.0359	-0.0338	0.1522	0.1065	0.1099	-0.0076	-0.5336
12	0.5867	-0.2901	0.1759	0.1495	-0.0873	0.0529	-0.0228	-0.0302	-0.0213	-0.0313	-0.0269	0.1284	0.0707	0.0839	-0.0089	-0.6897
13	0.0021	-0.0012	-0.0009	0.0004	0.0041	-0.0036	-0.0012	-0.0017	0.7940	-0.2575	-0.5374	-0.1036	0.0217	0.0567	0.0023	-0.0006
14	-0.0006	0.0006	-0.0002	0.0000	0.0008	-0.0004	-0.0011	-0.0010	0.1614	-0.7714	0.6079	-0.0236	0.0636	-0.0683	-0.0012	-0.0007
15	-0.4677	-0.4692	0.6737	0.2532	0.2076	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	0.0000	0.0000	-0.0000

TABLE 10.6: Principal components. Montreal, Toronto and Vancouver are referred to using initials. Columns: pairwise cosine distances (`cos`); semantic variation scores (`avg`, `diff`, `ratio`); target word's FrWaC frequency (`fr`); context FrWaC frequency (`fr_win`); frequency per subcorpus (`freq`); specificity scores in individual corpora (`sage`); context variability score (`context`); mean cosine distance for shuffled corpus (`shuff`).

gional trends driven by those two cities. The importance of the remaining components is very limited, as the proportion of explained variance decreases significantly.

In general terms, this analysis is coherent with the initial impression that different types of information may best describe the patterns observed in the data. From a practical standpoint, component 1 is notable because it groups together a range of potential sources of noise in the output of the models, with orthogonality allowing us to circumvent their impact. In looking for contact-induced semantic shifts in Quebec English, my focus is on English words which are semantically different in Montreal and which resemble French words. They may also be more frequent in Montreal than elsewhere, under the hypothesis that the acquisition of a new sense might lead to their use in a wider range of contexts. As a result, I more closely focus on components 2 and 3, which are associated with these features, and are further examined below.

10.2.3 Areas of interest

The information provided by components 2 and 3 was used to examine vocabulary-level patterns. Specifically, individual words were projected in a two-dimensional space defined by the values of components 2 and 3. An interactive setup implemented using the `plotly` library[5] was used to visually explore the entire 20,000-word vocabulary. For illustration, a small subset of this projection is plotted in Figure 10.4.

Manual inspection identified an area of interest corresponding to words that have high semantic difference scores (component 2), as well as a high Montreal specificity score and a high French frequency (component 3). Here we find previously described semantic shifts (e.g. *exposition* 'exhibition', *souvenir* 'memory'), as well as other words which exist in English but have formally similar French equivalents (e.g. *ambiance*, *bureau*). Relevant words are not at a neutral point (i.e. the center) of the two-dimensional space. Neither do they correspond to

[5] `https://plot.ly`

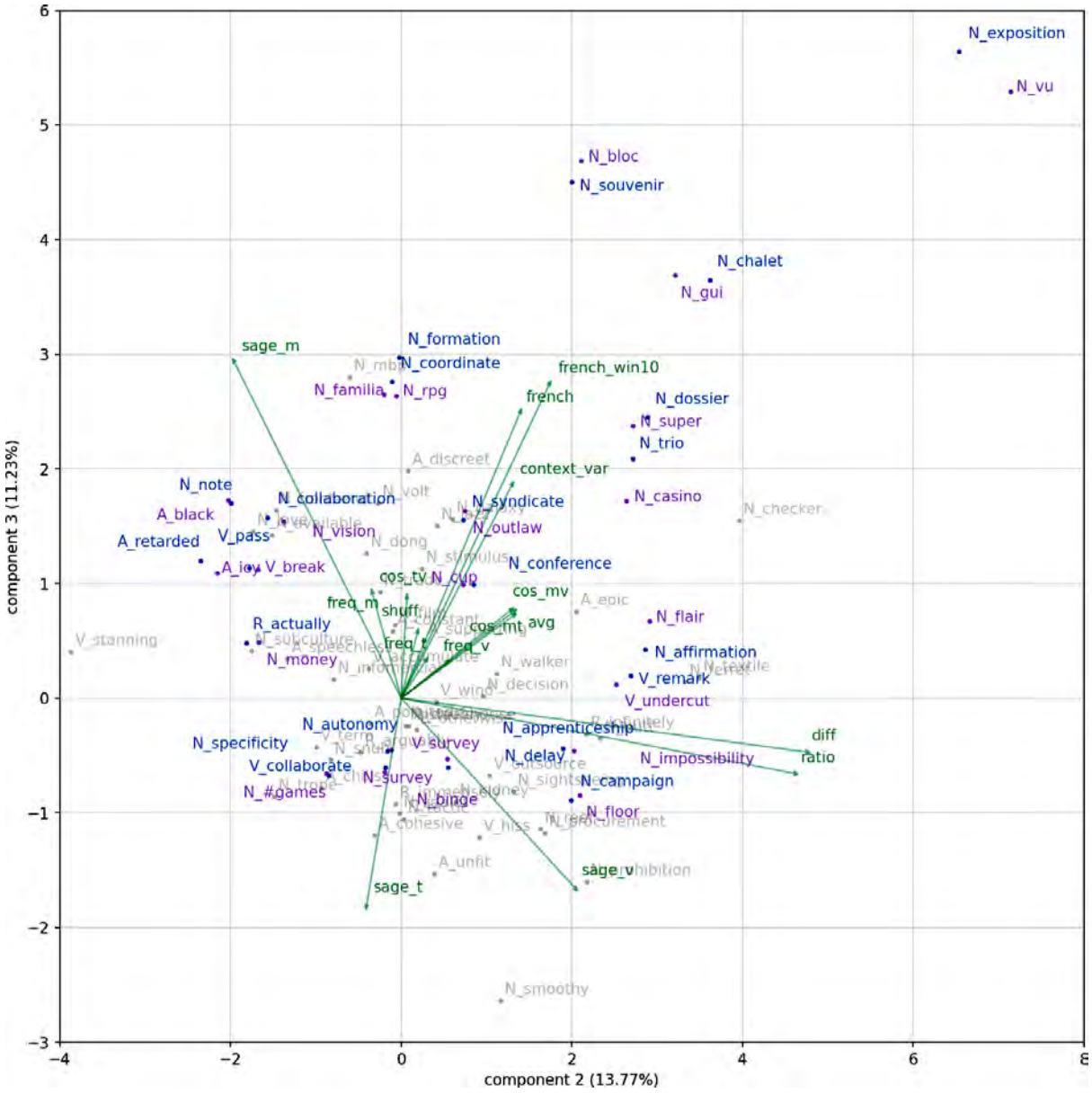


FIGURE 10.4: PCA biplot based on components 2 and 3. Words in blue: a sample of previously described semantic shifts, chosen based on the fact that they have the most extreme values alongside the two dimensions shown here. Words in purple: nearest neighbors of the previously described cases based on the Euclidean distance calculated on components 1 through 3. Arrows in green indicate the impact of original variables. A random sample of 50 additional words, plotted in gray, is provided for context.

the most extreme cases, which are generally associated with noise-related phenomena (French homographs, locally-specific referents).

Additionally, the set of 52 semantic shifts described in the sociolinguistic literature was projected onto this space (a sample is plotted in blue in the figure above). It was assumed that they might provide an indication as to the most relevant areas of the vocabulary. These cases fit into a relatively limited section of the two-dimensional space; like before, they are mostly found in a central area of the space rather than at its extremes. This area does not fully coincide with the one observed through manual inspection, but this may be reflective of the different nature of some of the semantic shifts (e.g. distinctly new sense vs. difference in register).

Two types of attempts were made at exploiting these items to detect additional instances of semantic shifts. First, I visually inspected their neighboring items in the two-dimensional space, aiming to identify those for which patterns of cross-linguistic influence could be posited, and then assessing their occurrences in the corpus. However, in addition to its inherent subjectivity, this approach was hampered by the fact that the previously described examples did not define a compact area in the two-dimensional space, but were interspersed with numerous other items.

The second approach builds on the same underlying assumptions, but aims to explore them more systematically. For each initial semantic shift, I computed its nearest neighbors in the reduced space using the Euclidean distance for the first three principal components. The neighbors were expected to behave similarly to the initial items in terms of the features with which these dimensions are associated, i.e. the stability of their representations, regional semantic differences, frequency in French data, and specificity to Montreal. The top neighbors for a sample of items are plotted in purple in the figure above. Manual inspection of the tweets in which these items are attested points to trends such as the following:

- potential instances of semantic shifts (*impossibility* used in a wider range of contexts, reflecting Fr. *impossibilité*);
- locally specific proper nouns (*bloc* as in *Bloc Québécois*, a political party);
- topical variation (*gui* as in *graphical user interface*, likely due to Montreal's IT industry);
- French homographs (*vu* as in *déjà vu*, but also in longer codeswitched spans of French).

Overall, this application of the PCA facilitated the identification of contact-induced semantic shifts, but its efficiency was limited by the presence of lexical items for which a link with contact could not be established based on their use in the corpus. In addition, the types of noise affecting this approach did not differ substantially from those reported for the previously implemented methods. This may not be surprising since some of the input variables were derived from the very methods affected by noise. However, it is significant that the issues persisted even after combining multiple types of information that were expected to reflect distinct trends in the vocabulary.

More generally, continued in-depth exploration of corpus data once again emphasized a clear presence of multiple senses per item, only some of which might be associated with language contact. This points to the need for a more precise token-level analysis, which is introduced in the next section.

10.3 Leveraging token-level models to facilitate data exploration

The vector models used so far produce type-level representations, i.e. a single vector is used to represent the meaning of all of the word's occurrences. This section introduces token-level representations, whereby each individual occurrence is represented by a slightly different vector, informed by its immediate linguistic context. This approach is implemented with the aim of automatically identifying tweets used in similar contexts and quantifying sense distributions, focusing on both regional and user-level patterns.

10.3.1 Experimental setup

This experiment examines previously identified lexical items of interest by first producing token-level representations for their individual occurrences, and then automatically grouping them together into clusters which are expected to reflect similar immediate linguistic contexts (and thereby similar uses of the target word). This allows for a more efficient manual analysis of the full range of uses exhibited by a lexical item: for instance, the fact that similar occurrences are grouped together means that it is not necessary to disambiguate them one at a time.

Token-level vector representations. This analysis was conducted using BERT (Devlin et al., 2019), a pretrained deep neural network based on the Transformer architecture. This model was previously introduced in Section 5.2.1.2, which provides a more extensive overview. Recall that although it is mainly applied to complex NLP tasks, the vectors produced by BERT can also be used as token-level semantic representations. I specifically used the HuggingFace implementation (Wolf et al., 2020) of `bert-base-uncased`, a 12-layer, 768-dimension version of the model pretrained on generic English data. No fine-tuning was performed given its computational cost and the assumption that word senses are reflected by differences in immediate linguistic context, which the pretrained model should be able to capture.

For each analyzed word, I extracted the tweets in which it appears in all three regional subcorpora. In order to limit processing and memory requirements, I retained no more than 1,000 total occurrences per word, and used a random sample for more frequent items. I fed each tweet as a single sequence into BERT, which then produced context-informed vectors for each token in the tweet. In fact, the model outputs multiple vector representations per token, each corresponding to a different hidden layer in the neural network architecture. Similarly to other recent studies (e.g. Laicher et al., 2021), I averaged over the last 4 hidden layers to obtain a single token-level vector. BERT's tokenizer splits some words into subparts with known representations; when this occurred, I averaged over the subparts to produce a single vector.

Clustering. I identified similar uses of a word by clustering its token-level vectors using affinity propagation, an algorithm which performed well in other semantic change studies (e.g. Martinc et al., 2020b). It does not require a predetermined number of clusters, and it produces clusters of variable size. These properties are well-suited for studying the senses with which a

given word is used, since both the number of senses and the number of occurrences per sense vary depending on the word. I used the `scikit-learn` implementation (Pedregosa et al., 2011) with default parameters, which is based on the negative squared Euclidean distance.

Data from the three regional subcorpora were clustered at the same time, meaning that a single cluster may contain tweets all three cities. As a result, regional distribution can be examined on the level of individual clusters. Only clusters containing at least 5 tweets were retained for analysis.

10.3.2 Qualitative analysis: types of variation

Token-level representations were first used to qualitatively explore the range of senses with which previously identified lexical items are associated in the corpus. An illustrative example is provided by the case of *deception*, which is typically used in English to refer to the action of deceiving (cheating) someone. In Quebec, we might expect it to be used with the sense associated with the phonologically similar French lexical item *déception* 'disappointment'.

The token-level analysis of this item is based on 923 occurrences grouped into 58 clusters, ranging in size from 5 to 62 tweets (median = 14). The original occurrences are roughly equally distributed across the regions, which is contrasted by the geographic distribution of tweets across different clusters, presented in Figure 10.5.

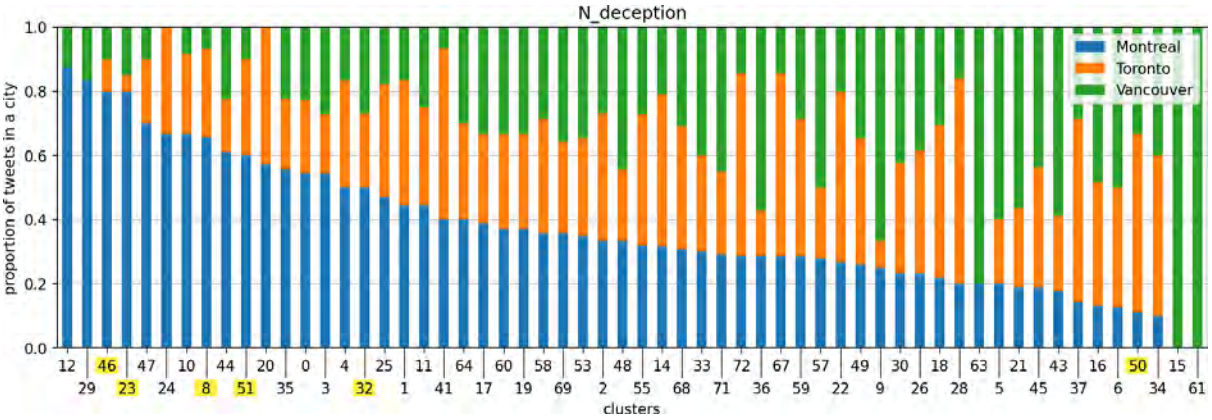


FIGURE 10.5: Distribution of tweets from different cities across clusters for *deception*. The clusters are sorted by proportion of tweets from Montreal. The numbers on the X-axis are used to identify the clusters; those that are highlighted in yellow correspond to the examples below.

The plot shows that the uses attested in some clusters are predominant in Montreal, others in Toronto and/or Vancouver, with a fairly large intermediate area. Intuitively, we might expect to find the contact-related sense in the clusters dominated by tweets from Montreal, and the typical English sense in the geographically mixed clusters. Consider the following set of examples illustrating the uses attested in a subset of the clusters (the number in square brackets indicates the cluster plotted above; all cited tweets were attested in Montreal).

(14) This press release is full of falsehoods and **deceptions**, I don't know where to start. [32]

(15) That moment of enormous **deception** when you realize all the beautiful clothes you left @ the store are unavailable online. [23]

(16) The new song **Deception** Bay, from Milk & Bone's second album, is out! [46]

The expected cases occur in some clusters. Example (14), appearing in a regionally balanced cluster, arguably contains the conventional sense 'act of deceiving, cheating', as suggested by the cooccurrent *falsehoods*. Example (15), attested in a Montreal-specific cluster, involves the contact-related sense 'disappointment'; the alternative interpretation is incoherent with the situation described in the tweet. However, example (16) also appears in a Montreal-specific cluster, and yet it is not related to a consequence of language contact but rather to a local referent: it discusses Deception Bay, a song by the Montreal band Milk & Bone. Note moreover that this is a proper noun, but it was tagged as a common noun, highlighting the fact that even limited shortcomings in corpus postprocessing can impact subsequent analyses.

Other cases are ambiguous, highlighting difficulties in the post-hoc determination of the attested meaning. This is even more pronounced in Twitter data due to formal constraints, in particular the limit of 280 characters per tweet.

(17) Canadian **deception**.. It's not nearly as warm outside as the sun would suggest [50]

(18) The great **deception** [51]

(19) Hrm, @<username> **deception** n'est-ce pas ? [8]
Hrm, @<username> deception isn't it?

The context of tweet (17) allows for multiple readings: the speaker may be saying that the unexpectedly cool weather is disappointing, or that the sunshine is misleading given the temperature. The context of tweet (18) is insufficient to provide any reliable determination. Finally, another common issue is illustrated by example (19): the tweet is written in French. It is difficult to determine if the target word constitutes a codeswitched span of text written in English, or if it is attested in French in a misspelled form (without the acute accent). Whatever the case, issues such as this one underscore the complexities of working with empirically occurring data.

Much like the previously implemented methods, the token-level analysis provides some relevant results, but it is also affected by the now familiar types of noise. Its key advantage lies in automatically grouping together similar occurrences of a given lexical item. This is vital in facilitating semi-automatic analyses of larger amounts of data, as shown in the next section.

10.3.3 Quantitative analysis: effects of bilingualism

The second experiment conducted using token-level vector representations aimed to assess their applicability to a broader analysis of quantitative patterns underlying the use of contact-induced semantic shifts. It focused on four lexical items whose contact-related use had been ascertained through manual use of the corpus. In addition to the already discussed example of *deception*, they include two cases described in previous studies, *souvenir* (McArthur, 1989, p. 25) and *terrace* (Fee, 2008, p. 179), as well as *definitively*, picked up in the first vector-based analysis

presented in Chapter 9. Sample tweets and discussion of the potentially contact-related use are presented below.

(20) January 2018. Such a great **souvenir**!! I'll always remember your devotion to my learning, my lovely friend

The conventional English sense of *souvenir* corresponds to 'memento, keepsake', whereas example (20) illustrates the sense 'memory, recollection'. Both are associated with the corresponding French lexical item *souvenir*. The description in the OED includes the second sense, but it is marked as "chiefly literary", suggesting that its use in a manifestly informal tweet could involve a motivating factor, such as the influence of French.

(21) Dear restos with **terraces**: when it's this hot outside, please close your windows & turn on the AC.

Example (21) illustrates the use of *terrace* to refer to the outdoor seating area of a bar or a restaurant, as opposed to the more general sense 'level, paved area next to a building'. The dining-related sense is associated with the French lexical item *terrasse*. Note that this item is also attested as a borrowing preserving its original form, both in previous sociolinguistic studies (e.g. Boberg, 2012, p. 497) and in lexicographic sources (e.g. in the OED, where it is marked as related to France).

(22) Pouring coffee beans in the water tank... I **definitively** need coffee!!!

The adverb *definitively* is mainly used with the narrow sense 'conclusively, in a definitive manner', often modifying verbs such as *prove* or *know*. In example (22), it is used as a generic intensifier, much like *definitely*. The only formally similar French adverb is *définitivement*; in Quebec French, it is attested with the sense 'definitely', which Usito describes as influenced by English. This might in turn have provided a direct pathway for *définitivement* to influence *definitively*. A more indirect route, potentially involving a higher salience of *definitively* for bilingual speakers, can also be posited.

In analyzing these lexical items, my aim was to understand whether their use with the conventional or contact-related senses could be associated with explanatory factors. I particularly focused on the degree of bilingualism, whose role in the use of semantic shifts has been previously noted (McArthur, 1989; see also its use as an explanatory factor in variationist sociolinguistics, discussed in Chapter 6). This experiment corresponds to a practical implementation of the notion of semasiological sociolinguistic variable, introduced in Chapter 5. Here, each lexical item corresponds to a variable, whose variants are the conventional and the contact-related sense. Their use is analyzed in terms of the linguistic profile exhibited by the authors of the tweets. This estimation is estimated based on the proportion of tweets they posted in English, out of the total of English and French tweets; this metric was presented in Chapter 8, and is taken to roughly approximate the degree of bilingualism.

Starting with the automatically generated clusters of tweets, I manually identified those which unambiguously contained either the conventional or the contact-related sense, as defined above for the four lexical items. I then extracted the linguistic information on the users from

Montreal who had posted the tweets in the retained clusters. The distribution of their linguistic profiles is plotted in Figure 10.6.

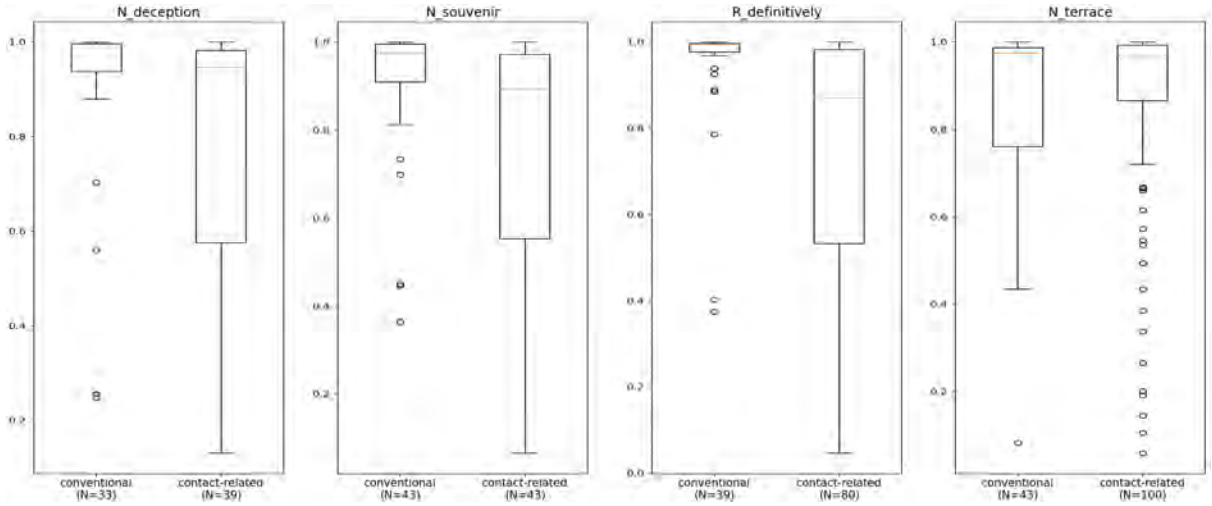


FIGURE 10.6: Distribution of users' linguistic profiles for conventional (left) and contact-related (right) meanings. Higher values on the y-axis indicate a larger proportion of tweets in English per user.

The results are strikingly similar for three of the four lexical item (*deception*, *souvenir*, and *definitively*). The speakers using the contact-related meaning post a noticeably lower proportion of tweets in English (i.e. a higher proportion of tweets in French) than those using the conventional meaning. However, the trend is different in the case of *terrace*, where no such distinction is apparent. Two main suggestions emerge from these observations. First, a strong association with the use of French for some semantic shifts indicate that these cases constitute variations in usage driven by individual bilingualism, rather than fully established regional semasiological variants. Second, the distinct trends in terms of the impact of bilingualism may reflect different degrees of diffusion of semantic shifts in the speech community. At this stage, however, these are only tentative hypotheses which require further testing.

10.4 Summary

Building on the methodological issues first observed in the initial exploratory experiments, this chapter aimed to provide a better understanding of the mechanisms underpinning the comparisons of vector representations across regional corpora. The first experiments in this chapter analyzed general patterns in type-level vector space models, examining 18 model configurations and three measures of regional semantic variation. It highlighted the questionable quality of some vector representations, as reflected by their instability in the control condition. It also underscored the differences introduced by model configurations and variation measures, both in terms of the vocabulary-level patterns that the models capture and the semantic shift candidates that they identify. Frequency was shown to be particularly important as it was strongly correlated with multiple measures reflective of unstable vector representations.

The relationship between different types of information characterizing individual lexical items was further explored using principal component analysis. In addition to corroborating

the central methodological role of frequency and the metrics associated with it, the analysis led to a better identification of the area of vocabulary the most likely to exhibit contact-related influence. This approach also facilitated the identification of additional semantic shifts, but noise remained present in the results.

The final experiment consisted in an implementation of token-level vector representations, which were used to automatically group together similar occurrences of a lexical item of interest. The resulting clusters were explored both to qualitatively examine the uses of a lexical item, linking them with regional characteristics, as well as to set the basis for a more extensive quantitative assessment of explanatory factors. This experiment also allowed me to formulate tentative hypotheses regarding the status of contact-induced semantic shifts, with a likely important role of the degree of bilingualism.

While I have begun putting these methods to good descriptive use, the same types of noise were repeatedly noted. The recurrence of methodological issues across the tested approaches, related both to their inherent characteristics and to the structure of the corpus, calls into question their practical value in descriptive research; this claim must be further investigated. These experiments have also provided additional support for a revised version of the high-level hypothesis that Montreal-specific usage likely reflects the influence of French, which was already called into question in the last chapter. The examined examples suggest that contact-related uses are indeed considerably more frequent in Montreal than in the other two cities; however, they generally only represent a fraction of all the uses attested in Montreal, making their discovery more challenging.

In order to address some of the open questions, this analysis should be conducted on a larger number of lexical items. Building on the progress made so far, it remains necessary to provide both a more extensive descriptive account as well as a more systematic evaluation of the implemented methods. We turn to these issues in the next chapter.

Chapter 11

Evaluating the descriptive contribution of vector space models

The computational experiments conducted so far have underscored that, in addition to descriptive potential, the implemented approaches display recurrent methodological issues. This calls for a further investigation of their utility in descriptively-oriented research, in particular by systematically evaluating their performance on a larger number of lexical items. This is the research direction pursued in the present chapter. It more generally builds on the view that existing computational work on semantic change detection has often focused on generic research questions and datasets, using them as a training ground for proof-of-concept studies (Boleda, 2020, p. 218; see also Chapter 5). Here, I draw on the precisely defined descriptive issue at the heart of this dissertation to explicitly address the descriptive potential of these methods.

In order to facilitate a systematic evaluation, I first developed a test set for semantic shift detection (Section 11.1). I then evaluated different type-level models and semantic variation measures, introduced in the previous chapter, in order to find the best-performing model and then deploy it on the discovery of new semantic shifts (Section 11.2). A token-level analysis, coupled with a qualitative annotation, was then used to further characterize the use of semantic shifts and explain some of the issues affecting type-level models (Section 11.3). The chapter concludes with an overview of the main results (Section 11.4).

11.1 Creating a test set for contact-induced semantic shifts

This section introduces a new test set allowing for a systematic evaluation of semantic shift detection in the context of English–French language contact. The role of a test set is to provide reliable information against which a computational system can be evaluated; on the evaluation of semantic change detection in general, and the use of test sets in particular, see Section 5.2.2.3.

Similarly to recent shared tasks on diachronic data (Basile et al., 2020; Schlechtweg et al., 2020), I formulate the task of contact-induced semantic shift detection as a binary classification problem. In this perspective, computational systems are asked to classify a lexical item in a binary manner, i.e. either as being semantically stable or as corresponding to a semantic shift. The items in the test set are labeled accordingly. It includes 80 items; this is comparable to the

diachronic test sets of which I am aware, containing between 18 and 100 items (Basile et al., 2020; Del Tredici et al., 2019; Gulordava and Baroni, 2011; Schlechtweg et al., 2020).

Note that most recent diachronic test sets were created through crowd-sourced annotation campaigns, whereas I rely on expert judgment, similarly to some existing work (e.g. McGillivray et al., 2019; Perrone et al., 2019). This is a viable approach because the phenomenon under study is more specific than general semantic change over time, and its existence can be reliably established based on the sociolinguistic literature, lexicographic sources, and observation of corpus data. In addition, the test set introduced here only uses binary labels, so the underlying decisions are comparatively straightforward.

11.1.1 Identifying shifting lexical items

In identifying a set of positive examples, I relied on semantic shifts previously reported in the literature on Quebec English (Boberg, 2012; Fee, 1991, 2008; Grant, 2010; Josselin, 2001; McArthur, 1989; Rouaud, 2019b), as well as the qualitative exploration of the Twitter corpus, presented in the previous two chapters. Consistently with the minimum frequency used in training vector space models, items with fewer than 100 occurrences per subcorpus were excluded. A concordance-based analysis was used to determine if the items presented at least one contact-related occurrence in the Montreal subcorpus; those that did not were also excluded.

In establishing the potential contact-related use, the existing descriptions and corpus-based observations were complemented with lexicographic evidence; for a discussion of the lexicographic sources, see Section 5.1.4. This process resulted in a list of 40 semantic shifts, whose mean frequency in the entire corpus is 5,268 (min = 345, max = 97,188). The list of target lexical items and the posited contact-related senses is summarized in Table 11.1. In order to more precisely illustrate the linguistic mechanisms at play and the scope of the manual analysis, two examples are discussed in more detail below. They are representative of different degrees of distinction between the conventional English senses and those hypothesized to be related to French; this in turn translates to variable difficulty in determining the relevance of individual lexical items.

11.1.1.1 A clear-cut distinction: *resume*

An example of a relatively straightforward analysis is provided by the verb *resume*. The impact of French on its use in Quebec English, via the verb *résumer*, is suggested by McArthur (1989, p. 25). He illustrates it with the following introspectively produced example: “He spoke for two hours, then carefully *resumed* the main points.”

In determining the conventional (i.e. non-contact-related) meaning of *resume* in Canadian English, we can begin by looking at the COD. It describes the following senses:

> **1** *transitive & intransitive* begin again or continue after an interruption. **2** *transitive* recover, occupy again (*resume a lifestyle*; *resume a political position*).

This is consistent with sense 1. a. (a) provided by the OED:

affirmation	claim, statement	hesitate	deliberate (between two options)
ambiance	atmosphere, vibe	laureate	winner
animator	activity leader, team leader	local (n)	room, site, premises
availability	availability (sg.), available times	manifestation	protest, demonstration
boutique	shop, store	merit (v)	deserve, be worthy of
chalet	summer cottage	militant	activist, campaigner
circulation	traffic	nomination	appointment to a role
coordinates	contact details, name and address	occasion	chance, opportunity
deceive	disappoint	pass by (v)	stop by
deception	disappointment	permit	driver's license
definitively	definitely, certainly	population	the people, general public
deputy	member of parliament	portable	cell phone, laptop
dossier	question, issue, (minister's) portfolio	proposition	suggestion, proposal
entourage	family circle, relatives, group of friends	prudent	careful
exchange (v)	talk with someone	remark (v)	notice
exploration	study (of a little known phenomenon)	reparation	repairs (of a device, appliance etc.)
exposition	art exhibition	resume	summarize
formation	training, course	souvenir	memory
formidable	great, terrific	terrace	outdoor eating area, patio
grave (adj)	highly important	trio	sandwich-fries-drink menu, combo

TABLE 11.1: List of target lexical items and posited contact-related senses

> To begin again or continue (a practice, occupation, course, etc.) after interruption.

However, the OED also includes the following senses:[1]

> **3.** a. *transitive*. To recapitulate or summarize (facts, etc.). Cf. RÉSUMÉ *v*.[2] Now *rare*. [...] † b. *transitive*. To repeat (a sentence or word). *Obsolete*. [...] † c. *intransitive*. To give a résumé or summary. *Obsolete*.

Note that the first sense references the English verb *résumé*, defined as 'to give a résumé of; to summarize', but also marked as "U.S. rare". It is this sense that is attested in McArthur's example. It is also described in French, with the TLFi providing the following definitions for sense I. A:

> **1.** Condenser (un texte, un discours) en peu de mots, en ne donnant que les informations principales [...] **2.** Rendre compte de façon succincte.

Drawing on these descriptions of the verb *resume*, we can identify a conventional sense 'continue following interruption' and a contact-induced sense 'summarize'. They are clearly distinct, but not to such an extent to warrant a description in terms of homonymy (i.e. two lexical items) rather than that of polysemy. In semantic terms, both senses include the idea of repetition. They are also linked etymologically, since *resume* traces its origin back to Middle French *résumer* and Latin *resūmere* (as per the OED).

One might argue that the extent of French influence may be limited in this case, because the posited contact-induced sense is attested in the OED. However, it is marked as rare or obsolete, as further confirmed by its absence from the COD; this constitutes evidence that its use requires an external "push", which might be provided by cross-linguistic influence. The conventional

[1] Elided information in dictionary definitions corresponds to examples.

and contact-induced senses are respectively attested in the following two examples from the Twitter corpus:

(23) I will **resume** birthday celebrations next weekend at someone else's party

(24) @<username> @<username> @<username> That **resumes** my whole Adobe experience recently. Awesome it is. #Cough

11.1.1.2 A subtler distinction: *formidable*

Comparatively more complex analyses are illustrated by the case of the adjective *formidable*. It was identified during exploratory analyses presented in Chapter 10, specifically while inspecting the two-dimensional space produced using the principal component analysis. Its close similarity – orthographic identity – with the French adjective *formidable* spurred further investigation of this example.

As before, let us begin by reviewing the definitions provided by the COD.

> **1** inspiring fear or dread. **2** inspiring respect or awe. **3** likely to be hard to overcome, resist, or deal with.

This largely corresponds to the main definition in the OED.

> That gives cause for fear or alarm; fit to inspire dread or apprehension. Now usually (with some obscuration of the etymological sense): Likely to be difficult to overcome, resist, or deal with; giving cause for serious apprehension of defeat or failure.

Both definitions highlight the idea of fear, which occupies a central position in the meaning of this adjective. The COD also includes a broader sense corresponding to 'awe-inspiring'; however, I would argue that even in this case there is a tendency towards a negative connotation, given the implicit link with the notion of superiority. While the TLFi notes the existence of a corresponding fear-related sense in French, it is marked as dated or literary. Contrast it with the definitions provided for sense C:

> **1.** [En parlant d'une chose, notamment dans le langage affectif ou publicitaire] Très beau ou excellent, admirable, très remarquable, extraordinaire. [...] **2.** [En parlant de pers.] **a)** Très sympathique, très serviable, etc. [...] **b)** Extraordinairement doué. [...] **3.** *Très fam.* [En parlant de qqn ou de qqc. qui surprend, pour exprimer étonnement, insatisfaction, impatience] Étonnant, surprenant.

Senses 1 and 2 express appreciation of an object or a person, and are strongly positively connoted. Sense 3 can be applied to an unsatisfactory situation, but it evokes surprise or astonishment, rather than the ideas of fear, respect, or awe, included in the English definitions. Moreover, WordReference indicates that the positively connoted French usage corresponds to English adjectives such as *terrific* and *great*. But it also suggests that a feature incorporated into

TLFi sense 1, that of being extraordinary, can correspond to the English adjective *formidable*; however, this usage arguably retains a link with the idea of superiority.

Compared to the previously discussed case of *resume*, this example outlines a fuzzier picture. The French and English adjectives partly overlap, but much of the strongly positively connoted French usage is absent from the examined English definitions. It is this aspect that presents an interest for cross-linguistic influence. Consider the following examples from the corpus of tweets:

(25) Can the Lakers ever win against this **formidable** opponent? They always seem to come short of the target!!

(26) I saw @starisbornmovie tonight and @ladygaga was **FORMIDABLE**! What a performance! She's the soul of the movie. I am here for the #GagaActress chapter.

(27) **Formidable**, thanks! (I'm always happy to double check colors if you need)

(28) Eeet izz, how you say, **formidable**! Absolutment! Zuh most fragrant and piquant leadership. Like a bowl of ripe fromage. Mmwwah! I love eet!

The conventional English sense is illustrated by tweet (25). Example (26) can be interpreted as both 'excellent' and 'awe-inspiring', highlighting the closeness of the two senses; it is difficult to reliably determine if it can be attributed to the influence of French. The potential scope of this influence is more clearly illustrated by example (27), which is reminiscent of the English adjective 'great'. This occurrence can be seen as a further step beyond the most immediate influence of French, since it corresponds to a partly semantically bleached exclamation. Example (28) lends additional support to the idea of a French-specific sense, as it includes *formidable* in a tweet mocking the use of English by native French speakers.

In summary, contact-related use of *formidable* mainly rests upon the connotation associated with the adjective. While its isolation from other uses is more complex, it can nevertheless be observed in the data, and its existence is further supported by metalinguistic commentary. This type of contact-related influence is moreover consistent with the definition of semantic shifts outlined in Chapter 3. Other instances of similarly fine-grained distinctions revolve around issues such as different degrees of generality and specification (e.g. *boutique* 'store' rather than 'small, fashionable or specialized, store') and differences in syntactic patterns which are additionally associated with a semantic distinction (e.g. *exchange with someone* 'talk with someone' rather than *exchange something with someone* 'reciprocally give and receive'; cf. Fr. *échanger*).

11.1.2 Identifying stable lexical items

Once the 40 semantic shifts were identified, I selected 40 stable lexical items to be included in the test set. These are the control items that computational models would be expected to classify as *not* corresponding to a semantic shift. It was important to limit the presence of items with formally similar French equivalents, as they are more likely to be involved in contact-related use. It was also important to control for frequency, which, as we have seen in Chapter 10, has a significant effect on vector-based measures of semantic variation.

I therefore started from a list of 3,231 English lexical items of Anglo-Saxon origin,[2] around half of which meet the frequency threshold (100 per subcorpus). For each of the 40 semantic shifts, I identified a lexical item in the list which was of the same part of speech and was the closest to it in terms of frequency measured on the whole corpus. Using a sample of occurrences, I then checked that the words were not affected by meaning variation across subcorpora or other issues which could bias subsequent analyses (e.g. homography, use in proper names etc.). If necessary, the items were replaced with the one which was the next closest in frequency. Sample stable items attested in tweets from the Montreal subcorpus are presented below.

(29) What a **blatant** example of corruption! There is no justice left in the world!!

(30) Should I continue my paper now or just **cram** it all in tomorrow and panic bc it's due on Tuesday

(31) Late evening **errands** as a #dad of a toddler are often composed of: milk... and a bottle of wine. #Parenting101

(32) yeah, I'm gonna go back tomorrow. I'm sure I can exchange it, it's just the **hassle** of having to commute all the way back.

11.1.3 Structure of the test set

The final test set contains 80 lexical items, split into two balanced classes of shifting and stable items. A sample of the test set is presented in Table 11.2. The entire test set is included in Appendix A, and it is also publicly available.[3]

Sem. shift	Fr. sense	Freq.	Stable item
formidable	'terrific'	1.48	damp
circulation	'traffic'	2.12	campfire
deceive	'disappoint'	2.98	cram
souvenir	'memory'	3.11	hassle
resume	'summarize'	4.91	arise

TABLE 11.2: Sample semantic shifts, with frequency per million words and corresponding stable words in the test set (same POS and closest in frequency)

Having constituted the test set, I deployed it to evaluate the performance of type-level vector space models. This experiment is presented in the next section.

11.2 Evaluating type-level vector space models

As we have seen in Chapter 10, different type-level vector space models exhibit comparable general characteristics, in that they capture broadly similar patterns within the whole vocabulary. However, the representations that they produce are not identical; this is particularly evident when analyzing the top semantic shift candidates output by the models. Moreover, the different

[2] https://en.wikipedia.org/wiki/List_of_English_words_of_Anglo-Saxon_origin
[3] http://redac.univ-tlse2.fr/corpora/canen.html

measures implemented to detect semantic shifts are not all strongly correlated, suggesting that they too reflect different trends in the data. These issues are addressed by first conducting an evaluation on the previously introduced 80-item test set to identify the best performing model configuration, and then more closely analyzing the top candidates output by that configuration.

11.2.1 Experimental setup

The setup used in this experiment is closely reminiscent of that introduced in the last chapter; the same models were used in this experiment. Details regarding their implementations can be found in Section 10.1; here, the discussion will be limited to a brief summary of the parameters I experimented with, as well as clarifications regarding the features that were introduced in addition to the previously investigated ones.

Vector representations. As before, I used two model architectures, the count-based PPMI models and the `word2vec` SGNS models. I experimented with different window sizes (2, 5, 10) and, for SGNS models, different vector dimensions (100, 300). Comparisons of meaning representations corresponding to different regions were based on two approaches. AL models involve the training of separate models for each region, which are then aligned using column intersection (for PPMI models) or the Orthogonal Procrustes approach (for SGNS models). SR (Spatial Referencing) models are trained on the entire corpus; target words are tagged so as to be specific to the subcorpus in which they appear, while context words are the same across the subcorpora. This results in regionally specific meaning representations, but they occupy a shared vector space and are therefore directly comparable.

Measuring differences in meaning. Like in the previous experiments, the basic measure of semantic difference was the cosine distance (CD). It was used to derive three semantic variation metrics, aiming to prioritize the lexical items whose meaning is the most different in Montreal compared to the other two cities. Recall that they are computed as follows:

$$avg(w) = \frac{CD(\vec{w}_m, \vec{w}_t) + CD(\vec{w}_m, \vec{w}_v)}{2}$$

$$diff(w) = avg(w) - CD(\vec{w}_t, \vec{w}_v) \qquad ratio(w) = \frac{avg(w)}{CD(\vec{w}_t, \vec{w}_v)}$$

with the word w represented by its vectors corresponding to the Montreal ($\vec{w}_m$), Toronto ($\vec{w}_t$), and Vancouver ($\vec{w}_v$) subcorpora. Summarizing, `avg` corresponds to the mean of the Montreal-Toronto and Montreal-Vancouver distances. It is further used to compute `diff` and `ratio`, which correspond to the difference and the ratio, respectively, between `avg` and the Toronto-Vancouver distance.

Three SGNS models were trained for each configuration, in order to control for the instability of vector representations which is inherent to this method. While in the previous

experiments this information was only used to examine the extent to which the representations vary across the runs, here I also computed average variation scores across the three runs. Specifically, the cosine distance for a word w in subcorpora a and b was computed as follows:

$$CD(w_a, w_b) = \frac{\sum_{i=1}^{n} CD(\vec{w}_{a_i}, \vec{w}_{b_i})}{n}$$

for $n = 3$ runs of the SGNS model, where $\vec{w}_{a_i}$ is the word's vector corresponding to the subcorpus a in the i^{th} run. It is this average cosine distance that was then used to compute the three derived semantic variation scores.

11.2.2 Finding the best performing model

I begin by evaluating the overall performance of model configurations on the previously introduced test set. The overarching aim is to tune the models and validate their performance relative to the results reported on other similar tasks.

Given the focus on the general patterns captured by the models, I used a simple classification based on the median variation score: I computed the score for the 80 words in the test set and considered that the 40 words with the highest score represented semantic shifts, whereas the others were stable. Admittedly, this evaluation method reflects the split of positive and negative items in the test set, which may introduce a bias. However, this approach allows for a simple and efficient comparison of different sets of parameters before subsequent qualitative analyses of the best performing model.

The best performing configuration (SGNS, Orthogonal Procrustes, window size of 5, 100-dimensional vectors, cosine distance averaged over 3 runs, `diff` score) obtains an accuracy score of 0.8. This is an improvement of 0.4 points compared to the worst result (PPMI, column intersection, window size of 10, `diff` score). It also represents an improvement of 0.225 points compared to the worst-performing SGNS configuration overall, and of 0.175 points compared to the worst-performing SGNS configuration using the `diff` score. These results confirm the interest of dataset-specific model tuning on this task. Further details are presented in Table 11.3.

Several key takeaways emerge regarding model configurations. (i) PPMI models are strongly outperformed by SGNS models, with the difference in mean accuracy reaching 0.2 points for the `diff` score. (ii) The alignments are roughly comparable, with AL models obtaining the best individual result and SR models the higher mean score. (iii) Smaller window sizes perform somewhat better, in line with results reported for synchronic semantic variation in German (Schlechtweg et al., 2019). (iv) The 100-dimension models systematically outperform the 300-dimension models, in line with Pražák et al.'s (2020) results in diachrony. (v) For SGNS models, using cosine distance averaged over multiple runs is beneficial. Although the increase in mean accuracy is limited compared to individual runs, the resulting scores are not only more robust, but they can also improve on the best performing individual run.

As for the three semantic variation measures, `diff` score is the best-performing. Its highest accuracy represents an improvement of 0.1 points on the best `avg` result, and of 0.025 points on the best `ratio` result. The latter difference may be limited, but `diff` exhibits consistently

	Model type		Alignment		Window size			Dims		Run	
	PPMI	SGNS	AL	SR	2	5	10	100	300	avg	rand
AVG											
mean	.508	**.635**	**.630**	.613	.621	**.629**	.614	**.644**	.627	**.635**	**.635**
min	.475	**.575**	**.475**	**.475**	.475	**.550**	.500	**.600**	.575	**.600**	.575
max	.550	**.700**	**.700**	.675	**.700**	.675	.675	**.700**	.675	**.700**	.675
DIFF											
mean	.500	**.703**	.667	**.694**	.690	**.693**	.657	**.718**	.688	**.706**	.701
min	.400	**.625**	.400	**.425**	**.575**	.500	.400	**.650**	.625	**.625**	**.625**
max	.575	**.800**	**.800**	.775	.775	**.800**	.775	**.800**	.775	**.800**	.775
RATIO											
mean	.508	**.684**	.655	**.675**	**.681**	.674	.640	**.696**	.673	**.688**	.683
min	.425	**.600**	**.425**	**.425**	**.575**	.500	.425	**.625**	.600	**.625**	.600
max	.625	**.775**	**.775**	**.775**	.750	**.775**	.725	**.775**	**.775**	.750	**.775**

TABLE 11.3: Accuracy across model configurations using different parameters and semantic variation measures. AL: separately trained and subsequently aligned models; SR: Spatial Referencing; Dims: vector dimensions; Run: avg corresponds to the use of a cosine distance averaged over three SGNS runs, rand corresponds to cosine distance from individual SGNS runs. Vector dimensions and runs are only applicable to SGNS models. Underlined values are the highest across the three semantic variation measures.

stronger performance. More generally, these results confirm that it is beneficial to subtract the Toronto-Vancouver distance from the mean Montreal cosine distance. As we have seen in Chapter 10, `diff` is uncorrelated with frequency (mean $\rho = -0.01$ across configurations for the entire vocabulary), unlike `avg` (mean $\rho = -0.78$). I hypothesize that the inclusion of the Toronto-Vancouver distance in the variation score might function as a control condition limiting the impact of background noise in the models, similarly to the use of shuffled corpora by Dubossarsky et al. (2017).

The highest accuracy score I obtained is comparable to the state of the art on similar semantic change detection tasks on diachronic data. In addition to indicating the best individual configuration, this validates the general experimental setup that I adopted, confirming that an observable regional distinction is present in the data in relation to the semantic influence of French. Using the best performing model configuration, I now turn to the discovery of semantic shift candidates from the whole vocabulary.

11.2.3 Deploying the model

This step of the experiment aims to systematically assess the performance of the model on the task of discovering new semantic shifts candidates. Automated approaches to this evaluation have been proposed, including the use of synthetic corpora (Shoemark et al., 2019) and dictionary attestation dates (Basile and McGillivray, 2018). However, my objective is to better understand how the model performs on empirically occurring data which might reflect the presence of contact-induced semantic shifts; I therefore manually analyzed the contexts in which the candidates were attested.

I calculated the `diff` score for the whole vocabulary (open classes only, i.e. nouns, verbs, adjectives, and adverbs), and selected the 50 words with the highest score; they are presented in Appendix B. I verified for each word (i) whether it presented a regionally specific use in the Montreal subcorpus, and (ii) whether this use could be explained by the influence of French, and specifically the presence of an equivalent sense in a formally or semantically similar French word, as established by lexicographic evidence. The same range of sources was used as in Section 11.1.

Only one candidate was found to clearly correspond to a contact-induced semantic shift; this translates to a precision score of 0.02. The positive example is the noun *exposition* (`diff` = 0.22, ranked 26th). As noted in the extensive discussion in Chapter 9, it usually refers to an art exhibition in the Montreal subcorpus, and to narrative structure in the two other subcorpora. These uses are respectively illustrated by examples (33), attested in Montreal, and (34), attested in Toronto:

(33) I really want to go to an art museum or an art **exposition**

(34) I found the first 2 episodes a little slow, but it does pick up once the **exposition** is done with

The contact-related sense is typical of the French homograph *exposition* 'art exhibition'. This use has been previously described (Fee, 1991, p. 14), and it is included in the test set.

The contrast between the accuracy on the test set and the precision on the discovery task is striking. Figure 11.1 shows that the problem lies in the fact that the lexical items of interest, like those in the test set, are not the ones with the most extreme variation scores. They tend to have higher variation scores than stable words, showing that the model does capture meaningful trends. But relevant results are ultimately obscured by other types of variation.

In an echo of the discussion from the previous two chapters, the false positives in the 50 word sample include proper nouns denoting local referents (*plateau* referring to Plateau-Mont-Royal, a Montreal neighborhood); topical variation (*detached* limited to real estate in Toronto and Vancouver, which have notoriously tight housing markets); French homographs in code-switched tweets (*pour* 'for'); misspellings indicative of an imperfect command of English (*trough* 'through'). While some of these issues have been reported in previous diachronic studies (e.g. referential effects in Del Tredici et al., 2019), these results underscore that they are highly widespread even when model configurations are carefully tuned. It is tempting to say that they could easily be avoided using basic data filtering, such as the exclusion of the words attested in French corpora or the use of additional frequency thresholds. But things are more complicated than that: for instance, homography also affects many longstanding borrowings (*bureau*) and targeted semantic shifts (*exposition*); higher relative frequency in Montreal may reflect noise as well as increased use of a word that has undergone semantic change. It is also not viable to keep extending the list of top candidates: there are on average 78 words in the list between each of the top 10 positive examples from the test set; this increases to 476 if all 40 positive examples are considered.

Overall, the evaluation of type-level models has shown that the experimental setup I have adopted – based on detecting regional semantic differences to isolate the effects of language

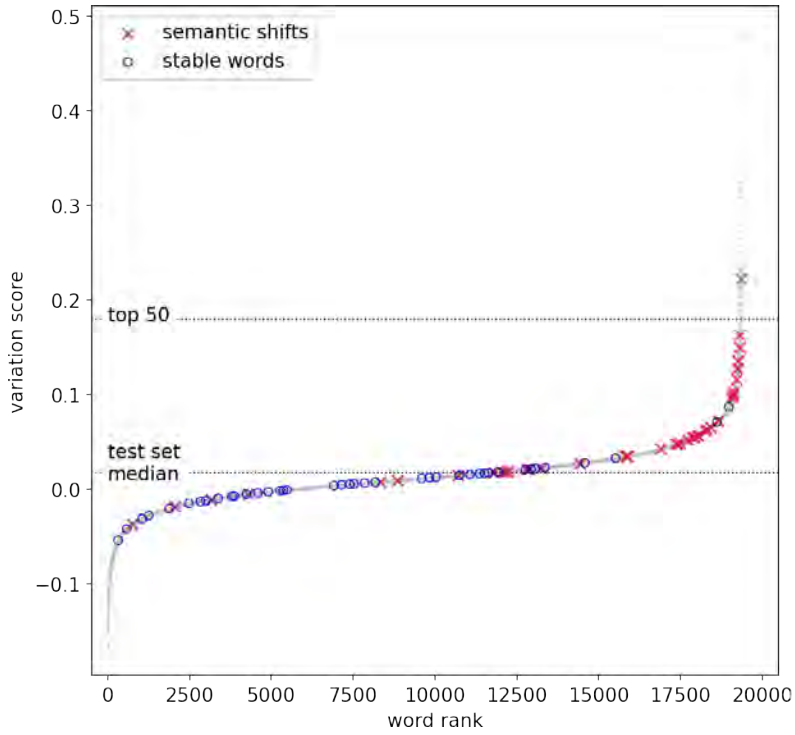


FIGURE 11.1: Variation scores for the whole vocabulary, with the position of semantic shifts and stable words from the test set. Horizontal lines indicate the cutoff score for the top 50 candidates and the test set median.

contact – is viable, at least when it comes to distinguishing between shifting and stable lexical items in a carefully constructed test set of limited size. It has also pointed to methodological considerations, largely in line with previous work; I have additionally underscored the importance of averaging over different SGNS runs, as well as that of considering different cosine-based variation measures. However, a systematic qualitative inspection of the top semantic shift candidates output by the best performing model reaffirmed a key conclusion from the previous two chapters: these models struggle to discover new semantic shifts, despite being able to discern meaningful general patterns in the data. I now turn to token-level representations with the aim of providing a finer-grained analysis.

11.3 Characterizing semantic shifts in context

As I have noted in the previous two chapters, the lexical items of interest tend to exhibit a limited number of contact-related occurrences, most of them being associated with the conventional sense or with noise-related phenomena. Moreover, the target occurrences are expected to be attested in similar immediate contexts, but to be dispersed throughout the corpus. These characteristics, coupled with the previously outlined issues with type-level analyses, suggest that manual inspection of corpus data remains necessary to reliably identify semantic shifts.

In order to facilitate this process, I extended the implementation of the token-level analysis introduced in Chapter 10. My aim was to automatically group semantically similar occurrences of a target lexical item and identify the uses that are specific to Montreal. This enabled me to analyze batches of occurrences all at once, rather than examining them one at a time, which

streamlined both the discovery of contact-related uses and the exclusion of false positives. In the remainder of this section, I will present the experimental setup, a sample analysis of clustered data, and the overall patterns for 40 semantic shifts that were manually annotated using this approach. This will clarify some of the methodological as well as descriptive issues raised so far.

11.3.1 Experimental setup

The experimental setup is nearly identical to the one introduced in Chapter 10. It consists in producing token-level vector representations, automatically grouping them into clusters, and then manually analyzing their use. This section briefly reviews the key methodological decisions; additional information is provided for the steps that were added to the previously presented approach.

Token-level vectors. As before, I produced token-level vector representations using BERT, a pretrained deep neural network, specifically by averaging over the last four hidden layers that it outputs. These vector representations are context-informed; as such, they are expected to reflect the sense with which each individual occurrence is used. Affinity propagation was used to group them into cluster so as to facilitate manual exploration of the data. For further details on this setup, see Section 10.3.

Analyzing regional use. In analyzing the output of the analysis, I considered the clusters containing at least five tweets, and retained them if more than half of the tweets were published in Montreal. This is because of the focus on the senses which are clearly more frequent in Montreal than elsewhere, but which may occasionally appear in other regions. Up to 10 such clusters were retained for each lexical item, starting with those with the highest proportion of Montreal tweets. I then manually annotated the data for the 40 semantic shifts included in the test set. I used binary labels, and established if a cluster presented a contact-related sense based on the majority usage in it.

More specifically, a target word's use was annotated as contact-related if it fulfilled the key criteria underlying the previously presented analyses: it was required to be regionally specific to Montreal and potentially explained by the influence of French. The clusters affected by the amply discussed noise-related issues – referential effects, topical variation, French homographs – were not considered as contact-related. Neither did I annotate as contact-related the clusters involving structural patterns (e.g. the target word being used with different senses but systematically appearing in the tweet-initial or tweet-final position) or those where no reliable determination could be made (e.g. short or ambiguous tweets). A 15-word sample was annotated by two annotators in order to test the reliability of the general procedure, obtaining a reasonably high Cohen's kappa coefficient of 0.55.

11.3.2 Exploring clusters of tweets

On average, 8 clusters per word (min = 3, max = 10) were retained for annotation. The mean average number of tweets per cluster stands at 13 (min = 8, max = 20).

The sample clusters for *manifestation* shown in Table 11.4 illustrate several types of usage that are frequently grouped together. In English, *manifestation* is typically used in to signify 'instance, display'. Cluster 1 contains straightforward examples of contact-related usage, which refers to protests or demonstrations; this is the sense associated with the corresponding French lexical item *manifestation*. Cluster 2 corresponds to the conventional English sense. Cluster 3 reflects noise in the results: *manifestation* is attested as its French homograph in code-switched tweets (in which most text is in English, explaining why they were tagged as English and retained during corpus creation). Additional sample clusters are provided in Appendix C.

	There was a	**manifestation**	in Montreal against the proposed religious
(1)	Montreal's	**manifestation**	protesting against loi 21 banning of « religious
	in Quebec's history . This walk is the biggest	**manifestation**	for this week . And 52 more towns in the province
	This is the most visual	**manifestation**	of patriarchal privilege . That's why it's especially
(2)	Probably the best	**manifestation**	of the benefits of physical/digital retail integration
	the the fact that Disneyland is the physical	**manifestation**	of 1950s American exceptionalism and right-wing
	Giving a Voice to the Voiceless — attending	**Manifestation**	Contre Projet De Loi 128 , Protest Against Bill
(3)	the streets this afternoon in Montréal Grande	**Manifestation**	contre la haine et le racisme . Demonstrators
	having a brownout the night before the	**manifestation**	pour le climat here in Montreal . How odd it is .

TABLE 11.4: Sample clusters for *manifestation*

As these examples illustrate, the clusters are largely homogeneous. Although some are occasionally difficult to interpret, e.g. due to the influence of orthographic information on BERT's representations, this is overall rare. The utility of this approach is confirmed by the fact that it led to the identification of at least one contact-related cluster for each of the 40 target items. From a practical standpoint, using cluster-level annotations was an order of magnitude faster than analyzing individual tweets. This is due to the lower number of required decisions and the comparative ease in determining the meaning of a larger number of similar examples appearing together.

11.3.3 Patterns of semantic variation

Let us now turn to general trends in the annotated data in order to determine if they are related to the variation scores established using type-level models. I specifically focused on two issues that may limit the performance of type-level models: (i) if a contact-related use concerns a minority of occurrences, it is unlikely to be captured by type-level models; (ii) if it is frequent but not regionally specific, it will not be reflected by the variation score. Two corresponding measures were computed: (i) the proportion of tweets, out of all annotated tweets, which appear in clusters that are tagged as contact-related; and (ii) the proportion of tweets, out of all tweets in the clusters that are tagged as contact-related, which originate from the Montreal subcorpus. The obtained values are plotted in Figure 11.2.

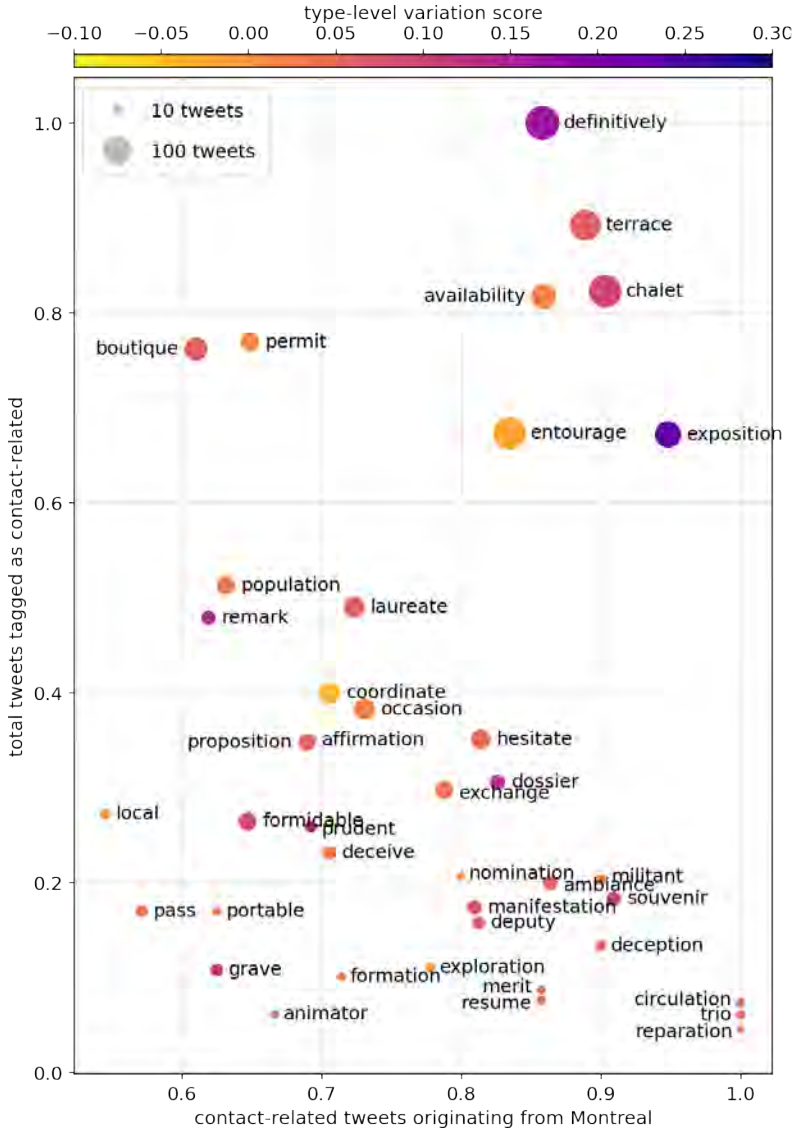


FIGURE 11.2: Scatter plot of annotated words. Y-axis: proportion of tweets that were tagged as contact-related (out of all annotated tweets). X-axis: proportion of tweets from the Montreal subcorpus (out of all tweets tagged as contact-related). Marker size reflects the total number of tweets that were tagged as contact-related. Color coding indicates the variation score computed on the best performing type-level model.

The results point to two overarching tendencies. On the one hand, several lexical items (*definitively* 'definitely', *exposition* 'exhibition' etc.) are characterized by a high proportion of contact-related tweets and, among those, a high proportion of tweets from Montreal. This is indicative of overwhelming contact-related influence which is moreover regionally specific. On the other hand, a larger number of examples (*circulation* 'traffic', *animator* 'group leader' etc.) present limited contact-related usage which additionally varies in terms of regional specificity.

The annotation-based measures are weakly correlated with the type-level variation score ($\rho = 0.23$ for both measures), as well as with one another ($\rho = -0.13$). This reflects contrasting patterns, like in the case of *entourage*, which has a large number of contact-related tweets and a low type-level variation score (-0.02, ranked 39th out of the 40 items). This is related to the relatively small difference between the conventional sense 'people attending an important person' and the contact-induced sense 'group of friends, family'. The distinction is immedi-

ately apparent to the annotator, but it is often underpinned by referential knowledge rather than differences in distributional contexts. Compare this with the adjective *grave*, which has a high type-level variation score (0.12, ranked 6th), but appears in few contact-related clusters. This is due to most of its clusters being excluded because they involve its French homograph, as in the expression *ce n'est pas grave* 'it doesn't matter'. However, the use of French implies drastic distributional differences which are easily captured by type-level models.

While these observations have important methodological implications, they are also significant from a descriptive standpoint. In particular, the range of characteristics exhibited by different lexical items, both in terms of regional specificity and the extent of contact-related usage, is indicative of different degrees of diffusion within the speech community. I already alluded to this issue in Chapter 10, in discussing the use of four lexical items in relation to the degree of bilingualism, as reflected by the proportion of English and French tweets posted by the authors of annotated examples.

We can now extend this analysis to the 40 manually annotated lexical items. Similarly to the exploratory analysis, for each lexical item I calculated the mean proportion of tweets in English (out of tweets in English in French) posted by the users who used the contact-related sense, as indicated by the clusters tagged as such. This value is not correlated with either the type-level variation score ($\rho = -0.06$) or the proportion of tweets tagged as contact-related ($\rho = 0.02$). However, it is moderately correlated with the proportion of contact-related tweets that were posted in Montreal ($\rho = -0.53$). This link is explored in more detail in Figure 11.3.

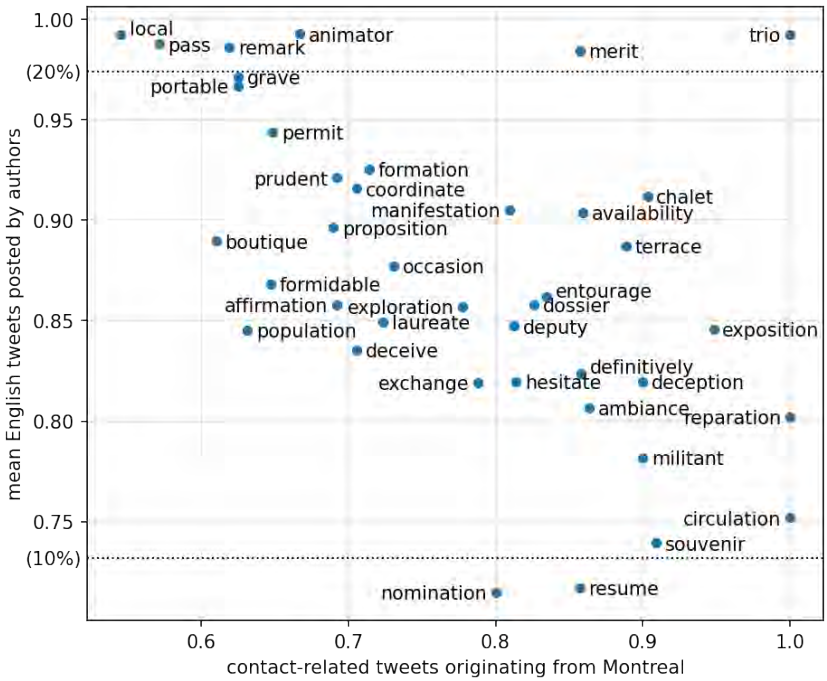


FIGURE 11.3: Scatter plot of annotated words and their relationship with the degree of bilingualism. Y-axis: mean proportion of tweets posted in English (out of tweets in English and French) by users who posted contact-related tweets. X-axis: proportion of tweets from the Montreal subcorpus (out of all tweets tagged as contact-related). Dotted lines show the 10th and 20th percentile for the proportion of tweets in English, for all users in the corpus.

The plotted results indicate that contact-related shifts which are more regionally specific

(i.e. attested in Montreal to a higher extent) are also more directly related to the effects of bilingualism (i.e. a lower proportion of English, and hence a higher proportion of French, tweets). A typical example (bottom right) is the previously mentioned case of *circulation*, attested with the sense of 'traffic', which is associated with the corresponding French homograph. All of its tweets from clusters tagged as contact-related come from Montreal; moreover, the mean proportion of English tweets stands at 0.75 per user. This may appear to be a relatively high value, but it is in fact just above the 10th percentile for all users in the corpus (0.73); at least within this dataset, this is suggestive of a comparatively important influence of bilingualism.

It is also relevant to look at the outliers from the general trend. For instance, in the case of *trio* 'sandwich-fries-soda menu; combo' (cf. QF *trio*; upper right in the plot above) all contact-related tweets similarly come from Montreal. However, the mean proportion of English tweets is higher, at 0.99 per user. This is indicative of a use which is regionally specific, but is widespread in the local linguistic community, including among monolingual speakers. This is further supported by existing descriptions which have shown it to be typical of the speech of native English-speaking Quebecers (Boberg, 2005b, p. 36; Boberg and Hotton, 2015, p. 307). The contact-related uses of both examples are attested in tweets such as the following:

(35) City and provincial police will be on hand to try and improve **circulation** as best as possible during rush hour, @<username> says.

(36) I'd like a Big Mac **trio** with 6 nuggets extra, and can you make the drink an iced coffee instead? I'll pay the difference.

These observations indicate that, barring some exceptions, the more regionally specific the contact-related use is, the more strongly it is associated with knowledge of French. However, these conclusions must be taken with a grain of salt. The manual annotation was conducted on the level of clusters, rather than individual tweets, meaning that some non-contact-related occurrences may have been included in the counts. This issue is not expected to be widespread, but a more precise analysis is needed to make more definitive claims. Moreover, the information on the use of French has the benefit of being empirically grounded in the attested use of languages by individual Twitter users, but it is only a very rough approximation of their linguistic profiles; for instance, there is no reliable way to determine their native language. These issues motivate a shift of perspective, leading to the face-to-face sociolinguistic survey presented in Part IV: it builds on the hypotheses developed through the computational analyses, further investigating the same set of lexical items, but it also provides a finer-grained description of individual speakers and their sociolinguistic behaviors.

11.4 Summary

This chapter drew upon the observations developed throughout the previously described experiments investigating the application of vector space models to contact-induced semantic shifts. Its central aim was to more systematically formulate and verify the claims emerging from the initial analyses, particularly focusing on the descriptive contributions of vector-based

representations. In order to do so, I first developed an 80-item test set for the detection of semantic shifts in English-French contact situations. I then used it to evaluate type-level models, observing robust performance on a standard classification task and very low precision on the discovery of new semantic shifts, confirming the initial intuitions. I then extended the previously implemented token-level analysis, using it to accelerate manual annotation of corpus data for 40 lexical items. This in turn allowed me to provide a more precise account of the issues impacting type-level methods, as well as to formulate descriptive hypotheses regarding the use and diffusion of contact-induced semantic shifts.

As already suggested, the analyses presented in this chapter represent a formalization of a series of methodological intuitions, developed over more than two years of using vector space models and related methods to investigate the data in the Twitter corpus. But these experiments also have more general implications, which reaffirm the central role of evaluation practices and corpora in advancing computational analyses of semantic change (cf. Hengchen et al., 2021). The comparison of evaluations on the test set and on the discovery task underscored the stark difference between the two approaches. This should be taken into account when choosing evaluation methods, especially where the aim is to establish practical usability. And while some reported issues are specific to my corpus, similar problems may affect other semantic change studies, as noisy datasets and complex sense distributions are not unique to this work. Finally, the comparison of type-level and token-level analyses highlighted diverging trends in the data which indicate that semantic shifts involve multiple dimensions of variation. Future computational work should therefore aim to identify different types of semantic change in addition to quantifying its presence.

As for the sociolinguistic objective pursued in this dissertation, these analyses have provided the first quantitative corpus-based account of phenomena which have often been described only anecdotally. While this is an important step in its own right, it remains vital to better understand the constraints on contact-induced semantic shifts and the representations associated with them. It is also essential to determine to what extent estimates derived from computational analyses reflect real-life sociolinguistic behaviors. These issues are at the center of Part IV.

Part IV

Sociolinguistic inquiry

The chapters in this part of the dissertation present sociolinguistic interviews conducted with a group of speakers from Montreal in order to further assess the use of contact-induced semantic shifts. Chapter 12 introduces the interview protocol and the recruitment procedure, as well as the main principles guiding the subsequent data analysis. Chapter 13 outlines the composition of the participant sample, focusing on a range of sociodemographic characteristics and attitudes towards language use. Chapter 14 draws on this description to investigate more closely the use of contact-induced semantic shifts. It highlights distinct patterns of synchronic variation, as reflected by quantitative acceptability ratings and qualitative remarks, which moreover indicate a potential pathway for their diachronic diffusion. It also identifies a core group of speakers who appear to be leading this linguistic practice. Finally, Chapter 15 takes a broader view at the analyses conducted over the course of this dissertation. It contrasts the descriptive contributions of corpus-based approaches and sociolinguistic interviews, underscoring their complementary nature which provides a promising way of investigating a wide range of issues.

Chapter 12

Interview protocol and participant recruitment

The corpus-based analyses presented in Part III led to the definition of a set of 40 lexical items affected by the semantic influence of French, as observed in the Quebec English data from the Twitter corpus. After analyzing the linguistic contexts in which they appear and correlating their use with broad quantitative estimates of linguistic profiles, I suggested that the examined lexical items varied in terms of diffusion across speech communities and association with knowledge of French. While large-scale analyses were instrumental in formulating these hypotheses, I now address them through a more focused, face-to-face sociolinguistic survey. It is limited to a comparatively small number of speakers, but it yields finer-grained descriptions which provide important interpretative context for the general overview obtained from corpus data.

This chapter addresses the methodological considerations underpinning the design and implementation of the sociolinguistic interviews conducted in this dissertation. Section 12.1 presents the structure of the interview protocol, discussing both standard tasks and a novel semantic perception test. Section 12.2 addresses the way in which sociolinguistic interviews were conducted and analyzed based on this protocol. Section 12.3 summarizes this discussion. Note that this chapter is limited to a presentation of the protocol implemented in this dissertation. Broader discussion of data collection in variationist sociolinguistics, including through face-to-face interviews, can be found in Chapter 4.

12.1 Devising a variationist protocol to study semantic shifts

A central methodological component of this dissertation is the validation of computational results from a variationist sociolinguistic perspective. The choice of face-to-face interviews – rather than, for example, written questionnaires – is underpinned by the following objectives:

(1) obtaining as detailed a description as possible of individual speaker profiles;
(2) accounting for cross-linguistic phonological similarity, as empirically observed in recorded speech production;

(3) addressing qualitative issues such as the representations associated with the examined lexical items and the social meaning that they convey;
(4) ensuring comparability with existing variationist sociolinguistic studies of Quebec English, including to validate the reported results.

More generally, these requirements target the aspects that are the least well-addressed by the computational analyses. The sociolinguistic interviews I conducted incorporate the core of the sociophonological protocol developed within the PAC research program. It is complemented with a perception test, designed to investigate contact-induced semantic shifts in a more controlled manner. The structure of the protocol is presented in more detail below.

12.1.1 Common PAC-LVTI protocol

The central elements of the protocol used in this study were developed within the PAC research program (The Phonology of Contemporary English: usage, varieties and structure).[1] It pursues a multifaceted description of spoken English based on the analysis of data collected using a shared protocol across a range of survey locations (Durand and Przewozny, 2012; Przewozny et al., 2020). The initial approach, based on well-established variationist sociolinguistic principles, was extended through the LVTI project (Language, Urban Life, Work, Identity). It crucially introduced a thematic questionnaire informed by sociological research, which is used as a basis for semi-structured interviews. They allow for a more detailed description of the speakers' life – and language use – in urban contexts, which constitute primary points of investigation in sociolinguistic research (Przewozny-Desriaux, 2016, pp. 59–71). Moreover, the protocol can be adapted or extended depending on the speech community and linguistic phenomena at hand.

The remainder of this section presents the tasks comprising the common protocol and their adaptations to the Montreal context. The next section introduces an additional task, developed to investigate contact-induced semantic shifts.

12.1.1.1 Core interview structure

The standard PAC-LVTI protocol is based on the classical variationist approach pioneered by William Labov, and previously presented in Chapter 4. Specifically, it is composed of two reading tasks – involving two word lists and a text – and of two conversations. Taken together, this structure is aimed at eliciting both spontaneous speech production and detailed background information on the informant. The tasks are also reflective of different degrees of formality; in the interviews conducted in this dissertation, they were administered in decreasing order of formality, and were followed by the semantic perception test. This order was particularly useful in establishing a rapport with the informants prior to the final task, setting the stage for a relaxed discussion of semantic shifts attested in corpus examples. The full range of protocol materials is presented in Appendix D; their key features are discussed below.

[1] The program is currently coordinated by Sophie Herment, Sylvain Navarro, Anne Przewozny-Desriaux, and Cécile Viollain. For more details, see https://www.pacprogramme.net/.

Word lists. The protocol includes two word lists, which are used to elicit the pronunciation of isolated lexical items. Word list 1 contains 129 items targeting vocalic features, including those typical of Canadian English, such as the low-back merger (e.g. *pause*, *calm*, *knot*) and *r*-conditioned mergers (e.g. *merry*, *marry*, *Mary*). Word list 2 contains 64 lexical items targeting consonantal features, allowing for an analysis of characteristics such as *t*-flapping (e.g. *betting*, *little*, *carter*) and the /w/ ∼ /ʍ/ opposition (e.g. *witch*, *which*). The items are shuffled so as to limit the proximity of minimal pairs.

Text. The informants are further asked to read a text, which simulates a less controlled communicative situation and enables the study of phenomena occurring in connected speech. The text, entitled *A Christmas interview*, formally resembles a newspaper article; it is roughly one page long. In this study, it was principally used to complement the analysis of phonological features associated with Canadian English, such as Canadian Raising (e.g. *south*, *out* for /aʊ/; *polite*, *like* for /aɪ/) and yod-dropping (e.g. *avenue*, *during*). As in the case of the word lists, the aim of this task is to obtain speech productions that are comparable across the interviewed speakers and, more broadly, other studies using the same protocol. Together with the word lists, this task also controls for style shifting, i.e. the tendency for speakers to adapt their language to different communicative contexts.

Formal conversation. The informants take part in a one-on-one conversation with the interviewer, whose aim is to obtain detailed background information as well as speech production in a more relaxed context. The interview begins with a predefined set of questions which the interviewer asks so as to fill in the PAC information sheet, a form used to systematically collect data on the speaker's basic sociodemographic characteristics; linguistic profile; residential, educational, and professional history; and so forth.

An additional set of questions, defined in the LVTI thematic questionnaire, elicits the speaker's personal view on life in their city, their professional experiences, and their use of languages. The self-reported information provided throughout this conversation, coupled with the displayed linguistic behaviors, is used to establish a detailed sociolinguistic profile for each respondent. Questions regarding the city and the job include:

- Do you feel that you're a true Montrealer?
- Is there another city you would prefer to live in Quebec or in Canada?
- Could you tell me about the things you regularly do in your work?
- Do you think you have a good work-life balance? Could you give me your reasons?

The part of the thematic questionnaire focusing on language use was adapted to the Montreal context; it is presented in more detail in the next section. Note however that a particularly wide range of information on language use is obtained, both throughout the formal conversation and by analyzing the linguistic characteristics of the participant's speech production. This specifically includes:

- overtly elicited information on the age and manner of acquisition of all languages spoken by the informant;

- self-reported degree of proficiency and frequency of use for all languages;
- indirect information on the contexts of use and passive exposure to the languages, obtained through questions on immediate family members, education, professional experience, leisure activities, integration into the neighborhood, and so forth;
- representations associated with the languages and varieties under study;
- observable linguistic behavior, which in this case provides direct information on the speaker's degree of proficiency in English, as well as indirect information on their use of French (e.g. native French realization of segmental features in borrowing and codeswitching, as discussed in Section 2.2.2).

As shown below, this information can be used to produce a quantitative score reflecting the speaker's degree of bilingualism. The resulting continuous variable is identical in form to the scores produced for users in the Twitter corpus, based on the number of tweets that they post in English and French (cf. Chapter 8). However, the information underlying the score derived from the interviews is both considerably more detailed and more reliable.

Informal conversation. The informants are also asked to participate in a conversation without the presence of the interviewer. The interlocutor is usually a person with whom the participant extensively interacts in other contexts, such as a friend or a family member; this person is not required to participate in the remainder of the protocol. The aim of this task is to obtain a sample of more spontaneous language use. Although the use of a recording device may introduce a degree of self-awareness, the absence of the interviewer is expected to diminish the impact of the observer's paradox (previously discussed in Chapter 4).

In a concrete illustration of this effect, one of the participants in this study produced essentially monolingual English utterances throughout the formal conversation, despite being highly proficient in both English and French, and being aware of the interviewer's own English–French bilingualism. She subsequently recorded an informal conversation with a close friend, who is likewise highly proficient in both languages; a high rate of codeswitching is present throughout the 12-minute discussion.

12.1.1.2 Adapting the protocol to the study of English in Montreal

As mentioned before, the structure of the interview tasks is usually adapted to the speech community under study. This includes the linguistic stimuli as well as the questions asked in the formal conversation. In designing the present protocol, I benefited from the work conducted by Julie Rouaud; she had carried out a PAC survey in Montreal in 2016 and 2017, similarly focusing on contact-related phenomena in Quebec English (Rouaud, 2019b). She introduced additional reading tasks and a modified thematic questionnaire to the standard protocol; they are briefly reviewed below.

Rouaud used an additional word list to target phonological features typical of North American English, such as *yod*-dropping, which is not represented in the standard lists. She also created an additional text, which crucially contains 20 French borrowings, allowing for a systematic study of their phonological realizations. Although these modifications enable a more

precise description of Quebec English features, I decided not to include them. This was based on two considerations: (i) the protocol that I was developing was already considerably lengthened by extending the thematic questionnaire and introducing the semantic perception test (see below); (ii) some phonological information targeted by the additional word list can be derived from the standard text, while the pronunciation of French-origin items is elicited in the semantic perception test.

As for the section of the thematic questionnaire which addresses the use of languages in Montreal, I largely drew on the version proposed by Rouaud. The questions include:

- Do you consider yourself a Canadian, a Quebecer, a Montrealer, or a [West Islander, Westmounter, NDGer,[2] etc.]? If so, in which order? Why?
- Can you make the distinction between yourself and American speakers? What about other Canadian people? Do you speak differently from Ontarians for instance?
- Do you think there are any movies, TV shows, podcasts etc. that accurately reflect the way people speak English in Montreal? If so, which ones? If not, why do you think that is the case?
- What is it like living in an officially Francophone province?
- Do you think French influences the way you speak English? In what way?
- What would you say it means to be bilingual for someone living in Montreal? Would you describe yourself as bilingual?
- If you were walking around Montreal and needed to ask for directions, which language would you use?

A notable question I added to the previous version addresses the language used to ask for directions. It turned out to be particularly useful, providing a concise summary of the persons' attitudes towards language use, as reflected by their (admittedly self-reported) communicative behaviors. The question was inspired by the longstanding series of studies on language choice in Montreal (Bourhis et al., 2007).

In addition, given the importance of language use in social media for the computational analyses presented in Part III, I included a set of questions on this issue. They were instrumental in providing a better understanding of the differences between online and offline sociolinguistic behaviors and perceptions. Some of the questions asked in this part of the conversation include:

- Would you say that you are an active user of social media sites such as Twitter?
- In which language do you tweet most often? Do you tweet in other languages as well? If so, under what circumstances? If not, why not?
- Would you say that your choice of languages on Twitter is similar to the way you use them in real life?
- Do you think you would be able to determine if someone is Canadian or American based on their tweets?

The parts of the protocol presented so far allow for a precise and comparable description of phonological features, including those that are typical of Canadian English, as well as a detailed

[2] An inhabitant of the neighborhood of Notre-Dame-de-Grâce.

sociolinguistic profile of individual speakers which takes into account the specifics of language use in Montreal. Let us now turn to the final interview task, which more directly addresses contact-induced semantic shifts.

12.1.2 Creating a task for semantic shifts

This section presents the development of the semantic perception test used to assess contact-induced semantic shifts. It discusses the choice of question types, the choice of examples containing the target lexical items, and the structure of the final task.

12.1.2.1 Potential question types

In order to determine the structure of the interview task that could the most adequately address the use of semantic shifts, I reviewed several previously implemented types of questions. They were initially used in research examining semasiological variation, either as a central object of study or within a wider focus on lexical phenomena; they are reviewed more extensively in Chapter 5. The main types of questions that they implemented are briefly summarized below:

- referent elicitation, where the informant is asked to provide a referent for the target lexical item (Robinson, 2012a);
- open-ended interpretation, eliciting a definition of a sentence containing the target lexical item (Dollinger, 2017);
- multiple choice interpretation, where the speaker is asked to choose one of several potential senses with which the target item is used in a sentence (Chambers, 2007b);
- acceptability rating, in which the target item attested in a sentence is scored for acceptability using a numerical scale, whose extremes respectively correspond to values such as *awkward* and *natural* (Bailey and Durham, 2020);
- community reporting, i.e. eliciting information on the use of the item by other members of the speech community rather than by the respondent. It can be applied to all previous types of questions, and is often used in addition to self-reporting (e.g. Chambers, 2007b; Dollinger, 2017).

Recall that the chosen question type is to be used to examine 40 contact-induced semantic shifts in the context of a face-to-face interview. Bearing this key requirement in mind, the choice of question type was guided by a set of criteria highlighting both descriptive and methodological concerns. The chosen question type should:

- allow for explicit assessment of contact-induced senses;
- be scalable, i.e. allow for an efficient development as well as deployment of questions for a large number of different lexical items;
- be easy to implement in conversation (as opposed to a written questionnaire);
- be fairly simple for the respondents;
- provide results that are reasonably simple to post-process.

In light of these requirements, I decided to implement acceptability ratings. They are

elicited for a specific example, so targeting contact-related use is fairly straightforward (i.e. it should be attested in the provided example), as is extending them to a large number of varied lexical items. While acceptability ratings are usually used in writing, they can theoretically be elicited in speech; alternatively, respondents can provide written answers in the presence of the interviewer. They represent a comparatively simple task: for instance, it is arguably easier to provide a numerical score than to enumerate potential referents of a lexical item. They require no post-processing, directly providing a numerical score which can be readily used in subsequent analyses.

That being said, a key criterion in excluding the remaining question types was the variety of the examined lexical items. The other question types are much more difficult to apply in cases such as abstract senses: for instance, it is more challenging to come up with a convincing question eliciting the referent or the definition for an item such as *definitively* 'definitely' or *deceive* 'disappoint'. On a different note, it should also be underscored that acceptability ratings reflect the perception of a lexical item rather than its observed use by a given speaker. This entails a conceptual difference with respect to the corpus-based analyses conducted in Part III.

12.1.2.2 Choosing examples of target lexical items

The choice of examples evaluated by the respondents plays a central role in implementing acceptability ratings. The examples were drawn from the Twitter corpus, based on the cluster-level annotation of 40 target lexical items presented in Chapter 11. Specifically, the tweets which were included in clusters tagged as contact-related, and which originated from the Montreal corpus, were retained as potential examples.

These tweets underwent a new round of annotation; in addition to myself, it included two other expert annotators. For a given lexical item, each annotator was asked to choose three potential examples among the retained tweets. The median number of potential tweets per lexical item was 15 (min = 4, max = 115).

In addition to taking into account the idiomaticity of the tweet – prioritizing those that most closely align with native English usage – the key requirement guiding the decisions was that the tweet should clearly reflect the posited contact-induced sense. This criterion is in turn related to the clarity of the immediate context in which the target lexical item is attested, as shown by the following examples (the first example was included in the protocol; the second was not retained by any annotator):

(37) Nothing will change with drunk or stoned drivers, or drivers texting, until they lose their **permit** for a minimum of 6 months for the first offense, and for good if they are stupid to try that stunt a 2nd time. Education? Really? At taxpayers expense. No! Hit them hard.

(38) Still havent received my **permit** :(

The examples target the use of *permit* 'driver's licence', reflecting Fr. *permis* (*de conduire*). In the first example, this sense is abundantly clear from the context; in the second, the occurrence might in theory refer to any other type of permit.

Following the annotation, all retained examples were pooled together. For each lexical item, the example receiving the most votes was retained. Since in most cases no consensus was reached (i.e. there were multiple items with the same number of votes), a reconciliation process was used to determine the most appropriate example. One example per lexical item was retained.

12.1.2.3 Structure and implementation of the task

While the task is based on the elicitation of an acceptability rating, it is also complemented with a range of other information. In particular, the participants are asked to:

- read the tweet out loud (providing phonological information on the pronunciation of the target lexical item in connected speech);
- rate the acceptability of the lexical item in the context of the tweet;
- provide a lexical item which could replace the target item without changing the meaning of the tweet (so as to ensure that the lexical item was interpreted with the posited sense);
- provide any further observations regarding the use of the tweets (eliciting representations and facilitating community reporting).

As concerns specifically the acceptability rating, it applies to the lexical item as it is used in the example, rather than the example as a whole; this is intended to limit the impact of elements occurring in the immediate context of the target lexical item. The rating is provided on a scale whose values range from 1 to 6. Participants are instructed to interpret the value of 1 as "very unnatural, awkward, you would never say something like that", and the value of 6 as "completely natural, just like something you might say". It is suggested that they should attempt to follow their initial instinct rather than overthink the decision. It is further made clear that the reference point in providing the rating is their own use of the lexical item in question (as opposed to the way they think it should be used or they hear others use it). In practice, however, there is little way of ensuring that all participants interpret the same instructions in the same way; this is an inherent limit of the acceptability rating approach. As stated above, I conducted the task in the final part of the interview, after the formal conversation. It was implemented as an online questionnaire using LimeSurvey, as shown in Figure 12.1.

The participants were asked to fill it in in the presence of the interviewer, enabling them to spontaneously express opinions on the use of the examined items. Each question was displayed on a separate page, with the main instructions repeated for every question as a reminder. All personally identifiable information in the examples, including Twitter user handles and hashtags, was redacted.

Having described the entire structure of the protocol used in the sociolinguistic interviews, I now turn to the way in which it was put to use.

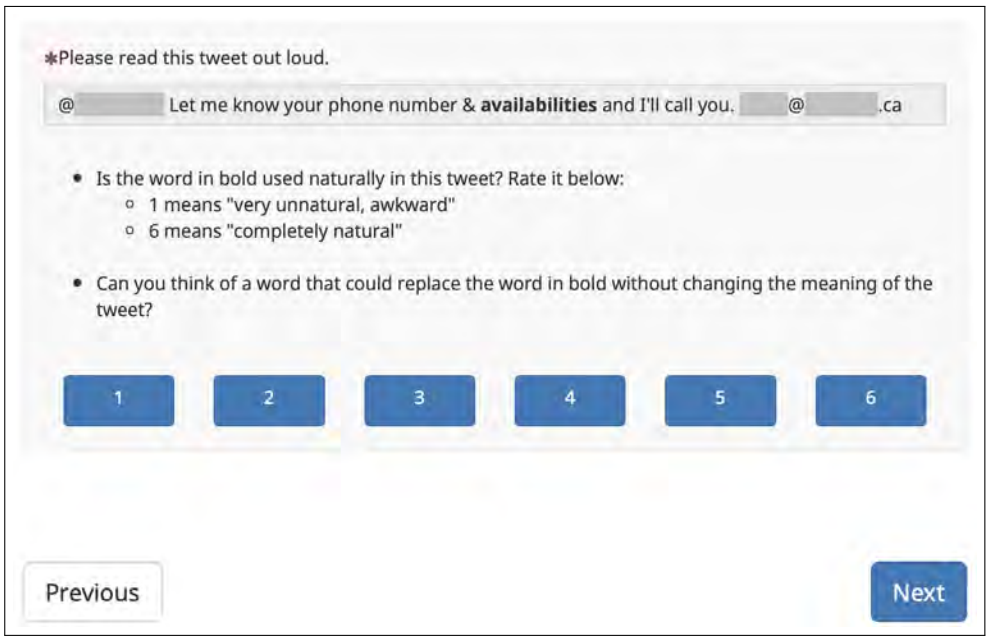


FIGURE 12.1: Screenshot of a semantic perception question on LimeSurvey

12.2 Deploying the protocol

This section presents the general context in which data collection in Quebec took place, as well as the approach I adopted in recruiting the participants, recording the interviews, and analyzing the obtained data.

12.2.1 General context of the fieldwork

The research protocol used in this dissertation received approval of the Ethics Research Board of the University of Toulouse (project number 2021-396). The sociolinguistic interviews were conducted in between mid-January and mid-February 2022. They were carried out as part of a research internship which I undertook at the Université de Sherbrooke, and which was supported by a grant from the Fonds de recherche du Québec. As such, this study was also subject to evaluation by, and received approval of, the Ethics Research Board of the Université de Sherbrooke (project number 2022-3289).

On a more practical note, it is important to underscore that the fieldwork was postponed multiple times due to the Covid-19 pandemic. Despite all reasonable precautions, my research stay in Quebec coincided with the introduction of reinforced public health restrictions in the province. During most of the period set aside for the interviews, access to universities and libraries was limited, businesses such as cafés and restaurants were closed, and private gatherings were severely restricted. In addition, the public health context required that precautions be taken in face-to-face interaction with potential participants so as not to expose them to any undue risks. These events are not anecdotal; rather, they strongly impacted my ability to both recruit potential participants and conduct in-person interviews. These issues are further referenced below.

12.2.2 Recruiting the participants

The number of participants in studies conducted within the PAC-LVTI framework ranges from an initial recommendation to recruit between 10 and 20 speakers (Durand and Przewozny, 2012, p. 27) to subsequent surveys including over 60 (e.g. Chatellier, 2016); in her earlier Montreal study, Rouaud (2019b) recruited 15 speakers. The participants should belong to dense social networks (Milroy, 1987), i.e. close-knit communities exhibiting a higher degree of linguistic stability. The sample is usually limited to speakers having completed most of their education in the community under study, it should ideally be balanced for gender, and include three age groups (Durand and Przewozny, 2012, p. 27). This echoes common sampling criteria applied in sociolinguistic studies in general, previously discussed in Chapter 4.

Potential participants were approached through a range of strategies: contacting student associations and university instructors, who then relayed the information on the study through their networks in Montreal; putting up posters advertising the study in busy public places (e.g. libraries, corner stores), mainly in highly bilingual neighborhoods such as Notre-Dame-de-Grâce; and contacting personal acquaintances. Active in-person recruitment was not practicable given the public health restrictions at the time. The information provided to potential participants included a brief explanation of the interview protocol and its general aims, without revealing the precise object of study. They were invited to express interest in participation by email; in response, they were provided with an informed consent form that they were asked to sign and return by email before the interview. The form explained the structure of the protocol in more detail, and outlined privacy-related safeguards (e.g. data anonymization, secure storage, right to withdraw from the study). Taking into consideration the public health context, the participants were free to choose between in-person participation (subject to legal feasibility and health protection measures) and remote participation via Zoom. A total of 15 participants were recruited; all but one decided to participate remotely.

While the present study follows common PAC-LVTI practice in terms of sample size, the sampling criteria were modified. This is related to two main reasons. The first has to do with the definition of Quebec English adopted in this dissertation and presented in Chapter 2. I consider this variety to encompass all use of English in the province, independently of the specific profile of the speaker. Consequently, the sample was not restricted in terms of the speaker's place of origin, native language, or broader linguistic profile. Recruitment materials specifically stated that eligibility extended to "all Montrealers aged 18 or over who are able to conduct a conversation in English". The rather broad formulation was voluntary: in addition to reflecting my view of the speech community, it is parallel to the criteria driving the identification of users of interest for the Twitter corpus, i.e. those for whom at least one English tweet had been collected, and who stated that they lived in Montreal. This choice was aimed at facilitating a comparison of the results produced using the two data collection methods.

The second reason behind modifying the sampling criteria is related to practical challenges in recruiting participants. The response rate to most of the approaches to recruitment was very low or null. I suspect that this might be related to the context of a pandemic surge in which the fieldwork took place. This issue, coupled with limited time to conduct the interviews, led

me to constitute a convenience sample without attempting to control for sociodemographic characteristics. A sample balanced for key features such as the speakers' linguistic profiles would have been preferable; given the practical constraints, it was not feasible. As the detailed description of the sample in Chapter 13 will show, its structure is too heterogeneous to provide a basis for quantitative observations that could be expected to generalize to the wider speech community. However, it includes a wealth of sociolinguistic profiles and behaviors which can be used to formulate and refine hypotheses, as well as to provide a detailed qualitative account of contact-induced semantic shifts. To this extent, the interview protocol fulfills its main purpose of complementing the computational analyses presented in Part III.

12.2.3 Recording the data

As stated above, most interviews were conducted using the Zoom video conferencing platform. They were recorded using Zoom's built-in recording functionality; only the audio recording was retained. Recording quality varied depending on the informant's microphone, as well as both my and their internet connection. As a result, the recordings are overall of a lower quality compared to those produced using standard recording devices during in-person interviews. Nevertheless, the quality is sufficient to observe general phonological trends, establish the informants' sociolinguistic profiles, and analyze the use of semantic shifts. The material used for the reading tasks was shown to the participants using the screen sharing function in Zoom. For readability, both the word lists and the text were split over multiple slides, unlike in the printed version. The reading tasks were followed by the formal conversation, which was in turn followed by the semantic perception test.

For the final task, the participants were provided with a link to the online LimeSurvey interface. They were asked to leave Zoom running in the background, so as to ensure the continuation of the video call, and to respond to the questionnaire in parallel. This setup proved to be surprisingly efficient, routinely leading to ample qualitative remarks regarding the use of the tested semantic shifts. Note however that not all participants interpreted the instructions for the task in the same way. In particular, the requirement to provide an alternative for the target lexical item was not universally respected. In order to avoid putting undue pressure on the informants, I only repeated this instruction a limited number of times. Where that was not efficient, I asked for clarifications regarding the examples which had previously proven to be difficult to interpret.

Following this task, the participants were briefed on the precise object of study, in line with ethics requirements. It was also at this stage that I discussed the informal conversation with another interlocutor. Given the previously mentioned difficulties in recruitment, this task was presented as optional; moreover, the considerable length of the formal conversations provided ample – albeit imperfect – data reflecting spontaneous speech production. In addition, over the course of the initial interviews, it became apparent that the very final part of the conversation (following the semantic task) was highly conducive to relaxed and direct discussions of the rated examples as well as bilingual behaviors in general. In the subsequent interviews, I aimed to actively foster these exchanges. They were central in identifying the representations

associated with the semantic shifts as well as validating the experimental setup.

A total of 18 hours and 40 minutes of interviews were recorded; the mean recording duration is 1 hour and 15 minutes (min = 56 min; max = 1 h 37 min). Only one interview was conducted in person. In this case, the reading materials were printed out, but the same LimeSurvey interface was used for the semantic perception test. Note moreover that three speakers declined to take the test due to time constraints, but they subsequently filled it in online, without the presence of the interviewer. As no phonological or qualitative information on the semantic shifts was recorded in these cases, only the numerical scores were retained for analysis. They were limited to the lexical items for which the targeted (contact-related) interpretation was reported by all remaining participants.

A note is also due on my own role as an interviewer. I have already discussed my position as a speaker external to the community under study regarding the analysis of the Twitter corpus (see Chapter 9). For the same reason, precautions also apply to my participation in face-to-face interviews. My position as external to the community might affect some sociolinguistic behaviors, particularly if they are subject to accommodation phenomena. It is unclear to what extent that was the case in the interviews. Nevertheless, it is important to be aware of this potential confound in interpreting the results, even though it is attenuated by the fact that it applies to all participants in the study.

12.2.4 Analyzing the data

Following standard PAC-LVTI practice, recorded audio files were segmented according to the interview tasks, and renamed using anonymized speaker codes coupled with suffixes corresponding to each task (see Appendix D). The semantic perception test was marked using the suffix *x*, and the final part of the conversation following that task was marked using the suffix *y*. The information provided in the formal conversation was entered into a spreadsheet corresponding to the full set of sociolinguistic descriptors elicited in the protocol. Key parts of the interviews were orthographically transcribed to allow for a further analysis; the same applied to samples of informal conversations.

Phonological information was analyzed perceptually in order to establish the informants' general profiles. Similarly, the target lexical items in the semantic perception test were analyzed so as to determine if they were fully integrated into the speaker's English phonological system. For each lexical item, I also noted the synonym provided by the informant and qualitative remarks regarding their use (if any). This information was used to ensure that only the correctly interpreted examples were retained for analysis, as well as to identify the representations associated with their use; I will come back to these issues in Chapter 14.

In analyzing the speakers' linguistic profiles, several numerical scores were derived from the qualitative information provided throughout the interviews. The way in which they are calculated is defined below; the scores are deployed to describe the participants in Chapter 13.

12.2.4.1 Degree of bilingualism

The score used to estimate the degree of bilingualism directly replicates the procedure introduced by Rouaud (2019b), who used the same core protocol for the same speech community. It is based on rating a wide range of information related to the use of languages which is elicited in the interview. The scores attributed to different characteristics reflect the importance that they are expected to have on bilingual language use. This way of formalizing qualitatively obtained background information is well-established in variationist sociolinguistics; Rouaud's score is similar in nature to the Language Use Index developed in the Dialect Topography Project (Chambers and Heisler, 1999), to give one example.

The complete range of information taken into account in calculating the score is presented in Table 12.1. It includes two types of self-reported information: the speaker's proficiency in and frequency of use of the language in question. This is complemented with overtly stated information on the age and manner in which the language was acquired; the score prioritizes early acquisition in natural contexts. For all four categories, the maximum score is retained. Finally, two types of information are indirectly inferred over the whole conversation: the domains in which the speaker actively uses the language and their passive exposure to it. In this case, all individual instances are summed together; the different weight attributed to them is reflective of the underlying importance that they are expected to exert on language use. For a more complete discussion of the evidence underpinning the formulation of the scoring system, see Rouaud (2019b, pp. 204–208).

Proficiency		**Age of acquisition**		**Domains of use**	
Basic	1	Early infancy	4	Home	4
Intermediate	5	Childhood	3	Extended family	3
Fluent	10	Teen age	2	Friends, colleagues, classmates	3
		Adulthood	1	Work, school	3
				Other	1
Maximum	10	Maximum	4	Maximum	14
Frequency of use		**Mode of acquisition**		**Passive exposure**	
Rarely	1	Home	6	Parent(s)	5
Monthly	5	French kindergarten, school	5	Partner	5
Daily	10	French immersion	4	Extended family	4
		French classes at school	3	Friends, colleagues, classmates	4
		CEGEP, university	2	Neighbors	3
		Work	2	Media	3
		Other	1	Other	2
Maximum	10	Maximum	6	Maximum	26

TABLE 12.1: Scoring system for language use, adapted from Rouaud (2019b, pp. 205–206).

The maximum value that can be produced by the scoring system is 70; this is then normalized to a range between 0 and 1. Note that the score was originally developed in a study that only recruited native English speakers who were additionally proficient in French to varying degrees. To that extent, it was used as a score assessing French proficiency, rather than more

varied bilingual profiles.

Since the speakers I recruited display a wider range of linguistic backgrounds, I used the same procedure outlined above two calculate an English and a French score. As a result, I was better able to account for profiles including native French speakers and Allophone speakers. I further computed a composite bilingualism score by subtracting the French score from the English score. It theoretically ranges from -1, corresponding to a monolingual French speaker, to 1, corresponding to a monolingual English speaker. A score of 0 is indicative of a roughly comparable use of both languages.

12.2.4.2 Socioeconomic status

A scoring system was used to assess the informants' socioeconomic status. I replicated the system already developed for the Montreal context by Rouaud (2019b, p. 188). It is based on five types of information: the informants' occupation, the occupation of their parental bread-winner, education, housing type, and neighborhood. Each of these is scored on a scale from 1 to 6; the maximum theoretical score is 30. Note that I slightly modified the system with respect to the scores attributed to the neighborhood, which was originally linearly estimated based on the proximity to the downtown core. I instead directly adapted it to the neighborhoods of the participants included in the sample based on differences in median household income. The scoring system is summarized in Table 12.2.

Occupation		**Parents' occupation**	
unemployed	1	unemployed	1
blue-collar unskilled worker	2	blue-collar unskilled worker	2
blue-collar skilled worker	3	blue-collar skilled worker	3
white-collar (sales, administrative assistant)	4	white-collar (sales, administrative assistant)	4
white-collar (managing role, engineering)	5	white-collar (managing role, engineering)	5
entrepreneur, owner of a large company	6	entrepreneur, owner of a large company	6
Education		**Housing type**	
grade school	1	no permanent residence	1
high school	2	renting	2
Cegep-level	3	own apartment	3
Cegep graduate	4	own house	4
university graduate	5	own two residences	5
postgraduate, professional school	6	own large estate	6
Neighborhood			
(other)	2		
Saint-Hubert	3		
Hampstead	5		
Westmount	6		

TABLE 12.2: Scoring system for socioeconomic status, adapted from Rouaud (2019b, p. 188).

12.2.4.3 Attitudes towards language policies and language use

Following Rouaud (2019b, pp. 208–209), I scored the speakers' attitudes toward language policies in Quebec, which are elicited throughout the formal interview. These are expected to potentially influence bilingual language use. The scoring system is based on the following three categories:

- negative attitude (score = 0), corresponding to an overt expression of discontent with French language policies;
- neutral attitude (score = 1), in the case of a factual description of the practical effects of the policies without openly stating an opinion;
- positive attitude (score = 2), corresponding to an understanding of the role played by the policies or of the concerns motivating their use.

I further applied the same scoring system to attitudes towards language use in general, in an attempt to capture the extent to which the informants adopt a prescriptive view. This information is important in contextualizing the acceptability ratings provided in the final task. The following categories were used:

- negative attitude (score = 0), if the speaker overtly expresses negative value judgments regarding specific linguistic features;
- neutral attitude (score = 1), if the speaker does not provide sufficiently explicit metalinguistic information to determine their attitude;
- positive attitude (score = 2), if the speaker overtly indicates acceptance of non-standard linguistic variants or of non-native proficiency.

12.2.4.4 Regionality index

The speakers recruited in this study present a range of geographic origins. In order to account for this information in a concise way, I adopted the Regionality Index developed by Chambers and Heisler (1999). Starting from a base value of 1, it assigns a score to (i) the place where the speaker was born; (ii) the place where the speaker was raised; and (iii) the place where the speaker's parents were born. Adapting the system to the Montreal context, the following scores are used for all three places:

- Montreal region = 0
- elsewhere in Quebec = 1
- outside of Quebec = 2
- outside of Canada = 3

Given the heterogeneous structure of the participant sample, I added an additional level to the original scoring procedure to explicitly account for speakers born outside of Canada. The minimum score is 1, corresponding to the informants who were born, raised, and continue to live in the Montreal region. The maximum score is 10, corresponding to the informants who live in Montreal, but were born and raised outside of Canada, as were their parents.

12.3 Summary

This chapter presented the variationist sociolinguistic protocol that I used to conduct face-to-face interviews with speakers from Montreal. I first reviewed the tasks comprising the protocol, starting with the core structure developed in the PAC-LVTI framework. Parts of the standard protocol – particularly the thematic questionnaire – were adapted to the local context and the object of study. In addition, the protocol was extended using a novel task designed to assess a large number of contact-induced semantic shifts in an interview setting. I outlined the motivations behind the structure of the task, the choice of examples used in it, and its practical implementation.

I then presented the way in which this protocol was deployed to collect data in Montreal. Building on the general context in which the fieldwork took place, I discussed the strategies that led to the recruitment of 15 participants, as well as the practical choices made in running the interviews and analyzing the recorded data. This included the discussion of three quantitative scores addressing the central issues in explaining the sociolinguistic behaviors observed in the data: the speakers' degree of bilingualism, their attitudes towards language policies and language use in Quebec, and their geographic origin.

While this chapter has highlighted a range of difficulties in data collection, the presented approach nevertheless enabled me to obtain qualitatively rich data, directly applicable to the study of contact-induced semantic shifts, and produced by a diverse group of speakers who are reflective of the wide range of linguistic profiles in Montreal. The data are analyzed in more detail in the remainder of this dissertation, starting with a description of the recruited sample in the next chapter.

Chapter 13

Establishing sociolinguistic profiles

Building on the discussion of the sociolinguistic interview protocol implemented in this dissertation, the present chapter provides a description of the participant sample. Section 13.1 presents the structure of the sample in terms of key sociodemographic characteristics. Section 13.2 highlights the main ideas reflective of the participants' identity and their view of language use in Montreal. Drawing on the whole range of available information, Section 13.3 introduces a multidimensional analysis, helping to identify distinct speaker profiles in the sample. Section 13.4 summarizes the main findings.

The key characteristics of the sample identified through this general overview will play a central role in accounting for variability in the perception of contact-induced semantic shifts, as shown in Chapter 14. Note moreover that the coming discussion will focus on the description of the interviewed speakers. The importance of key sociodemographic characteristics for the variationist sociolinguistic theory, including in the context in Quebec English, is outlined in more detail in Chapter 6. The most relevant parts of that discussion will be referenced throughout the present chapter.

13.1 Sociodemographic characteristics

This section breaks down the structure of the sample in terms of the main sociodemographic characteristics of the interviewed speakers. It specifically presents their age and gender, geographic origin, language use, and socioeconomic status. It concludes with a brief discussion of their participation in social networks.

13.1.1 Age and gender

As previously discussed, age and gender are extensively used to account for patterns of language variation. In previous studies on Quebec English, a common approach to variation across age has consisted in splitting the speakers into groups born before and after the passage of Bill 101 in 1977. The assumption here is that the younger age group is likely to exhibit stronger effects of contact with French due to it being directly affected by the effects of the bill (e.g. restrictions on access to English schools, presence of French public signage etc.; see Chapter 2

for an extensive overview). As for gender, it has mainly been applied to phonological variation, with women leading several changes in progress, although it often interacts with other factors. For a more general discussion of age and gender in variationist sociolinguistic studies, see Chapter 6.

The structure of the participant sample in terms of age and gender is presented in Table 13.1. The median age in the sample is 27; it ranges from 19 to 70. As suggested in the earlier discussion of the recruitment process, the sample is strongly skewed towards younger and female participants. This lack of balance, particularly in terms of gender, is frequently reported in sociolinguistic studies (e.g. Boberg and Hotton, 2015, pp. 285–286).

	Gender		
Age group	**Female**	**Male**	**Non-binary**
Pre-Bill 101	2	2	—
Post-Bill 101	8	2	1
Total	10	4	1

TABLE 13.1: Cross-tabulation of age and gender for the participant sample

In analyzing the results of the interviews, younger speakers are expected to present a higher rate of acceptability of semantic shifts due to their stronger exposure to contact with French. When it comes to the potential effects of gender, we might hypothesize that semantic shifts – to the extent that they are seen as nonstandard usage and as such are subject to a degree of stigmatization – are more readily accepted by men, as predicted by the classical formulation of Labov's gender paradox. However, given the highly uneven distribution of gender in the sample, caution is required in using it as an explanatory variable.

13.1.2 Geographic origin and current neighborhood

In terms of geographic patterns, two types of information will be taken into account: the informants' geographic origin, i.e. the place in which they were born and raised, as well as the geographic origin of their parents; and the part of Montreal in which they currently live. The first can be seen as a reflection of their representativeness of well-established local speech trends; the second is indicative of their everyday communicative context, including the likelihood of being exposed to the use of French.

The participants' geographic origin is formalized by calculating the Regionality Index (RI). The score ranges from 1, for participants who were born and raised in Montreal, as were their parents, to 10, for participants who were born and raised outside of Canada, as were their parents; for the full scoring procedure, see Chapter 12. The histogram in Figure 13.1 shows that the whole range of values is present in the sample. They allow for a balanced division of informants into two groups: those with scores ranging from 1 to 4 (N = 8), who were all born and grew up in Quebec, most of them in Montreal; and those with scores ranging from 7 to 10 (N = 7), who were all born outside of Quebec, and in most cases grew up outside of the province as well.

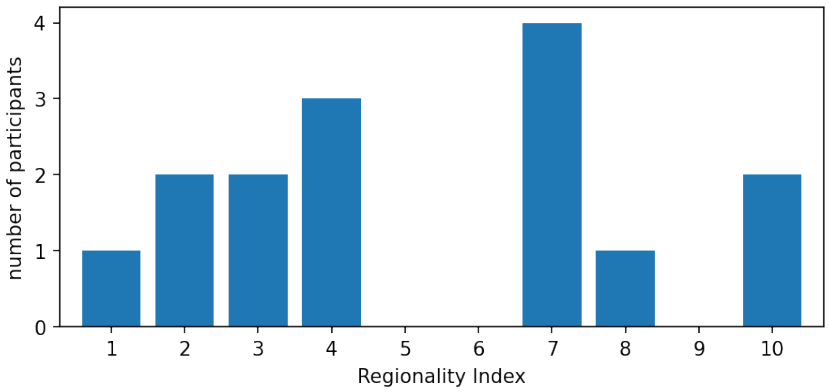


FIGURE 13.1: Distribution of participants in terms of the Regionality Index

The specific speaker profiles within the two general groups should also be highlighted. The speakers with strong local ties (RI < 5) include those born into well-established Anglophone families in Montreal; second-generation and third-generation immigrants born in Montreal; and Anglophone as well as Francophone Quebecers born elsewhere in the province. The speakers with weak local ties (RI > 6) include individuals who were born and grew up in other Canadian provinces; and first-generation immigrants, arriving both at an early age (and subsequently growing up in Montreal) and as adults.

A breakdown of the neighborhoods in which the participants live is presented in Table 13.2; their position is indicated on a map of Montreal in Figure 13.2. They can be roughly grouped into three categories based on their linguistic profile. The adjoining neighborhoods of Hampstead, Notre-Dame-de-Grâce, and Westmount, located just west of downtown Montreal, are comparatively the most English-speaking; the reported rate of knowledge of French is around 80%. An intermediate category is constituted by the similarly adjoining areas of Ville-Émard, Verdun, and Little Burgundy. Located in Montreal's southwest, and just south of Notre-Dame-de-Grâce, they exhibit a somewhat higher rate of French knowledge, at around 90%. Among the reported neighborhoods, the most strongly French-speaking are Little Italy and the wider Rosemont area, to the east of the downtown core, as well as Saint-Hubert, a borough of the city of Longueuil on Montreal's South Shore. French knowledge is reported by well above 95% of residents in these areas.

		Neighborhoods	**N**
more French	↑	Little Italy, Rosemont, Saint-Hubert	3
		Little Burgundy, Verdun, Ville-Émard	6
less French	↓	Hampstead, Notre-Dame-de-Grâce, Westmount	6

TABLE 13.2: Participants' neighborhoods in Montreal, grouped by degree of exposure to French. The total number of participants is provided for each group.

The differences between these neighborhoods may appear to be negligible, but they are reflective of observably different linguistic experiences. While they are all part of the same

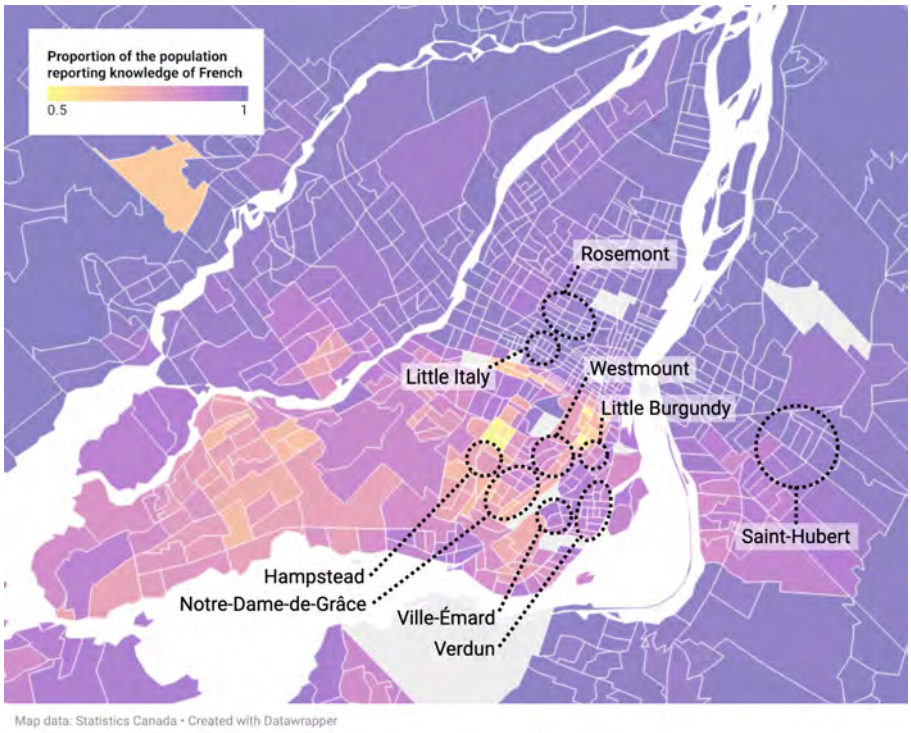


FIGURE 13.2: Approximate location of the informants' neighborhoods superimposed onto a map of the Montreal Island and part of the surrounding area. Color coding reflects the proportion of the population reporting knowledge of French (including jointly with English). The map was created based on the data from the 2016 Census (Statistics Canada, 2017d). Multiple speakers reported that they live in Notre-Dame-de-Grâce (N = 4); Ville-Émard, Verdun, and Little Burgundy (N = 2 for each neighborhood).

metropolitan area,[1] a walk around the neighborhoods of the west and the east end of Montreal suffices to observe – or more precisely, hear – clearly distinct trends in terms of language choice. This impression of mine is corroborated by the informants' reports on language use in different neighborhoods, further discussed below. It is significant in accounting for potential effects of language contact because it indicates potentially different degrees of exposure to French on a daily basis.

An impact of both geographic variables can be hypothesized with regard to contact-induced semantic shifts. Higher acceptability ratings are expected (i) for informants with a lower Regionality Index, corresponding to a higher degree of integration in the local community; (ii) for informants living in neighborhoods with greater exposure to French. The first assumption reflects the view of contact-induced semantic shifts as a regionally specific phenomenon; the second is more directly related to the effect that individual bilingualism, including passive exposure, may have on this phenomenon.

13.1.3 Language use

The impact of individual bilingualism mentioned above has also been measured more directly. The participants provided precise self-reported information on the degree of proficiency, fre-

[1] In administrative terms, most of these neighborhoods are also part of the same city – the City of Montreal. The exceptions are Westmount and Hampstead, which constitute towns in their own right, and the previously mentioned case of Saint-Hubert, which is part of the city of Longueuil.

quency of use, age and manner of acquisition of all languages that they speak. They also provided extensive indirect information on the domains of use and passive exposure to the languages.

In general terms, it is important to note that no speakers in the sample are monolingual. They all have at least some knowledge of and exposure to both English and French. Most (N = 12) reported knowledge of additional languages, up to six in total for two speakers. The languages beyond English and French are often heritage languages, i.e. those spoken in the country to which the respondents' family traces its origins. Knowledge of additional language is relevant to note because they may also be involved in processes of cross-linguistic influence.

That said, the focus here remains on English–French bilingualism. As explained in Chapter 12, each participant was graded for their knowledge of and exposure to both languages in order to produce a language use score; it theoretically ranges from 0, for a complete lack of use of a language, to 1, for native-like proficiency and use of a language in all contexts. A composite bilingualism score was also computed by subtracting the French score from the English score in order to estimate the relative importance of the two languages. (For the full scoring procedure, see Section 12.2.4.1.) The plot in Figure 13.3 presents the scores for all speakers in the sample.[2]

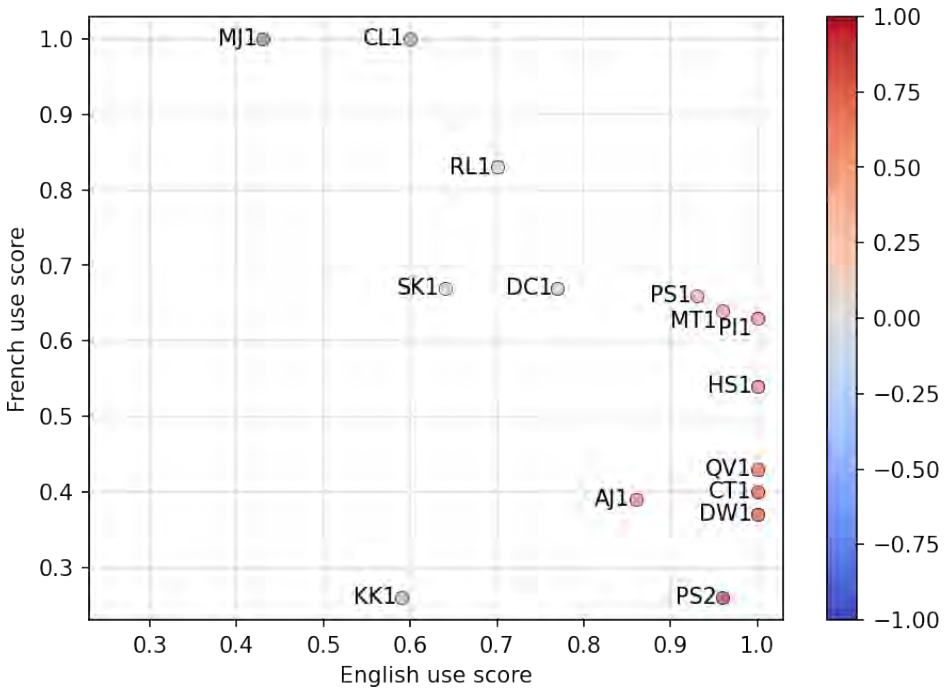


FIGURE 13.3: Distribution of informants in terms of language use scores. The scores for individual languages are plotted on the X-axis (for English) and on the Y-axis (for French). The composite bilingualism score is color-coded; negative values indicate predominance of French, and positive values indicate predominance of English.

The plot points to several distinct profiles of bilingual speakers. (i) The first clearly identifiable group is located in the lower right corner of the plot, with an English use score above 0.9. These informants – constituting one half of the sample – are all native English speakers.

[2]In line with standard anonymization practices in the PAC protocol, individual speakers are designated using a code based on their initials.

The relative difference in their English use scores is mainly related to specific areas of their lives in which they lack exposure to English. In contrast, the differences in the French use score are indicative of substantially different ability and extent of French use. Interestingly, the speakers located higher in the plot – i.e. those who are more proficient in French – also tend to be younger; we will come back to this interaction between sociodemographic factors at the end of the chapter.

Several smaller groups of speakers can also be identified. (ii) The two participants with the maximum French use score are native French speakers who are also proficient in English. (iii) The group of three individuals forming a triangle roughly in the middle of the plot corresponds to speakers who are highly proficient in both English and French. They are first-generation immigrants arriving at an early age or, in the case of SK1, a second-generation immigrant. Their less-than-maximum English and French scores reflect the use of a heritage language in some areas of their lives. (iv) The two remaining speakers, located in the bottom part of the plot, are both Allophone first-generation immigrants who arrived in Montreal as adults. The difference in their language scores mainly comes from the language used at home: KK1 speaks a heritage language, and AJ1 speaks English.

This overview points to the presence of a variety of bilingual profiles in the sample, not unlike those that are characteristic of the wider Montreal population. It also underscores the importance of assessing the proficiency and use of both languages involved in a contact situation,. It further highlights the fact that balanced bilingualism – indicated here as a composite bilingualism score of around 0 – does not translate to an identical and native-like use of all languages in all situations; this is in line with Grosjean's complementarity principle in bilingual language use, introduced in Chapter 1.

In terms of the potential relationship between bilingualism and contact-induced semantic shifts, it is reasonable to expect higher acceptability ratings among speakers (i) who use French to a greater extent, (ii) whose English is less well-entrenched, and (iii) who use French more than English. This specifically corresponds to (i) speakers with a higher French use score, (ii) speakers with a lower English use score, and (iii) speakers with a lower composite bilingualism score. That said, this discussion has reaffirmed the fact that a wide variety of qualitative patterns are subsumed by general quantitative scores; this should also be taken into account in interpreting the results.

13.1.4 Socioeconomic status

In estimating socioeconomic status, I used a scoring system which takes into account the informants' occupation and education, their parents' occupation, their housing, and the neighborhood in which they live. As described in more detail in Section 12.2.4.2, the maximum resulting score is 30; it is split into five categories, each of which spans six points. The categories are taken to correspond to social classes, the most relevant for us being the middle class (score 13–18) and the upper middle class (score 19–24).

The participants present relatively limited divergence in terms of their socioeconomic status. The scores range from 15 to 23; the median score stands at 19, right at the transition

between the middle class and the upper middle class. An important source of difference is the participants' neighborhood, with the median income in Hampstead and Westmount roughly two times higher than in most other reported neighborhoods of residence. Other reasons driving the differences are related to the degree of education, occupation, and housing type. However, this is likely explained by the correlation of socioeconomic status with age ($\rho = 0.67$), which indicates that the main differences captured by the score may largely be a reflection of the gradual development of economic capacity over the lifespan. Coupled with the overall homogeneity, this would suggest a relatively limited importance of socioeconomic status in the present sample. This impression closely reflects the results reported by Rouaud (2019b, p. 189) for the West Island of Montreal.

13.1.5 Social networks

In discussing the PAC-LVTI protocol in Chapter 12, I introduced the general principle according to which participants should be recruited in dense social networks; consequently, it is often possible to establish fairly precise links between the informants in a sample. As previously stated, the practical constraints impacting recruitment in this study did not allow for that criterion to be closely followed, but basic trends can nevertheless be outlined.

Given the variety of recruitment approaches implemented in this study, it is unlikely that most participants are familiar with the others. A clear exception to this rule is constituted by two informants: KK1 and her daughter SK1. More generally, close to half of the participants (AJ1, CL1, DC1, HS1, MT1, PS1, and RL1) were recruited through the same student association at Concordia University. While it is unclear if they are directly familiar with one another, it can be assumed that they partly participate in similar patterns of interaction. However, given the lack of precise information on social networks in the participant sample, their use in interpreting the results will remain limited.

I have so far outlined the general structure of the participant sample recruited for this study. In terms of age and gender, the sample is skewed towards younger and female speakers. The informants exhibit a variety of geographic profiles, both in terms of their origin and of their reported neighborhoods of residence. All informants speak both English and French, but the specific relationship between the two languages is variable, ranging from predominantly English-speaking to predominantly French-speaking individuals. The informants are of a relatively similar socioeconomic status. The impact of social network structure on the behaviors observed in this study is moreover expected to be limited. Building on this factual description of the sample, the next section provides more interpretative context by discussing the identity and attitudes reported by the speakers.

13.2 Identity and attitudes

This section reviews the qualitative evidence reported by the speakers regarding their individual sense of identity, as well as their view of Montreal and of communicative patterns in the

city. These observations draw on the formal conversation based on the thematic questionnaire, presented in Chapter 12.

13.2.1 Individual sense of identity

The informants were asked a series of questions regarding their own identity in relation to the city and the wider Canadian context, which is essential in interpreting their sociolinguistic behaviors and the social meanings that they might aim to convey. We can begin to understand their profiles by examining the answers to the following two questions:

- Do you consider yourself a Canadian, a Quebecer, a Montrealer, or a [West Islander, Westmounter, NDGer, etc.]? If so, in which order? Why?
- Do you feel that you're a true Montrealer? What does that mean for you?

Key aspects of the answers, together with main sociolinguistic descriptors provided for context, are presented for all informants in Table 13.3.

					Montrealer	
Speaker	**Age**	**RI**	**Biling.**	**Identity**	**Self**	**Definition**
PS1	33	3.5	0.27	Canadian	yes	—
HS1	27	7.0	0.46	Canadian	no	bilingualism
QV1	32	7.0	0.57	Canadian	yes	diversity, tolerance
KK1	54	10.0	0.33	Canadian	yes	bilingualism
PS2	70	2.5	0.70	Anglo-Quebecer	—	—
CT1	45	4.0	0.60	English Quebecer	yes	diversity, regional French
MJ1	27	3.5	-0.57	Quebecer	yes	integration, diversity
DC1	22	7.0	0.10	Romanian Quebecer	yes	diversity, tolerance
MT1	24	1.0	0.31	Montrealer	yes	integration
DW1	70	2.5	0.63	Montrealer	yes	French culture
SK1	19	4.0	-0.03	Montrealer	yes	bilingualism, joie de vivre, tolerance
CL1	23	4.0	-0.40	Montrealer	yes	diversity, regional French
PI1	65	8.0	0.37	NDGer[a]	no	integration
RL1	25	7.0	-0.13	other	unsure	integration
AJ1	26	10.0	0.47	other	no	bilingualism, diversity

TABLE 13.3: Summary of speaker profiles and key self-reported identity information. RI: Regionality Index (lower values indicate more local origin); biling.: composite bilingualism score (negative values indicate predominance of French, positive values indicate predominance of English); identity: the highest-ranked identity descriptor; Montrealer: answer to the question "Do you feel that you're a true Montrealer?". (a) NDGer: an inhabitant of Notre-Dame-de-Grâce.

As shown in the table, there is no unifying trend which could immediately characterize this sample. However, all speakers provided well thought-through and clearly articulated answers to identity-related questions, suggesting a strong relevance of these issues. While the above summary is useful in distinguishing between potentially coherent subgroups of speakers, it is underpinned by more complex expressions of identity.

The term *Canadian* is geographically and conceptually the broadest. It is cited as the most important identity term by a well-established Montrealer (PS1), two speakers who were born and raised elsewhere in Canada (HS1, QV1), and a first-generation immigrant (KK1). For the last three speakers, its use also conveys a hesitance to assert a link with Quebec. Commonly invoked reasons are presented particularly clearly by QV1, who was born in Prince Edward Island but has lived in Montreal for close to 10 years:

> I don't think I would ever identify myself as Québécois. Just because I feel like, like, as much as they feel Anglos like coming in and speaking English in their spaces is like appropriative, I feel like claiming a Québécois identity is super appropriative, like 'cause I don't have any of the Québécois culture. Like I feel like Montreal culture is sort of outside of that. Or it's like a Venn diagram where they're just overlapping. And yeah, there's also just like a lot of negative connotations with the Québécois identity.

Other participants are more at ease with the term *Quebecer*. This group includes MJ1, a native French speaker who self-identifies as "a well-assumed separatist", overtly distancing herself from a Canadian identity. The term is also used by native English speakers born and/or raised in Quebec, who qualify it in order to more clearly position themselves. For instance, CT1 describes herself as "an English Quebecer with a tri-cultural background living in Montreal". She further clarifies the central role of language in Quebec identity:

> I'm very proud of the fact that I'm an English Quebecer. [...] I speak French, but it's not my culture, which I'm ok with it.

Four speakers in the sample identify the most immediately with Montreal, and another with the neighborhood of Notre-Dame-de-Grâce. Like in the previous two groups, the speakers represent the full range of profiles in terms of age and geographic origin. All of them also identify as Canadians and/or Quebecers. Those who discuss their unwillingness to assert a Quebec identity note its association with provincial politics to which they do not generally subscribe. Finally, two speakers – who are both first-generation immigrants – avoid using any of the Canadian terms, prioritizing those that are associated with their country of birth, but they nevertheless express appreciation of the local community.

Overall, the informants appear to more readily claim the broad Canadian identity or the highly local Montreal identity, both of which they associate with diversity and acceptance. The provincial identity appears to be more politically charged and to involve more decisive positioning. That said, the participant sample is too small, and the trends it captures too complex, to derive any reliable generalizations about specific expressions of identity; it is however clear that in Montreal they are closely associated with language use. Moreover, most participants (11 out of 14 who answered the question) would describe themselves as "true Montrealers". Although the precise definitions associated with this descriptor are somewhat variable, the general trend is indicative of a high degree of perceived integration in the local community. This issue is further addressed in the next section.

13.2.2 Life and language in Montreal

This section summarizes the views expressed by the participants throughout the formal interview regarding three broad topics: their experience of life in Montreal; bilingualism and the related issue of language policies in Quebec; and the way English is spoken in Montreal, both in terms of linguistic features and patterns of interaction.

13.2.2.1 A general view of the city

All participants, without exception, express highly positive views of Montreal. This importantly extends to informants with intermediate and weak local ties. They describe Montreal as "incredibly welcoming" (QV1), "this lovely big city that, you know, I'm privileged to live in" (PI1); in a word, "Montreal feels like home" (CT1). In discussing the specific characteristics that they appreciate, some speakers point to the general atmosphere in the city, a *je ne sais quoi* that is unique to it. In the words of SK1:

> There is this thing that Montreal has, I'm not quite sure what it is, but other places in Canada don't necessarily have.

Others are more precise in pinpointing the positive aspects, which are often related to multiculturalism, as DC1 explains:

> We're a lot more accepting of whoever comes here. [...] If you come here as an immigrant you can immediately try to relate to a certain group and from then being in that group, since you're next to all these other little groups, it's almost impossible not to mingle. So you definitely feel more open.

This is particularly appreciated by people who arrived in Montreal only recently. Take for example the view expressed by AJ1, who moved to the city form her home country of Brazil around two years ago.

> Everyone's from somewhere else, everyone has like a different way of doing things and different habits and different cultures, so you're just like one more, you know, like the weird person.

These remarks overall reflect the features associated with being a "true Montrealer" (summarized in Table 13.3). Multiculturalism and acceptance of diversity – linguistic and otherwise – are seen as defining characteristics of the city; many participants overtly take pride in them. More generally, this is indicative of the relative ease with which individual speakers, including those arriving as adults from other countries, can establish links with local communities.

13.2.2.2 Bilingualism and language policy

All participants express overall positive views towards bilingualism in general and the use of French in particular. In discussing bilingualism, most speakers (N = 11) express views coherent with the definition adopted in this dissertation (cf. Chapter 1). For them, being bilingual

generally corresponds to being able to communicate in two languages, or potentially to work in two languages. The remaining four speakers recognize multiple types of bilingualism, often drawing a distinction between basic communicative ability and using languages in specific contexts. Three informants (out of 14 who answered the question) would not describe themselves as bilingual due to a perceived lack of French knowledge. Interestingly, two of them are not among the speakers with the lowest French use scores overall, pointing to likely linguistic insecurity.

When it comes to language policies used to promote the use of French in Quebec, most participants express neutral views. Two note the potential for rare issues to arise – described as "backlash" (MT1) or "hostility" (PS1) – related to not being Francophone. Another three participants discuss language policies as politically challenging, but having little negative impact on their day-to-day life in Montreal. Perhaps the most critical view is expressed by QV1, a native English speaker who is also critical of their own limited French proficiency.

> I feel like the, the steps that they take in order to like protect the French language are typically a lot less about protecting French and more about eliminating English. [...] They're not trying to make it easier for me to learn French. They're trying to make it harder for me to speak English.

The most supportive position on French language policy is taken by MJ1. She is a native Francophone Quebecer who actively advocates for a more widespread use of French, which she feels is threatened by the majority status of English in the wider North American context.

> Every time we try to have like linguistic policies to protect French, it's like, "Oh yeah, but, but what about the English minorities in Quebec?" Yeah but you're a minority in Quebec, but you're an, overall you're not.

Note that this view is grounded in a strong link between language and identity.

> At least in my case, the language is like the center, like a very, very strong center of like my identity. That, that's what I am, and if you take apart the French part of it, I think that I would be less myself I guess.

Other speakers also attach a similar degree of importance to their languages. Take for example KK1, who immigrated to Canada from the former Yugoslavia in the 1990s, together with her partner. In discussing the transmission of their native Serbian to their two children born in Canada, she notes:

> Keeping Serbian was really important for us. We don't go to the church, we don't go, we don't keep any tradition. I mean, we are not big traditionalists, but we think that language is the most important, that's for our, us religion, you know.

More generally, like most other informants in the sample, she expresses a positive view of the obligation (or possibility, depending on the profile of the family) for her children to

be educated in French. In these discussions, the practical value of bilingualism, particularly in the workplace, is routinely noted. In the same line of thought, some of the older participants express regret at not being more proficient in French themselves. Echoing the theoretical discussion in Chapter 1, the remarks on bilingualism overall confirm a strong link between language and identity for the informants in the sample.

13.2.2.3 Speaking English in Montreal

In discussing their perception of language variation, most speakers (N = 12, out of 14 who answered this set of questions) claim that they can distinguish broad geographic distinctions, like the ones between Canadian and American speakers; this is often based on stereotypical phonological differences. The same speakers also report that they are able to distinguish the use of English in Montreal from the rest of Canada to some extent. Several informants, both native to Montreal and recently arrived from other Canadian provinces, suggest that features typical of Canadian English pronunciation are less prominent in Montreal. These specifically include Canadian Raising (*aboot*, instead of *about*, being the universally cited example) and lexically-specific resistance to *r*-conditioned vowel mergers (e.g. *sorry* preserving a rounded realization). The same applies to the discourse marker *eh*, a marker of Canadian English.

When it comes to identifying linguistic features specific to the way English is spoken in Montreal, five informants suggest that a variety typical of the whole city is unlikely to exist given the widespread sociodemographic diversity. This is further supported by occasional discussion of linguistic features specific to smaller communities. For instance, MT1, a third-generation immigrant of Italian and Greek descent, associates neighborhoods with large Italian and Greek populations with specific ways of speaking English. The distinctive character of English used by native French speakers is also noted, although several informants suggest that the stereotypical French accent is less pronounced in Montreal than elsewhere in Quebec.

That being said, lexical influence of French is nearly universally reported; it is usually illustrated with the example of *depanneur* ‘corner store’. Another widely discussed phenomenon is codeswitching. Older speakers tend to associate it with younger Montrealers, whose behavior is described by PS2 as follows:

> I hear them on the street! They’re speaking a little bit English, speaking a little bit of French, speaking a little bit English.

This trend is corroborated by younger speakers themselves. Take for example SK1, a highly bilingual 19-year-old:

> The way I speak English is not one hundred percent English. Yes. When I lose my words in English or in French I will replace it with a word of the corresponding language.

She also indicates awareness of constraints on codeswitching behavior (which she terms “Franglais”), in line with well-established empirical evidence on this process (cf. Section 1.3.1).

> When we speak Franglais or stuff, there's like a, we, there's a way of doing it, there's like a, we have a grammar in our heads that we know which words can be in English and can't when we speak.

The widespread nature of this tendency is further confirmed by speakers who arrived in Montreal more recently. After a year living in the city, HS1, originally from British Columbia, observes:

> Someone will forget a word and they'll use a French word and that's more accepted obviously than in other parts because generally we all know what's happening.

More generally, all participants report growing up and/or living in multilingual environments. Their integration in the local neighborhoods is variable, ranging from those who are largely unfamiliar with their neighbors to informants who are close friends with all those living nearby. Whatever the case, the interactions in the local context are either mostly French or bilingual.

In discussing the language they would be the most likely to use in addressing an unknown interlocutor in the street, five informants state that they would default to English. The others would either start in French or adapt to the specific situation. All would be ready to accommodate based on contextual factors including the neighborhood, overhearing the other person speak in a different language, or by asking for the preferred language. This parallels the view of the ability to choose a language appropriate to the interlocutor as "the Montreal thing" (PI1). The mechanisms used to reconcile English and French are particularly well illustrated by DC1:

> I would use both. So I would say, let's say, "Where is the Starbucks?", I would be like, "Oh, excusez-moi, désolé, sorry, do you know where the Starbucks is?"

Summarizing, these observations suggest that, on a metalinguistic level, the speakers in the sample recognize some influence of French on the way English is spoken in Montreal, mainly in terms of codeswithcing and borrowing. They also report extensive participation in and exposure to interactions in both languages. For most speakers, the use of both English and French is an everyday occurrence, likely facilitating cross-linguistic effects.

The participant sample has so far been described using a range of sociodemographic characteristics and expressed attitudes. The next section brings together the full range of information to provide a more comprehensive overview of the informants' profiles.

13.3 Identifying sociolinguistic profiles

Drawing on all the perspectives deployed to describe the participant sample, we can now outline a general overview of their sociolinguistic profiles, exploring more clearly how different characteristics interact. In order to do so, I conducted a principal component analysis; this method was previously discussed in Section 10.2 to analyze the patterns in the whole vocabulary based

on a range of quantitative information. While the perspective here shifts from lexical items to individual speakers, the reasoning behind the approach is the same. It enables an efficient analysis of the patterns captured by a series of input variables by producing principal components – new variables formed through linear combinations of the initial ones – which are uncorrelated (orthogonal) to one another. This is helpful in identifying complementary trends in the data.

Input variables included the whole range of information discussed in this chapter:

- age;
- gender (coded as follows: 0 = male, 1 = non-binary, 2 = female);
- Regionality Index, reflecting the speaker's geographic origin (higher values indicate a less local origin);
- the number of years spent living in Montreal;
- exposure to French in the neighborhood (based on the split of neighborhoods into three categories in Section 13.1.2, with higher values indicating more exposure to French);
- English use score;
- French use score;
- composite bilingualism score (negative values indicate predominance of French, positive values indicate predominance of English);
- socioeconomic status score;
- language used to ask for directions (coded as follows: 0 = French, 1 = mixed, 2 = English);
- attitude towards language policy (coded as follows: 0 = negative, 1 = neutral, 2 = positive);
- attitude towards nonstandard language use (same coding).

The input variables were mean-centered and scaled to unit variance before the analysis, which was conducted using the `statsmodels` implementation (Seabold and Perktold, 2010). The first two principal components were retained; together they explain 63% of variance. The position of individual speakers and input variables with respect to these two dimensions is plotted in Figure 13.4.

A key distinction in the data captured by component 1 (variation along the horizontal dimension) is related to the influence of age. The speakers in the left-hand half of the plot were all born after the passage of Bill 101; with two exceptions (AJ1 and CT1), those in the right-hand half were born in the period preceding it. Age is in turn related to language use: it is negatively correlated with the French score ($\rho = -0.72$) and positively with the English score ($\rho = 0.49$) and the English-dominated bilingualism score ($\rho = 0.70$).[3] This points to a key global trend in the sample indicating a higher rate of French use among younger speakers. As might be expected, positive attitude to French language policy and neighborhood exposure to French are associated with the same direction as the French use score; at the other end of the plot, the socioeconomic status points in a similar direction as age, likely reflecting the previously discussed correlation between the two variables.

Variation along the vertical dimension (component 2) involves the informants' geographic origin. Those with the lowest Regionality Index (born and raised in Montreal, in locally-established families) are located in the bottom third of the plot. Those with the highest Re-

[3]The critical value of Spearman's ρ for N = 15 observations is 0.52 at the 0.05 level of significance.

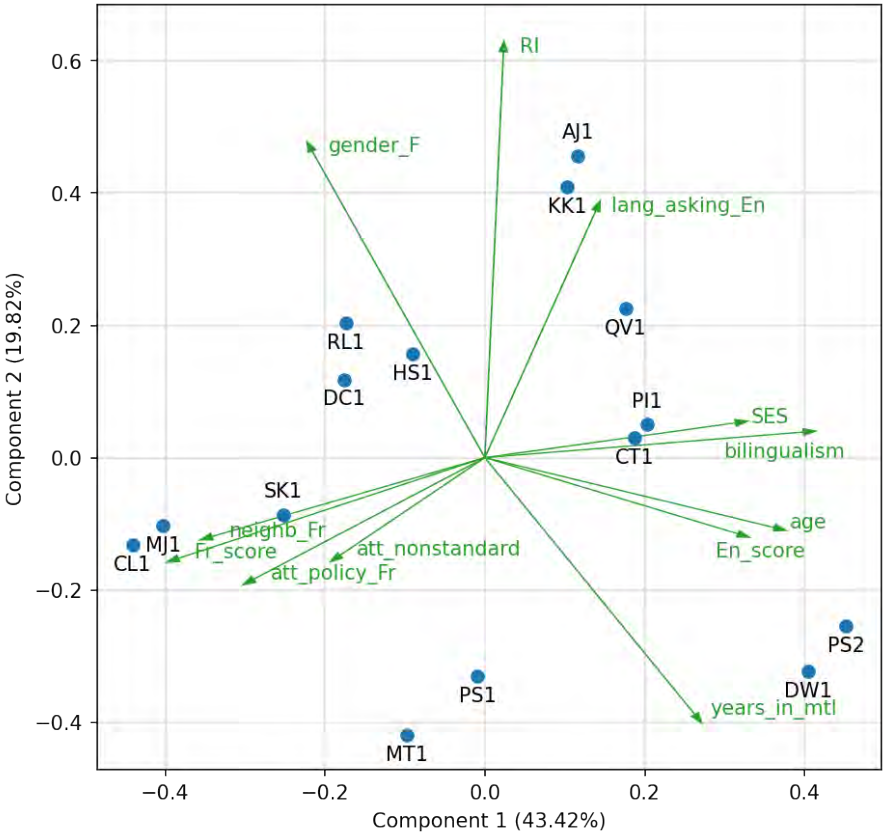


FIGURE 13.4: PCA biplot of informants and input sociodemographic variables. Abbreviations clockwise from top left: `gender_F`: gender, the highest coded value is female; `RI`: Regionality Index; `lang_asking_En`: language used to ask for directions, the highest coded value is English; `SES`: socioeconomic status; `En_score`: English use score; `years_in_mtl`: number of years spent living in Montreal; `att_nonstandard`: attitude towards nonstandard language use; `att_policy_Fr`: attitude towards French language policy; `Fr_score`: French use score; `neighb_Fr`: exposure to French in the neighborhood.

gionality Index (e.g. first-generation immigrants) are mostly located in the top third of the plot. The remaining central area corresponds to intermediate speaker profiles (e.g. those born elsewhere in Quebec or Canada; second-generation immigrants). The vertical dimension also involves gender, with all four male participants located in the lower third of the plot; however, this may simply reflect the chance fact that they were all born and raised in Montreal.

Note moreover that attitude towards nonstandard language use is correlated with a range of variables, the most strongly with socioeconomic status ($\rho = -0.74$) and age ($\rho = -0.50$). This suggests that younger speakers are more accepting of nonstandard linguistic features; incidentally, they are also the ones who speak French more. Finally, the tendency to initiate conversations in English exhibits a moderate positive correlation with the Regionality Index ($\rho = 0.43$) as well as a negative one with the French use score ($\rho = -0.43$). In line with intuitive expectations, this would indicate that the speakers who are not proficient in French and/or do not have a local origin are the ones most likely to initiate conversations in English.

Summarizing, this analysis points to several overarching trends in the data. One distinguishes younger, more French-speaking participants from older, more English-speaking ones. Another opposes speakers with strong local ties to those with weak local ties. Smaller, highly similar groups – often pairs – of informants can also be identified:

- MT1 and PS1 (bottom center) are younger, male, highly bilingual speakers with strong local ties;
- DW1 and PS2 (bottom right) are older, male, English-dominant speakers with strong local ties;
- CL1 and MJ1 (middle left) are younger, female, French-dominant speakers with intermediate local ties; and so forth.

Given the limited sample size, these observations only reflect the characteristics of the recruited participants; it is unclear to what extent they are representative of the general population of Montreal. However, they are useful in identifying potential speaker profiles in the sample at hand, clarifying the links between different sociolinguistic descriptors.

13.4 Summary

This chapter has provided a general overview of the participants in the sample recruited for sociolinguistic interviews. I first discussed their sociodemographic characteristics, highlighting a skew in the sample towards younger and female participants; a high degree of variability in terms of geographic origin and linguistic profiles; and relative homogeneity in terms of socioeconomic status. While the overall diversity limits the generalizability of the final results, it provides a much needed means of exploring the perception of semantic shifts by speakers with clearly distinct, reliably described, sociolinguistic profiles.

I then summarized the participants' qualitative remarks concerning their individual identity, as well as life and language use in Montreal. While the specific ways in which they define their identity are variable, most of them see themselves as typical inhabitants of Montreal, and all express highly positive views of the city. They underscore the central role of bilingualism in characterizing Montreal in general, the way English is spoken there in particular, as well as their own identity. The reported patterns of interaction and exposure to languages provide a plausible pathway for cross-linguistic influence.

The chapter concluded with a multidimensional analysis bringing together different types of information in order to discern more comprehensive trends in the data. It suggests that the main distinction between the participants in the sample is related to their age, which is in turn associated with differences in bilingualism and specifically in knowledge of French. Another important dimension of variation is related to different local ties with Montreal. The potential of the examined sociodemographic and attitudinal variables to account for the perception of contact-induced semantic shifts will be put to the test in the next chapter.

Chapter 14

Status and diffusion of semantic shifts

This chapter analyzes the use of contact-induced semantic shifts, as reflected by the acceptability ratings collected using the semantic perception test as well as the qualitative comments formulated by the informants. It explores the variability between different lexical items and between different speakers, principally in order to discern the external (social) constraints on this sociolinguistic behavior and its diffusion within the speech community. Note that the impact of internal (linguistic) factors and the relationship with computationally-derived measures of variation is more extensively addressed in Chapter 15.

Section 14.1 presents acceptability ratings for individual semantic shifts, focusing on general trends in their distribution and their key linguistic characteristics. Section 14.2 analyzes the distinctions between different semantic shifts, using a multidimensional analysis to jointly explore the whole range of lexical items and the impact of sociodemographic and attitudinal variables on their use. Section 14.3 focuses on the differences between individual speakers, identifying similar behaviors and interpreting them in terms of their potential role in the diffusion of semantic shifts. Section 14.4 provides a summary of the main observations.

14.1 An overview of semantic shifts

This section presents an initial overview of the acceptability ratings associated with the examined semantic shifts. It discusses the procedure used to validate the retained items, their global distribution in terms of acceptability ratings, and their main linguistic characteristics.

14.1.1 Items retained for analysis

As noted in the discussion of the semantic perception test in Chapter 12, the informants were asked to provide a range of information for each tested lexical item: its phonetic realization; an acceptability rating on a scale from 1 to 6; an alternative lexical item which would not modify the meaning of the example; and any qualitative comments. The whole range of information was used in analyzing the collected data. Recall in particular that the synonym provided for the target item was used to confirm that the participants interpreted the examples with the posited

contact-related sense. This ensured that comparisons of different acceptability ratings were limited to the targeted sense.

As a general rule, individual ratings were retained only if the expected interpretation was provided. However, some speakers did not systematically offer an alternative for the target lexical items. In an attempt to find a balance between the amount of retained data and the risk of incorrect interpretation, the following principles were applied:

- if no synonym was provided for an individual lexical item, the acceptability rating was nevertheless retained if all the remaining speakers provided the same interpretation for that item, pointing to a largely unambiguous example;
- similarly, for the three informants who took the perception test online, after the face-to-face interview was concluded, I only retained the answers for which all the remaining speakers provided the same interpretation.

As a result of this filtering step, three out of 40 lexical items were excluded from further analysis because only three ratings were retained for each of them. The impacted items, and the posited contact-related senses, are: *deception* 'disappointment', *laureate* 'winner', and *local* 'room, site, premises'. In the first two cases, the participants provided multiple interpretations, some of which were vague and could apply to both the conventional and contact-related sense; in the third case, most participants were unable to interpret the retained example. More than half of the ratings were retained for all remaining lexical items; all were retained for 29 of them. Any impact of missing values on the conducted analyses will be noted as needed.

14.1.2 Distribution of acceptability ratings

Mean acceptability ratings (on a scale from 1 to 6), calculated by averaging over those provided by individual informants, are plotted for all lexical items in Figure 14.1. The value ranges from 1.9 to 5.7; it stands on average at 3.8, just above the threshold corresponding to a symmetrical split of the rating scale into unacceptable and acceptable uses.

The spread of item-level ratings across the whole range of values points to clearly distinct degrees to which they are accepted by the informants, reflecting different degrees of exposure to these items and/or their active use. The degree of acceptability in turn appears to be related to different linguistic characteristics of the items in questions. For instance, the six lexical items with the highest mean acceptability rating – above 5 – include two semantic shifts involving a clear difference in referential meaning, which are also described in the existing literature:

- *terrace* 'restaurant patio' (Fee, 2008, p. 179; see also Section 10.3.3); and
- *pass by* 'stop by' rather than 'continue past' (Boberg, 2012, p. 498).

These examples also include two cases involving finer-grained semantic distinctions:

- *population* 'general public, community' rather than 'inhabitants' in a demographic sense, also described in the literature (Fee, 2008, p. 181; Grant, 2010, p. 186); and
- *boutique* 'store' rather than 'small, fashionable or specialized, store', one of the examples identified through the corpus analyses.

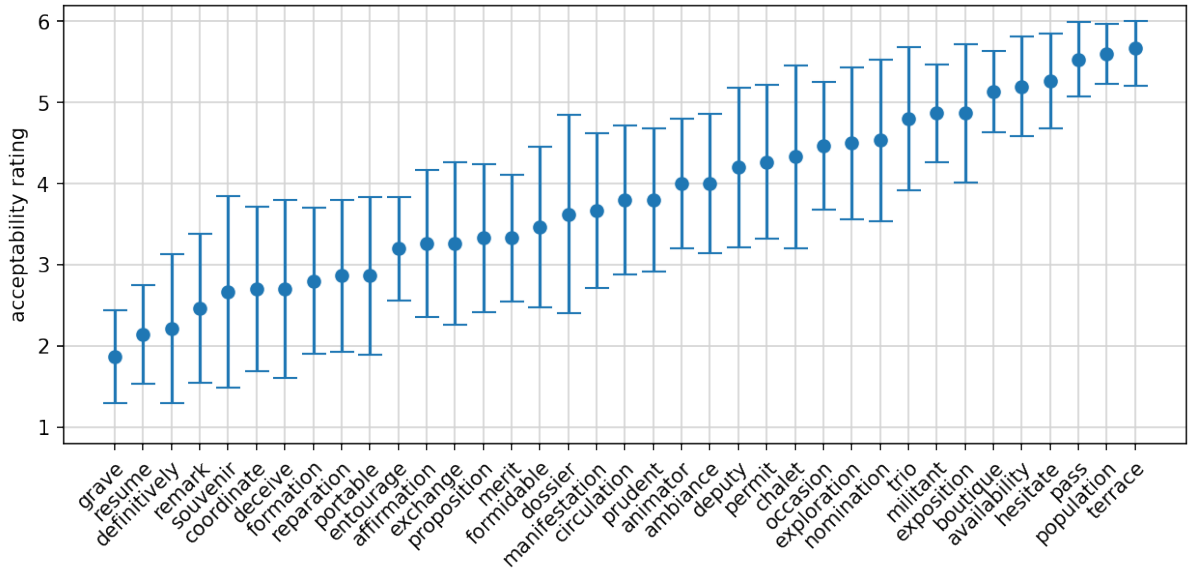


FIGURE 14.1: Mean acceptability ratings for individual lexical items. Error bars represent 95% confidence intervals.

Finally, this group also includes two newly identified examples where the influence of French is primarily observed on the morphosyntactic level, but is arguably also underpinned by subtle semantic distinctions:

- *hesitate*, used in the structure *hesitate between X and Y* 'be undecided between X and Y' (cf. Fr. *hésiter*); and
- *availability*, used in the plural form *availabilities* 'available times' (cf. Fr. *disponibilités*).

Turning to the same number of lexical items at the other end of the plot, the lowest mean acceptability ratings are mostly related to cases with stark differences between the conventional English sense and the contact-related sense. This extends to four cases previously described in the literature:

- *resume* 'summarize' rather than 'continue' (cf. Fr. *résumer*) (McArthur, 1989, p. 25; see also Section 11.1.1);
- *remark* 'notice' rather than 'comment upon' (cf. Fr. *remarquer*) (McArthur, 1989, p. 25);
- *souvenir* 'memory' rather than 'memento' (cf. Fr. *souvenir*) (McArthur, 1989, p. 25; see also Section 10.3.3);
- *coordinate*, usually in plural, meaning 'contact information' rather than 'position on a map' (cf. Fr. *coordonnées*) (Grant, 2010, p. 187).

As for the two examples picked up during the computational analyses, the general trend clearly applies to the first, but not to the second of them:

- *definitively* is used as a generic intensifier, much like 'definitely', rather than with its conventional narrow sense 'conclusively, in a definitive manner' (see Section 10.3.3);
- *grave*, which is principally characterized by a syntactic difference, specifically a tendency towards predicative use (e.g. "it is not very grave") which additionally appears to involve

a semantic generalization (cf. Fr. *grave*).

These examples suggest that higher acceptability ratings tend to be associated with cases where a difference introduced through contact with French is semantically limited, although various types of semantic shifts are present across the range of acceptability ratings. Let us now turn to another linguistic characteristic: the observed phonetic characteristics of the lexical items under study.

14.1.3 Phonetic realization of semantic shifts

Phonetic realizations of the target lexical items, attested in the context of the tweets used as stimuli for the semantic perception test, were perceptually analyzed. This was guided by the assumption that, given their frequent formal similarity with French lexical items, some semantic shifts might be phonetically realized as their cross-linguistic equivalents, and that this might in turn impact their acceptability. I noted all instances in which the target items were phonetically gallicized. These observations are summarized in Table 14.1.

	Speakers					
Variable	**CL1**	**RL1**	**DC1**	**SK1**	**MJ1**	**PI1**
chalet	X	X	X	X	X	
ambiance	X	X	X	X		X
entourage	X	X	X			
terrace	X	X	X			
dossier	X	X				
Fr. score	1.00	0.83	0.67	0.67	1.00	0.63
En. score	0.60	0.70	0.77	0.64	0.43	1.00

TABLE 14.1: Phonetic gallicization of target lexical items, with French and English use scores provided for context.

Phonetic realizations typical of French are overall limited in the recorded data. When it comes to individual lexical items, the most affected are *chalet* and *ambiance*, both of which are gallicized by a third of all informants (5 out of 15). From the perspective of individual speakers, the highest rate of gallicization stands at 12.5%, corresponding to 5 out of 40 lexical items.[1] Moreover, it is important to note that the set of impacted items is compact; all but one have orthographically identical French equivalents. The one exception – *terrace* – is compatible with French orthography, and it would be pronounced in the same way as the actual equivalent, *terrasse*.

It is also important to underscore the variable ease in determining if a given item is pronounced with French characteristics. Specifically, *entourage* and *terrace* entail easily perceptible differences in *r*-realizations, whereas *chalet* and *dossier* can be recognized based on the

[1]The 40 lexical items are likely not equally susceptible to gallicization given their variable degree of formal similarity with French lexical items.

final segment, corresponding to /eɪ/ when adapted to English. In contrast, one of the realizations of *ambiance* reported in the COD is the French-like /ɑ̃mbiˈɑ̃s/. This suggests its wider acceptance in Canadian English, and it also potentially explains its use by PI1, a predominantly English-speaking informant. Apart from this case, the remaining five informants who exhibit phonetic gallicization either predominantly speak French or use the two languages to a comparable extent. This is broadly consistent with earlier reports of higher rates of gallicization among native English speakers who are more highly proficient in French (Rouaud, 2019b, pp. 250–256).

Although the limited extent of this phenomenon in the participant sample precludes any conclusive analyses, distinct potential trends can be identified regarding its link with the use of the target lexical items. For example, RL1 produces acceptability ratings of gallicized lexical items that are on average 1.25 points *lower* than her mean acceptability rating of 4.0. For CL1, on the other hand, the ratings of gallicized items are on average 1.1 points *higher* than her mean rating of 3.5. Potential explanations for this divergence can be found in their spontaneous comments collected in the interviews. In discussing the example of *dossier*, RL1 notes:

> Most words in English that have a very obvious French nature, I would not use it, them because I feel like it would, I don't know, because I feel like I would accidentally start speaking French when I apply them (laughter). So I try to avoid them when, when I write or when I speak.

A similar view, suggesting an avoidance of formally similar lexical items in order to limit cross-linguistic interference, is also expressed by CL1 during a discussion at the end of the semantic perception test.

> When I started learning English, we were, they, they taught us to not translate our speech from French to English, because some words obviously don't make sense. [...] So ambiance for me, I know it's, it's used in English, but in my head it's in French, so I wouldn't say it. But if I hear it, it's okay, it's fine 'cause it's an English word as well.

However, CL1 also acknowledges active use of some formally similar lexical items. She suggests that their realizations are always gallicized; she moreover perceives them as French even when she uses them in an otherwise English utterance.

> But then again, if I say, "Oh, we're gonna go to the chalet /ʃaˈlɛ/", kind of being a hypocrite there 'cause I'm using /ʃaˈlɛ/ as a French word but I'm saying it as a French word as well, I'm not saying it like /ʃəˈleɪ/ or, I'm saying /ʃaˈlɛ/. Or terrace /tɛˈʁas/, I'm not saying /ˈtɛɹəs/, I'm saying /tɛˈʁas/, like a more Québécois way.

In summary, it appears that French phonetic realizations of semantic shifts may reflect two distinct views of these lexical items. For RL1, they are associated with a degree of cross-linguistic similarity which is too high for her to use them comfortably in English, leading to

lower acceptability ratings. For CL1, French realizations are precisely what confers usability to these lexical items, which is in turn reflected by higher acceptability ratings. To what extent these trends generalize across other speakers, and with what social factors they are associated, remains to be seen in future work on larger samples. But these observations reaffirm the relevance of the links between phonetics and the lexicon, including in understanding contact-induced lexical semantic phenomena.

A range of social factors may also explain more general differences in the perception of semantic shifts. This issue is further explored below.

14.2 Accounting for variability between semantic shifts

In accounting for the variable degree of acceptability of different semantic shifts, I drew on the full range of speaker-level descriptors obtained through the PAC interview protocol. I specifically conducted a principal component analysis (PCA), with input data including both sociodemographic and linguistic variables (i.e. item-level acceptability ratings). The aim of this approach was to jointly explore the similarities between different lexical items as well as their association with potential explanatory variables. For a further discussion of PCA and a full list of included sociodemographic variables, see Section 13.3.

As noted earlier, some of the 37 retained linguistic variables contain missing values. In order to include all lexical items in the PCA, the missing values were imputed using the expectation maximization (EM) algorithm. Only the first two principal components were computed when using this approach. For comparison, a PCA was also run on the smaller set of variables which do not contain any missing values. The patterns reflected by the first two components obtained in that way were highly similar and would not modify the interpretation of the results. The analysis conducted on the full set of variables, with imputed missing values, was therefore retained.

The contribution of individual variables to the first two principal components is plotted in Figure 14.2, while Spearman's correlation coefficients for a subset of acceptability ratings and the explanatory variables are presented in Table 14.2.[2] In interpreting the PCA plot, note that the variables pointing in the same direction tend to be positively correlated with one another; those that are at a right angle are likely uncorrelated; and those that point in opposite directions tend to be negatively correlated. It should also be borne in mind that an input variable may exhibit a weaker positive correlation with a variable pointing in a similar direction and a stronger negative correlation with a variable pointing away from it. More generally, the principal components represented in the plot jointly explain 38.6% of variance, meaning that they only illustrate part of the patterns in the data. They are nevertheless useful in distinguish different profiles of linguistic variables, which roughly correspond to the four quadrants defined by the coordinate axes. It must be emphasized that the quadrants are not entirely homogeneous and that the use of lexical items contained within each of them is not explained by identical

[2]The critical value of Spearman's ρ for $N = 15$ observations, as in the case of acceptability ratings without missing values, is 0.52 at the 0.05 level of significance.

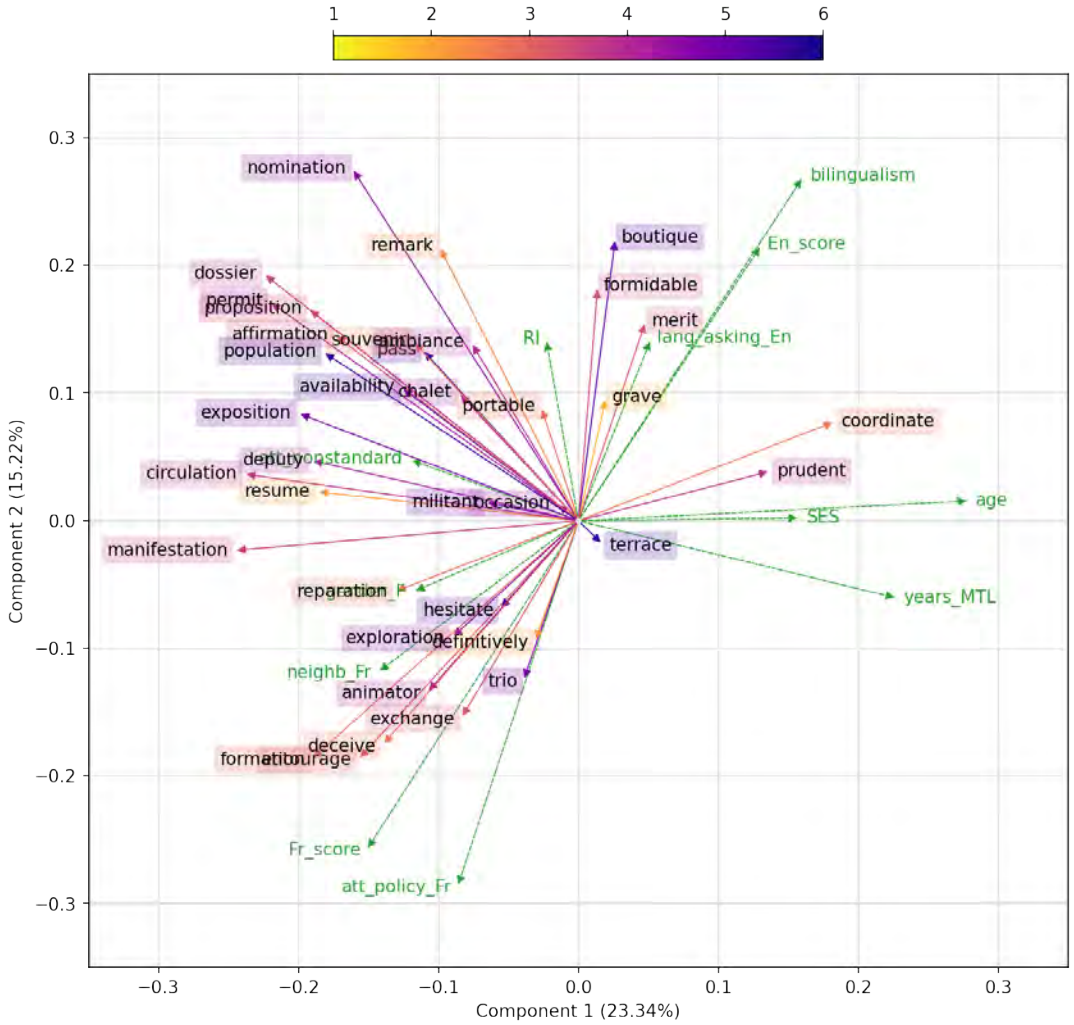


FIGURE 14.2: PCA plot of linguistic and sociodemographic variables based on acceptability ratings. Color coding indicates the mean acceptability rating for each lexical item. Sociodemographic variables are plotted in green. Abbreviations clockwise from top left: `RI`: Regionality Index; `lang_asking_En`: language used to ask for directions, the highest coded value is English; `En_score`: English use score; `SES`: socioeconomic status; `years_MTL`: years spent living in Montreal; `att_policy_Fr`: attitude towards French language policy; `Fr_score`: French use score; `neighb_Fr`: exposure to French in the neighborhood; `gender_F`: gender, the highest coded value is female; `att_nonstandard`: attitude towards nonstandard language use.

factors. However, they are indicative of important general distinctions, which can be further interpreted based on the spontaneous comments expressed by the informants.

14.2.1 Local specificity and influence by French

Let us begin by examining the lower left quadrant, which corresponds to the direction of the French use score, as well as several other variables indirectly reflecting the use of French (e.g. exposure to French in the neighborhood, attitude towards French linguistic policies). Correlation coefficients confirm that the acceptability of lexical items in this area is principally associated with the use of French, as well as, to varying degrees, the variables reflecting strong local ties.

For instance, *formation* 'training' (cf. Fr. *formation*; mean acceptability rating = 2.8) is the

	Lexical items						
Variable	**formation**	**trio**	**circulation**	**permit**	**grave**	**prudent**	**terrace**
age	-0.525	-0.170	**-0.598**	-0.446	0.048	**0.617**	0.151
En_score	-0.438	0.045	-0.145	-0.283	0.379	0.253	-0.117
Fr_score	**0.656**	0.467	0.379	-0.029	0.046	-0.413	-0.163
bilingualism	-0.620	-0.262	-0.295	-0.134	0.103	0.335	0.115
neighb_Fr	0.489	0.042	0.176	0.104	0.185	-0.169	**-0.350**
lang_asking_En	-0.326	**-0.792**	0.062	0.040	0.066	-0.071	-0.250
RI	-0.193	-0.551	0.057	0.373	-0.364	-0.061	-0.230
years_MTL	-0.278	0.360	-0.498	**-0.526**	0.211	-0.107	0.297
gender_F	0.196	-0.180	-0.035	0.343	**-0.619**	-0.122	-0.273
SES	-0.203	-0.154	-0.331	-0.111	-0.370	0.167	0.320
att_nonstandard	0.104	0.134	0.491	-0.001	0.369	-0.160	-0.006
att_policy_Fr	0.475	0.391	-0.103	-0.081	-0.372	-0.262	0.117

TABLE 14.2: Correlation between acceptability ratings for a subset of lexical items and sociolinguistic descriptors. Values in bold indicate the highest absolute Spearman's ρ for the lexical item in question.

most strongly correlated with the French use score ($\rho = 0.66$); consequently, it is also negatively correlated with the English-dominated bilingualism score ($\rho = -0.62$). It moreover exhibits a moderate negative correlation with age ($\rho = -0.52$). This is indicative of greater acceptability among speakers who are more proficient in French than in English, a group which in this participant sample also tends to be younger. The informants overtly perceive this case as typical of French, whether in terms of their own reported use of the item (DC1, MJ1, SK1), that of other native French speakers (QV1), or their ability to understand the item when it is used in English (MT1). A potential regional link is also present, as PS1 suggests in the discussion of a potential alternative:

> I would say training or training session, but obviously living in Montreal that is something I've heard quite a bit.

This local specificity is more pronounced in other cases, such as the well-known example of *trio* 'sandwich-fries-drink menu; combo' (cf. QF *trio*; mean acceptability rating = 4.8). It preserves a trend towards a positive correlation with the French use score ($\rho = 0.47$) and a negative one with age ($\rho = -0.41$). However, its association is stronger with the Regionality Index ($\rho = -0.55$; lower values of the Regionality Index indicate stronger local ties) and with the use of English to ask for directions ($\rho = -0.79$). As discussed in Chapter 13, the latter variable can be seen as an additional, indirect indicator of local ties.

The local link also emerges from a qualitative standpoint, with this use described as "a very Montreal thing" (PS1). A similar observation is offered by MT1:

> Yeah, this is exactly how you use it. You wouldn't hear somebody say the word combo in I think Montreal English, never hear that.

In the same line of thought, the only informant who explicitly expressed uncertainty over the interpretation was HS1, a native British Columbian who has lived in Montreal for around a year. Overall, this example complements the previous one in illustrating the joint but variable importance of French use and local specificity for a subset of lexical items.

14.2.2 Growing diffusion in the local community

Let us now turn to the upper left quadrant of the PCA plot. It contains a large number of lexical items, which roughly follow the same direction as the attitude towards nonstandard language use (higher values indicating a more positive attitude); they also point away from other key sociodemographic variables, such as age and the time spent living in Montreal. They are also roughly orthogonal to the variables related to the use of languages, suggesting that these have a more limited impact. However, the specific ways in which different factors interact are variable.

Take for example the case of *circulation* 'traffic' (cf. Fr. *circulation*; mean acceptability rating = 3.8). It is the most strongly correlated with age ($\rho = -0.60$), followed by the time spent living in Montreal ($\rho = -0.50$). It also exhibits a trend towards a correlation with several other variables, such as the attitude towards nonstandard language ($\rho = 0.49$) and the French use score ($\rho = 0.38$). The stronger correlations interestingly point to younger speakers who have not necessarily spent a long time in Montreal; more generally, the picture is more neutral compared to the previously discussed examples, without a strong association with language use or local ties. This impression is also reflected by the variety of the informants' comments, which include an association with French (PS1); an association with text genres, specifically newspaper articles (HS1); and a perceived neutral connotation compared to the synonymous lexical item *traffic* (QV1).

A somewhat different example is that of *permit* 'driver's license' (cf. Fr. *permis* (*de conduire*); mean acceptability rating = 4.3). It is entirely uncorrelated with the attitude towards nonstandard language and French use, but it presents comparable correlations with age ($\rho = -0.45$) and time spent in Montreal ($\rho = -0.53$). While this would suggest a lack of importance of the local community, the informants indicate a clear association of this lexical item with Montreal, particularly in terms of its use by other speakers (PI1, PS1). Two informants who are not native to Montreal – QV1 and HS1 – note that they have grown accustomed to the use of the item during the time spent in the city. But QV1 also potentially associates it with bilingualism, as reflected by their remark following this example:

> This is starting to make me recognize some very specific Montreal things or specific English-second-language things.

In summary, the constraints on these two lexical items – like the others present in the same area of the PCA projection – are not clear-cut. The informants qualitatively associate them with regional variation, as well as bilingualism and stylistic variation. The most consistent quantitative links are reflected by the negative correlation with age and the amount of time spent in Montreal; however, the precise interpretation of these patterns is not straightforward. The fact that younger speakers find these lexical items more acceptable might indirectly reflect the

tendency for this group to use French more extensively. But if that were the main explanation, we would expect a stronger association with the French, English, or bilingualism scores, neither of which exceeds the critical ρ value for any of the 19 lexical items in this set. It might then be the case that the trend towards younger speakers reflects a language change in progress, with the contact-induced use of these items spreading past the initial stage of cross-linguistic influence directly arising from individual bilingualism. Their broader diffusion is further supported by the negative correlation with the time spent in Montreal. This trend suggests that these items are adopted – or at least deemed acceptable – relatively quickly by speakers arriving in the city and entering the local community, independently of their specific linguistic profile. This hypothesis should however be examined on a larger participant sample.

14.2.3 Limited effect of language contact

Moving on to the upper right quadrant in the PCA plot, we find several variables which broadly appear to be associated with older speakers in the sample and the key characteristics they exhibit, including a tendency to be more English-dominant and to have a higher socioeconomic status. But, as before, the precise explanations at play appear to be more complex.

For instance, the example of *grave* 'serious' was tested in the previously discussed context of "it is not grave" (see discussion in Section 14.1.2; mean acceptability rating = 1.9). Its acceptability is negatively correlated with gender ($\rho = -0.62$; the highest coded value corresponds to female speakers). The informants' comments cover a range of perceived phenomena, including a higher degree of formality (QV1); French influence (MJ1, MT1, PS1); and uncertainty over having heard the item used in that context (DC1, PI1). Two speakers (QV1, RL1) claim that they would not use the item, but are unable to find an alternative. Another tendency is illustrated by *prudent* 'careful' (cf. Fr. *prudent*; mean acceptability rating = 3.8), which is correlated with age ($\rho = 0.62$). Like in the previous case, the qualitative observations are fairly disparate, pointing to a higher degree of formality (QV1), frequent use by others (PS1), and rare use by others (DC1).

While the individual trends are heterogeneous and the associations between variables are often weak to moderate, the speakers who perceive this group of items as more acceptable tend to be older, male, and less proficient in French. Recall however that the sample is highly skewed in terms of both age and gender, and that their apparent impact might in turn be related to other factors, including the strength of local ties, English proficiency, and socioeconomic status (see Chapter 13). With the clear exception of *boutique* 'store', which multiple speakers associate with the way English is spoken in Quebec, the other items are vaguely described as awkward or unusual, sometimes due to a higher degree of formality. This stands in stark contrast with the previous two categories, where perceived regional or French-related usage was clearly noted by the participants. It might then be the case that these items are not actively favored by cross-linguistic influence, perhaps due to the fact that the posited contact-related uses are mostly related to connotational rather than clear referential differences in meaning. As such, these items may be more strongly related to a highly proficient use of English, involving a good command of a range of registers, rather than the impact of French.

14.2.4 Near-universal acceptance

The lower right quadrant in the PCA plot contains a single lexical item: *terrace* 'restaurant patio' (cf. Fr. *terrasse*; mean acceptability rating = 5.7). In addition to being the most highly rated out of all lexical items, it is distinguished by a lack of significant correlation with any sociodemographic variables. This reflects the fact that all but two informants chose the highest acceptability rating for it. The remaining speakers still recognized its currency in the local speech community (HS1; rated 3) and even suggested that they would be unable to find an alternative (RL1; rated 4). More generally, other informants repeatedly described the use of *terrace* as typical of Montreal, confirming its regional character. This is indicative of a semantic shift that has largely completed its spread through the local speech community.

It might be expected that other highly rated semantic shifts which were also previously described in the literature – suggesting longstanding use – would exhibit similar patterns. Potential candidates include *population* 'the people' and *pass by* 'stop by', for which acceptability is only slightly less decisive (mean acceptability rating = 5.6 and 5.5, respectively). The picture is slightly different for other emblematic examples, including *exposition* 'exhibition', *trio* 'combo', and *chalet* 'cottage'. They enjoy broad acceptance overall (mean acceptability rating = 4.9, 4.8, and 4.3, respectively), but they all received multiple low scores, including outright rejection. To that extent, they appear to be on their way to full diffusion within the community, but clear pockets of resistance remain. Their trajectory might ultimately lead to the same position as that of *terrace*; time will tell if that is indeed the case.

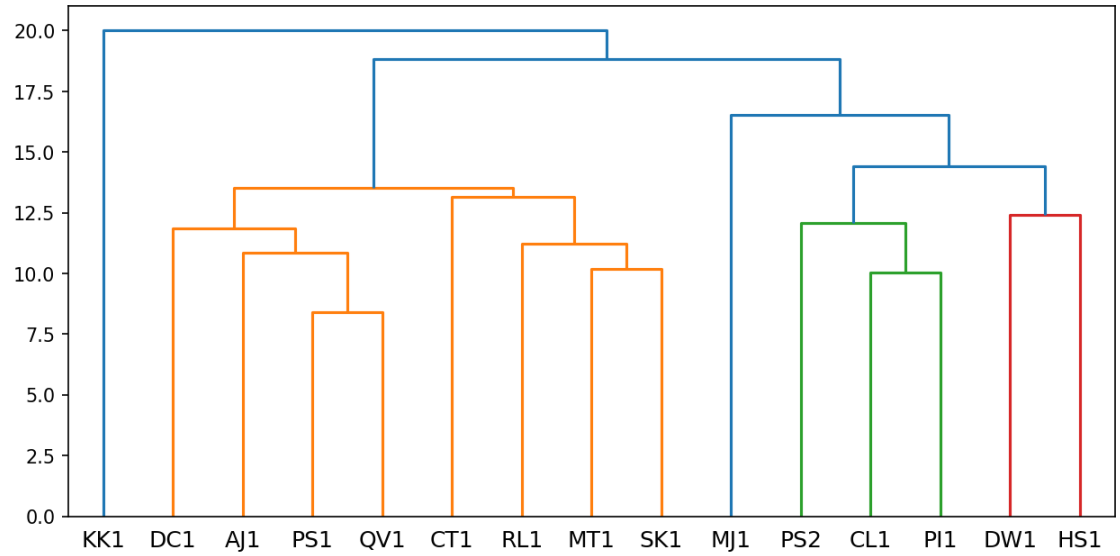
In summary, the analysis of the links between different lexical items – among themselves and with sociodemographic variables – has pointed to several broad patterns of language variation. Some examples (*grave*, *prudent*) do not appear to be directly related to the influence of language contact despite the existence of phonologically and semantically similar French lexical items. But that is not the case in most other instances, ranging from semantic shifts which are associated with both the active use of French and local geographic origin (*formation*, *trio*), to those which are increasingly used by a diverse subset of speakers (*circulation*, *permit*), to those that are nearly universally accepted (*terrace*). As suggested throughout the preceding discussion, these examples correspond to the most clearly identifiable trends in the data; divergent cases and potentially alternative explanations exist in most of the subgroups. Further investigation, relying on a larger sample and more rigorous statistical testing, is required to validate these conclusions. But even at this preliminary stage, they arguably constitute an important contribution: to the best of my knowledge, this is only the second study, after McArthur (1989), to explicitly address the diffusion of different categories of contact-induced semantic shifts in Quebec English; it is the first variationist sociolinguistic study to do so.

The analysis has so far focused on the patterns of variation characterizing individual lexical items. In order to more fully understand the mechanisms operating within the speech community, we should also take a look at the differences among speakers. This is the focus of the next section.

14.3 Accounting for variability between speakers

The speakers in the participant sample present a degree of divergence in terms of their overall tendency to perceive a semantic shift as acceptable. On average, the speaker-level mean acceptability rating stands at 3.8, ranging from 2.7 to 4.6. The ratings attributed by individual speakers tend to cover the full range of values; the mean standard deviation stands at 1.8. The only significant correlation between mean speaker-level acceptability ratings and sociodemographic characteristics is that with age (ρ = -0.52), suggesting that younger speakers tend to perceive semantic shifts as more acceptable. I also inspected pairwise correlations between the individual acceptability ratings produced by different speakers. Spearman's rho ranges from -0.07 to 0.69, with the mean value at 0.29. This indicates that the informants tend to perceive the relative acceptability of different items in a broadly similar way, but the extent to which this is the case for any two speakers is highly variable.

In order to analyze the differences between the speakers' responses more comprehensively, I ran hierarchical agglomerative clustering using the `scikit-learn` implementation (Pedregosa et al., 2011). This approach computes pairwise distances between all observations – in this case, between the full range of scores produced by a pair of speakers. Then, starting from the pair of speakers who are the closest to one another, it links them together into clusters, progressively increasing in distance. In this case, the algorithm was implemented using complete linkage based on the Euclidean distance adapted for missing values. In this case, an initial distance is computed based on the present values, and then scaled up proportionately to the number of missing values that need to be compensated. The result is plotted in Figure 14.2.



	KK1	DC1	AJ1	PS1	QV1	CT1	RL1	MT1	SK1	MJ1	PS2	CL1	PI1	DW1	HS1
score	3.7	4.1	4.6	4.3	4.5	4.2	4.0	4.4	4.3	3.8	2.7	3.5	2.7	3.3	3.3
age	54	22	26	33	32	45	25	24	19	27	70	23	65	70	27
En.	0.6	0.8	0.9	0.9	1.0	1.0	0.7	1.0	0.6	0.4	1.0	0.6	1.0	1.0	1.0
Fr.	0.3	0.7	0.4	0.7	0.4	0.4	0.8	0.6	0.7	1.0	0.3	1.0	0.6	0.4	0.5
RI	10	7	10	3	7	4	7	1	4	3	2	4	8	2	7

FIGURE 14.3: Dendrogram of speakers reflecting the differences between their acceptability ratings. Additional data are provided for context: `score`: mean acceptability score per speaker; `En.`: English use score; `Fr.`: French use score; `RI`: Regionality Index (higher values indicate weaker local ties).

The plot indicates the presence of similar behaviors for a group of speakers (orange cluster) which is also largely coherent in sociodemographic terms. They tend to be younger compared to the remaining informants, as well as proficient in both English and French. They also exhibit higher mean acceptability ratings. As for the remaining speakers, they tend to be older and exhibit a less balanced, often English-dominant, bilingualism. Two important exceptions to this trend are MJ1 and CL1, who are younger and more proficient in French. Another marked exception is KK1, whose acceptability ratings are clearly distinct from all other speakers. This likely reflects her status as a first-generation immigrant who extensively uses her heritage language, potentially entailing weaker links with the wider English-speaking community. However, this trend does not apply to all first-generation immigrants: for example, AJ1, who arrived to Montreal around two years ago, is located within the central cluster based on her acceptability ratings. But she entertains stronger links with locally-established native English speakers, who include her partner and close friends.

Further evidence indicative of the relevance of this group of speakers is provided by the characteristics of their pronunciation on the segmental level; this analysis is presented in detail in Appendix E. The speakers present occasional realizations which can be classified as reading errors – with clear differences relative to the expected phonemic features – including due to a potential influence of French. However, this trend is overall marginal, affecting less than 3% of examined realizations. While some variability exists, all speakers generally exhibit the main features typical of Canadian English pronunciation (cf. Section 2.3). This includes, on the vocalic level, near-categorical presence of Canadian Raising in expected syllabic contexts, as well as categorical low-back merger; on the consonantal level, categorical rhoticity and loss of /ʍ/, as well as variable *yod*-dropping. These general observations are based on the patterns emerging from an auditory analysis conducted by two annotators; a future acoustic analysis will provide more precise insights. But already at this stage, it constitutes additional evidence of the target speakers' clear integration into the wider speech community, which is consonant with their potential role in diffusing a linguistic pattern.

On the whole, the clustering analysis suggests the presence of a core group of speakers – younger, actively bilingual, and well-integrated into the local community although not necessarily native to it – who are driving the use of contact-induced semantic shifts, at least as reflected by the acceptability ratings. The remaining informants, who present both overall lower and comparatively more distinct acceptability ratings, are variably external to this group based on their age, linguistic profiles, and ties with the local community. This is coherent with their position as external with respect to the use of semantic shifts exhibited by the core group. As was the case with the earlier analyses, the generalizability of these trends remains to be confirmed. That said, the differences in sociolinguistic behaviors distinguishing the informants in the sample point to underlying dynamics within the speech community which complement the item-level explanation of the diffusion of semantic shifts.

14.4 Summary

In this chapter, I presented a series of analyses conducted on the data collected in sociolinguistic interviews, focusing on the degree of acceptability of semantic shifts and the representations that are associated with them. In general terms, the acceptability of different lexical items is highly variable, ranging from those that are outright rejected to those that are universally accepted. This might be partly related to the nature of the posited semantic shifts. And while most examined lexical items are fully integrated into the informants' English phonological system, a subset of speakers produce French realizations, which, I have suggested, might serve very different roles in their use of the affected items.

A multidimensional analysis was then deployed in order to explore the potential links between the perception of different semantic shifts and the informants' sociodemographic and attitudinal characteristics. I proposed four global patterns for the examined lexical items: (i) a lack of direct influence of language contact; (ii) regionally specific use that is principally related to individual bilingualism; (iii) regionally specific use that has become adopted by a more diverse group of speakers, having lost a direct link with bilingualism; (iv) a near-universal acceptance in the local community. These patterns of synchronic variation reflect a potential diachronic pathway for the diffusion of contact-induced semantic shifts, with their use likely starting at stage (ii) and gradually moving towards stage (iv).

I moreover explored inter-speaker variability, using a clustering analysis to automatically identify groups of speakers who produced similar acceptability ratings. Drawing on this analysis, I suggested that the use of contact-induced semantic shifts is driven by a relatively coherent group of speakers, who tend to be younger and proficient in both English and French. Together with the item-level analysis, this provides a solid starting point for further investigation into the diffusion of contact-induced semantic shifts. Another, more general question that also requires attention is the link between the interview-based observations, presented here, and the previously discussed computational analyses. This issue is addressed by the next chapter.

Chapter 15

Contrasting Twitter-based analyses and real-life sociolinguistic behaviors

The series of analyses presented over the last seven chapters has addressed the use and perception of contact-induced semantic shifts from different angles. The computational approaches discussed in Part III deployed vector space models to identify a set of semantic shift candidates in a large corpus of tweets, and to broadly characterize their use based on the available metadata. The variationist sociolinguistic approach described in the preceding chapters of Part IV further investigated the same set of lexical items through face-to-face interviews, obtaining detailed background information on the speakers, as well as quantitative and qualitative information on their perception of semantic shifts. This final chapter aims to bring together the outcome of the two approaches, contrasting the information they provide and clarifying their contributions.

The general issue of the relationship between Twitter-based and face-to-face communication is addressed in Section 15.1, drawing on the qualitative comments on this issue collected during the interviews. Descriptions of individual semantic shifts across the range of approaches are explored in Section 15.2, focusing on the relationship between acceptability ratings obtained in sociolinguistic interviews and a range of corpus-based estimates of variation. The overall descriptive contributions of both approaches are discussed in Section 15.3. The chapter concludes with a brief summary in Section 15.4.

15.1 Reported and observed communication on Twitter

One of the broadest distinctions between the analyses conducted in Parts III and IV is the type of data on which they are based. Beyond the technical matters related to collecting and processing these types of information, a more general and fundamentally important issue is the extent to which social media communication reflects real-life, face-to-face interactions. This in turn has implications for the solidity of linguistic conclusions drawn from patterns of language variation in corpora composed of social media posts.

In the sociolinguistic interviews conducted in this dissertation, the informants' self-reported practices and perception of language use on social media were elicited through a set of ques-

tions at the end of the formal interview; the questions were previously discussed in Chapter 12 and are presented in full in Appendix D. Out of the 15 participants, 12 reported relatively regular ongoing use of social media. Most of them (N = 7) tend to use social media passively, i.e. to read other users' content rather than produce posts of their own. The most frequently cited platforms are Facebook and Instagram, with only five participants reporting a Twitter account. This is interesting to note with respect to the representativeness of Twitter-based corpora and the demographic skew that they entail (cf. Chapter 4), even though no reliable conclusions can be drawn given the limited size of the sample. Moreover, since the discussion of language use on social media largely focused on general trends, these can be expected to apply across social media sites. In the remainder of the section, two issues are addressed in more detail: language choice and language variation.

15.1.1 Language choice

The languages that the participants most often use on social media were initially addressed through an overt question, usually at the beginning of this part of the interview. After a discussion of their social media use, the participants were also asked if they would characterize those patterns as comparable to their language choice in real life, providing an additional, less direct assessment.

Out of the 12 informants who actively use social media, six stated that their degree of interaction and exposure to different languages was roughly equivalent to that in real life; for the remaining six, exposure to English was higher on social media. This symmetrical split aligns closely with their linguistic profiles. All informants reporting higher exposure to English on social media have higher French use scores than any of the remaining participants; the mean French use score stands at 0.80 and 0.48 for the two groups, respectively.

An example of comparable behaviors in face-to-face communication and on social media is provided by HS1, who is a native English speaker with intermediate proficiency in French. She describes the language used in her own social media content as follows:

> Any of my posts or anything are probably going to be in English, occasionally if it's just like a small event then maybe I'll use French like to say, I don't know, "Happy New Year" and stuff like that.

When it comes to interacting with other users, for example by commenting on someone else's post, she describes convergence to the interlocutor's language.

> Whatever language they post in, I might try to reply to them in their language.

She also draws a clear parallel between this behavior and language choice in real life.

> I don't like that English is like the unconscious default everywhere, even though I'm not doing much to counter that by speaking English primarily. But I don't like people thinking that I expect them to switch to English for me, so if I can, I usually try and speak whatever language I've heard them use.

Turning to the informants who reported more exposure to English on social media than in real life, a typical example is provided by SK1, who is highly proficient in both languages. After explicitly confirming that her social media use is more English-dominant than her everyday interactions, she further notes:

> And it's also very funny to me to see people that their native language isn't English or, I don't know, maybe it's Spanish, but they all will use English as a, as a like medium language between their social media consumers. The other, the other day I saw my friend whose main language is Spanish and most of the people they, they speak with are Spanish and the caption was in English. That was like very intriguing to me that most people use, on social media, use English as a, as a kind of neutral language.

This impression of English being used by a wide range of speakers on social media echoes similar comments by several other informants. More generally, the reported discrepancy between online and offline language use – both active and passive – highlights a lack of reliable information on individual linguistic profiles in social media corpora. This has significant implications for their construction, as even complex data collection and filtering pipelines, such as the one presented in Chapter 8, are likely to include language content produced by speakers with a variety of linguistic backgrounds. And while aggregate analyses of regional variation in Twitter-based corpora produce results that are broadly comparable to traditional dialectological surveys (cf. Chapter 6), finer-grained linguistic descriptions, including in the context of language contact, may be more strongly affected by this issue. Bearing in mind the vast amount of available data on social media, the precise extent to which this is the case remains an open question requiring further investigation.

15.1.2 Language variation

Another issue explored during the interviews is the participants' perception of language variation on social media. While the previously discussed matter of language choice is important from the standpoint of corpus construction, in this case the focus is on understanding the link between patterns of variation that can be identified using corpus-based methods and the speakers' intuitive views of this mechanism. This in turn contributes to validating the protocol I implemented.

In terms of the characteristics of language use on social media, the informants mainly underscore relative informality, as well as the presence of medium-specific features such as abbreviations and emojis. They are generally not aware of language variation on social media; for those that are, this is limited to broad geographic differences, such as those opposing British and North American speakers, or in limited cases American and Canadian speakers. When prodded for specific linguistic features that might be useful in recognizing a speaker's geographic origin, they generally cite pragmatic factors – such as more enthusiastic messages posted by American than by Canadian social media users – as well as topical differences. Only one speaker raises

the potential presence of regional linguistic expressions, but is unable to provide specific examples. Another speaker is the only to suggest that codeswitching on social media might be indicative of speakers from Montreal.

These observations are at odds with the corpus-based analyses in Part III, which have shown that clear regional trends can be observed in social media data. This discrepancy might be explained by the fact that an average social media user does not pay close enough attention to language use so as to be able to spontaneously report regional linguistic differences. However, it is unlikely that this explains the whole issue given that all informants were able to discuss highly specific linguistic behaviors observed in other areas of their lives. Another potential explanation is the fact that the data I collected on Twitter is several orders of magnitude larger than the number of tweets seen by any given user. This, coupled with the use of quantitative analyses, likely allows for observation of phenomena that are rarer than those noticed by individual speakers.

Crucially, the reverse argument – that the patterns of variation captured through computational analyses of social media data are spurious – does not hold. As extensively discussed in Chapter 14, the informants routinely recognized the semantic shifts attested in tweets as typical of Montreal or of native French speakers, in line with initial expectations. More generally, the vast majority of tweets were interpreted with the posited contact-related senses, further confirming the relevance of computational analyses. Following the semantic perception test, the informants frequently discussed the extent to which the examples capture language use typical of Montreal. For instance, PS1, a native English speaker born in the city who is also highly proficient in French, described his acceptability ratings in these terms:

> I didn't, I don't think I might have put 1 for one or two tweets, but usually, even the ones where I wouldn't say it, I put 3 or 4, because it is something that I'll hear pretty frequently around the city. So although it's a little bit maybe grammatically strange for me, it's someth– it barely affects, you barely notice it because you hear these things all the time.

For some speakers, similar observations extend to their own linguistic practices, seemingly brought to the fore by the large number of examined examples. That was the case of MT1, a highly bilingual native Montrealer of Italian and Greek descent:

> Definitely reading the tweets I noticed more so than when you asked the question about "do you feel like your, your English is influenced by French", I definitely see that more now. Like thinking about the words like chalet and affirmation and stuff, that's where they're coming from. [...] I guess I had this perception that it's mostly just like that Italian East End upbringing that kind of shaped my language more so but I definitely like, there's definitely a lot of that sort of aspect in it as well.

Summarizing, individual speakers do not routinely notice language variation phenomena, including those related to language contact, in the course of their day-to-day use of social media. However, examples of language variation identified through large-scale corpus-based

analyses are near-universally recognized as such. This confirms that social media corpora used with proper precautions can provide informative analyses of language variation, including to describe poorly understood phenomena which are overall rare in spontaneous speech, as is the case with contact-induced semantic shifts.

Beyond these global observations on the nature of different data sources, the resulting sociolinguistic descriptions can also be affected by more precise methodological choices, such as the metrics used to quantify the extent of language variation. This issue is addressed below.

15.2 Comparing the description across methods

This section explores the relationship between the acceptability ratings obtained using the semantic perception test and a series of corpus-based measures. It first presents an overview for the full range of previously introduced measures, and it then focuses in more detail on the link between acceptability ratings and key type-level and token-level variation scores.

15.2.1 Overview of corpus-based measures

Correlation between mean acceptability ratings, computed by averaging over the ratings provided by individual informants, and corpus-based information is presented in Table 15.1.[1] The metrics include information derived from type-level vector space models; frequency and specificity-based information; and the metrics based on the token-level vector analysis and the subsequent manual annotation. For more background on their calculation, see Section 10.2 and, for cluster-based measures, Section 11.3. Type-level variation scores are based on the top model from the evaluation in Chapter 11 (SGNS architecture, window size 5, vector dimensions 100, Orthogonal Procrustes alignment, average of cosine distances from three runs).

Out of the frequency and specificity-based measures, acceptability ratings are the most strongly correlated with the Montreal subcorpus frequency ($\rho = 0.40$). Correlation is similar for frequency in the Toronto and Vancouver subcorpora, with the three frequencies also highly correlated with one another (mean pairwise $\rho = 0.93$). Coupled with a lack of significant correlation between acceptability ratings and the SAGE specificity scores, this points to an association between higher acceptability and a lexical item's overall frequency rather than, say, regional differences in frequency. A possible interpretation of this trend is the claim that semantic shifts which are more widely used – as reflected by their frequency – are also perceived as more acceptable. However, it is important to note that acceptability ratings apply to the contact-induced sense attested in a single example, whereas the corpus-based measures cover all occurrences of a given lexical item, likely diluting the impact of the target use. More precise quantitative information, ideally isolating the frequency of the target sense, would be necessary to further validate this trend.

As for the scores used to directly assess regional semasiological variation, they point in different directions. The type-level variation scores – raw cosine distances as well as the three

[1] The critical value of Spearman's ρ for N = 37 observations, corresponding to the number of retained lexical items, is 0.33 at the 0.05 level of significance.

cos_mt	-0.390	freq_fr	0.115
cos_mv	-0.422	freq_fr_win	-0.199
cos_tv	-0.438	sage_m	-0.109
avg	-0.409	sage_t	0.009
diff	-0.149	sage_v	0.065
ratio	-0.055	charsim	-0.040
freq_m	0.396	context	0.271
freq_t	0.373	clust_contact	0.246
freq_v	0.383	clust_m	0.094
		clust_biling	0.020

TABLE 15.1: Spearman's correlation coefficients for mean acceptability ratings and corpus-based metrics. Montreal, Toronto and Vancouver are referred to using initials. Variables: pairwise cosine distances (`cos`); semantic variation scores (`avg`, `diff`, `ratio`); frequency per subcorpus (`freq`); target word's FrWaC frequency (`freq_fr`); context FrWaC frequency (`freq_fr_win`); specificity scores in individual subcorpora (`sage`); orthographic similarity (`charsim`); context variability score (`context`); proportion of tweets annotated as contact-related (`clust_contact`); proportion of contact-related tweets posted in Montreal (`clust_m`); mean bilingualism score (proportion of English tweets per user) for contact-related tweets from Montreal (`clust_biling`).

derived scores, `avg`, `diff`, and `ratio` – are negatively correlated with acceptability ratings. This suggests that lexical items with more regionally divergent distributional profiles – and therefore potentially more pronounced differences in meaning – have attained a lower degree of diffusion in the local speech community. Recall however that these measures, and raw cosine distances in particular, are strongly correlated with frequency (cf. Chapter 10), which may explain at least part of this trend. This is further supported by the considerably lower correlation of acceptability ratings with `diff` and `ratio` scores, which are calculated in a way that may implicitly neutralize some frequency-related effects.

Correlation between acceptability ratings and cluster-based scores is not significant. It is the highest for the proportion of tweets annotated as contact-related out of all tweets retained at the end of the token-level analysis ($\rho = 0.25$). Although this is only a trend, it is important to note that the correlation coefficient is positive, suggesting that this metric captures a different dimension of variation compared to type-level scores. The way in which this discrepancy, and its relationship with acceptability ratings, impacts the description of individual lexical items is further addressed in the next section.

15.2.2 Type-level and token-level variation scores

In order to more closely explore the relationship between acceptability ratings, on the one hand, and the general trends captured by type-level and token-level analyses, on the other, I retained two corpus-based scores: `avg`, the average of Montreal–Toronto and Montreal–Vancouver cosine distances produced in type-level models; and the proportion of tweets appearing in clusters that were annotated as contact-related following the token-level analysis. They are plotted in Figure 15.1, with the regression lines clearly indicating the divergent trends suggested by correlation coefficients.

In the left-hand plot, the relationship between `avg` scores and acceptability ratings shows

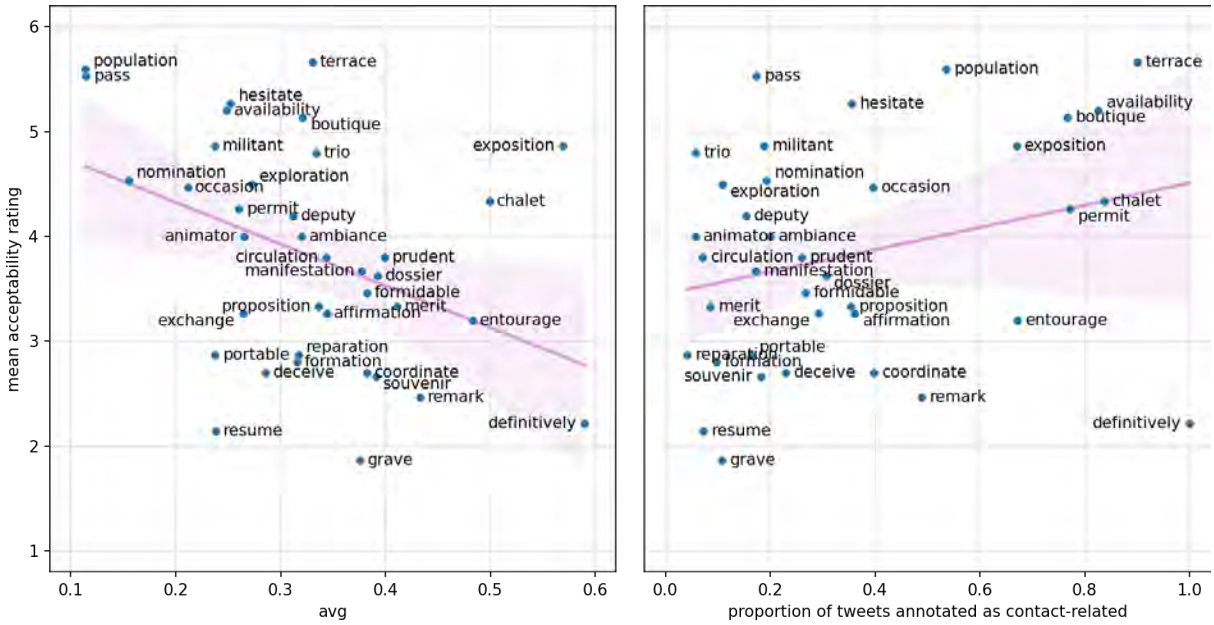


FIGURE 15.1: Comparison of semantic shift acceptability ratings and corpus-based variation scores. Y-axis: mean acceptability rating. X-axis (left): avg, calculated as the mean of the Montreal–Toronto and Montreal–Vancouver cosine distances based on type-level vectors. X-axis (right): proportion of tweets appearing in clusters that were tagged as contact-related, based on the token-level analysis.

that some lexical items whose semantic shifts are seen as highly acceptable are associated with low cosine distances, suggesting a relative stability of their distributional environment across regional subcorpora. This is the case of *population* 'the people' (cf. Fr. *population*) and *pass by* 'stop by' (cf. Fr. *passer*) (upper left). While the first example involves a relatively subtle difference compared to the conventional demographic sense, as in 'inhabitants', the second presents a contact-related sense that is clearly opposed to the expected 'go past'. But these are also the two most frequent lexical items in the entire set retained for analysis, meaning that in both cases the occurrences of their other senses likely drown out the contact-related uses, leading to very similar regional type-level representations.

However, some of the other examples that were highly rated for acceptability also display a high avg score, such as *exposition* 'exhibition' (cf. Fr. *exposition)* and *chalet* 'cottage' (cf. QF *chalet*) (upper right). Compared to the previous examples, they are both less frequent and feature clear differences with respect to their conventional senses (e.g. *exposition* 'opening section in fiction'; *chalet* 'ski lodge'). As discussed throughout Part III, they are also occasionally attested as their French homographs in codeswitched tweets, leading to dramatic differences in distributional patterns. This is reflected by high cosine distances, which facilitate their detection by type-level approaches to regional variation. Although these cases might seem spurious, their relevance is confirmed by high acceptability ratings.

But that is not the only trend at play. Take for example *definitively* 'definitely' (cf. QF *définitivement*) (lower right), which similarly has a high avg score but is rated very weakly in terms of acceptability. It has no orthographically identical equivalent in French, so its use in the corpus is limited to English contexts. In the sociolinguistic interviews, however, the vast majority of informants described it as highly unusual or associated with French speakers. Some

even failed to read it at first, spontaneously replacing it with the more common lexical item *definitely*. Even though a corpus-based description such as this one may seem more convincing than that of *exposition* or *chalet* – given a similarly high variation score but considerably less noise arising from codeswitching – it does not automatically translate to wide diffusion in the speech community.

Looking at the proportion of tweets appearing in clusters that were annotated as contact-related (right-hand plot), the first apparent difference is the overall distribution of values along the horizontal axis, which more clearly splits the lexical items into two groups. Most of them have low cluster-based scores, which is reflected by the reversed direction of the correlation with acceptability ratings. Note however that cluster-based measures are fundamentally different compared to type-level scores, as they assess the uses previously determined to be regionally specific. The proportion of contact-related tweets can therefore be seen as a more direct estimate of the diffusion of the target use within the local community. It is coherent with the acceptability ratings in some cases, such as *terrace* 'restaurant patio' (upper right), but other issues persist. For instance, cases such as *definitively* would once again be described as widespread in the community based on corpus information, contrary to their acceptability ratings. At the other end of the spectrum, highly accepted semantic shifts – including emblematic examples such as *trio* 'sandwich-fries-drink menu; combo' – appear in a limited number of regionally-specific clusters of tweets, which obfuscates their importance.

In summary, this analysis has shown that high type-level scores – devised as an indicator of regional semasiological variation – are generally related to low acceptability ratings. This suggests that limited contact-induced semantic differences tend to be more readily accepted, but the trend is not universal. It is particularly complicated by noise-related phenomena such as codeswitching, which may in fact facilitate the detection of relevant lexical items. As for the cluster-based scores, they more directly reflect the diffusion of semantic shifts in the local community, but once again this is not a general trend; both false positives and false negatives are likely to be picked up by the score.

Although the acceptability ratings and corpus-based scores are not entirely comparable – as previously noted, acceptability ratings are limited to a single occurrence – the discrepancy between them is important to note. In an echo of the evaluation presented in Chapter 11, it highlights the fact that isolated occurrences of target phenomena, or even large numbers of similar occurrences produced by a variety of speakers, should not be taken as directly representative of the diffusion of a linguistic phenomenon in the speech community. The aspects of sociolinguistic description to which they can more reliably contribute are discussed below.

15.3 Sources of descriptive contributions

Taking a broader view of the analyses presented throughout this dissertation, this section addresses the respective contribution of corpus-based methods and sociolinguistic interviews to the description of contact-induced semantic shifts. Three specific issues are discussed: the isolation of patterns of semasiological variation in empirically occurring data; the explanation of

these patterns using sociolinguistic factors; and the interpretation of the social meaning conveyed by semantic shifts.

15.3.1 Modeling semasiological variation

In identifying semantic shifts in empirically occurring data, I deployed corpus-based methods both in a bottom-up approach, using type-level vector representations to detect instances of regional semasiological variation, and in a top-down approach, further examining a predefined set of semantic shift candidates using token-level vector representations. These approaches require large amounts of data, which is a methodological challenge in its own right; however, it also represents an advantage, as the large number of occurrences of individual lexical items allows for a study of phenomena which would not otherwise be observable in spontaneous communication. Computational approaches such as the ones I implemented are systematic and can in theory be extended to an unlimited number of linguistic variables. In practice, however, their descriptive use still requires manual data exploration relying on the linguist's analytical skills. Overall, these analyses were instrumental in detecting previously undescribed semantic shifts, facilitating manual corpus exploration in order to further characterize them, and formulating hypotheses regarding the constraints on their use. The last point in particular highlighted the variable regional specificity and association with bilingualism across a set of semantic shifts.

The identified semantic shifts were further examined using a variationist sociolinguistic interview protocol. Given the practical constraints on data collection in face-to-face interviews, approaches such as these can only be implemented in a top-down approach; the amount of collected data is generally insufficient for meaningful quantitative analyses of lexical features in spontaneous speech. I specifically introduced a semantic perception test, which ensured the systematicity of the resulting description and provided ample additional information on the speakers and their perception of the examined items. This approach refined the view of the initially identified semantic shifts, providing a clearer picture of their diffusion within the speech community. It also included additional types of information, such as the phonetic realization of the targeted lexical items, providing tentative evidence of its interaction with cross-linguistic semantic influence. The data collected using this approach more generally contributed to an evaluation of the patterns captured by the computational methods, discussed in the previous section.

15.3.2 Accounting for patterns of variation

In order to explain the constraints on the use of semantic shifts, I similarly relied on both types of methods. In particular, corpus-based analyses were designed so as to capture regional semasiological variation, providing a way of both identifying semantic shift candidates and characterizing their use. These analyses also drew on publicly available Twitter metadata to characterize individual users, particularly in terms of their linguistic profiles. These two types of information – geographic origin and degree of bilingualism – were derived for the entire

corpus of tweets. Their strength is the vast amount of data, meaning that any conclusions are based on the behavior of tens of thousands of individuals. This is counterbalanced by the uncertainty over their reliability. These two types of information nevertheless enabled me to formulate the previously discussed hypothesis on the variable and potentially joint impact of regional specificity and French knowledge on the use of contact-induced semantic shifts, providing a clear direction for the subsequent analyses.

The variationist interview protocol involved the participation of a comparatively very limited number of participants – four orders of magnitude smaller than that included in the Twitter corpus. However, the description of their sociodemographic background is both considerably more detailed and more reliable, providing finer-grained distinctions between speaker profiles. In addition, the presence of specific variables – in particular age, length of time spent in Montreal, and detailed information on both English and French proficiency – was instrumental in analyzing the diffusion of semantic shifts in the speech community. Although the conclusions of this analysis broadly reflect the corpus-based hypothesis regarding the importance of bilingualism and regional specificity, they are considerably more precise and crucially introduce a potential explanation of diachronic processes.

15.3.3 Interpreting the social meaning of variation

In understanding the social meaning conveyed through the use of contact-induced semantic shifts, corpus-based analyses provided indications mainly in terms of qualitatively analyzed metalinguistic comments and interactions between different Twitter users. This issue was addressed more explicitly through sociolinguistic interviews, where representations associated with different semantic shifts were actively elicited. This information was vital in understanding the way in which semantic shifts are perceived by speakers of different linguistic profiles, and it facilitated the interpretation of more general patterns regarding their use and diffusion. I have also experimented with the use of token-level representations to automatically analyze lexical items based on the representations that are associated with their occurrences. A more systematic implementation of this approach, potentially including the knowledge obtained from the interviews, is a promising direction of future work.

15.4 Summary

In this chapter, I reviewed the descriptive contributions of corpus-based analyses and face-to-face interviews. Focusing on different types of data used in this dissertation, I first summarized the informants' comments regarding language use on social media. They indicate a tendency for French-dominant speakers to actively use or be more exposed to English on social media than in real life, with potential implications for the construction of social media corpora. Moreover, the informants are generally unaware of patterns of language variation on social media. However, the examples that corpus-based analyses identified as specific to Montreal were near-systematically recognized as such, validating the overall approach adopted in this dissertation.

Next, the acceptability ratings obtained in sociolinguistic interviews were compared with a range of corpus-based estimates of semasiological variation. This analysis underscored the fact that type-level variation measures, used to detect semasiological variation between different regions, and token-level variation measures, used to further characterize the diffusion of regionally-specific uses, exhibit inverse relationships with acceptability ratings. This trend, coupled with the fact that the correlation between acceptability ratings and all other quantitative estimates is weak to moderate, suggests that they capture different types of information regarding the use of semantic shifts.

In concluding this chapter, I presented the more general contributions of corpus-based methods and sociolinguistic interviews to the various stages of description pursued in this dissertation. This discussion underscored their complementary nature, with corpus-based analyses providing systematic large-scale overviews covering vast amounts of data and numbers of speakers, and sociolinguistic interviews allowing for an in-depth investigation focusing on speaker profiles and linguistic variables of particular interest. This interdisciplinary setup was instrumental in producing a systematic and comprehensive description of contact-induced semantic shifts in Quebec English.

Conclusion

In conclusion, let us take a step back to review the full range of analyses implemented to investigate contact-induced semantic shifts in Quebec English. Recall once again that they are grounded in a clearly delimited view of the central notions of bilingualism and language contact; of the historical and sociodemographic context of Quebec; and of the linguistic phenomenon under study (Part I). They further draw on an interdisciplinary methodological setup, bringing together approaches typically used in variationist sociolinguistics and natural language processing. Potential contributions and limitations of both disciplines have been reviewed in terms of collecting linguistic data; isolating patterns of semasiological variation in that data; and accounting for those patterns using a range of factors (Part II). A more comprehensive summary of this research background and further discussion of the methodological setup I proposed can be found in Chapter 7.

I now turn more directly to the analyses that I implemented, which are presented in full in Parts III and IV. I will first outline a summary of the individual stages in this research effort, more closely discuss the contributions they provided, and briefly present possible directions of future work.

A summary of implemented analyses

The starting point for corpus-based exploration of contact-induced semantic shifts was the construction of a corpus of tweets, presented in Chapter 8. It contains 1.3 billion tokens, corresponding to 79 million tweets published by 196,000 users from Montreal, Toronto, and Vancouver. Its construction was based on a carefully devised data collection and filtering pipeline, aiming to ensure its usability in descriptions of regional language variation and, more broadly, the reliability of the linguistic information that it contains. The corpus is publicly available in the form of a list of tweet IDs, which can be used to collect the original data.

Chapter 9 presented an exploratory analysis of lexical specificity, which confirmed the presence of regional variation in the dataset, including with respect to contact-related phenomena in the Montreal subcorpus; it also validated the comparability of the collected data. Likewise, the initial analysis of regional semasiological variation based on type-level vector space models identified both previously described and new examples of contact-induced semantic shifts. However, it also highlighted important methodological challenges related to this method.

The nature and extent of these issues were explored more thoroughly through a subsequent series of experiments in Chapter 10. The first of them examined the performance of 18 model

configurations across experimental and control condition corpora. It indicated considerable differences between the configurations and underscored the instability of some vector representations, particularly in relation to the impact of word frequency. The second experiment used a multidimensional analysis, providing further evidence of interactions between different quantitative estimates of lexical semantic variation. Importantly, it also facilitated manual exploration of corpus data. The final experiment implemented token-level vector representations, leading to an initial overview of their potential descriptive applications. They were used to automatically cluster target occurrences, providing a pathway to more efficient and comprehensive analyses of target lexical items.

In Chapter 11, the utility and shortcomings of type-level and token-level representations were more systematically evaluated. In order to do so, I first created an 80-item test set for semantic shift detection in the context of English–French language contact. I then used it to evaluate type-level models on a standard binary classification task, obtaining results comparable to the state of the art on other similar tasks. However, they were counterbalanced by very poor performance on the discovery of new semantic shifts, due to several types of noise which have been identified and described. The 40 semantic shifts in the test set were then analyzed using token-level vector representations, followed by a manual cluster-level annotation. These results provided further clarity on the methodological issues affecting type-level models, which were mainly related to complex sense distributions and noise in the data. The annotations were also used to more extensively characterize the semantic shifts in the Twitter corpus, indicating that their use is variably associated with regional specificity and active use of French.

These analyses were complemented with sociolinguistic interviews so as to provide a finer-grained descriptive account and a better understanding of the contributions provided by different data sources and methodological approaches. The protocol and the recruitment procedure used in this study were introduced in Chapter 12. Building on the standard PAC-LVTI framework, I developed a semantic perception test, focusing on the 40 semantic shifts analyzed using corpus-based methods and attested in examples from the corpus of tweets. A sample of 15 participants from Montreal was recruited for this study.

The structure of the sample was presented in Chapter 13, highlighting a variety of reliably described sociolinguistic profiles. Beyond a recruitment-related skew in age and gender, the distinctions between participants are mostly related to the languages they speak, with variable degrees of proficiency in English and French, as well as the strength of their local ties and the amount of time they have spent in Montreal. Although they define their identity in different ways, most of them see themselves as typical inhabitants of Montreal. They associate bilingualism with the city in general and their own identity in particular, confirming the relevance of the sample for the study of contact-related linguistic practices.

The use of contact-induced semantic shifts was analyzed in Chapter 14, based on the informants' quantitative acceptability ratings and extensive qualitative remarks. I suggested that varying correlations between lexical items and a range of sociodemographic factors can be explained by four synchronic patterns of variation: (i) lack of contact-related influence; (ii) regionally-specific use related to bilingualism; (iii) regionally-specific use spreading to a wider group of speakers; (iv) near-universal acceptance. This analysis also provides a pathway for a

diachronic diffusion of semantic shifts, likely from stage (ii) to (iv). Moreover, inter-speaker variability suggests that the use of contact-induced semantic shifts is driven by a core group of speakers, who in this sample tend to be younger and proficient in both English and French.

Twitter-based analyses and sociolinguistic interviews were compared in Chapter 15. Summarizing the informants' comments, I first underscored a divergence between their reported inability to notice language variation on social media and the validity of the corpus-based analyses of these patterns. I then compared acceptability ratings and corpus-based estimates of semasiological variation, with contrasting patterns of correlation suggesting that they capture different trends in the data. Finally, I assessed the descriptive contributions of the full range of corpus-based approaches and sociolinguistic interviews, highlighting the complementarity of systematic large-scale overviews and more focused in-depth investigations.

Main contributions

The research conducted over the course of this dissertation – in the succession of steps outlined above – has led to a series of contributions, to which the initial summary has alluded. They more specifically include:

- a range of data sources: a 1.3-billion-token corpus of tweets for regional variation in Canadian English, an 80-item test set for binary classification of semantic shifts, cluster-level annotations for 40 semantic shifts, and audio recordings of face-to-face interviews with 15 Montrealers;
- a workflow for corpus-based analyses of semantic shifts, which includes the optimal setup for type-level models, a comprehensive implementation for token-level models, and an inventory of other methods, sources of information, and precautions;
- a coherent variationist sociolinguistic protocol, including a novel interview task directly targeting the use of contact-induced semantic shifts;
- a quantitative and qualitative description of 40 semantic shifts attested in empirically occurring data, around half of which were not previously reported in the reviewed literature;
- an analysis of patterns of variation and diffusion of semantic shifts, based on corpus features, sociodemographic factors, and representations reported by local speakers;
- a direct comparison of the computational and variationist sociolinguistic approaches.

Key resources produced as part of this work – including the corpus of tweets, the test set for semantic shift detection, and the code used for the analyses – are publicly available at the following address: `http://github.com/FilipMiletic/QuebecEnglish`.

In order to take a closer look at these contributions and bring together complementary results produced at different stages of the dissertation, let us revisit the high-level aims and hypotheses formulated at the outset (Chapter 7). In doing so, I will first summarize the main descriptive findings, and then formulate a series of methodological recommendations. The sections of the dissertation providing evidence for these takeaways will be referenced.

Descriptive findings

The first of the three initially defined descriptive aims was to **determine the diffusion and status of contact-induced semantic shifts in Quebec English**. This issue can be addressed on different levels.

In terms of vocabulary-level trends, the results confirm the high-level hypothesis according to which the diffusion of semantic shifts is wider than previously indicated. This is supported by the identification of previously described and newly identified semantic shifts in corpus data (see in particular Chapters 10 and 11) as well as by the ease with which local speakers interpreted a wide range of items used with a contact-related meaning (Section 14.1.2). In terms of community-level trends, the results are indicative of strong diffusion of semantic shifts among Quebec English speakers. Their currency in local speech is evidenced by broad regional distinctions captured by vector space models (Section 11.2) and by familiarity with semantic shifts reported by speakers of very different sociolinguistic profiles (Section 14.1.2).

I also hypothesized that diffusion would vary across individual semantic shifts as well as individual speakers. Both claims are borne out by the data, but the precise patterns are different than those posited initially. Specifically, I claimed that the diffusion of semantic shifts could be analyzed as ranging from a strong association with French-dominant speakers to widespread regional use typical of Quebec. Local speakers associate both of these values with semantic shifts (see below), but they are not mutually exclusive. In fact, the semantic shifts that are strongly associated with a higher rate of French use also tend to be strongly regionally specific, as first shown in a corpus-based analysis (Section 11.3.3). Interview data clarified that this was likely the starting point of diffusion. Semantic shifts may then become more widespread in the local community, losing the direct link with bilingualism, and in the final step become near-universally accepted (Section 14.2).

The second high-level aim was to **establish the sociolinguistic factors influencing the use of contact-induced semantic shifts**. The initial hypothesis posited an overarching link with bilingualism, reflected by both internal and external factors; the data confirm this broad assumption, but also point to more complex patterns.

In terms of internal (linguistic) factors, a multidimensional corpus-based analysis highlighted a facilitating role of formal cross-linguistic similarity; together with estimates of regional specificity, this information was central in discovering new semantic shifts (Section 10.2). Interview data suggested a potentially parallel role of semantic cross-linguistic similarity: acceptability ratings tend to be higher for lexical items where the contact-induced (French) sense is closer to the conventional (English) sense (Section 14.1.2). Higher frequency of semantic shifts might facilitate their use, as indicated by a trend towards positive correlation with acceptability ratings; however, this pattern is not straightforward because frequency interacts with other corpus-based measures (Section 15.2.1). As for the effect of phonetic gallicization of contact-induced semantic shifts, there is insufficient data to provide a definitive answer. The available observations suggest that this behavior may have both a facilitating and an inhibitory effect on the use of semantic shifts, which is likely mediated by other speaker-level factors (Section 14.1.3).

In terms of external (social) factors, corpus analyses pointed to a potential role of the degree of bilingualism and regional specificity (Section 11.3.3). Broadly confirming those observations, interview data further identified a more specific subsection of the community that seems to be driving the use of semantic shifts. It corresponds to younger and more strongly bilingual speakers; in diachronic terms, this would further indicate an increase in the use of semantic shifts over time (Section 14.3). The potential role of these factors has been additionally described in the above discussion on the diffusion of semantic shifts. More generally, these trends must be validated on a more robust participant sample.

The final descriptive aim was to **identify the social meanings conveyed through the use of contact-induced semantic shifts**. Initial indications were obtained through corpus-based analyses, mainly from occasional metalinguistic commentary highlighting perceived links with French use (e.g. Section 11.1.1.2). The interviews provided a much clearer picture, pointing to subjective associations with both English–French bilingualism and regional specificity (Section 14.2). Since these two values are not mutually exclusive – i.e. both can be attributed to the same item – they do not appear to be respectively associated with a different sociolinguistic status of semantic shifts; this is contrary to my initial hypothesis. On a more general note, these results provide further evidence of the high symbolic value of contact-induced lexical variants in Quebec English.

Methodological recommendations

The broad methodological aim pursued by this dissertation was to implement an approach that could provide a systematic description of contact-induced semantic shifts in Quebec English, relying on a combination of computational and sociolinguistic methods to obtain the most comprehensive outcome. The concrete steps that were initially defined correspond to the implementation of different methods, summarized in general terms at the beginning of this conclusion. I now revisit some of the major decisions in more detail in order to provide an overview of the most robust methodological choices and other general recommendations.

We begin with corpus-based experiments, and specifically the creation of the corpus of tweets (Chapter 8). It was central to the remainder of the dissertation: it enabled large-scale computational analyses and provided qualitative evidence on contact-induced semantic shifts, which were then examined in face-to-face interviews. However, there are vital precautions when using this type of data in linguistic research. One important issue is the strongly irregular distribution of data across users; another is the presence of noisy posts. In a striking illustration of the potential impact of these issues, we have seen that corpus-based characterizations of a lexical item can be strongly skewed by a single highly productive account (Section 9.2.3). This and other problems I encountered suggest that it is ill-advised to use any corpus of tweets without balancing the distribution of tweets across users, for example by subsampling the initially collected data. Further filtering decisions, such as exclusion of near-duplicate content, are also highly relevant.

Turning to implementations of vector space models, the results of systematic evaluations have provided clear indications of the best performing approaches – at least on the test set used

here (Chapter 11). In terms of type-level VSMs used to detect semantic shifts based on regional variation, the results show that it is beneficial to use:

(i) neural (`word2vec`) rather than count-based (PPMI) models;
(ii) 100-dimension rather than 300-dimension models;
(iii) a composite semantic variation score which incorporates information from a control region unaffected by contact (`diff`), rather than a score which only focuses on the region of interest (`avg`);
(iv) cosine distances averaged over multiple runs of the same model configuration, as a way of limiting the inherent instability of the model;
(v) smaller window sizes and model alignment based on Orthogonal Procrustes rather than Spatial Referencing (although evidence is less conclusive regarding these two choices).

In terms of token-level models, it has been shown that an implementation based on meaning representations extracted from pretrained BERT, followed by clustering using affinity propagation, can be readily applied to analyses of regional variation. This was used to identify individual occurrences of the target phenomenon as well as to assess the relationship between its use and corpus-derived sociolinguistic descriptors.

These implementations have broadly responded to the initially defined methodological aims of identifying lexical items and their individual occurrences most affected by contact, as well as uncovering patterns behind variation in their use. However, the performance of the implemented methods is not faultless. This is especially true of the bottom-up discovery of new semantic shifts, which was strongly affected by noise in the data and the models. I would therefore suggest that these methods are more fruitfully used in other ways. Type-level models – which are easier to implement and run on the whole vocabulary – are well-suited to hypothesis-driven top-down analyses which can benefit from their quantitative power. Token-level models are particularly useful in facilitating manual linguistic analyses as well as quantifying sense-level patterns for lexical items of interest. The experiments conducted in this dissertation suggest that successful discovery of new semantic shifts requires taking into account additional information, beyond vector-based semantic representations, as well as linguistic expertise. Methods which simplify the task but retain the human in the loop, such as principal component analysis, represent a potential way forward.

As for the sociolinguistic survey, the central methodological challenge was to design a semantic perception task which could be integrated within the standard interview task and would allow for elicitation of comparable information (Chapter 12). The implemented solution, based on dialect questionnaires and used in a face-to-face setup, proved well-suited to those objectives. Note however that each individual subtask – reading the target word in context out loud; rating its acceptability; providing a synonym; and commenting on its use – provided vital information in interpreting the results. Moreover, data analysis was faced with the challenge of a comparatively small and heterogeneous sample; the exploratory multivariate approach I implemented illustrates an efficient way of exploring patterns in this type of data.

Finally, as previously stated, the corpus-based and interview-based results were found to be complementary (Chapter 15). This confirms the interest of data sources such as Twitter

and of computational methods such as VSMs in sociolinguistic research. From a different perspective, it also suggests that none of the implemented approaches is likely to provide the full picture when it is taken in isolation; we are more likely to get to it by combining large-scale corpus-based evidence, input from members of the speech community under study, and linguistic expertise.

Future work

Looking forward, the work conducted in this dissertation can be pursued in several directions. In terms of computational analyses, the use of different types of data provides a potential alternative to the approach adopted here. In particular, a multilingual experimental setup – including French as well as English data – might provide a more direct way of identifying the consequences of contact-related influence. Other types of model implementations can also be considered, including in conjunction with external linguistic knowledge. For instance, the use of dictionary definitions may provide a way of automatizing token-level analyses. This work also raises more general issues of interpretability of neural language models, with significant implications for the reliability of linguistic descriptions.

As for the sociolinguistic interviews, the data collected over the course of this dissertation can be more fully exploited. This involves a more systematic transcription of the recordings and an acoustic analysis of the key features. From an analytical standpoint, a real-time component could be introduced, for example by including the partly comparable observations from McArthur's (1989) study. More generally, together with the data previously collected by Rouaud (2019b) using the same protocol in Montreal, these recordings provide a solid basis for an analysis of as yet poorly described characteristics of Quebec English, such as suprasegmental features. The description of contact-induced semantic shifts would in turn benefit from further data collection, aiming to obtain a larger and more balanced participant sample. This would provide a means to more reliably assess the posited patterns of variation and the key role of a specific group of speakers.

Concluding remarks

The work presented in this dissertation leaves me with two convictions. The first is that contact-induced semantic shifts, and lexical semantic phenomena in general, can and should be described from a variationist sociolinguistic perspective. The results presented here show that they are implicated in familiar patterns of orderly heterogeneity, whose importance is further confirmed by their demonstrable symbolic salience. My second conviction is that interdisciplinary research of the kind presented here, for all its challenges, represents a very promising way forward. It crucially imposes a clearer view of methodological decisions, leading to a more robust assessment of the deployed tools, over and above its descriptive potential. This paves the way for continued contributions to linguistic methods and descriptions.

Extended summary in French
Résumé étendu en français

Nous proposons ici un résumé de la thèse en français. Il s'agit plus précisément d'un aperçu global des différentes étapes présentées plus en détail dans le corps de la thèse. Nous abordons d'abord la problématique de la thèse, détaillons ensuite la structure des chapitres originaux et proposons enfin un résumé synthétique des principales contributions de la thèse. La numérotation des section et sous-sections correspond respectivement à celle des parties et des chapitres de la thèse. Des renvois précis vers les parties complètes, rédigées en anglais, sont systématiquement fournis.

Introduction

L'anglais parlé au Québec est caractérisé, entre autres, par l'utilisation de mots préexistants avec un sens qui ne leur est pas habituellement associé, mais qui est en revanche typique d'un mot français sémantiquement et/ou formellement similaire. A titre d'exemple, le mot anglais *exposition* est attesté au Québec avec le sens propre du français – dénotant un événement artistique – qui n'est pas typiquement utilisé en anglais. La prévalence de ce phénomène, habituellement appelé *glissement de sens*, s'explique par le contact entre l'anglais et le français, ce dernier étant parlé par une grande majorité des Québécois. Bien que diverses sources descriptives attestent de son existence, nous disposons de peu d'informations systématiques sur ce comportement sociolinguistique, et plus particulièrement en ce qui concerne sa diffusion au sein de la communauté linguistique, les contraintes linguistiques et sociales sur son utilisation et la signification sociale qu'il véhicule. Ce sont les questions que nous nous proposons d'aborder dans cette thèse, notamment dans une perspective sociolinguistique variationniste.

Or toute tentative de poursuivre cette description est confrontée à une série de défis. D'un point de vue théorique, la sociolinguistique variationniste peut s'appuyer sur une tradition de plusieurs décennies pour étudier des phénomènes phonologiques et morphosyntaxiques. En revanche, son traitement du lexique, et tout particulièrement de la sémantique lexicale, est moins bien établi. Ceci a également des implications du point de vue méthodologique. Les méthodes habituellement utilisées pour collecter des données, telles que l'entretien sociolinguistique, impliquent des contraintes pratiques qui se traduisent par des corpus de taille trop limitée pour permettre une étude systématique de la variation lexicale. D'autres approches, telles que les enquêtes dialectologiques à partir de questionnaires écrits, contournent ce prob-

lème en obtenant des informations directement comparables auprès d'un plus grand nombre d'informateurs. Cependant, elles fournissent des informations sociodémographiques plus limitées, sont détachées de la communication spontanée et se limitent à l'étude d'un ensemble de mots prédéfini.

Une solution potentielle provient du domaine du traitement automatique des langues (TAL), où les modèles sémantiques vectoriels – des représentations computationnelles du sens des mots – sont utilisés pour étudier les changements sémantiques. Ils permettent une évaluation systématique et quantitative de l'évolution du sens des mots au fil du temps ou à travers d'autres dimensions. Ces analyses peuvent être étendues au lexique entier, permettant potentiellement un repérage spontané d'exemples qui n'ont pas été décrit dans les travaux existants. Cependant, elles présentent à leur tour d'importants défis méthodologiques. Il se pose tout d'abord la question des données : pour que ces analyses soient fiables, elles sont menées sur de très grands corpus, avec un volume de données nettement supérieur à celui des corpus habituellement créés par des entretiens sociolinguistiques. Ensuite, il faut choisir l'architecture des modèles, les hyperparamètres et les mesures de variation à mettre en œuvre ; de plus, la fiabilité des résultats produits reste à ce jour incertaine. Le premier problème pourrait être résolu en utilisant les vastes quantités de données issues des réseaux sociaux, qui sont géolocalisées et publiquement disponibles, mais cela entraîne à son tour une incertitude supplémentaire quant à la proximité des descriptions résultantes à la communication observée dans la vie de tous les jours. Le deuxième problème peut être résolu par une évaluation systématique des méthodes de repérage des changements sémantiques, mais il n'existe aucun jeu d'évaluation disponible pour les changements sémantiques induits par le contact en anglais québécois.

Puisqu'il ne semble pas y avoir de solution simple au problème posé, nous avons adopté une perspective interdisciplinaire. Notre objectif est de produire une description exhaustive en tirant parti des aspects complémentaires des deux types d'approches, tout en contournant leurs défauts. Cela permet d'évaluer les contributions descriptives des méthodes computationnelles mises en œuvre, tant sur le plan quantitatif que qualitatif, ainsi que d'évaluer la fiabilité des corpus issus des réseaux sociaux dans les descriptions sociolinguistiques. Plus précisément, nous avons utilisé des modèles vectoriels créés à partir d'un nouveau corpus de tweets pour obtenir un aperçu systématique à grande échelle et une caractérisation initiale des glissements de sens induits par le contact en anglais québécois. L'ensemble des mots identifiés par ces analyses a ensuite été examiné plus finement au moyen d'entretiens sociolinguistiques avec 15 locuteurs montréalais. Le résultat conjoint de ces deux approches a permis d'identifier des facteurs à l'origine de l'utilisation des glissements de sens, les représentations qui leur sont associées, et de fournir une analyse systématique de leur diffusion au sein de la communauté linguistique. Il a également démontré le rôle prometteur des méthodes computationnelles à grande échelle pour faciliter le travail descriptif, tout en mettant en évidence des défis significatifs ainsi que l'importance de l'expertise linguistique dans ce type d'analyse.

Partie I. Effets sémantiques du contact de langues en anglais québécois : un aperçu

Les chapitres de la partie I donnent un aperçu général des mécanismes qui peuvent expliquer l'émergence des glissements de sens induits par le contact, le contexte spécifique dans lequel ils sont utilisés et la vision théorique adoptée pour les décrire. Le chapitre 1 introduit les notions centrales de bilinguisme et de contact de langues. Il se concentre à la fois sur les caractéristique du bilinguisme du point de vue des locuteurs indiviudels, ainsi que sur leurs implications pour les pratiques langagières au niveau de la communauté linguistique. Le chapitre aborde en outre les effets linguistiques du bilinguisme individuel, notamment les mécanismes qui peuvent faciliter l'émergence de l'influence sémantique dans une situation de contact de langues. Le chapitre 2 présente le contexte sociohistorique et les variétés linguistiques qui sont au centre de cette thèse. Il décrit l'évolution historique de la société québécoise et son profil démolinguistique, qui constituent une base pour la situation actuelle de contact de langues. Le chapitre décrit ensuite certaines des principales caractéristiques du français et de l'anglais québécois, en se concentrant sur celles qui sont impliquées dans les processus liés au contact ou qui sont directement pertinentes pour cette thèse. Nous présentons notamment un résumé des descriptions existantes des glissements de sens en anglais québécois, qui sont abordées plus en détail dans le chapitre 3. Celui-ci fournit une définition plus précise de notre objet d'étude et introduit les principes théoriques qui orientent son analyse dans cette thèse. Nous posons ainsi les bases pour le développement de la méthodologie proposée dans cette thèse.

Chapitre 1. Bilinguisme et contact de langues

Le point de départ du chapitre 1 est la considération traditionnelle selon laquelle deux ou plusieurs langues sont considérées comme étant en contact si elles sont utilisées de manière alternée par les mêmes personnes (Weinreich, 1953, p. 1). Déjà dans le travail de Weinreich, l'intérêt pour le contact de langues est motivé par l'étude des *interférences*, ou "déviations des normes de l'une ou l'autre langue qui se produisent dans le discours des bilingues en raison de leur familiarité avec plus d'une langue" (p. 1). De la même manière, nous considérons le changement linguistique induit par le contact comme étant "le produit des innovations que les locuteurs multilingues individuels introduisent en discours dans un contexte multilingue" (Matras, 2009, p. 5). C'est pourquoi, afin de comprendre les mécanismes qui sous-tendent ce type de changement linguistique, nous commençons par introduire les spécificités des locuteurs bilingues.

Ce chapitre aborde dans un premier temps l'acquisition et l'utilisation de plusieurs langues du point de vue du locuteur individuel (section 1.1). Il présente ensuite le développement des communautés bilingues et leur lien avec l'identité (section 1.2). Enfin, il décrit les principales manifestations du bilinguisme dans le discours des locuteurs individuels, en soulignant le lien entre ces patterns et les changements qui s'opèrent au sein des communautés linguistiques (section 1.3). Compte tenu de notre intérêt général pour les glissements de sens induits par le contact, ce chapitre illustre les mécanismes qui facilitent leur émergence ainsi que les facteurs

qui peuvent conditionner leur utilisation.

De manière générale, nous considérons que le bilinguisme correspond à l'utilisation de deux ou plusieurs langues dans la vie quotidienne ; dans le contexte québécois, cela correspond aux locuteurs qui utilisent régulièrement (au moins) l'anglais et le français. Dans cette thèse, aucun type de bilinguisme, degré de bilinguisme ou mode d'acquisition bilingue ne constituera un critère d'inclusion. Au contraire, tous les bilingues sont considérés comme des membres à part entière de leurs communautés linguistiques ; de plus, les bilingues simultanés et successifs, par exemple, peuvent atteindre des degrés comparables de compétence linguistique et présenter des interférences linguistiques similaires. Cela correspond également au point de vue adopté dans la définition des communautés linguistique du Québec, présentée dans le chapitre 2. Dans notre travail descriptif, cette vision volontairement largé est complémentée par une description détaillée de tous les aspects de l'histoire linguistique d'un bilingue ; cette démarche fournit un ensemble de variables explicatives dans les analyses menées. Une attention particulière est également portée aux facteurs contextuels qui peuvent influencer l'interaction bilingue. La manière précise dont nous mettons en œuvre cette approche est présentée dans le chapitre 8, pour les analyses sur corpus, et dans le chapitre 12, pour les entretiens sociolinguistiques.

Au niveau sociétal, il est souligné que le développement du bilinguisme est lié à une série de facteurs sociohistoriques, et que le statut des communautés linguistiques auxquelles participent les bilingues est dynamique et souvent précaire. Pour cette raison, entre autres, le bilinguisme est étroitement lié à l'identité. Ces aspects sont pris en compte pour définir le statut global de la communauté anglophone du Québec (voir le chapitre suivant). Ils sont également utilisés pour interpréter les pratiques langagières des locuteurs individuels qui constituent cette communauté ; en ce qui concerne les entretiens sociolinguistiques, les avis des locuteurs sur leur identité sont présentés dans le chapitre 13.

En plus d'être associées à des facteurs sociolinguistiques, les pratiques de communication bilingue se manifestent par une variété de mécanismes linguistiques. En particulier, nous avons présenté les façons dont des éléments de différentes langues peuvent être utilisés dans un seul énoncé, en nous concentrant notamment sur les phénomènes de codeswitching et d'emprunt. Nous avons également souligné que ces pratiques linguistiques, mises en oeuvre par des locuteurs individuels, peuvent donner lieu à des phénomènes de variation et de changement linguistiques. Ces questions sont approfondies dans la partie II.

La suite de ce travail se concentre sur une pratique linguistique particulière, sous-tendue par une influence sémantique qui s'opère entre deux langues. Pour l'aborder, nous nous appuyons sur l'idée selon laquelle les langues présentes dans le cerveau des locuteurs bilingues sont en interaction constante. Le lexique des locuteurs bilingues n'est donc pas spécifique à une langue donnée ; les représentations sémantiques sont partagées entre les langues. Les sens associés aux mots des langues différentes peuvent ainsi interagir dans le lexique mental. Dans le chapitre 3, nous abordons ce phénomène du point de vue des tendances observées dans les communautés linguistiques, décrites par la sociolinguistique et d'autres disciplines connexes, dans le but de définir la notion de glissements de sens induits par le contact. Mais nous nous nous intéressons tout d'abord au contexte plus général dans lequel ce comportement a lieu, et qui peut nous aider à mieux comprendre son importance.

Chapitre 2. Contact de langues au Québec

Comme indiqué précédemment, les mécanismes de communication bilingue sont ancrés dans le contexte spécifique de la communauté linguistique en question. Cette thèse porte sur l'usage de l'anglais québécois ; il est donc essentiel de mieux comprendre le profil sociohistorique du Québec ainsi que les principales caractéristiques des langues qui y sont parlées. Le chapitre 2 commence par un bref aperçu de l'histoire du Québec : cela permet de mettre en lien les étapes historiques clés et le développement des communautés linguistiques, puis d'illustrer la situation démolinguistique actuelle (section 2.1). Le chapitre s'appuie ensuite sur les recherches sociolinguistiques existantes pour présenter les principales caractéristiques du français québécois (section 2.2) et de l'anglais québécois (section 2.3) ; dans ce dernier cas, un résumé des travaux portant sur les glissements de sens est également fourni. Nous terminons par un bref résumé (section 2.4).

La discussion de la situation sociohistorique et démographique a tout d'abord souligné l'arrivée initiale de la population francophone, avec l'établissement de colonies permanentes en Nouvelle-France au début du XVII^{e} siècle. Un tournant a été marqué par la conquête britannique en 1763, qui a entraîné l'arrivée d'une importante population anglophone. Malgré son statut de minorité, celle-ci a occupé une position de pouvoir politique, social et économique jusqu'à la moitié du XX^{e} siècle environ. Le vent commence à tourner dans les années 1960, avec l'affirmation croissante des Québécois francophones dans la vie politique de la province. Cette tendance se renforce à partir de 1977, avec l'adoption de la Loi 101, qui a transformé l'équilibre des pouvoirs entre les deux communautés linguistiques. Le statut minoritaire de la communauté anglophone, son exposition intense au français, ainsi qu'un taux élevé de bilinguisme chez les Québécois en général, constituent des facteurs facilitant l'émergence de pratiques langagières induites par le contact de langues.

Ce chapitre a également examiné certaines des principales caractéristiques des deux langues officielles du Canada, telles qu'elles sont parlées dans la province. Le français québécois se caractérise par l'affirmation d'une norme endogène, c'est-à-dire qu'il est désormais clairement défini par rapport à son usage au Québec. Il est caractérisé de manière distinctive sur les plans phonologique, morphosyntaxique et lexical. La réalisation variable de ces traits et par ailleurs souvent révélatrice du degré de bilinguisme des locuteurs anglophones ; cela témoigne plus globalement d'une interaction intense entre les deux communautés linguistiques, y compris en termes de participation au changements linguistiques en cours.

Quant à l'anglais québécois, il est principalement décrit dans ce chapitre en termes de ses spécificités régionales dans le contexte plus général de l'anglais canadien. Sur le plan de la prononciation, il se caractérise par un inventaire phonémique typiquement nord-américain, avec d'autres caractéristiques typiques de l'anglais canadien en général, ainsi que certains traits qui le distinguent des autres variétés canadiennes. Sa morphosyntaxe et, dans une plus large mesure encore, son lexique comportent des effets évidents du contact avec le français. Sur le plan lexical en particulier, cette influence constitue un aspect majeur de la spécificité régionale du Québec, qu'elle s'opère par transfert lexical direct (*dépanneur*), calque (*all-dressed*), glissement de sens (*chalet*), ou autres procédés. De manière plus générale, tant pour le français que

pour l'anglais québécois, nous avons proposé une vision large des communautés linguistiques. Notre définition s'étend ainsi à tous les individus qui sont capables de parler les langues de la province.

Enfin, les descriptions existantes du principal objet d'étude de cette thèse – les glissements de sens induits par le contact de langues – ont également été présentées. Ces descriptions fournissent des indications convaincantes de l'influence sémantique lexicale du français sur l'anglais québécois. Toutefois, une analyse plus complète de la diffusion de ces pratiques langagières, des contraintes sociales et linguistiques qui influent sur leur utilisation et de la signification sociale qu'elles véhiculent n'est toujours pas disponible. Ceci s'explique par des défis théoriques et méthodologiques propres aux études sociolinguistiques variationnistes sur le plan de la sémantique lexicale. Nous abordons la première de ces deux questions dans le chapitre suivant.

Chapitre 3. Glissements de sens induits par le contact de langues

Dans les deux chapitres précédents, nous avons vu que la capacité de parler plusieurs langues peut être à l'origine de différents types d'interférences linguistiques du point de vue des locuteurs individuels, et que ces pratiques langagières peuvent ensuite donner lieu à des phénomènes de variation et de changement à l'échelle des communautés linguistiques. Un type d'influence associé au contact de langues, y compris dans le contexte de l'anglais québécois, est la présence de glissements de sens. Ceux-ci constituent le principal objet d'étude de cette thèse. Cependant, leur définition et leur analyse posent de nombreux défis théoriques ; ce chapitre présente la position que nous adoptons à l'égard de ces questions.

La section 3.1 s'appuie sur les études existantes des glissements de sens pour fournir une définition consolidée de ce phénomène. La section 3.2 identifie une série de questions – principalement théoriques – qui peuvent avoir un impact sur les analyses linguistiques que nous prévoyons de mener ; nous présentons donc la manière dont nous aborderons ces questions. La section 3.3 fournit un bref résumé.

Nous avons d'abord proposé une définition de notre objet d'étude à partir des recherches existantes sur le changement sémantique diachronique, la variation sémantique synchronique et les effets sémantiques du contact de langues. Nous avons proposé une vision large des glissements de sens induits par le contact en anglais québécois, correspondant à l'utilisation d'un mot anglais préexistant avec un sens donné qui s'explique par l'existence d'un mot français, formellement et/ou sémantiquement similaire, auquel le sens en question est habituellement associé. De manière plus précise, nous avons suggéré que ce phénomène peut comporter des effets sur les niveaux dénotationnel, connotationnel ou collocationnel.

A partir de cette définition et des exemples discutés, nous avons ensuite brièvement présenté plusieurs questions qui ont des incidences sur la description des glissements de sens. Plus précisément, nous avons introduit une vision empirique du sens des mots. Celle-ci est basée sur l'idée selon laquelle la plupart des mots sont polysémiques, leurs occurrences individuelles fournissent un point de départ pour identifier leurs sens, et le résultat de ce processus n'est pas immuable mais dépend plutôt de la perspective adoptée et des données utilisées pour l'analyse.

Nous avons ensuite présenté plus en détail la notion de sémantique distributionnelle, qui formalise certains de ces principes généraux et servira de base aux analyses sur corpus menées dans cette thèse. Nous avons ensuite souligné l'importance de distinguer entre différents types d'indétermination affectant le sens d'un mot, ainsi qu'entre des mots formellement identiques mais sémantiquement ou grammaticalement différents. Enfin, nous avons établi une distinction entre les perspectives onomasiologiques et sémasiologiques en sémantique lexicale.

Dans la suite de cette thèse, nous aborderons le phénomène général des glissements de sens induits par le contact principalement à travers une analyse de la variation sémasiologique synchronique, en contrastant les sens typiques de l'anglais québécois avec ceux utilisés dans d'autres variétés régionales ou par des groupes de locuteurs spécifiques. L'interprétation de ces patterns sera complétée au besoin par une perspective diachronique (analyse de l'émergence des sens spécifiques à l'anglais québécois au fil du temps), ainsi que par des considérations onomasiologiques (analyse de l'utilisation d'autres mot ayant un sens similaire). Le contexte méthodologique de cette approche est présenté dans la partie II.

Partie II. Une approche interdisciplinaire

La partie I a fourni un ensemble d'informations générales sur l'objet d'étude de cette thèse : il s'agit d'étudier les glissements de sens induits par le contact de langues, vus ici comme un effet du bilinguisme et spécifiquement étudiés dans le contexte de l'anglais québécois. Afin d'examiner ce comportement de manière empirique et systématique à l'échelle de la communauté linguistique d'intérêt, nous nous appuyons sur des méthodologies développées dans deux disciplines distinctes : la sociolinguistique variationniste et le traitement automatique des langues (TAL).

Les chapitres suivants examinent en détail la motivation et les modalités selon lesquelles des approches issues de ces deux disciplines peuvent être réunies pour répondre à la question descriptive définie au départ. Le chapitre 4 présente les critères et les pratiques pour constituer différents types de corpus reflétant des phénomènes de variation linguistique. Le chapitre 5 présente les stratégies permettant d'isoler ces phénomènes, notamment au niveau sémasiologique, dans les données collectées. Le chapitre 6 présente différentes manières de rendre compte de la variation linguistique observée, à la fois en termes des facteurs qui la motivent et de la signification sociale qu'elle revêt. Enfin, le chapitre 7 définit plus précisément nos objectifs de recherche et expose les principaux éléments de l'approche mise en œuvre dans la suite de l'étude.

Chapitre 4. Données pour la variation linguistique

Si la sociolinguistique variationniste et le traitement automatique des langues s'appuient sur des méthodes très différentes, ces deux disciplines partagent néanmoins une orientation fortement empirique. Ce chapitre passe en revue quelques-unes des principales méthodes que les deux champs proposent pour collecter des données linguistiques d'origine naturelle afin d'étudier des phénomènes de variation au sein et à travers des communautés linguistiques diverses. La section 4.1 aborde cette question du point de vue de la sociolinguistique variationniste. La discussion porte principalement sur la structure de l'entretien sociolinguistique classique, mais évoque aussi d'autres méthodes de collecte de données couramment utilisées dans la discipline. La section 4.2 examine en revanche la constitution de corpus issus des données des réseaux sociaux, en se concentrant en particulier sur Twitter. Ce type de communication a suscité un certain intérêt sociolinguistique en tant que variété linguistique à part entière. Cependant, nous le considérerons principalement comme une source de données alternative, qui partage d'importantes similarités avec la communication en face-à-face, et qui facilite notamment la constitution de très grands corpus linguistiques. Comme nous le verrons dans le chapitre suivant, ce dernier point constitue une exigence pratique essentielle pour les méthodes de TAL mises en œuvre dans cette thèse. Enfin, la section 4.3 fournit un résumé des points principaux.

Il convient de noter que les sections 4.1 et 4.2 privilégient les questions méthodologiques considérées comme centrales dans leurs disciplines respectives, mais elles abordent globalement le même ensemble de problèmes : les caractéristiques des communautés de locuteurs ciblées et leurs comportements langagiers ; le processus pratique de collecte et de filtrage des

données ; et les limites de chacune des approches. Notons également que ce chapitre est limité à une vue d'ensemble des approches courantes de collecte de données focalisées sur la variation linguistique. Les manières dont cette variation peut être modélisée et expliquée sont respectivement présentées dans les chapitres 5 et 6. La collecte de données effectuée dans le cadre de cette thèse est discutée dans les chapitres chapter 8 (pour les données de Twitter) et 12 (pour les entretiens sociolinguistiques).

De manière générale, la discussion menée dans le présent chapitre illustre deux approches distinctes – mais complémentaires – permettant de collecter des données qui reflètent des phénomènes de variation linguistique. En sociolinguistique variationniste, l'objectif principal est de décrire les pratiques langagières d'une communauté linguistique et d'obtenir une caractérisation fine des locuteurs recrutés afin de pouvoir expliquer les phénomènes de variation observés. Le processus de collecte des données repose sur une réflexion approfondie concernant le choix des locuteurs, la conception de la méthode de collecte et le traitement des données. Cela exige à la fois des efforts et des compétences considérables, notamment en ce qui concerne l'interaction directe avec les participants, démarche complexe visant à les mettre à l'aise et à faciliter une production de parole spontanée.

Cependant, les données collectées de cette manière sont quantitativement insuffisantes pour une étude systématique des phénomènes lexicaux. L'utilisation de données issues des réseaux sociaux, telles que celles disponibles sur Twitter, constitue une réponse potentielle à ce problème. Twitter est particulièrement adapté car il permet un accès relativement facile à de grandes quantités de données linguistiques géolocalisées, qui sont souvent de nature informelle et préservent des informations sur les utilisateurs individuels. Les pratiques de communication typiques, y compris celles des communautés bilingues, sont largement représentées sur Twitter, tout comme des schémas d'interaction complexes. Bien que les données démographiques disponibles soient considérablement plus limitées, les corpus issus de Twitter sont généralement plusieurs ordres de grandeur plus grands que les corpus sociolinguistiques traditionnels. C'est un avantage important lorsqu'il s'agit d'étudier la variation lexicale ; un autre aspect positif est le fait que l'utilisation de Twitter évite entièrement le paradoxe de l'observateur.

Il serait sans doute intuitif de considérer que les divergences entre ces deux sources de données constituent le principal résultat de cette comparaison, mais il est également important de souligner les aspects qui les réunissent. D'une part, les deux types de corpus partagent un certain nombre de problèmes, même si ces deniers les affectent différemment. Il s'agit notamment de problèmes de représentativité causés par des biais démographiques dans les échantillons respectifs, ainsi que des limitations juridiques et éthiques en termes de la distribution et de la reproductibilité des données. D'autre part, la sociolinguistique et le traitement automatique des langues sont des disciplines empiriques qui reposent toutes les deux sur l'utilisation de données linguistiques attestées. Tout comme dans le cas de la collecte de données, elles fournissent également des méthodes différentes mais complémentaires pour analyser la variation entre les communautés de locuteurs. C'est le point abordé par le chapitre suivant.

Chapitre 5. Modélisation de la variation sémasiologique

Nous avons vu que différentes sources de données potentielles peuvent fournir des informations sur la variation linguistique. Le chapitre 5 aborde donc la question de l'utilisation de ces données pour modéliser la variation linguistique, en se focalisant sur les différentes stratégies permettant d'isoler des instances de variation sémasiologique. La section 5.1 présente les solutions méthodologiques potentielles qui existent en sociolinguistique variationniste et dans d'autres disciplines connexes, ainsi que les principaux défis théoriques posés par l'étude de la variation sémasiologique dans ce cadre. La section 5.2 passe en revue les recherches connexes menées en TAL sur les modèles computationnels de la variation et du changement sémantiques. La section 5.3 résume les points clés.

Cette discussion met en évidence un large éventail de considérations méthodologiques, et certaines considérations théoriques, liées à l'extraction et à l'analyse des patterns de variation sémantique à partir de données linguistiques attestés. Du point de vue de la sociolinguistique variationniste, ce type de variation est rarement étudié, en grande partie à cause d'un manque de méthodes adéquates. Cette situation est aggravée par la nature des variables sémasiologiques, qui n'est pas directement alignée à tous les aspects de la théorie variationniste. Bien que les études existantes en dialectologie et en sociolinguistique variationniste soient limitées en nombre, elles fournissent des solutions méthodologiques possibles, tout en démontrant l'intérêt descriptif de la variation sémasiologique.

Or aussi pertinentes que soient ces études, il est clair qu'elles ne peuvent pas rendre compte de patterns quantitatifs à grande échelle. C'est pourquoi nous avons également discuté d'études computationnelles de la variation et du changement sémantiques, en soulignant notamment l'utilisation de différents types de modèles vectoriels. Les méthodes développées dans ce cadres sont nombreuses, prometteuses, et devraient permettre une étude systématique de la variation sémasiologique. Mais il s'agit également d'un domaine relativement récent, avec de nombreuses questions sur la mise en œuvre optimale et l'évaluation des modèles, ainsi que sur leur utilité dans la recherche descriptive.

En résumé, les choix méthodologiques possibles ressemblent à un exercice d'équilibrisme. Les approches sociolinguistiques ne sont généralement applicables qu'à un nombre limité de variables sémasiologiques, mais elles peuvent fournir des informations sociodémographiques détaillées sur les locuteurs, qui ensuite permettent une interprétation des résultats à la lumière de la théorie variationniste établie. Les approches computationnelles permettent des analyses systématiques, potentiellement à l'échelle du lexique entier, mais ne fournissent que peu ou pas d'informations sur les locuteurs, et comportent des incertitudes quant à leur validité descriptive. Compte tenu de ces constats, le choix fait dans cette thèse est de mettre en œuvre les deux types d'approches de manière complémentaire ; ceci est décrit en détail dans le chapitre 7. Mais tout d'abord, nous abordons la question de l'explication des phénomènes de variation observées dans les données.

Chapitre 6. Explication de la variation linguistique

Dans le chapitre précédent, nous avons vu que différentes d'approches, opérant à des échelles très variées, peuvent être utilisées pour identifier des phénomènes de variation sémasiologique dans des données linguistiques. Certaines analyses procèdent de manière *top-down*, en examinant en détail un ensemble prédéfini de variables sémasiologiques ; d'autres adoptent une approche *bottom-up*, visant à découvrir spontanément des traces de variation sémasiologique. Quoi qu'il en soit, le fait que ces méthodes repèrent des cas potentiellement intéressants ne constitue pas en soi une description sociolinguistique exhaustive. Les comportements linguistiques observés doivent également être expliqués ; en d'autres termes, il faut rendre compte des contraintes qui les influencent.

Cette question est au cœur du chapitre 6. La section 6.1 introduit des critères pour déterminer si un cas de variation linguistique peut être attribué au contact de langues. La section 6.2 décrit les facteurs internes (linguistiques) et externes (sociaux) qui peuvent conditionner la variation linguistique, en se concentrant plus particulièrement sur l'utilisation des glissements de sens en anglais québécois. Cette section s'appuie sur les principes généraux de la théorie variationniste ainsi que sur des études antérieures de l'anglais canadien et québécois. La section 6.3 aborde la question de la signification sociale que pourraient revêtir l'utilisation des glissement de sens. Passant des analyses variationnistes aux analyses computationnelles, la section 6.4 présente les méthodes qui peuvent être utilisées pour étudier les facteurs sociolinguistiques standard dans des corpus de taille importante. La section 6.5 fournit enfin les principales conclusions de cette discussion.

Nous avons tout d'abord proposé que les effets du contact de langues puissent être établis sur la base d'un ensemble de critères stricts, tout en tenant compte de différences d'usage relatives – plutôt que catégoriques – entre les communautés linguistiques et les périodes historiques. Nous avons ensuite passé en revue une série de facteurs internes – tels que la fréquence, les propriétés sémantiques et phonologiques – ainsi que des facteurs externes – notamment l'âge, le genre et le profil linguistique – qui pourraient fournir des explications quantitatives systématiques concernant l'utilisation des glissements de sens induits par le contact. En complément de cette vue d'ensemble, nous avons illustré l'utilité du concept d'indexicalité pour expliquer les interactions entre le sens lexical et le sens social dans l'utilisation des glissements de sens, avec d'autres implications potentielles concernant les mécanismes de communication et le statut de la variation observée. Enfin, d'un point de vue plus pratique, nous avons abordé l'estimation des facteurs externes à partir des corpus de tweets, en soulignant une série de problèmes et de solutions potentielles.

Dans l'ensemble, cette discussion reflète les conclusions déjà formulées concernant les sources de données et les modèles de la variation linguistique : étant donné la spécificité des glissements de sens induits par le contact, une description exhaustive et fiable de ce phénomène ne peut être produite qu'en mettant ensemble différentes approches. Notre objectif est de tirer profit des avantages de chacune d'entre elles tout en contournant leurs limites respectives. L'implémentation spécifique de cette stratégie est présentée dans le chapitre suivant.

Chapitre 7. Aperçu de la méthode

Les chapitres précédents ont présenté le cadre général de cette thèse d'un point de vue théorique, méthodologique et descriptif. Nous avons passé en revue différentes approches ; leur rôle spécifique dans la méthode globale que nous proposons, ainsi que les liens entre elles, sont maintenant abordés de manière explicite. Le chapitre 7 résume donc la position théorique et méthodologique adoptée (7.1), présente une série d'objectifs et d'hypothèses globaux (7.2), et décrit les principales étapes des analyses computationnelles (7.3) et sociolinguistiques (7.4).

De manière globale, cette thèse adopte un double objectif : descriptif et méthodologique. Sur le plan descriptif, nous poursuivons les objectifs spécifiques suivants.

(1) Déterminer la diffusion et le statut des glissements de sens induits par le contact en anglais québécois.

La diffusion peut être considérée en termes de phénomènes internes à la langue, c'est-à-dire en identifiant la portion du lexique affectée par cette pratique langagière, ainsi qu'en termes de la communauté linguistique, c'est-à-dire en caractérisant le sous-ensemble de locuteurs de l'anglais québécois qui présentent ce comportement.

Par ailleurs, le statut sociolinguistique fait référence ici à l'étendu de l'utilisation des glissements de sens au sein de la communauté linguistique. Nous nous attendons à observer une échelle allant d'une forte association avec une maîtrise imparfaite de l'anglais, notamment par des locuteurs francophones, à un usage régional établi, typique du Québec en général. D'un autre point de vue, le statut sociolinguistique peut être analysé en termes de stabilité diachronique d'un phénomène de variation observé en synchronie.

(2) Établir les facteurs sociolinguistiques qui influent sur l'utilisation des glissements de sens induits par le contact.

Il s'agit à la fois des facteurs internes – liés aux caractéristiques inhérentes d'un mot, comme sa fréquence ou son degré de similarité à un équivalent français – et les facteurs externes, y compris les variables standard comme l'âge, le genre et le degré de bilinguisme.

(3) Identifier la signification sociale véhiculée par l'utilisation des glissements de sens induits par le contact.

Cet objectif est fondé sur une analyse de l'indexicalité, en partant de l'hypothèse que ce processus est de nature interactive : certaines significations sociales peuvent être transmises consciemment par les locuteurs, mais d'autres peuvent découler de la perception du comportement d'un locuteur par son interlocuteur.

Quant aux objectifs méthodologiques, ceux-ci peuvent être résumés par un objectif global : la mise en œuvre d'une approche pouvant fournir une description systématique des glissements de sens induits par le contact en anglais québécois, comme décrit ci-dessus. Bien entendu, cela implique plusieurs composantes, partant de l'idée selon laquelle une combinaison de méthodes computationnelles et sociolinguistiques peut fournir le résultat le plus complet. Plus précisément, les méthodes computationnelles devraient :

- identifier, dans l'ensemble du lexique, les mots les plus susceptibles d'être influencés par l'utilisation du français ;
- mettre en évidence les facteurs globaux qui sous-tendent cette variation, tels que reflétés par les données du corpus ;
- pour les mots les plus affectés par le contact, isoler les occurrences individuelles qui reflètent directement l'influence du français.

Ces méthodes comportent des objectifs supplémentaires liés aux données qui sont nécessaires à leur mise en œuvre :

- constituer un corpus qui est (i) suffisamment grand pour assurer la fiabilité des méthodes computationnelles ; (ii) diversifié d'un point de vue régional, pour permettre une approche comparative ; et (iii) contient suffisamment d'informations sur les locuteurs pour permettre une description sociolinguistique globale ;
- créer un jeu d'évaluation pour valider systématiquement les méthodes computationnelles.

En bref, l'utilisation des méthodes computationnelles devrait permettre une analyse de type *bottom-up*, à grande échelle, qui devrait repérer un ensemble de mots affectés par le contact de langues, ainsi que d'occurrences individuelles dans lesquelles l'usage lié au contact peut être observé. Ces résultats représenteront le point de départ de l'analyse sociolinguistique variationniste, qui examinera les mêmes mots de manière plus fine, notamment dans le cadre d'un entretien. Les objectifs spécifiques sont les suivants :

- développer une tâche spécifique pour étudier les glissement de sens dans le cadre d'un entretien, qui devrait à la fois fournir des information quantitatives sur leur utilisation et faire émerger les représentations qui leur sont associées ;
- analyser les données pour identifier les facteurs sociolinguistiques qui conditionnent l'utilisation des glissements de sens, ainsi que les significations sociales qu'ils véhiculent ;
- utiliser les résultats obtenus pour évaluer davantage la validité descriptive des méthodes computationnelles.

Partie III. Analyses sur corpus

Les chapitres dans la partie III présentent les analyses sur corpus menées pour étudier les glissements de sens induits par le contact de langues en anglais québécois. Le chapitre 8 décrit la constitution d'un grand corpus de tweets permettant une comparaison régionale des pratiques langagières, et plus précisément l'identification des caractéristiques spécifiques à Montréal et dans cette mesure potentiellement liées au contact avec le français. Le chapitre 9 introduit deux analyses exploratoires des données recueillies, respectivement axées sur le repérage des mots et des sens spécifiques à Montréal. Ceci fournit une première indication du caractère régionalement spécifique et comparable des données. Cette analyse met en outre en évidence des problèmes méthodologiques liés à l'utilisation des modèles vectoriels statiques dans ce contexte. À partir de ces observations, le chapitre 10 vise une meilleure compréhension des phénomènes langagiers repérés par nos données et modèles. Dans une série d'expériences, nous soulignons plus clairement les limites des modèles statiques, mettons en œuvre une analyse multidimensionnelle pour faciliter une exploration plus approfondie des données et introduisons les modèles vectoriels contextuels afin d'accélérer cette analyse. Enfin, le chapitre 11 aborde plus formellement certains des défis méthodologiques observés. Nous introduisons un jeu d'évaluation pour le repérage des glissements de sens et l'utilisons pour évaluer systématiquement les modèles statiques. Un ensemble de mots d'intérêt est ensuite analysé à l'aide de modèles contextuels, fournissant une caractérisation initiale de leur utilisation. Ceci constitue la base de l'enquête sociolinguistique présentée dans la partie IV.

Chapitre 8. Constitution d'un corpus de tweets pour la variation régionale

Comme établi dans la partie II, les analyses computationnelles menées dans cette thèse nécessitent un corpus très spécifique. Certains critères sont déterminés par l'objectif descriptif consistant à examiner les glissements de sens induits par le contact, en particulier en observant des phénomènes de variation sémasiologique régionale au Canada ; d'autres critères sont liés aux méthodes utilisées pour identifier ces phénomènes. Une solution potentielle pour répondre aux deux types d'exigences – la seule solution qui nous était facilement accessible – consiste à constituer un corpus de tweets, et ce, au moyen d'un procédé de collecte et filtrage de données soigneusement défini.

Cette démarche est décrite dans le chapitre 8. La motivation pour cette approche est clarifiée dans la section 8.1, qui passe en revue les corpus existants à la lumière de nos critère de constitution de corpus. La méthode adoptée pour la collecte des données est décrite dans la section 8.2, et les étapes de filtrage sont présentées dans la section 8.3. La structure du corpus constitué est décrite dans la section 8.4. Un résumé de cette discussion est fourni dans la section 8.5. Notons également que ce chapitre est limité à la collecte et au filtrage des données implémentés dans ce travail. Il s'appuie sur la discussion plus large concernant les données issues de Twitter dans le chapitre 4, qui fournit un contexte général pour les décisions méthodologiques présentées ici.

Comme évoqué ci-dessus, nous exposons tout d'abord une série de critères précis auquel

notre corpus doit répondre. Ceux-ci découlent de l'intersection des objectifs descriptifs et des exigences méthodologiques. Étant donné l'absence d'un corpus existant qui pourrait répondre à ces critères, nous avons introduit une série d'étapes pour collecter et filtrer des données issues de Twitter, visant à trouver un équilibre entre efficacité et fiabilité.

Plus précisément, une première collecte de données a permis d'identifier un ensemble d'utilisateurs de Montréal, Toronto et Vancouver qui avaient envoyé au moins un tweet en anglais. Afin de remédier à la distribution irrégulière des données en fonction des régions et des utilisateurs, ainsi que d'étendre la quantité d'informations disponibles concernant ces derniers, nous avons implémenté une deuxième collecte à partir des profils individuels. Les données récoltées ont ensuite été filtrées en fonction de la localisation et de la langue attestée ; les quasi-doublons ont été automatiquement supprimés. Cette démarche a entraîné une diminution non négligeable de la quantité de données disponibles, qui est néanmoins justifiée par l'amélioration de l'utilité descriptive du corpus.

L'ensemble des expérimentations décrites, menées pendant 11 mois, ont abouti à un corpus contenant 1,3 milliard de tokens, soit 78,8 millions de tweets postés par 196 000 utilisateurs. Ce corpus répond aux critères initialement définis : il est suffisamment grand pour les méthodes de modélisation à grande échelle ainsi que pour des analyses fines des utilisateurs individuels, avec une meilleure répartition des données à travers les utilisateurs et les régions. De plus, il reflète les spécificités nationales et régionales de l'anglais canadien, comme nous le verrons dans le prochain chapitre.

Chapitre 9. Aperçu exploratoire de la variation régionale

La procédure de collecte de données et la structure du corpus décrits dans le chapitre précédent – avec une distinction géographique entre Montréal, Toronto et Vancouver – sont basées sur l'hypothèse selon laquelle les comportements linguistiques qui distinguent Montréal des deux autres villes pourraient refléter l'influence du contact avec le français. Le chapitre 9 introduit deux expériences qui constituent la première étape de vérification de cette hypothèse ; l'accent est mis sur la spécificité et la comparabilité régionales. L'objectif de ces expériences n'est pas de répondre de manière définitive à nos questions de recherche globales, mais plutôt de mettre en évidence les principales tendances dans les données et ainsi fournir une base solide pour des analyses plus systématiques.

La section 9.1 présente une analyse de spécificité, visant à identifier les mots les plus surreprésentés dans les données de Montréal et à identifier ainsi différents types de variation. La section 9.2 se concentre davantage sur la sémantique lexicale à travers l'implémentation exploratoire d'une méthode basée sur l'utilisation de modèles vectoriels, appliqués ici au repérage de la variation sémasiologique régionale. Des considérations pratiques concernant l'analyse des données issues de Twitter sont présentées dans la section 9.3, et un bref résumé clôt le chapitre (section 9.4).

La première expérience présentée dans ce chapitre confirme que le sous-corpus de Montréal est caractérisé par des traits langagiers induits par le contact, tels que les emprunts et les glissements de sens, ainsi que d'autres pratiques typiques de la communication bilingue,

comme le codeswitching. Cette expérience a également mis en évidence une forte variabilité dans l'utilisation de caractéristiques typiques de ce mode de communication : à titre d'exemple, l'utilisation des abréviations informelles semble suivre des régularités régionales et être contrainte par une série de facteurs. Nos observations confirment plus globalement la spécificité régionale et la comparabilité des données collectées, ainsi que la pertinence des corpus basés sur Twitter dans l'étude des phénomènes de variation linguistique.

Dans la deuxième analyse, nous avons mis en œuvre une méthode computationnelle développée dans les études de changements sémantiques diachroniques ; cela nous a permis de montrer que cette approche peut être utilisée pour repérer des exemples de variation sémasiologique régionale observée en synchronie. Notre analyse a permis de repérer des glissements de sens induits par le contact (certains déjà connus, d'autres identifiés pour la première fois), ce qui confirme l'intérêt de l'approche pour cette thèse. Mais une série de problèmes potentiels sont également apparus. D'une part, même lorsque l'usage spécifique à Montréal présente des liens clairs avec le contact de langues, les patterns sous-jacents sont souvent beaucoup plus complexes. Ils impliquent d'habitude la présence des sens conventionnels à Montréal, ainsi que celle des sens liés au contact dans les deux autres villes, mais à des degrés différents. D'autre part, tous les cas de variation sémasiologique régionale ne sont pas liés au contact de langues : différents types de bruit affectent les résultats de l'analyse.

Dans l'ensemble, ces résultats indiquent que nous avons affaire à un phénomène sociolinguistique complexe qu'il n'est pas facile de repérer de manière immédiate à l'aide d'outils déjà disponibles. Il semble nécessaire d'introduire de nombreux ajustements méthodologiques, mais nous avons à ce jour peu d'indications quant aux meilleures pratiques. En effet, les études computationnelles sur des données diachroniques – sur lesquelles repose notre implémentation des modèles vectoriels – adoptent souvent une vision très large du changement sémantique. En revanche, une description sociolinguistique de la variation sémantique en synchronie nécessite une analyse plus précise des variantes linguistiques coexistantes (dans ce cas, des sens différents), la variante d'intérêt apparaissant potentiellement dans un nombre très limité d'occurrences. Le choix des modèles vectoriels, et notamment leur capacité à représenter différents sens, est donc crucial. Il semble également que l'hypothèse qui sous-tend la conception méthodologique proposée – selon laquelle les traits langagiers qui distinguent Montréal de Toronto et de Vancouver seraient liés au contact de langues – ne se vérifie pas dans une version forte, puisque de nombreuses autres sources de variation sont également repérées dans les données. Cependant, il pourrait être possible de caractériser ces types de bruit et de réduire leur impact sur l'analyse globale. Ce chapitre a également évoqué l'exploration manuelle du corpus ; celle-ci continuera à jouer un rôle important, compte tenu notamment du manque de jeu d'évaluation pour les glissements de sens en anglais québécois. Les descriptions sociolinguistiques existantes peuvent nous aider à évaluer les mots potentiellement affectés par le contact, mais cela passe par une exploration manuelle des données.

Chapitre 10. Vers une meilleure compréhension de la variation dans les modèles et les données

La première expérience sur l'utilisation des modèles vectoriels pour le repérage des glissements de sens a montré que cette méthode est prometteuse, comme indiqué par sa capacité à identifier des exemples pertinents, mais elle pose également des défis méthodologiques, reflétés par divers types de bruit dans les résultats. Afin de mettre en œuvre cette méthode de manière plus efficace, et ce avec un objectif descriptif, il est important de mieux comprendre les mécanismes en jeu, notamment lors de la comparaison des représentations vectorielles entre plusieurs modèles. Il est également important d'évaluer si ce type d'information sémantique interagit avec d'autres caractérisations empiriques du lexique, et si des approches alternatives peuvent aider à combler les lacunes de la méthode initiale. Ces points sont abordés par le chapitre 10.

La performance des modèles vectoriels statiques est abordée dans la section 10.1, qui compare une série d'implémentations différentes ; le dispositif expérimental visant à identifier les spécificités régionales est par ailleurs complété par une condition de contrôle. Une analyse multidimensionnelle, présentée dans la section 10.2, est ensuite utilisée pour explorer la contribution de différents types d'informations linguistiques et circonscrire davantage les caractéristiques des mots susceptibles de présenter un intérêt descriptif. La section 10.3 introduit les modèles contextuels, utilisés pour produire une analyse plus fine d'un ensemble de mots précédemment identifiés. La section 10.4 fournit un bref résumé.

Les premières expériences de ce chapitre portent sur les tendances générales observées dans les modèles vectoriels statiques, en examinant 18 configurations et trois mesures de variation sémantique régionale. Les résultats ont mis en évidence la faible qualité de certaines représentations vectorielles, comme en témoigne leur instabilité dans la condition de contrôle. Nous avons également souligné les différences introduites par les différentes configurations et mesures de variation, et ce, en termes des tendances globales observées dans le lexique entier ainsi qu'en termes des mots que ces modèles identifient comme étant influencés par le contact. Pour ce qui est des caractéristiques empiriques des mots, la fréquence s'est avérée particulièrement importante : elle était fortement corrélée à plusieurs mesures typiques de représentations vectorielles instables.

Nous avons ensuite exploré le rapport entre les différents types d'informations caractérisant le lexique à l'aide d'une analyse en composantes principales. En plus de corroborer le rôle central de la fréquence et des mesures qui lui sont associées, cette analyse a permis de mieux identifier la zone du lexique qui est la plus susceptible aux influences liées au contact de langues. Cette approche a également facilité l'identification de glissements de sens supplémentaires, même si les résultats étaient toujours affectés par le bruit.

Dans la dernière expérience, nous avons mis en œuvre un modèle vectoriel contextuel. Cela nous a notamment permis de regrouper automatiquement les occurrences similaires d'un mot d'intérêt. Les clusters ainsi obtenus ont été explorés pour examiner qualitativement les usages d'un mot donné ainsi que pour relier ces derniers aux patterns régionaux. Ces étapes constituent la base d'une analyse quantitative plus poussée portant sur les facteurs explicatifs. Cette expérience nous a également permis de formuler des hypothèses provisoires concernant

le statut des glissements de sens induits par le contact, notamment par rapport au rôle du degré de bilinguisme.

Malgré des résultats descriptifs prometteurs, les mêmes types de bruit ont été noté à toutes les étapes. La récurrence des problèmes méthodologiques, liés à la fois à la structure du corpus et aux caractéristiques inhérentes des méthodes déployées, remet en question la valeur pratique de ces dernières dans la recherche descriptive ; cette observation doit être approfondie. Comme dans le dernier chapitre, les résultats de ces expériences sont cohérents avec une version faible de l'hypothèse globale selon laquelle l'usage spécifique à Montréal pourrait refléter l'influence du français. Les exemples examinés suggèrent que les usages liés au contact sont effectivement beaucoup plus fréquents à Montréal que dans les deux autres villes ; cependant, ils ne représentent généralement qu'une fraction de tous les usages attestés à Montréal, ce qui rend leur découverte plus difficile.

Afin d'aborder certaines des questions en suspens, cette analyse devrait être menée sur un plus grand nombre de mots. Cela permettra de fournir un aperçu descriptif plus étendu ainsi qu'une évaluation plus systématique des méthodes mises en œuvre.

Chapitre 11. Évaluation des contributions descriptives des modèles vectoriels

Les analyses computationnelles menées jusqu'à présent ont souligné que, outre leur potentiel descriptif, les approches mises en œuvre présentent des problèmes méthodologiques récurrents. Il est donc nécessaire de poursuivre l'étude de leur utilité dans la recherche descriptive, notamment en évaluant systématiquement leurs performances sur un plus grand nombre de mots. C'est la direction de recherche poursuivie dans le chapitre 11, qui part plus globalement du fait que les travaux computationnels sur le repérage des changements sémantiques se concentrent généralement sur des questions de recherche et des corpus génériques. Leur objectif est souvent de valider les capacités potentielles d'une méthode computationnelle donnée ; nous nous partons en revanche d'une question descriptive précisément définie – celle qui est au cœur de cette thèse – pour aborder explicitement la contribution descriptive de ces méthodes.

Afin de faciliter une évaluation systématique, nous avons tout d'abord constitué un jeu d'évaluation pour le repérage des glissements de sens (section 11.1). Nous avons ensuite évalué différents modèles statiques et différentes mesures de variation, introduits dans le chapitre précédent, afin de trouver le modèle le plus performant, puis de le déployer sur le repérage de nouveaux glissements de sens (section 11.2). Une analyse basée sur les modèles contextuels, associée à une annotation qualitative, a ensuite été utilisée pour caractériser davantage l'utilisation des glissements de sens et expliquer certains problèmes affectant les modèles statiques (section 11.3). Un résumé des principaux résultats clôt le chapitre (section 11.4).

Comme indiqué plus haut, l'objectif central de ces analyses était de vérifier plus systématiquement les observations émergeant des analyses exploratoires. Pour ce faire, nous avons d'abord développé un jeu d'évaluation, comportant 80 items, pour le repérage des glissements de sens dans des situations de contact anglais-français. Nous l'avons ensuite utilisé pour évaluer les modèles statiques : nous avons constaté des résultats robustes sur une tâche de classifica-

tion standard et une précision très faible sur le repérage de nouveaux glissements de sens ; cela valide de manière plus formelle nos intuitions initiales. Nous avons ensuite étendu l'analyse aux modèles contextuels, que nous avons utilisés pour accélérer l'annotation manuelle des données du corpus, et ce, pour un ensemble de 40 mots. Cela nous a permis de décrire plus précisément les problèmes affectant les modèles statiques, ainsi que de formuler des hypothèses descriptives concernant l'utilisation et la diffusion des glissements de sens induits par le contact de langues.

Ces analyses nous ont permis de formaliser une série d'intuitions méthodologiques, développées au cours de plus de deux ans d'utilisation de modèles vectoriels et de méthodes connexes pour étudier les données issues de Twitter. Mais nos constats ont aussi des implications plus générales, qui réaffirment le rôle central des approches choisies pour constituer les corpus et évaluer les modèles dans le cadre d'analyses computationnelles des changements sémantiques. Ceci est illustré par la forte différence entre les résultats sur le jeu d'évaluation et la tâche de repérage de nouveaux glissements de sens. Cette tendance devrait être prise en compte lors du choix des méthodes d'évaluation, notamment lorsque l'objectif est d'établir la valeur pratique des méthodes étudiées. Par ailleurs, si certains problèmes repérés sont spécifiques à notre corpus, des problèmes similaires peuvent affecter d'autres études de changement sémantique, notamment en ce qui concerne le bruit dans le corpus et les distributions de sens complexes. Enfin, la comparaison des modèles statiques et contextuels a mis en évidence des tendances divergentes dans les données, ce qui indique que les glissements de sens impliquent de multiples dimensions de variation. Une piste pour les travaux futurs consisterait à qualifier les changements sémantiques en plus de quantifier leur présence.

Quant à l'objectif sociolinguistique poursuivi dans cette thèse, ces analyses ont fourni la première description quantitative basée sur corpus de notre objet d'étude. Il s'agit là d'un résultat à part entière, mais il est également essentiel de mieux comprendre les contraintes sur l'utilisation des glissements de sens et les représentations qui leur sont associées. Il est également important de déterminer la mesure dans laquelle les caractérisations issues des analyses computationnelles reflètent les comportements sociolinguistiques observées dans la communication spontanée. Ces questions sont au centre de la partie IV.

Partie IV. Enquête sociolinguistique

Les chapitres dans la partie IV présentent les entretiens sociolinguistiques menés auprès d'un groupe de locuteurs montréalais afin d'évaluer plus précisément l'utilisation des glissements de sens induits par le contact. Le chapitre 12 présente le protocole sociolinguistique et la procédure de recrutement, ainsi que les grands principes orientant l'analyse des données recueillies. Le chapitre 13 décrit la composition de l'échantillon recruté, notamment en termes d'une série de caractéristiques sociodémographiques et attitudinales. Le chapitre 14 s'appuie sur cette description pour étudier plus finement l'utilisation des glissements de sens induits par le contact de langues. Cette analyse fait émerger différents patterns de variation synchronique, tels que reflétés par les scores d'acceptabilité et les remarques qualitatives ; ces tendances pourraient à leur tour refléter des processus de diffusion en diachronie. Notre analyse identifie également un groupe de locuteurs qui semble être particulièrement impliqué dans ces pratiques langagières. Enfin, le chapitre 15 présente une réflexion plus globale sur les analyses menées au cours de cette thèse. Nous mettons en contraste les contributions descriptives des approches basées sur corpus et des entretiens sociolinguistiques, soulignant notamment leur nature complémentaire.

Chapitre 12. Protocole d'entretien et recrutement des participants

Les analyses sur corpus présentées dans la partie III ont abouti à la définition d'un ensemble de 40 mots caractérisés par une influence sémantique du français, qui est par ailleurs attestée dans les données en anglais québécois issues de notre corpus de tweets. Après avoir analysé les contextes linguistiques dans lesquels ces mots apparaissent et quantitativement caractérisé leur utilisation en termes de facteurs sociolinguistiques dérivés du corpus, nous avons pointé une certaine variabilité concernant la diffusion de ces mots au sein de la communauté linguistique et leur association avec l'utilisation du français. Nos analyses à grande échelle ont été déterminantes dans la formulation de ces hypothèses, que nous abordons maintenant de manière plus ciblée au moyen d'entretiens sociolinguistiques. Notre enquête est limitée à un nombre relativement faible de locuteurs, mais elle aboutit à des descriptions fines qui facilitent l'interprétation des tendances globales observées sur corpus.

Le chapitre 12 aborde les considérations méthodologiques qui sous-tendent la conception et la mise en œuvre des entretiens sociolinguistiques menés dans cette thèse. Plus particulièrement, la section 12.1 présente la structure de notre protocole sociolinguistique en s'attardant notamment sur les tâches standard ainsi qu'un nouveau test de perception sémantique. La section 12.2 aborde la manière dont les entretiens sociolinguistiques ont été menés et analysés à partir de ce protocole. La section 12.3 résume cette discussion. Notons que ce chapitre se limite à une présentation du protocole mis en œuvre dans cette thèse. Pour une discussion plus large sur la collecte de données en sociolinguistique variationniste, y compris par le biais d'entretiens, voir le chapitre 4.

Les tâches composant le protocole s'inspirent du protocole standard développée dans le cadre du programme de recherche PAC-LVTI. Certaines parties de ce protocole ont été directement reprises ; d'autres – en particulier le questionnaire thématique – ont été adaptées au

contexte local et à notre objet de l'étude. En outre, le protocole a été étendu à travers une nouvelle tâche conçue pour évaluer un grand nombre de glissements de sens dans le cadre d'un entretien. Nous avons notamment exposé les motivations qui sous-tendent la structure de la tâche, le choix des exemples utilisés et sa mise en œuvre pratique.

Nous avons ensuite présenté la manière dont ce protocole a été déployé pour recueillir des données sociolinguistiques à Montréal. Partant du contexte général dans lequel s'est déroulé ce travail de terrain, nous avons discuté des stratégies qui ont permis le recrutement de 15 participants, ainsi que les choix pratiques effectués pour mener les entretiens et analyser les données enregistrées. Nous avons notamment discuté de trois scores quantitatifs liés à une série de questions centrales pour expliquer les comportements sociolinguistiques observés dans les données : le degré de bilinguisme des locuteurs, leurs attitudes à l'égard des politiques linguistiques et du bilinguisme au Québec, et leur origine géographique.

Si ce chapitre a mis en évidence une série de difficultés pratiques dans la réalisation de l'enquête, l'approche présentée nous a néanmoins permis d'obtenir des données qualitativement riches, directement applicables à l'étude des glissements de sens et produites par un groupe diversifié de locuteurs qui reflètent la grande variété de profils linguistiques typique de Montréal. Les données sont analysées plus en détail dans la suite de cette thèse, en commençant par une description de l'échantillon dans le chapitre suivant.

Chapitre 13. Identification des profils sociolinguistiques

Le chapitre 13 fournit une description des participants recrutés. La section 13.1 présente la structure de l'échantillon en termes de ses principales caractéristiques sociodémographiques. La section 13.2 résume les avis exprimés par les participants sur leur identité et sur les pratiques langagières à Montréal. Afin de tirer profit de l'ensemble des informations disponibles, la section 13.3 introduit une analyse multidimensionnelle, permettant d'identifier des profils de locuteurs distincts dans l'échantillon. La section 13.4 résume les principaux résultats.

La caractérisation globale de l'échantillon obtenue à travers cette analyse initiale sera fondamentale pour expliquer les patterns de variabilité dans la perception des glissements de sens, comme le montre le chapitre 14. Notons par ailleurs que le présent chapitre se limite à une description des participants recrutés. L'importance des caractéristiques sociodémographiques centrales pour la théorie variationniste, y compris dans le contexte de l'anglais québécois, est exposée plus en détail dans le chapitre 6.

Cette caractérisation initiale de notre échantillon a mis en évidence un déséquilibre en termes de genre et d'âge, avec une présence plus importante de femmes et de locuteurs plus jeunes ; une forte variabilité en termes d'origine géographique et de profils linguistiques ; et une homogénéité relative en termes de statut socio-économique. Bien que la diversité globale limite la généralisation des résultats finaux, elle facilite l'exploration de la perception des glissements de sens : cette dernière peut être interprétée à la lumière de profils sociolinguistiques divers, décrits de manière claire et fiable.

Nous avons ensuite résumé les remarques qualitatives exprimées par les participants concernant leur identité individuelle, ainsi que la vie et les pratiques langagières à Montréal. Bien

que les façons spécifiques dont ils définissent leur identité soient variables, la plupart des participants s'identifient comme des Montréalais typiques, et tous expriment des opinions très positives sur la ville. Ils soulignent le rôle central du bilinguisme dans leurs caractérisations de Montréal en général, de la façon dont l'anglais y est parlé en particulier, ainsi que de leur propre identité. Les informations fournies sur les pratiques de communication et sur l'exposition aux langues parlées à Montréal reflètent un cadre propice aux influences induites par le contact de langues.

Le chapitre se termine par une analyse multidimensionnelle rassemblant différents types d'informations afin de discerner des tendances plus générales dans les données. Cette analyse suggère que la principale distinction entre les participants recrutés est liée à leur âge, qui est à son tour associé à des différences de bilinguisme et plus particulièrement de maîtrise du français. Une autre dimension de variation importante est liée aux différents degrés de liens locaux avec Montréal. Les variables sociodémographiques et attitudinales présentées jusqu'à présent seront déployées pour expliquer l'utilisation des glissements de sens dans le prochain chapitre.

Chapitre 14. Statut et diffusion des glissements de sens

Le chapitre 14 analyse l'usage des glissements de sens induits par le contact, tel que reflété par les scores d'acceptabilité recueillis à l'aide du test de perception sémantique ainsi que par les commentaires qualitatifs formulés par les participants. Il s'agit plus précisément d'explorer la variabilité observée entre les différents stimuli ainsi que celle entre les différents locuteurs. Notre objectif principal est de discerner les contraintes externes (sociales) sur ce comportement sociolinguistique et sa diffusion au sein de la communauté linguistique. Notons que l'impact des facteurs internes (linguistiques), ainsi que leur rapport avec les mesures de variation computationnelles, est abordé plus en détail dans le chapitre 15.

La section 14.1 présente les scores d'acceptabilité des glissements de sens individuels, en se concentrant sur leur distribution générale et sur les principales caractéristiques linguistiques qui pourraient expliquer les tendances observées. La section 14.2 explore les distinctions entre les différents glissements de sens, en utilisant une analyse multidimensionnelle pour examiner conjointement l'ensemble des mots et des variables sociodémographiques et attitudinales. La section 14.3 se concentre sur les différences entre les locuteurs individuels, en identifiant les comportements similaires et en les interprétant en fonction de leur rôle potentiel dans la diffusion des glissements de sens. La section 14.4 fournit un résumé des principales observations.

Cette série d'analyses a montré que, de manière générale, l'acceptabilité des différents mots est très variable, allant de ceux qui sont entièrement rejetés à ceux qui sont universellement acceptés. Cela pourrait être en partie lié à la nature des glissements de sens en question. Par ailleurs, si la plupart des mots examinés sont pleinement intégrés dans le système phonologique anglais des participants, un sous-ensemble de locuteurs produit des réalisations gallicisées, qui pourraient refléter des usages très différents des mots en question.

Une analyse multidimensionnelle a ensuite été déployée afin d'explorer les liens potentiels entre la perception des différents glissements de sens et les caractéristiques sociodémo-

graphiques et attitudinales des participants. Nous avons proposé quatre tendances globales pour les mots examinés : (i) une absence d'influence directe du contact de langues ; (ii) des usages régionalement spécifiques qui sont principalement liés au bilinguisme individuel ; (iii) des usages régionalement spécifiques qui sont adoptés par un groupe de locuteurs plus diversifié, et qui perdent ainsi leur lien direct avec le bilinguisme ; (iv) des usages acceptés de manière presque unanime dans la communauté locale. Ces tendances synchroniques reflètent à leur tour une explication potentielle pour la diffusion des glissements en diachronie : leur utilisation pourrait commencer au stade (ii) et évoluer progressivement vers le stade (iv).

Nous avons en outre exploré la variabilité entre les locuteurs à travers une classification hiérarchique ascendante. Cette analyse de clustering nous a permis d'identifier automatiquement des groupes de locuteurs ayant produit des scores d'acceptabilité similaires. D'après ces résultats, l'utilisation des glissements de sens semble particulièrement typique d'un groupe de locuteurs relativement cohérent, qui tend à être plus jeune et à maîtriser à la fois l'anglais et le français. L'ensemble des analyses produites dans ce chapitre fournissent un point de départ solide pour des études futures, encore plus approfondies, de la diffusion des glissements de sens induits par le contact. Une autre question plus générale qui mérite également notre attention est le lien entre les observations issues des entretiens, présentées ici, et les analyses computationnelles discutées précédemment. Cette question est abordée dans le chapitre final.

Chapitre 15. Comparaison des analyses basées sur Twitter et sur la communication spontanée

La série d'analyses présentées au cours des sept derniers chapitres a abordé l'utilisation et la perception des glissements de sens sous différents angles. Les approches computationnelles discutées dans la partie III ont utilisé des modèles vectoriels pour identifier un ensemble de glissements de sens potentiels à partir d'un corpus de tweets, et pour caractériser globalement leur utilisation sur la base des métadonnées disponibles. L'approche sociolinguistique variationniste décrite dans les chapitres précédents de la partie IV a été utilisée pour étudier le même ensemble de mots au moyen d'entretiens en face-à-face. Cela nous a permis d'obtenir des informations détaillées sur le profil sociolinguistique des locuteurs, ainsi que des informations quantitatives et qualitatives sur leur perception des glissements de sens. Le chapitre 15 propose maintenant une mise en commun des résultats issus de ces deux approches, en contrastant les informations qu'elles fournissent et en clarifiant leurs contributions.

La relation entre la communication basée sur Twitter et la communication dans la vie de tous les jours est abordée dans la section 15.1 à partir des commentaires qualitatifs sur cette question recueillis lors des entretiens. Les descriptions des glissements de sens produites par l'ensemble des approches déployées sont explorées dans la section 15.2, notamment en examinant le rapport entre les scores d'acceptabilité obtenus lors des entretiens sociolinguistiques et une série de mesures de variation dérivées du corpus de tweets. Les contributions descriptives globales des deux approches sont abordées dans la section 15.3. Le résumé dans la section 15.4 clôt le chapitre.

Les avis formulés par les participants concernant les pratiques langagières sur les réseaux

sociaux indiquent que les locuteurs plutôt francophones tendent à utiliser activement l'anglais et à y être passivement exposés davantage sur les réseaux sociaux que dans la vie réelle ; ce constat entraîne des implications potentielles pour la constitution de corpus issus des réseaux sociaux. De plus, les participants ne sont généralement pas conscients de phénomènes de variation linguistique sur les réseaux sociaux. Cependant, les exemples repérés par nos analyses computationnelles comme étant spécifiques à Montréal sont quasi-systématiquement reconnus comme tels, validant ainsi l'approche globale adoptée dans cette thèse.

Nous avons ensuite comparé les scores d'acceptabilité obtenus lors des entretiens sociolinguistiques à une série de mesures de variation sémasiologique dérivées du corpus de tweets. Cette analyse a mis en évidence le fait que les mesures basées sur les modèles statiques – utilisées pour repérer des phénomènes de variation sémasiologique entre différentes régions – et les mesures basées sur les modèles contextuels – utilisées pour caractériser davantage la diffusion des usages régionaux – sont associées aux scores d'acceptabilité de manières différentes. Cette tendance, associée au fait que la corrélation entre les scores d'acceptabilité et toutes les autres mesures quantitatives est faible à modérée, suggère que les informations déployées reflètent des aspects différents de l'utilisation des glissements de sens.

En conclusion de ce chapitre, nous avons présenté les contributions plus générales des méthodes basées sur corpus et des entretiens sociolinguistiques aux questions descriptives abordées par cette thèse. Cette discussion a mis en évidence leur nature complémentaire : les analyses sur corpus fournissent des analyses systématiques à grande échelle, couvrant de vastes quantités de données et un grand nombre de locuteurs ; les entretiens sociolinguistiques permettent une étude approfondie axée sur les profils des locuteurs et les variables linguistiques d'intérêt. La configuration interdisciplinaire que nous avons proposée a facilité une description systématique et exhaustives des glissements de sens induits par le contact en anglais québécois.

Résumé des principales contributions

Le travail de recherche présenté dans cette thèse – dans la succession des étapes décrites ci-dessus – a abouti à une série de contributions, en partie déjà évoquées. Nous avons plus précisément fourni les contributions suivantes :

- un ensemble de ressources : un corpus de tweets pour la variation régionale en anglais canadien, contenant 1,3 milliards de token ; un jeu d'évaluation de 80 items pour la classification binaire des glissements de sens ; les annotations manuelles au niveau de clusters pour 40 glissements de sens ; et les enregistrements d'entretiens sociolinguistiques avec 15 Montréalais ;
- une méthodologie exhaustive pour l'analyse des glissements de sens sur corpus : la configuration optimale pour les modèles statiques, une implémentation efficace des modèles contextuels, ainsi qu'un inventaire d'autres outils, sources d'information et précautions méthodologiques ;
- un protocole sociolinguistique variationniste cohérent, comportant une nouvelle tâche pour étudier plus directement l'utilisation des glissements de sens induits par le contact ;
- une description quantitative et qualitative de 40 glissements de sens attestés dans des données empiriques, dont environ la moitié n'avait pas été décrite dans les travaux existants que nous avons consultés ;
- une analyse des patterns de variation et de diffusion des glissements de sens, basée sur les caractéristiques dérivées du corpus, les facteurs sociodémographiques et les représentations exprimées par les locuteurs locaux ;
- une comparaison directe des approches computationnelle et sociolinguistique variationniste.

Les principales ressources produites dans le cadre de ce travail – dont le corpus de tweets, le jeu d'évaluation pour le repérage des glissements de sens et le code utilisé pour les analyses – sont diffusées à l'adresse suivante : http://github.com/FilipMiletic/QuebecEnglish.

Afin d'illustrer plus précisément ces contributions et de mettre en commun les résultats complémentaires obtenus à différentes étapes de la thèse, nous revenons maintenant sur les objectifs et hypothèses globaux formulés au départ (chapitre 7). Nous résumerons d'abord les principaux résultats descriptifs et formulerons ensuite des recommandations méthodologiques. Nous fournirons également des renvois précis aux sections originales rédigées en anglais.

Résultats descriptifs

Le premier des trois objectifs descriptifs initialement définis était de **déterminer la diffusion et le statut des glissements de sens induits par le contact en anglais québécois**. Cette question peut être abordée à différents niveaux.

Du point de vue du lexique, les résultats confirment l'hypothèse globale selon laquelle la diffusion des glissements de sens est plus importante qu'indiqué précédemment. Cette hypothèse est étayée par l'identification de glissements précédemment décrits et nouvellement identifiés dans notre corpus (voir notamment les chapitres 10 et 11) ainsi que par la facilité avec laquelle les locuteurs locaux ont interprété un important ensemble de mots utilisés avec un sens associé au contact (section 14.1.2). Du point de vue de la communauté linguistique, les résultats indiquent une forte diffusion des glissements de sens parmi les locuteurs de l'anglais québécois. Leur utilisation active au sein de cette communauté est confirmée par les distinctions régionales identifiées au moyen de modèles vectoriels (section 11.2) et par la familiarité avec les glissements de sens qu'on démontrée des locuteurs de profils sociolinguistiques très différents (section 14.1.2).

Nous avons également émis l'hypothèse selon laquelle le degré de diffusion varierait en fonction des glissements et des locuteurs individuels. Les deux points sont confirmés par les données, mais les patterns précis sont différents de ceux que nous avions envisagés. Plus précisément, nous avions proposé que la diffusion des glissements de sens pourrait être analysée comme allant d'une forte association avec les locuteurs francophones à un usage régional typique du Québec. Les locuteurs locaux associent les deux valeurs aux glissements de sens (voir ci-dessous), mais l'une n'exclut pas l'autre.

En fait, les glissements de sens qui sont fortement associés à l'utilisation du français sont souvent également caractérisés par une forte spécificité régionale, comme nous l'avons tout d'abord montré à travers une analyse sur corpus (section 11.3.3). Les entretiens ont permis de préciser que cela correspondait probablement au point de départ dans le processus de diffusion des glissements de sens. Ceux-ci peuvent ensuite se généraliser dans la communauté locale, en perdant le lien direct avec le bilinguisme, et dans l'étape finale, devenir presque universellement acceptés (section 14.2).

Le deuxième objectif descriptif global consistait à **établir les facteurs sociolinguistiques influençant l'utilisation des glissements de sens induits par le contact**. L'hypothèse initiale postulait un lien général avec le bilinguisme, reflété par les facteurs internes comme externes ; les données confirment cette hypothèse globale, mais mettent également en évidence des tendances plus complexes.

En ce qui concerne les facteurs internes (linguistiques), une analyse multidimensionnelle sur corpus a mis en évidence un rôle facilitateur de la similarité formelle entre le mots anglais et son équivalent français ; cette information a joué un rôle central dans le repérage de nouveaux glissements de sens (section 10.2). Les données issues des entretiens suggèrent un rôle potentiellement parallèle de la similarité sémantique : les taux d'acceptabilité sont en général plus élevés pour les mots dont le sens induit par le contact (français) est plus proche du sens conventionnel (anglais) (section 14.1.2). Une fréquence plus élevée des glissements de sens pourrait faciliter leur utilisation, comme l'indique une tendance vers la corrélation positive avec les scores d'acceptabilité. Cette tendance n'est cependant pas facile à interpréter car la fréquence interagit avec d'autres mesures dérivées du corpus (section 15.2.1). Quant à l'effet de la gallicisation phonétique, les données sont insuffisantes pour fournir une réponse définitive. Les observations disponibles suggèrent que ce comportement peut avoir à la fois un effet

facilitateur et inhibiteur sur l'utilisation des glissements de sens, et que cet effet pourrait être influencé par d'autres caractéristiques des locuteurs (section 14.1.3).

En ce qui concerne les facteurs externes (sociaux), les analyses sur corpus ont mis en évidence un rôle potentiellement important du degré de bilinguisme et de la spécificité régionale (section 11.3.3). Les entretiens ont confirmé ces observations globales, mais ils ont également permis d'identifier une sous-section plus spécifique de la communauté linguistique qui semble être à l'origine de l'utilisation des glissements de sens. Il s'agit notamment des locuteurs plus jeunes et plus fortement bilingues ; en termes diachroniques, cette tendance indiquerait une diffusion des glissements de sens au fil du temps (section 14.3). Le rôle potentiel de ces facteurs a également été décrit ci-dessus dans la discussion sur la diffusion des glissements de sens. Plus généralement, ces tendances doivent être validées sur un échantillon de participants plus robuste.

L'objectif descriptif final consistait à **identifier les significations sociales véhiculées par l'utilisation des glissements de sens induits par le contact**. Nous avons obtenu des indications initiales grâce aux analyses sur corpus, principalement à partir de commentaires métalinguistiques soulignant les liens perçus avec l'usage du français (p. ex. section 11.1.1.2). Les entretiens ont fourni des informations plus précises, avec un rôle particulièrement important des associations avec le bilinguisme anglais–français et le caractère régional de ces pratiques langagières (section 14.2). Comme ces deux valeurs ne s'excluent pas mutuellement – c'est-à-dire qu'elles peuvent toutes deux être associées au même mot – elles ne semblent pas correspondre à des statuts sociolinguistiques différents, ce qui est contraire à mon hypothèse initiale. De manière plus générale, ces résultats fournissent une preuve supplémentaire de la forte valeur symbolique des variantes lexicales induites par le contact de langues en anglais québécois.

Recommendations méthodologiques

L'objectif méthodologique général poursuivi par cette thèse était de mettre en œuvre une approche pouvant fournir une description systématique des glissements de sens induits par le contact en anglais québécois, et ce, en tirant parti de méthodes computationnelles et sociolinguistiques pour obtenir le résultat le plus exhaustif possible. Les étapes concrètes définies initialement correspondent à la mise en œuvre des différentes méthodes, résumées en termes globaux tout au long de ce résumé. Nous revenons maintenant plus en détail sur certaines décisions majeures afin de donner un aperçu des choix méthodologiques à privilégier et d'autres recommandations générales.

Nous commençons par les expériences menées sur corpus, et plus particulièrement la constitution du corpus de tweets (chapitre 8). Ce corpus a occupé un rôle central dans la suite de la thèse : il a permis des analyses computationnelles à grande échelle ainsi que des caractérisations qualitatives des glissements, qui ont ensuite été examinés dans les entretiens sociolinguistiques. Toutefois, l'utilisation de ce type de données dans le cadre d'une recherche linguistique nécessite des précautions. Parmi les problèmes majeurs, citons la distribution fortement irrégulière des données en fonction des utilisateurs et la présence du bruit dans les données. L'impact potentiel de ces problèmes est illustré, entre autres, par le fait que la caractérisation d'un mot

donnée peut être fortement biaisée par un seul utilisateur très actif (section 9.2.3). Nous avons constaté une variété de problèmes de ce type ; il semble donc peu judicieux d'utiliser un corpus de tweets sans rééquilibrer le nombre de tweets par utilisateur, par exemple en créant un sous-échantillon des données initialement collectées. D'autres décisions de filtrage, telles que l'exclusion des quasi-doublons, sont également très pertinentes.

En ce qui concerne les implémentations des modèles vectoriels, les résultats des évaluations systématiques ont fourni des indications claires sur les approches les plus performantes, du moins sur le jeu d'évaluation que nous avons utilisé (chapitre 11). En ce qui concerne les modèles statiques utilisés pour repérer les glissements reflétés par des phénomènes de variation régionale, les résultats montrent qu'il est préférable d'utiliser :

(i) les modèles neuronaux (`word2vec`) plutôt que les modèles à base de fréquences (PPMI) ;

(ii) pour `word2vec`, les vecteurs comportant 100 plutôt que 300 dimensions ;

(iii) un score de variation sémantique qui prend également en compte les informations issues d'une région de contrôle (`diff`), plutôt qu'un score focalisé uniquement sur la région d'intérêt (`avg`) ;

(iv) les distances de cosinus moyennes, calculées sur plusieurs exécutions de la même implémentation, notamment pour limiter l'effet de l'instabilité du modèle ;

(v) des fenêtres de mots plus faibles et un alignement des modèles basé sur l'analyse procustéenne (même si les résultats sont moins clairs sur ces deux points).

En ce qui concerne les modèles contextuels, nous avons montré l'utilité descriptive d'une implémentation basée sur des représentations sémantiques qui sont extraites d'un modèle BERT pré-entraîné puis regroupées dans des clusters en utilisant la méthode de propagation d'affinité. Cette approche nous a permis d'identifier des occurrences individuelles des glissements de sens ainsi que d'évaluer le rapport entre leur utilisation et des descripteurs sociolinguistiques dérivés du corpus.

Les méthodes mises en œuvre nous ont permis de répondre, de manière globale, aux objectifs méthodologiques initialement définis, à savoir le repérage des mots – et de leurs occurrences individuelles – les plus affectés par le contact, ainsi que l'analyse des phénomènes de variation sous-tendant leur utilisation. Cependant, le recours à ces méthodes comporte également des défis. Cela est particulièrement vrai pour le repérage spontané de nouveaux glissements de sens dans une perspective *bottom-up*, qui est fortement affecté par le bruit dans les données et les modèles. D'autres types d'utilisation semblent donc plus adéquats. Les modèles statiques – qu'il est plus facile de mettre en œuvre et d'appliquer au lexique entier – semblent particulièrement bien adaptés aux analyses *top-down*, qui permettent de valider des hypothèses précises grâce aux mesures quantitatives qui peuvent être obtenues. Les modèles contextuels sont particulièrement utiles pour faciliter les analyses linguistiques manuelles ainsi que pour quantifier l'utilisation des sens différents d'un mot donné. Les expériences menées dans cette thèse suggèrent que le repérage de nouveaux glissements de sens nécessite la prise en compte d'informations supplémentaires, au-delà des représentations sémantiques vectorielles, ainsi qu'une expertise linguistique. Les méthodes qui simplifient la tâche tout en intégrant le

jugement humain, comme l'analyse en composantes principales, représentent une perspective prometteuse.

Pour ce qui est de l'enquête sociolinguistique, le principal défi méthodologique consistait à concevoir un test de perception sémantique pouvant être intégré à l'entretien standard. La solution mise en œuvre – basée sur des questionnaires dialectaux mais utilisée dans le cadre d'un entretien en face-à-face – s'est avérée bien adaptée à nos objectifs. Il faut toutefois noter que chaque sous-tâche individuelle – lecture du glissement de sens en contexte ; évaluation de son acceptabilité ; identification d'un synonyme ; commentaires sur son utilisation – a fourni des informations essentielles pour l'interprétation des résultats. Par ailleurs, nous avons dû trouver une solution pour analyser systématiquement un échantillon hétérogène et de taille relativement réduite ; l'analyse exploratoire multivariée que nous avons mise en œuvre représente une manière efficace d'explorer les tendances dans ce type de données. Enfin, comme indiqué précédemment, les résultats obtenus à partir du corpus et des entretiens se sont révélés complémentaires. Cela confirme l'intérêt des sources de données et des méthodes computationnelles que nous avons implémentées pour la recherche sociolinguistique. D'un autre point de vue, aucune des approches mises en œuvre ne semble être capable de fournir une description exhaustive lorsqu'elle est utilisée de manière isolée : il semble donc préférable de mettre ensemble les indices à grande échelle basés sur corpus, les contributions des membres de la communauté linguistique étudiée, et l'expertise des linguistes.

References

Abitbol, J. L., Fleury, E., and Karsai, M. (2019). Optimal proxy selection for socioeconomic status inference on Twitter. *Complexity*, 2019:6059673.

Adsett, M. and Morin, M. (2005). Contact and attitudes towards bilingualism in Canada. In Cohen, J., McAlister, K. T., Rolstad, K., and MacSwan, J., editors, *ISB4: Proceedings of the 4th International Symposium on Bilingualism*, pages 1–17. Cascadilla Press, Somerville, MA.

Ajao, O., Hong, J., and Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, 41(6):855–864.

Al Zamal, F., Liu, W., and Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 387–390.

Aletras, N. and Chamberlain, B. P. (2018). Predicting Twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, pages 20–24, Baltimore MD USA. ACM.

Alimoradian, K. (2014). 'Makes me feel more Aussie': Ethnic identity and vocative *mate* in Australia. *Australian Journal of Linguistics*, 34(4):599–623.

Alvarez-Carmona, M. A., Pellegrin, L., Montes-y-Gómez, M., Sánchez-Vega, F., Escalante, H. J., López-Monroy, A. P., Villaseñor-Pineda, L., and Villatoro-Tello, E. (2018). A visual approach for age and gender identification on Twitter. *Journal of Intelligent & Fuzzy Systems*, 34(5):3133–3145.

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Anttila, A. (2002). Variation and phonological theory. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 206–243. Blackwell, Malden.

Armstrong, N. (1998). La variation sociolinguistique dans le lexique français. *Zeitschrift für romanische Philologie*, 114(3):462–495.

Avis, W. S. (1967). Introduction. In Avis, W. S., editor, *Dictionary of Canadianisms on Historical Principles*, pages xii–xv. Gage, Toronto.

Avis, W. S., Crate, C., Drysdale, P., Leechman, D., Scargill, M., and Lovell, C. (1967). *Dictionary of Canadianisms on Historical Principles*. Gage, Toronto.

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., and Kamps, J. (2017). Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, CIKM '17, pages 1509–1518, New York, NY, USA. Association for Computing Machinery.

Bailey, G. (2016). Automatic detection of sociolinguistic variation using forced alignment. *University of Pennsylvania Working Papers in Linguistics*, 22(2):3.

Bailey, L. R. and Durham, M. (2020). A cheeky investigation: Tracking the semantic change of *cheeky* from *monkeys* to *wines*: Can social media spread linguistic change? *English Today*, pages 1–10.

Baker, C. and Prys Jones, S. (1998). *Encyclopedia of Bilingualism and Bilingual Education*. Multilingual Matters, Clevedon.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *International Joint Conference on Natural Language Processing*, pages 14–18.

Baldwin, T., de Marneffe, M.-C., Han, B., Kim, Y.-B., Ritter, A., and Xu, W. (2015). Shared tasks of the 2015 Workshop on Noisy User-generated Text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Barbaresi, A. (2016). Collection and indexing of tweets with a geographical focus. In *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC)*, pages 24–27.

Barbaud, P. (1998). French in Quebec. In Edwards, J., editor, *Language in Canada*, pages 177–201. Cambridge University Press, Cambridge.

Barber, K., editor (2004). *Canadian Oxford Dictionary*. Oxford University Press, Don Mills.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1313–1316, Lisbon, Portugal. European Language Resources Association.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Barysevich, A. (2012). *Variation et changement lexicaux en situation de contact de langues*. PhD thesis, Western University, London, ON.

Basile, P., Caputo, A., Caselli, T., Cassotti, P., and Varvara, R. (2020). DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*.

Basile, P., Caputo, A., Luisi, R., and Semeraro, G. (2016). Diachronic analysis of the Italian language exploiting Google Ngram. *Third Italian Conference on computational Linguistics CLiC-it 2016.*

Basile, P. and McGillivray, B. (2018). Exploiting the web for semantic change detection. In Soldatova, L., Vanschoren, J., Papadopoulos, G., and Ceci, M., editors, *Discovery Science*, volume 11198, pages 194–208. Springer International Publishing, Cham.

Basile, V. and Nissim, M. (2013). Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia. Association for Computational Linguistics.

Beaton, M. E. and Washington, H. B. (2015). Slurs and the indexical field: The pejoration and reclaiming of favelado 'slum-dweller'. *Language Sciences*, 52:12–21.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual event, Canada. ACM.

Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on Twitter. In *Proceedings of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS)*, Redmond, USA.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., and Wilson, T. (2012). Language identification for creating language-specific Twitter collections. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pages 65–74.

Bernard Barbeau, G. (2018). 40 ans après, qu'en est-il de la loi 101 ? Représentations et discours conflictuels dans la presse québécoise. *Circula*, 7:52–69.

Bhatt, R. M. (1997). Code-switching, constraints, and optimal grammars. *Lingua*, 102(4):223–251.

Bigot, D. (2016). De la variation lexicale en franco-ontarien: les données du corpus de Casselman (Ontario). *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 61(1):1–30.

Bigot, D. and Papen, R. A. (2013). Sur la « norme » du français oral au Québec (et au Canada en général). *Langage et société*, 146:115–132.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, 9:81.

Birdsong, D. and Molis, M. (2001). On the evidence for maturational constraints in second-language acquisition. *Journal of Memory and Language*, 44(2):235–249.

Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Blodgett, S. L., Wei, J., and O'Connor, B. (2017). A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Blom, J.-P. and Gumperz, J. J. (1972). Social meaning in linguistic structure: Code-switching in Norway. In Gumperz, J. J. and Hymes, D., editors, *Directions in Sociolinguistics: The Ethnography of Communication*, pages 407–434. Holt, Rinehart and Winston, New York.

Blondeau, H. (2020). Pratiques langagières et diversité culturelle chez de jeunes Montréalais: le français dans la métropole. In Reinke, K., editor, *Attribuer un sens: La diversité des pratiques langagières et les représentations sociales*, pages 151–175. Les Presses de l'Université Laval, Québec, QC.

Blondeau, H., Nagy, N., Sankoff, G., and Thibault, P. (2002). La couleur locale du français L2 des anglo-montréalais. *Acquisition et interaction en langue étrangère*, 17.

Bloomfield, L. (1933). *Language*. Holt, Rinehart & Winston, New York.

Boberg, C. (2004a). The Dialect Topography of Montreal. *English World-Wide*, 25(2):171–198.

Boberg, C. (2004b). Ethnic patterns in the phonetics of Montreal English. *Journal of Sociolinguistics*, 8(4):538–568.

Boberg, C. (2004c). Real and apparent time in language change: Late adoption of changes in Montreal English. *American Speech*, 79(3):250–269.

Boberg, C. (2005a). The Canadian shift in Montreal. *Language Variation and Change*, 17:133–154.

Boberg, C. (2005b). The North American Regional Vocabulary Survey: New variables and methods in the study of North American English. *American Speech*, 80(1):22–60.

Boberg, C. (2008). Regional phonetic differentiation in Standard Canadian English. *Journal of English Linguistics*, 36(2):129–154.

Boberg, C. (2010). *The English Language in Canada: Status, History and Comparative Analysis*. Cambridge University Press, Cambridge.

Boberg, C. (2012). English as a minority language in Quebec. *World Englishes*, 31(4):493–502.

Boberg, C. (2014). Ethnic divergence in Montreal English. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 59(1):55–82.

Boberg, C. (2016). Newspaper dialectology: Harnessing the power of the mass media to study Canadian English. *American Speech*, 91(2):109–138.

Boberg, C. and Hotton, J. (2015). English in the Gaspé region of Quebec. *English World-Wide*, 36(3):277–314.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–234.

Boot, A. B., Tjong Kim Sang, E., Dijkstra, K., and Zwaan, R. A. (2019). How character limit affects language usage in tweets. *Palgrave Communications*, 5:76.

Bourhis, R. Y. (2001). Reversing language shift in Quebec. In Fishman, J. A., editor, *Can Threatened Languages Be Saved?*, pages 101–141. Multilingual Matters, Clevedon.

Bourhis, R. Y. (2012). Social psychological aspects of French-English relations in Quebec: From vitality to linguicism. In Bourhis, R. Y., editor, *Decline and Prospects of the English-Speaking Communities of Quebec*, pages 313–378. Canadian Heritage, Ottawa.

Bourhis, R. Y. and Landry, R. (2002). La loi 101 et l'aménagement du paysaqe linguistique au Québec. *Revue d'aménagement linguistique*, hors série (L'aménagement linguistique au Québec : 25 ans d'application de la Charte de la langue française):107–132.

Bourhis, R. Y. and Landry, R. (2012). Group vitality, cultural autonomy and the wellness of language minorities. In Bourhis, R. Y., editor, *Decline and Prospects of the English-Speaking Communities of Quebec*, pages 23–69. Canadian Heritage, Ottawa.

Bourhis, R. Y., Montaruli, E., and Amiot, C. E. (2007). Language planning and French-English bilingual communication: Montreal field studies from 1977 to 1997. *International Journal of the Sociology of Language*, 185:187–224.

boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10, Honolulu, HI. IEEE.

Bright, W. (1966). *Sociolinguistics. Proceedings of the UCLA Sociolinguistics Conference, 1964*. Mouton, Hague.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates.

Bucholtz, M. (2002). From 'sex differences' to gender variation in sociolinguistics. *University of Pennsylvania Working Papers in Linguistics*, 8(3):33–45.

Budinich, L. (2016). La variazione semasiologica di sottocodice: Un esempio di analisi lessicale corpus-based. *CHIMERA. Romance Corpora and Linguistic Studies*, 3(1):23–56.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Bušta, J., Herman, O., Jakubíček, M., Krek, S., and Novak, B. (2017). JSI Newsfeed corpus. In *The 9th International Corpus Linguistics Conference*, Birmingham, UK.

Butler, Y. G. (2012). Bilingualism/multilingualism and second language acquisition. In Bhatia, T. K. and Ritchie, W. C., editors, *The Handbook of Bilingualism and Multilingualism*, pages 109–136. Wiley-Blackwell, Malden.

Cajolet-Laganière, H. (2021). L'essor d'une norme endogène au Québec : l'exemple du dictionnaire Usito. *Gragoatá*, 26(54):105–138.

Cajolet-Laganière, H., Martel, P., Masson, C.-É., and Mercier, L. (2014). Usito. https://usito.usherbrooke.ca/.

Canada (1982). Canadian Charter of Rights and Freedoms. https://laws-lois.justice.gc.ca/eng/const/page-12.html.

Canada (1985). Official Languages Act. https://laws-lois.justice.gc.ca/eng/acts/o-3.01/.

Canada (2013). First Nations in Canada. https://www.rcaanc-cirnac.gc.ca/eng/1307460755710/1536862806124.

Caselles-Dupré, H., Lesaint, F., and Royo-Letelier, J. (2018). Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 352–356, Vancouver, British Columbia, Canada. ACM.

Cerruti, M. (2011). Il concetto di variabile linguistica a livello del lessico. *Studi italiani di linguistica teorica e applicata*, 40(2):211–231.

Cesare, N., Grant, C., and Nsoesie, E. O. (2017). Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv:1702.01807*.

Chambers, J. K. (1994). An introduction to dialect topography. *English World-Wide*, 15(1):35–53.

Chambers, J. K. (1995). The Canada-US border as a vanishing isogloss: The evidence of *chesterfield*. *Journal of English Linguistics*, 23(1-2):155–166.

Chambers, J. K. (1998). Social embedding of changes in progress. *Journal of English Linguistics*, 26(1):5–36.

Chambers, J. K. (2000). Region and language variation. *English World-Wide*, 21(2):169–199.

Chambers, J. K. (2002). Patterns of variation including change. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 349–372. Blackwell, Malden.

Chambers, J. K. (2007a). Geolinguistic patterns in a vast speech community. *Linguistica Atlantica*, 28:27–36.

Chambers, J. K. (2007b). A linguistic fossil: Positive *any more* in the Golden Horseshoe. In Reich, P., Sullivan, W. J., Lommel, A. R., and Griffen, T., editors, *Lacus Forum XXXIII: Variation*, pages 31–44. Linguistic Association of Canada and the United States, Houston, TX.

Chambers, J. K. (2010). English in Canada. In Gold, E. and McAlpine, J., editors, *Canadian English: A Linguistic Reader*, number 6 in Strathy Occasional Papers on Canadian English, pages 1–37. Queen's University, Kingston, ON.

Chambers, J. K. and Heisler, T. (1999). Dialect topography of Québec City English. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 44(1):23–48.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW).

Chatellier, H. (2016). *Nivellement et contre-nivellement phonologique à Manchester: étude de corpus dans le cadre du projet PAC-LVTI*. PhD thesis, Université Toulouse - Jean Jaurès, Toulouse.

Cheshire, J. (1987). Syntactic variation, the linguistic variable, and sociolinguistic theory. *Linguistics*, (25):257–282.

Childs, B., Van Herk, G., and Thorburn, J. (2011). Safe harbour: Ethics and accessibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory*, 7(1):163–180.

Chiu, B., Crichton, G., Korhonen, A., and Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany. Association for Computational Linguistics.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Clarke, S., Elms, F., and Youssef, A. (1995). The third dialect of English: Some Canadian evidence. *Language Variation and Change*, 7(2):209–228.

Clyne, M. (1987). Constraints on code switching: How universal are they? *Linguistics*, 25(4):739–764.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., and Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning - ICML '08*, pages 160–167, Helsinki, Finland. ACM Press.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates.

Cook, P., Han Lau, J., Rundell, M., McCarthy, D., and Baldwin, T. (2013). A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Proceeedings of eLex 2013*, pages 49–65.

Costa, A., Heij, W. L., and Navarrete, E. (2006). The dynamics of bilingual lexical access. *Bilingualism: Language and Cognition*, 9(2):137–151.

Côté, M.-H. and Remysen, W. (2019). L'adaptation phonologique des emprunts à l'anglais dans les dictionnaires québécois. In Dister, A. and Piron, S., editors, *Les discours de référence sur la langue française*, pages 173–195. Presses de l'Université Saint-Louis, Bruxelles.

Coto-Solano, R., Stanford, J. N., and Reddy, S. K. (2021). Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with DARLA. *Frontiers in Artificial Intelligence*, 4:662097.

Couture, C. (2021). Quebec. *The Canadian Encyclopedia*.

D'Arcy, A. (2004). Contextualizing St. John's youth English within the Canadian quotative system. *Journal of English Linguistics*, 32(4):323–345.

D'Arcy, A. (2007). Like and language ideology: Disentangling fact from fiction. *American Speech*, 82(4):386–419.

D'Arcy, A. (2017). *Discourse-Pragmatic Variation in Context: Eight Hundred Years of LIKE*. John Benjamins, Amsterdam.

Davies, M. (2011). N-grams data from the Corpus of Contemporary American English (COCA). http://www.ngrams.info.

Davies, M. (2013a). Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries (GloWbE). https://www.english-corpora.org/now/.

Davies, M. (2013b). Corpus of News on the Web (NOW): 3+ billion words from 20 countries, updated every day. https://www.english-corpora.org/now/.

Davies, M. (2018). The 14 Billion Word iWeb Corpus. https://www.english-corpora.org/iWeb/.

De Pascale, S. (2019). *Token-Based Vector Space Models as Semantic Control in Lexical Lectometry*. PhD thesis, KU Leuven, Leuven.

De Wolf, G. D. (1996). Word choice: Lexical variation in two Canadian surveys. *Journal of English Linguistics*, 24(2):131–155.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Del Tredici, M. and Fernández, R. (2017). Semantic variation in online communities of practice. In *IWCS 2017 – 12th International Conference on Computational Semantics – Long Papers*.

Del Tredici, M., Fernández, R., and Boleda, G. (2019). Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

Delais-Roussarie, E. and Post, B. (2014). Corpus annotation. Methodology and transcription systems. In Durand, J., Gut, U., and Kristoffersen, G., editors, *The Oxford Handbook of Corpus Phonology*, pages 46–88. Oxford University Press, Oxford.

Demszky, D., Sharma, D., Clark, J., Prabhakaran, V., and Eisenstein, J. (2021). Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Dendien, J. and Pierrel, J.-M. (2003). Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. *Traitement Automatique des Langues*, 44(2):11–37.

Denis, D. (2015). *The Development of Pragmatic Markers in Canadian English*. PhD thesis, University of Toronto, Toronto, ON.

Denis, D. (2017). The development of *and stuff* in Canadian English: A longitudinal study of apparent grammaticalization. *Journal of English Linguistics*, 45(2):157–185.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing*, pages 198–206.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dickinson, J. A. (2007). The English-speaking minority of Quebec: A historical perspective. *International Journal of the Sociology of Language*, 2007(185):11–24.

Diebold, A. R. (1961). Incipient bilingualism. *Language*, 37(1):97–112.

Dijkstra, T. and van Heuven, W. J. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175–197.

Dinu, G., Pham, N. T., and Baroni, M. (2013). DISSECT - DIStributional SEmantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia, Bulgaria. Association for Computational Linguistics.

Dollinger, S. (2010). New data for an English usage puzzle: The long history of spelling variation in Canadian English and its linguistic implications [handout]. In *Sixteenth International Conference on English Historical Linguistics (ICEHL-16)*, Pécs.

Dollinger, S. (2012). The western Canada-US border as a linguistic boundary: The roles of L1 and L2 speakers. *World Englishes*, 31(4):519–533.

Dollinger, S. (2015). *The Written Questionnaire in Social Dialectology: History, Theory, Practice*. John Benjamins, Amsterdam.

Dollinger, S. (2017). TAKE UP #9 as a semantic isogloss on the Canada-US border. *World Englishes*, 36(1):80–103.

Dollinger, S. (2020). English in Canada. In Nelson, C. L., Proshina, Z. G., and Davis, D. R., editors, *The Handbook of World Englishes*, pages 52–69. Wiley Blackwell, Hoboken, NJ.

Dollinger, S. (2022). Canadian English lexis and semantics: A historical-comparative resource in contrastive, real-time perspective, 1683–2016. In Kytö, M. and Siebers, L., editors, *Earlier North American Englishes*. John Benjamins, Amsterdam [forthc.].

Dollinger, S. and Brinton, L. J. (2008). Canadian English lexis: Historical and variationist perspectives. *Anglistik: International Journal of English Studies*, 19(2):43–64.

Dollinger, S. and Clarke, S. (2012). On the autonomy and homogeneity of Canadian English. *World Englishes*, 31(4):449–466.

Dollinger, S. and Fee, M. (2017). DCHP-2: The dictionary of Canadianisms on historical principles, second edition. http://www.dchp.ca/dchp2.

D'Onofrio, A. (2020). Personae in sociolinguistic variation. *WIREs Cognitive Science*, 11(6):e1543.

Donoso, G. and Sánchez, D. (2017). Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 16–25, Valencia, Spain. Association for Computational Linguistics.

Donovan, P. (2019). English-speaking Quebecers. *The Canadian Encyclopedia*.

Doyle, G. (2014). Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Dubossarsky, H., Hengchen, S., Tahmasebi, N., and Schlechtweg, D. (2019). Time-Out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

Dubossarsky, H., Tsvetkov, Y., Dyer, C., and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In *Proceedings of the NetWordS Final Conference*, pages 66–70.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2016). Verbs change more than nouns: A bottom-up computational approach to semantic change. *Lingue e linguaggio*, 15(1):5–25.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Durand, J., Gut, U., and Kristoffersen, G. (2014). *The Oxford Handbook of Corpus Phonology*. Oxford University Press, Oxford.

Durand, J. and Przewozny, A. (2012). La phonologie de l'anglais contemporain : usages, variétés et structure. *Revue française de linguistique appliquée*, XVII(1):25–37.

Durand, M. (2002). *Histoire du Québec*. Imago, Paris.

Durkin, P. (2012). Variation in the lexicon: The 'Cinderella' of sociolinguistics?: Why does variation in word forms and word meanings present such challenges for empirical research? *English Today*, 28(4):3–9.

Eckert, P. (2000). *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Blackwell, Malden.

Eckert, P. (2006). Communities of practice. *Elsevier Encyclopedia of Language and Linguistics*.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Eckert, P. (2019). The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776.

Eckert, P. and McConnell-Ginet, S. (1992a). Communities of practice: Where language, gender, and power all live. In Hall, K., Bucholtz, M., and Birch, M., editors, *Locating Power*, pages 89–99, Berkeley. Berkeley Women and Language Group.

Eckert, P. and McConnell-Ginet, S. (1992b). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, 21:461–490.

Edwards, J. (2012). Bilingualism and multilingualism: Some central concepts. In Bhatia, T. K. and Ritchie, W. C., editors, *The Handbook of Bilingualism and Multilingualism*, pages 5–25. Wiley-Blackwell, Malden.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of NAACL-HLT 2013*, pages 359–369.

Eisenstein, J. (2018). Identifying regional dialects in on-line social media. In Boberg, C., Nerbonne, J., and Watt, D., editors, *The Handbook of Dialectology*, pages 368–383. John Wiley & Sons, Hoboken, NJ.

Eisenstein, J., Ahmed, A., and Xing, E. P. (2011). Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048, Madison, WI, USA. Omnipress.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287.

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114.

Elchacar, M. and Salita, A. L. (2019). Étude diachronique du discours normatif sur les anglicismes dans les chroniques de langue au Canada francophone : d'Alphonse Lusignan à Guy Bertrand. *Circula*, 9:5–28.

Eleta, I. and Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Feagin, C. (2002). Entering the community: Fieldwork. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 20–39. Blackwell, Malden.

Fee, M. (1991). Frenglish in Quebec English newspapers. In *Papers of the Fifteenth Annual Meeting of the Atlantic Provinces Linguistic Association*, pages 12–23. Atlantic Provinces Linguistic Association.

Fee, M. (2008). French borrowing in Quebec English. *Anglistik: International Journal of English Studies*, 19(2):173–188.

Fiesler, C. and Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, 4(1):1–14.

Fink, C., Kopecky, J., and Morawski, M. (2012). Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 459–462.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford.

Fišer, D. and Ljubešić, N. (2018). Distributional modelling for semantic shift detection. *International Journal of Lexicography*, 32(2):163–183.

Flekova, L., Ungar, L., and Preoţiuc-Pietro, D. (2016). Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Francis, W. S. (2005). Bilingual semantic and conceptual representation. In Kroll, J. F. and De Groot, A. M. B., editors, *Handbook of Bilingualism. Psycholinguistic Approaches*, pages 251–267. Oxford University Press, Oxford.

Franco, K., Geeraerts, D., Speelman, D., and Van Hout, R. (2019). Concept characteristics and variation in lexical diversity in two Dutch dialect areas. *Cognitive Linguistics*, 30(1):205–242.

Franco, K. and Tagliamonte, S. A. (2021). Interesting *fellow* or tough old *bird*? 3rd person male referents in Ontario. *American Speech*, 96(2):192–216.

Frermann, L. and Lapata, M. (2016). A Bayesian Model of Diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Gardner, M. H., Denis, D., Brook, M., and Tagliamonte, S. A. (2021). *Be like* and the Constant Rate Effect: From the bottom to the top of the *S*-curve. *English Language and Linguistics*, 25(2):281–324.

Garí Soler, A. and Apidianaki, M. (2021). Let's play Mono-Poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Gauthier, M., Guille, A., Rico, F., and Deseille, A. (2015). Text mining and Twitter to analyze British swearing habits. In Levallois, C., Marchand, M., Mata, T., and Panisson, A., editors, *Handbook of Twitter for Research*, pages 27–44. EMLYON, Lyon.

Geeraerts, D. (1993). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics*, 4(3):223–272.

Geeraerts, D. (1997). *Diachronic Prototype Semantics: A Contribution to Historical Lexicology*. Clarendon Press, Oxford.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford University Press, Oxford.

Geeraerts, D. and Kristiansen, G. (2014). Cognitive linguistics and language variation. In Littlemore, J. and Taylor, J. R., editors, *The Bloomsbury Companion to Cognitive Linguistics*, pages 202–217. Bloomsbury Academic, London.

Geeraerts, D., Kristiansen, G., and Peirsman, Y. (2010). Introduction. Advances in Cognitive Sociolinguistics. In Geeraerts, D., Kristiansen, G., and Peirsman, Y., editors, *Advances in Cognitive Sociolinguistics*, pages 1–20. De Gruyter Mouton, Berlin.

Gilani, Z., Farahbakhsh, R., Tyson, G., and Crowcroft, J. (2019). A large-scale behavioural analysis of bots and humans on Twitter. *ACM Transactions on the Web*, 13(1):1–23.

Giles, H., Bourhis, R. Y., and Taylor, D. M. (1977). Towards a theory of language in ethnic group relations. In Giles, H., editor, *Language, Ethnicity and Intergroup Relations*, pages 307–348. Academic Press, London.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Giulianelli, M., Del Tredici, M., and Fernández, R. (2020). Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Gonçalves, B. and Sánchez, D. (2014). Crowdsourcing dialect characterization through Twitter. *PLoS ONE*, 9(11):e112074.

Gonen, H., Jawahar, G., Seddah, D., and Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Gorman, K., Howell, J., and Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics / Acoustique canadienne*, 39(3):192–193.

Gouws, S., Metzler, D., Cai, C., and Hovy, E. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29, Portland, Oregon, USA.

Graham, M., Hale, S. A., and Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, 66(4):568–578.

Grant, P. (2010). English usage in contemporary Quebec: Reflections of the local. In Gold, E. and McAlpine, J., editors, *Canadian English: A Linguistic Reader*, number 6 in Strathy Occasional Papers on Canadian English, pages 177–197. Queen's University, Kingston, ON.

Grant-Russell, P. (1999). The influence of French on Quebec English: Motivation for lexical borrowing and integration of loanwords. *LACUS Forum*, 26:473–486.

Grant-Russell, P. and Beaudet, C. (1999). Lexical borrowings from French in written Quebec English: Perspectives on motivation. *University of Pennsylvania Working Papers in Linguistics*, 6(2):17–33.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1(2):67–81.

Gries, S. T. (2006). Corpus-based methods and cognitive semantics: The many senses of *to run*. In Gries, S. T. and Stefanowitsch, A., editors, *Corpora in Cognitive Linguistics*, pages 57–99. De Gruyter Mouton, Berlin.

Gries, S. T. (2015). Polysemy. In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 472–490. De Gruyter Mouton, Berlin.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., and Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*, 2:11.

Grieve, J., Nini, A., and Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46(4):293–319.

Grosjean, F. (1997). The bilingual individual. *Interpreting. International Journal of Research and Practice in Interpreting*, 2(1-2):163–187.

Grosjean, F. (2008). *Studying Bilinguals*. Oxford University Press, Oxford.

Grosjean, F. (2010). *Bilingual: Life and Reality*. Harvard University Press, Cambridge, MA.

Grosjean, F. (2012). An attempt to isolate, and then differentiate, transfer and interference. *International Journal of Bilingualism*, 16(1):11–21.

Grosjean, F. (2013). Bilingualism: A short introduction. In Grosjean, F. and Li, P., editors, *The Psycholinguistics of Bilingualism*, pages 5–25. Wiley-Blackwell, Malden.

Grosjean, F. (2015). Bicultural bilinguals. *International Journal of Bilingualism*, 19(5):572–586.

Gruzd, A. and Mai, P. (2020). *The State of Social Media in Canada 2020*. Ryerson University Social Media Lab, Toronto, ON.

Gruzd, A., Wellman, B., and Takhteyev, Y. (2011). Imagining Twitter as an imagined community. *American Behavioral Scientist*, 55(10):1294–1318.

Gueunier, N. (2003). Attitudes and representations in sociolinguistics: Theories and practice. *International Journal of the Sociology of Language*, (160):4–62.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press, Cambridge.

Guo, D. and Chen, C. (2014). Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, 18(3):370–384.

Gupta, S. and DiPadova, A. (2019). Deep learning and sociophonetics: Automatic coding of rhoticity using neural networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 92–96, Minneapolis, Minnesota. Association for Computational Linguistics.

Haber, J. and Poesio, M. (2020). Assessing polyseme sense similarity through co-predication acceptability and contextualised embedding distance. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 114–124, Barcelona, Spain (Online). Association for Computational Linguistics.

Haber, J. and Poesio, M. (2021). Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hakuta, K. and D'Andrea, D. (1992). Some properties of bilingual maintenance and loss in Mexican background high-school students. *Applied Linguistics*, 13(1):72–99.

Hamilton, D. E. (1958). Notes on Montreal English. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 4(2):70–79.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Harris, B. P. (1975). *Selected Political, Cutlural, and Socio-Economic Areas of Canadian History as Contributors to the Vocabulary of Canadian English*. PhD thesis, University of Victoria, Victoria, BC.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hasan, R. (2009). *Semantic Variation: Meaning in Society and in Sociolinguistics*. Equinox, London.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26(2):210–231.

Haugen, E. (1969). *The Norwegian Language in America: A Study in Bilingual Behavior*. Indiana University Press, Bloomington.

Hengchen, S., Tahmasebi, N., Schlechtweg, D., and Dubossarsky, H. (2021). Challenges for computational lexical semantic change. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors, *Computational Approaches to Semantic Change*, pages 341–372. Language Science Press, Berlin.

Hoffman, M. F. and Walker, J. A. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*, 22(1):37–67.

Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–10, Waikoloa, Hawaii, USA. IEEE.

Hu, R., Li, S., and Liang, S. (2019). Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Hultgren, A. K. (2014). Lexical variation at the internationalized university: Are indexicality and authenticity always relevant? In Lacoste, V., Leimgruber, J., and Breyer, T., editors, *Indexing Authenticity: Sociolinguistic Perspectives*, pages 304–323. De Gruyter Mouton, Berlin.

Ilbury, C. (2020). "Sassy Queens": Stylistic orthographic variation in Twitter and the enregisterment of AAVE. *Journal of Sociolinguistics*, 24(2):245–264.

Ilson, R. F. (1992). Lexeme. In McArthur, T., editor, *The Oxford Companion to the English Language*, pages 599–600. Oxford University Press, Oxford, New York.

Insee (2022). Comparateur de territoire: France métropolitaine. https://www.insee.fr/fr/statistiques/1405599?geo=METRO-1.

Jauhiainen, M. (2020). 'Cool, cool cool cool': A diachronic corpus study on adjectives of positive evaluation in spoken British English. Master's thesis, Tampere University, Tampere, Finland.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network analysisDD '07*, pages 56–65, San Jose, California. ACM Press.

Jawahar, G., Sagot, B., and Seddah, D. (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Johnson, E. (1996). *Lexical Change and Variation in the Southeastern United States, 1930-1990*. University of Alabama Press, Tuscaloosa.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1):60–99.

Jones, T. (2015). Toward a description of African American Vernacular English dialect regions using "Black Twitter". *American Speech*, 90(4):403–440.

Jørgensen, A., Hovy, D., and Søgaard, A. (2015). Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China. Association for Computational Linguistics.

Joseph, K., Landwehr, P. M., and Carley, K. M. (2014). Two 1%s don't make a whole: Comparing simultaneous samples from twitter's streaming API. In Kennedy, W. G., Agarwal, N., and Yang, S. J., editors, *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 75–83. Springer International Publishing, Cham.

Josselin, A. (2001). *L'emprunt lexical en France et au Canada : le cas particulier des anglicismes et des gallicismes et leur traitement lexicographique*. DEA thesis, Université de Lyon II, Lyon.

Jung, S.-G., An, J., Kwak, H., Salminen, J., and Jansen, B. J. (2018). Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, pages 624–627.

Jurafsky, D. and Martin, J. H. (2022). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition draft.

Jurgens, D., Dimitrov, S., and Ruths, D. (2014). Twitter users #CodeSwitch hashtags! #MoltoImportante #wow. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 51–61, Doha, Qatar. Association for Computational Linguistics.

Jurgens, D., Finnethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 188–197.

Kachru, B. B. (1965). The *Indianness* in Indian English. *Word*, 21(3):391–410.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In Quirk, R. and Widdowson, H. G., editors, *English in the World:*

Teaching and Learning the Language and Literatures, pages 11–30. Cambridge University Press, Cambridge.

Kachru, B. B. (2008). World Englishes in World Contexts. In Momma, H. and Matto, M., editors, *A Companion to the History of the English Language*, pages 567–580. Wiley-Blackwell, Oxford.

Karami, A., Lundy, M., Webb, F., and Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. *IEEE Access*, 8:67698–67717.

Kastronic, L. (2011). Discourse *like* in Quebec English. *University of Pennsylvania Working Papers in Linguistics*, 17(2):105–114.

Kaufmann, M. (2010). Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, Kharagpur, India.

Kiesling, S. F. (2004). Dude. *American Speech*, 79(3):281–305.

Kiesling, S. F. (2009). Style as stance: Stance as the explanation for patterns of sociolinguistic variation. In Jaffe, A., editor, *Stance: Sociolinguistic Perspectives*, pages 171–194. Oxford University Press, Oxford.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. (2014). Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Knowles, R., Carroll, J., and Dredze, M. (2016). Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 108–113, Austin, Texas. Association for Computational Linguistics.

Koch, P. (2016). Meaning change and semantic shifts. In Juvonen, P. and Koptjevskaja-Tamm, M., editors, *The Lexical Typology of Semantic Shifts*, number 58 in Cognitive Linguistics Research, pages 21–66. De Gruyter Mouton, Berlin.

Koptjevskaja-Tamm, M. (2016). "The lexical typology of semantic shifts": An introduction. In Juvonen, P. and Koptjevskaja-Tamm, M., editors, *The Lexical Typology of Semantic Shifts*, number 58 in Cognitive Linguistics Research, pages 1–20. De Gruyter Mouton, Berlin.

Kreutz, T. and Daelemans, W. (2020). Streaming language-specific Twitter data with optimal keywords. In *Proceedings of the 12th Web as Corpus Workshop*, pages 57–64, Marseille, France. European Language Resources Association.

Kroll, J. F. and Ma, F. (2018). The bilingual lexicon. In Fernández, E. M. and Cairns, H. S., editors, *The Handbook of Psycholinguistics*, pages 294–319. John Wiley & Sons, Hoboken, NJ.

Kroll, J. F. and Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33:149–174.

Kulkarni, V., Al-Rfou, R., Perozzi, B., and Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635, Geneva. International World Wide Web Conferences Steering Committee.

Kulkarni, V., Perozzi, B., and Skiena, S. (2016). Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):615–618.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Labov, W. (1994). *Principles of Linguistic Change. Volume 1: Internal Factors*. Blackwell, Oxford.

Labov, W. (2001). *Principles of Linguistic Change. Volume 2: Social Factors*. Blackwell, Oxford.

Labov, W. (2004). Quantitative analysis of linguistic variation. In Dittmar, N., Mattheier, K. J., and Trudgill, P., editors, *Sociolinguistics: An International Handbook of the Science of Language and Society*, volume 1, pages 6–21. Mouton de Gruyter, Berlin.

Labov, W. (2006). *The Social Stratification of English in New York City*. Cambridge University Press, Cambridge.

Labov, W. (2010). *Principles of Linguistic Change. Volume 3: Cognitive and Cultural Factors*. Wiley-Blackwell, Oxford.

Labov, W., Ash, S., and Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. De Gruyter Mouton, Berlin.

Lackenbauer, P. W., Moses, J., Sheffield, R. S., and Gohier, M. (2010). *A Commemorative History of Aboriginal People in the Canadian Military*. National Defence, Ottawa.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., and Schulte im Walde, S. (2021). Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Laitinen, M., Lundberg, J., Levin, M., and Martins, R. (2018). The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. In *Digital Humanities in the Nordic Countries*.

Lamarre, P. (2013). Catching "Montréal on the move" and challenging the discourse of unilingualism in Québec. *Anthropologica*, 55(1):41–56.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology*, 60(1):44–51.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lau, J. H., Cook, P., McCarthy, D., Newman, D., and Baldwin, T. (2012). Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

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

Lave, J. and Wenger, E. (1991). *Situated Learning. Legitimate Peripheral Participation*. Cambridge University Press, Cambridge.

Leimgruber, J. R. E. and Fernández-Mallat, V. (2021). Language attitudes and identity building in the linguistic landscape of Montreal. *Open Linguistics*, 7(1):406–422.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Rivista di linguistica*, 20(1):1–31.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. John Wiley & Sons, New York.

Lepage, J.-F. (2017a). *English, French and Official Language Minorities in Canada*. Statistics Canada, Ottawa.

Lepage, J.-F. (2017b). *English–French Bilingualism Reaches New Heights*. Statistics Canada, Ottawa.

Lepage, J.-F. (2020). *Interpreting and Presenting Census Language Data*. Statistics Canada, Ottawa.

Levy, O. and Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Levy, O. and Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Li, P. (2013). Successive language acquisition. In Grosjean, F. and Li, P., editors, *The Psycholinguistics of Bilingualism*, pages 145–167. Wiley-Blackwell, Malden.

Lichtenberk, F. (1991). Semantic change and heterosemy in grammaticalization. *Language*, 67(3):475–509.

Lindberg, C. A. (2012). *Oxford American Writer's Thesaurus*. Oxford University Press, Oxford.

Linteau, P. A. (2017). *Une histoire de Montréal*. Boréal, Montréal.

Linteau, P.-A. (2021). Québec since Confederation. *The Canadian Encyclopedia*.

Liu, F., Weng, F., and Jiang, X. (2012). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pages 1035–1044. Association for Computational Linguistics.

Liu, F., Weng, F., Wang, B., and Liu, Y. (2011). Insertion, deletion, or substitution?: Normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, pages 71–76. Association for Computational Linguistics.

Liu, H., Bates, E., and Li, P. (1992). Sentence interpretation in bilingual speakers of English and Chinese. *Applied Psycholinguistics*, 13(4):451–484.

Ljubešić, N. and Fišer, D. (2016). Private or corporate? Predicting user types on Twitter. In *Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT)*, pages 4–12, Osaka, Japan.

Ljubešić, N., Fišer, D., and Erjavec, T. (2014). TweetCaT: A tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2279–2283, Reykjavik, Iceland. European Language Resources Association.

Lui, M. and Baldwin, T. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Lui, M. and Baldwin, T. (2014). Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden. Association for Computational Linguistics.

Lynn, T. and Scannell, K. (2019). Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology Workshop 2019*, pages 32–40.

Mackey, W. F. (1962). The description of bilingualism. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 7(2):51–85.

Macnamara, J. and Kushnir, S. L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior*, 10(5):480–487.

Malik, Z. and Haidar, S. (2020). Online community development through social interaction — K-Pop stan twitter as a community of practice. *Interactive Learning Environments*, pages 1–19.

Marian, V. and Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient". *Applied Psycholinguistics*, 42(2):527–548.

Martel, P. (2006). Le français standard en usage au Québec : question de normes et d'usages. *Revue belge de philologie et d'histoire*, 84(3):845–864.

Martinc, M., Kralj Novak, P., and Pollak, S. (2020a). Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020b). Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 343–349, New York, NY, USA. Association for Computing Machinery.

Mathieu, J. (2021). New France. *The Canadian Encyclopedia*.

Matras, Y. (2009). *Language Contact*. Cambridge University Press, Cambridge.

McArthur, T. (1989). *The English Language as Used in Quebec: A Survey*. Number 3 in Strathy Occasional Papers on Canadian English. Queen's University, Kingston, ON.

McArthur, T. (1992a). Semantic change. In McArthur, T., editor, *The Oxford Companion to the English Language*, page 912. Oxford University Press, Oxford.

McArthur, T. (1992b). Sense. In McArthur, T., editor, *The Oxford Companion to the English Language*, pages 917–918. Oxford University Press, Oxford.

McArthur, T. and Fee, M. (1992). Quebec. In McArthur, T., editor, *The Oxford Companion to the English Language*, pages 831–833. Oxford University Press, Oxford.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017*, pages 498–502. ISCA.

McCandless, M. (2014). Chromium compact language detector. https://code.google.com/p/chromium-compact-language-detector.

McCorriston, J., Jurgens, D., and Ruths, D. (2015). Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 650–653.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., and McCullough, D. (2012). On building a reusable Twitter corpus. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12*, pages 1113–1114, Portland, Oregon, USA. ACM Press.

McGillivray, B., Hengchen, S., Lähteenoja, V., Palma, M., and Vatri, A. (2019). A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*, 34(4):893–907.

Meney, L. (2017). *Le français québécois entre réalité et idéologie : Un autre regard sur la langue. Étude sociolinguistique*. L'espace public. Les Presses de l'Université Laval, Québec.

Mercier, L., Remysen, W., and Cajolet-Laganière, H. (2017). Québec. In Reutner, U., editor, *Manuel des francophonies*, pages 277–310. De Gruyter, Berlin.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, Scottsdale, Arizona, USA.

Miletic, F. (2016). *Lexical and Morphosyntactic Features of Canadian English on Twitter*. BA thesis, University of Genoa, Genoa.

Miletic, F. (2018). Contact-induced lexical and morphosyntactic phenomena in Quebec English. Master's thesis, University of Genoa, Genoa.

Milne, P. (2014). *The Variable Pronunciations of Word-Final Consonant Clusters in a Force Aligned Corpus Of*. PhD thesis, University of Ottawa, Ottawa.

Milroy, L. (1987). *Language and Social Networks*. Basil Blackwell, Oxford.

Milroy, L. (2002). Social networks. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 549–572. Blackwell, Malden.

Milroy, L. and Gordon, M. (2003). *Sociolinguistics. Method and Interpretation*. Blackwell, Malden.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 554–557.

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., and Mukherjee, A. (2015). An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Mitra, S., Mitra, R., Riedl, M., Biemann, C., Mukherjee, A., and Goyal, P. (2014). That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

Mohammady, E. and Culotta, A. (2014). Using county demographics to infer attributes of Twitter users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 7–16, Baltimore, Maryland. Association for Computational Linguistics.

Montariol, S., Martinc, M., and Pivovarova, L. (2021). Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Morgan-Lopez, A. A., Kim, A. E., Chew, R. F., and Ruddle, P. (2017). Predicting age groups of Twitter users based on language and metadata features. *PLoS ONE*, 12(8):e0183537.

Mott, B. and Laso, N. J. (2019). Semantic borrowing in language contact. In Grant, A. P., editor, *The Oxford Handbook of Language Contact*, pages 156–172. Oxford University Press, New York, NY.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford University Press, Oxford.

Nadasdi, T. (2005). Living in Ontario French. *Canadian Journal of Applied Linguistics / Revue canadienne de linguistique appliquée*, 8(2):167–181.

Nadasdi, T. and Mckinnie, M. (2003). Living and working in immersion French. *Journal of French Language Studies*, 13(1):47–61.

Nadasdi, T., Mougeon, R., and Rehner, K. (2004). Expression de la notion de « vehicule automobile » dans le parler des adolescents de l'Ontario. *Francophonies d'Amérique*, 17(1):91–106.

Nadasdi, T., Mougeon, R., and Rehner, K. (2008). Factors driving lexical variation in L2 French: A variationist study of automobile, auto, voiture, char and machine. *Journal of French Language Studies*, 18(3):365–381.

Nakatani, S. (2010). Language detection library [slides]. https://www.slideshare.net/shuyo/language-detection-library-for-java.

Nerbonne, J. (2018). Section 2 - Methods. Introduction. In Boberg, C., Nerbonne, J., and Watt, D., editors, *The Handbook of Dialectology*, pages 233–239. John Wiley & Sons, Hoboken, NJ.

Newman, J. (2016). Semantic shift. In Riemer, N., editor, *The Routledge Handbook of Semantics*, pages 266–280. Routledge, Abingdon.

Nguyen, D. (2021). Dialect variation on social media. In Zampieri, M. and Nakov, P., editors, *Similar Languages, Varieties, and Dialects. A Computational Perspective*, pages 204–218. Cambridge University Press, Cambridge.

Nguyen, D. and Cornips, L. (2016). Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86, Berlin, Germany. Association for Computational Linguistics.

Nguyen, D. and Doğruöz, A. S. (2013). Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA. Association for Computational Linguistics.

Nguyen, D., Gravel, R., Trieschnigg, D., and Meder, T. (2013). "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 439–448.

Nguyen, D. and Rosé, C. P. (2011). Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 76–85.

Nguyen, D., Rosseel, L., and Grieve, J. (2021). On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Nguyen, D., Trieschnigg, D., and Cornips, L. (2015). Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669.

OCOL (2022). *Official Languages Tracking Survey 2021*. Office of the Commissioner of Official Languages, Ottawa.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

OQLF (2019). *Rapport sur l'évolution de la situation linguistique au Québec*. Office québécois de la langue française, Montréal.

Otheguy, R. (2012). Concurrent models and cross-linguistic analogies in the study of prepositional stranding in French in Canada. *Bilingualism: Language and Cognition*, 15(2):226–229.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Papalexakis, E., Nguyen, D., and Doğruöz, A. S. (2014). Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 42–50, Doha, Qatar. Association for Computational Linguistics.

Papen, R. A. (1998). French: Canadian varieties. In Edwards, J., editor, *Language in Canada*, pages 160–176. Cambridge University Press, Cambridge.

Paris, C., Thomas, P., and Wan, S. (2012). Differences in language and style between two social media communities. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 539–542.

Patrick, P. L. (2002). The speech community. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 573–597. Blackwell, Malden.

Paul, H. (1891). *Principles of the History of Language*. Longmans, Green, and Co., London.

Pavalanathan, U. and Eisenstein, J. (2015a). Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Pavalanathan, U. and Eisenstein, J. (2015b). Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148, Lisbon, Portugal. Association for Computational Linguistics.

Pavlenko, A. and Blackledge, A. (2004). Introduction: New theoretical approaches to the study of negotiation of identities in multilingual contexts. In Pavlenko, A. and Blackledge, A., editors, *Negotiation of Identities in Multilingual Contexts*, pages 1–33. Multilingual Matters, Clevedon.

Pearson, B. Z. (2007). Social factors in childhood bilingualism in the United States. *Applied Psycholinguistics*, 28(3):399–410.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peersman, C., Daelemans, W., Vandekerckhove, R., Vandekerckhove, B., and Vaerenbergh, L. V. (2016). The effects of age, gender and region on non-standard linguistic variation in online social networks. *arXiv:1601.02431*.

Peirsman, Y., Heylen, K., and Geeraerts, D. (2010). Applying word space models to sociolinguistics. Religion names before and after 9/11. In Geeraerts, D., Kristiansen, G., and Peirsman, Y., editors, *Advances in Cognitive Sociolinguistics*, pages 111–137. De Gruyter Mouton, Berlin.

Penfield, W. and Roberts, L. (1959). *Speech and Brain Mechanisms*. Princeton University Press, Princeton, NJ.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., and McGillivray, B. (2019). GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26.

Pfeffer, J., Mayer, K., and Morstatter, F. (2018). Tampering with Twitter's Sample API. *EPJ Data Science*, 7:50.

Pienemann, M. and Keßler, J.-U. (2007). Measuring bilingualism. In Auer, P. and Wei, L., editors, *Handbook of Multilingualism and Multilingual Communication*, pages 247–275. De Gruyter Mouton, Berlin.

Pierrejean, B. and Tanguy, L. (2018). Predicting Word Embeddings Variability. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 154–159, New Orleans, Louisiana. Association for Computational Linguistics.

Pivovarova, L. and Kutuzov, A. (2021). RuShiftEval: A shared task on semantic shift detection for Russian. In *Computational Linguistics and Intellectual Technologies*, number 20, pages 533–545.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. *Linguistics*, 18:581–618.

Poplack, S. (1993). Variation theory and language contact. In Preston, D. R., editor, *American Dialect Research: An Anthology Celebrating the 100th Anniversary of the American Dialect Society*, pages 251–263. Benjamins, Amsterdam.

Poplack, S. (2008). Quebec English. *Anglistik: International Journal of English Studies*, 19(2):189–200.

Poplack, S. (2012). What does the Nonce Borrowing Hypothesis hypothesize? *Bilingualism: Language and Cognition*, 15(3):644–648.

Poplack, S. (2015). Code switching: Linguistic. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences*, volume 3, pages 918–925. Elsevier, Amsterdam.

Poplack, S. and Meechan, M. (1995). Patterns of language mixture: Nominal structure in Wolof-French and Fongbe-French bilingual discourse. In Muysken, P. and Milroy, L., editors, *One Speaker, Two Languages*, pages 199–232. Cambridge University Press, Cambridge.

Poplack, S. and Meechan, M. (1998). Introduction: How languages fit together in codemixing. *International Journal of Bilingualism*, 2(2):127–138.

Poplack, S., Robillard, S., Dion, N., and Paolillo, J. C. (2020). Revisiting phonetic integration in bilingual borrowing. *Language*, 96(1):126–159.

Poplack, S., Sankoff, D., and Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Poplack, S., Walker, J. A., and Malcolmson, R. (2006). An English "like no other"?: Language contact and change in Quebec. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 51(2/3):185–213.

Poplack, S., Zentz, L., and Dion, N. (2012a). Phrase-final prepositions in Quebec French: An empirical study of contact, code-switching and resistance to convergence. *Bilingualism: Language and Cognition*, 15(2):203–225.

Poplack, S., Zentz, L., and Dion, N. (2012b). What counts as (contact-induced) change. *Bilingualism: Language and Cognition*, 15(2):247–254.

Pratt, T. (1993). The hobgoblin of Canadian English spelling. In Clarke, S., editor, *Focus on Canada*, pages 45–64. Benjamins, Amsterdam.

Pražák, O., Přibáň, P., Taylor, S., and Sido, J. (2020). UWB at SemEval-2020 task 1: Lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 246–254, Barcelona (online). International Committee for Computational Linguistics.

Preoţiuc-Pietro, D., Lampos, V., and Aletras, N. (2015). An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association*

for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1754–1764, Beijing, China. Association for Computational Linguistics.

Preston, D. R. (2002). Language with an attitude. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 40–66. Blackwell, Malden.

Przewozny, A., Viollain, C., and Navarro, S. (2020). *The Corpus Phonology of English: Multifocal Analyses of Variation*. Edinburgh University Press, Edinburgh.

Przewozny-Desriaux, A. (2016). *De la phonologie de corpus à la sociolinguistique. Enjeux de définition de la communauté linguistique australienne*. HDR thesis, Université Toulouse - Jean Jaurès, Toulouse.

Purschke, C. and Hovy, D. (2019). Lörres, Möppes, and the Swiss. (Re)Discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography*, 7(2):113–134.

Pütz, M., Robinson, J. A., and Reif, M. (2014). The emergence of Cognitive Sociolinguistics: An introduction. In Pütz, M., Robinson, J. A., and Reif, M., editors, *Cognitive Sociolinguistics: Social and Cultural Variation in Cognition and Language Use*, pages 1–22. John Benjamins, Amsterdam.

Quebec (1977). Charte de la langue française. https://www.legisquebec.gouv.qc.ca/fr/document/lc/C-11.

Quirk, R. (1985). The English language in a global context. In Quirk, R. and Widdowson, H. G., editors, *English in the World: Teaching and Learning the Language and Literatures*, pages 1–6. Cambridge University Press, Cambridge.

Radice, M. (2000). *Feeling Comfortable? The Urban Experience of Anglo-Montrealers*. Presses de l'Université Laval, Sainte-Foy, QC.

Ramponi, G., Brambilla, M., Ceri, S., Daniel, F., and Di Giovanni, M. (2019). Vocabulary-based community detection and characterization. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1043–1050, Limassol Cyprus. ACM.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents – SMUC '10*.

Reddy, S. and Stanford, J. N. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard*, 1(1):15–28.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Remysen, W. (2016). Langue et espace au Québec: les Québécois perçoivent-ils des accents régionaux? *Lingue, culture, mediazioni*, (3):31–57.

Remysen, W. (2019). Les communautés francophones dans les provinces majoritairement anglophones du Canada : aperçu et enjeux. *Travaux de linguistique*, 78(1):15–45.

Remysen, W., Salita, A. L., and Barrière, M. (2020). Les accents régionaux au Québec : représentations et perceptions linguistiques dans la région de Beauce. *Cahiers de l'Association d'études en langue française*, 23(1):21–54.

Robinson, J. A. (2010). *Awesome* insights into semantic variation. In Geeraerts, D., Kristiansen, G., and Peirsman, Y., editors, *Advances in Cognitive Sociolinguistics*, pages 85–110. De Gruyter Mouton, Berlin.

Robinson, J. A. (2012a). A *gay* paper: Why should sociolinguistics bother with semantics? *English Today*, 28(4):38–54.

Robinson, J. A. (2012b). A sociolinguistic approach to semantic change. In Allan, K. and Robinson, J. A., editors, *Current Methods in Historical Semantics*, pages 199–232. De Gruyter Mouton, Berlin.

Robinson, J. A. (2014). Quantifying polysemy in Cognitive Sociolinguistics. In Glynn, D. and Robinson, J. A., editors, *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, pages 87–115. John Benjamins, Amsterdam.

Rodda, M. A., Lenci, A., and Senaldi, M. S. G. (2017). *Panta Rei*: Tracking semantic change with distributional semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, 3(1):11–24.

Rodríguez-Ordóñez, I. (2021). The role of social meaning in contact-induced variation among new speakers of Basque. *Journal of Sociolinguistics*, 25(4):533–556.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Romaine, S. (1984). On the problem of syntactic variation and pragmatic meaning in sociolinguistic theory. *Folia Linguistica*, 18(3-4):409–437.

Romaine, S. (1995). *Bilingualism*. Wiley-Blackwell, Oxford.

Romaine, S. (2012). The bilingual and multilingual community. In Bhatia, T. K. and Ritchie, W. C., editors, *The Handbook of Bilingualism and Multilingualism*, pages 445–465. Wiley-Blackwell, Malden.

Rosenfeld, A. and Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics.

Rosenfelder, I., Fruehwald, J., Evanini, K., and Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite. http://fave.ling.upenn.edu.

Rouaud, J. (2019a). French loanwords in Canadian English: A usage-based approach. *Anglophonia*, 28.

Rouaud, J. (2019b). *Lexical and Phonological Integration of French Loanwords into Varieties of Canadian English since the Seventeenth Century*. PhD thesis, Université Toulouse - Jean Jaurès, Toulouse.

Rudolph, M. and Blei, D. (2018). Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web – WWW '18*, pages 1003–1011, Lyon, France. ACM Press.

Rudra, K., Sharma, A., Bali, K., Choudhury, M., and Ganguly, N. (2019). Identifying and analyzing different aspects of English-Hindi code-switching in Twitter. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(3):1–28.

Russ, R. B. (2013). Examining regional variation through online geotagged corpora. Master's thesis, The Ohio State University, Columbus, OH.

Russell, P. (1996). An investigation of lexical borrowings from French in Quebec English. *LACUS Forum*, 23:429–440.

Sabourin, P. and Bélanger, A. (2015). The dynamics of language shift in Canada. *Population*, 70(4):727–757.

Sahlgren, M. (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, Stockholm.

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di linguistica*, 20(1):33–53.

Sankoff, D. and Laberge, S. (1978). The linguistic market and the statistical explanation of variability. In Sankoff, D., editor, *Linguistic Variation: Models and Methods*, pages 239–250. Academic Press, New York, NY.

Sankoff, D., Poplack, S., and Vanniarajan, S. (1990). The case of the nonce loan in Tamil. *Language Variation and Change*, 2(1):71–101.

Sankoff, D., Thibault, P., and Bérubé, H. (1978). Semantic field variability. In Sankoff, D., editor, *Linguistic Variation: Models and Methods*, pages 23–43. Academic Press, New York, NY.

Sankoff, G. (1980). Above and beyond phonology in variable rules. In Sankoff, G., editor, *The Social Life of Language*, pages 81–93. University of Pennsylvania Press, Philadelphia.

Sankoff, G. (1997). Deux champs sémantiques chez les anglophones et les francophones de Montréal. In Auger, J. and Rose, Y., editors, *Explorations du lexique*, pages 133–145. CIRAL, Québec.

Sankoff, G. (2002). Linguistic outcomes of language contact. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 638–668. Blackwell, Malden.

Sankoff, G. (2006). Age: Apparent time and real time. *Elsevier Encyclopedia of Language and Linguistics*.

Sankoff, G. and Thibault, P. (1977). L'alternance entre les auxiliaires *avoir* et *être* en français parlé à Montréal. *Langue française*, 34(1):81–108.

Saywell, J. (1996). *Canada: Pathways to the Present*. Stoddart, Toronto.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 486–493, Istanbul, Turkey. European Language Resources Association.

Scheffler, T. (2014). A German Twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2284–2289, Reykjavik, Iceland. European Language Resources Association.

Schindler, M., Legendre, G., and Mbaye, A. (2008). Violations of the PF interface condition in Urban Wolof. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 44(2):169–184.

Schlechtweg, D., Hätty, A., Del Tredici, M., and Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23. International Committee for Computational Linguistics.

Schlechtweg, D., Schulte im Walde, S., and Eckmann, S. (2018). Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings*

of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Schleef, E. (2014). Written surveys and questionnaires in sociolinguistics. In Holmes, J. and Hazen, K., editors, *Research Methods in Sociolinguistics: A Practical Guide*, pages 42–57. Wiley-Blackwell, Chichester.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the World*. Cambridge University Press, Cambridge.

SCOL (2012). *After the Roadmap: Toward Better Programs and Service Delivery*. Standing Committee on Official Languages, House of Commons, Ottawa.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Serratrice, L. (2012). The bilingual child. In Bhatia, T. K. and Ritchie, W. C., editors, *The Handbook of Bilingualism and Multilingualism*, pages 87–108. Wiley-Blackwell, Malden.

Shoemark, P., Kirby, J., and Goldwater, S. (2017a). Topic and audience effects on distinctively Scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68, Copenhagen, Denmark. Association for Computational Linguistics.

Shoemark, P., Kirby, J., and Goldwater, S. (2018). Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 1–6, Brussels, Belgium. Association for Computational Linguistics.

Shoemark, P., Liza, F. F., Nguyen, D., Hale, S., and McGillivray, B. (2019). Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., and Goldwater, S. (2017b). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1239–1248, Valencia, Spain. Association for Computational Linguistics.

Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4):193–229.

Sloan, L. and Morgan, J. (2015). Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS ONE*, 10(11):e0142209.

Sloan, L., Morgan, J., Burnap, P., and Williams, M. (2015). Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE*, 10(3):e0115545.

Squires, L. M. (2007). Whats the use of apostrophes? Gender difference and linguistic variation in instant messaging. *American University TESOL Working Papers*, 4.

Stammers, J. R. and Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, 15(3):630–643.

Statistics Canada (2016). Land and freshwater area, by province and territory. https://www150.statcan.gc.ca/n1/pub/11-402-x/2012000/chap/geo/tbl/tbl06-eng.htm.

Statistics Canada (2017a). Census Profile, 2016 Census. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/index.cfm?Lang=E.

Statistics Canada (2017b). Table 98-400-X2016052. Mother tongue, knowledge of official languages, language spoken most often at home, other language(s) spoken regularly at home, age and sex for the population excluding institutional residents of Canada, provinces and territories, census divisions and census subdivisions, 2016 Census – 100% data. https://www150.statcan.gc.ca/catalogue/98-400-X2016052.

Statistics Canada (2017c). Table 98-400-X2016053. Mother tongue, knowledge of official languages, language spoken most often at home, other language(s) spoken regularly at home, age and sex for the population excluding institutional residents of Canada, provinces and territories, census metropolitan areas and census agglomerations, 2016 Census – 100% data. https://www150.statcan.gc.ca/catalogue/98-400-X2016053.

Statistics Canada (2017d). Table 98-400-X2016057. Mother tongue, knowledge of official languages, age and sex for the population excluding institutional residents of census metropolitan areas, tracted census agglomerations and census tracts, 2016 Census – 100% data. https://www150.statcan.gc.ca/catalogue/98-400-X2016057.

Statistics Canada (2017e). Table 98-400-X2016346. Mother tongue, language spoken most often at home, other language(s) spoken regularly at home, knowledge of official languages, first official language spoken, age and sex for the population excluding institutional residents of Canada, provinces and territories, census metropolitan areas and census agglomerations, 2016 Census – 100% data. https://www150.statcan.gc.ca/catalogue/98-400-X2016346.

Statistics Canada (2019a). Table 15-10-0003-01. Population by mother tongue and geography, 1951 to 2016. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1510000301.

Statistics Canada (2019b). Table 15-10-0004-01. Population by knowledge of official languages and geography, 1951 to 2016. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1510000401.

Statistics Canada (2020). Table 15-10-0008-01. Population by language spoken most often at home and geography, 1971 to 2016. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1510000801.

Statistics Canada (2022). Census Profile, 2021 Census of Population. https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/.

Strathy Language Unit (2011). Strathy corpus of Canadian English. https://www.english-corpora.org/can/.

Stratton, J. M. (2020). Fiction as a source of linguistic data: Evidence from television drama. *Token: A Journal of English Linguistics*, 10:39–58.

Stratton, J. M. (2022). Tapping into German adjective variation: A variationist sociolinguistic approach. *Journal of Germanic Linguistics*, 34(1):63–102.

Strik, H. and Cucchiarini, C. (2014). On automatic phonological transcription of speech corpora. In Durand, J., Gut, U., and Kristoffersen, G., editors, *The Oxford Handbook of Corpus Phonology*, pages 89–109. Oxford University Press, Oxford.

Tagliamonte, S. and D'Arcy, A. (2004). *He's like, She's like*: The quotative system in Canadian youth. *Journal of Sociolinguistics*, 8(4):493–514.

Tagliamonte, S. and Hudson, R. (1999). *Be like* et al. beyond America: The quotative system in British and Canadian youth. *Journal of Sociolinguistics*, 3(2):147–172.

Tagliamonte, S. A. (2002). Comparative sociolinguistics. In Chambers, J. K., Trudgill, P., and Schilling-Estes, N., editors, *The Handbook of Language Variation and Change*, pages 729–763. Blackwell, Malden.

Tagliamonte, S. A. (2006). *Analysing Sociolinguistic Variation*. Cambridge University Press, Cambridge.

Tagliamonte, S. A. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, 45(1):1–32.

Tagliamonte, S. A. and Brooke, J. (2014). A weird (language) tale: Variation and change in the adjectives of strangeness. *American Speech*, 89(1):4–41.

Tagliamonte, S. A. and D'Arcy, A. (2009). Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 85(1):58–108.

Tagliamonte, S. A., D'Arcy, A., and Louro, C. R. (2016). Outliers, impact, and rationalization in linguistic change. *Language*, 92(4):824–849.

Tagliamonte, S. A. and Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1):3–34.

Tagliamonte, S. A. and Denis, D. (2010). The stuff of change: General extenders in Toronto, Canada. *Journal of English Linguistics*, 38(4):335–368.

Tagliamonte, S. A. and Pabst, K. (2020). A *cool* comparison: Adjectives of positive evaluation in Toronto, Canada and York, England. *Journal of English Linguistics*, 48(1):3–30.

Tahmasebi, N. (2013). *Models and Algorithms for Automatic Detection of Language Evolution*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover.

Tahmasebi, N., Borin, L., and Jatowt, A. (2021). Survey of computational approaches to lexical semantic change. In Tahmasebi, N., Borin, L., Jatowt, A., Xu, Y., and Hengchen, S., editors, *Computational Approaches to Semantic Change*, pages 1–91. Language Science Press, Berlin.

Tahmasebi, N. and Risse, T. (2017). Finding individual word sense changes and their delay in appearance. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, pages 741–749.

Takamura, H., Nagata, R., and Kawasaki, Y. (2017). Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Takhteyev, Y., Gruzd, A., and Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, 34(1):73–81.

Tan, L., Zhang, H., Clarke, C., and Smucker, M. (2015). Lexical comparison between Wikipedia and Twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661, Beijing, China. Association for Computational Linguistics.

Tang, X., Qu, W., and Chen, X. (2013). Semantic change computation: A successive approach. *Behavior and social computing*, pages 68–81.

Tao, K., Abel, F., Hauff, C., Houben, G.-J., and Gadiraju, U. (2013). Groundhog day: Near-duplicate detection on Twitter. In *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, pages 1273–1284, Rio de Janeiro, Brazil. ACM Press.

Taylor, J. R. (1992). How many meanings does a word have? *Stellenbosch Papers in Linguistics*, 25:133–168.

Tellez, E. S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., and Ortiz-Bejar, J. (2018). Gender identification through multi-modal tweet analysis using MicroTC and bag of visual words. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*.

Tjong Kim Sang, E. and van den Bosch, A. (2013). Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134.

Traugott, E. C. (2017). Semantic change. *Oxford Research Encyclopedia of Linguistics*.

Traugott, E. C. and Dasher, R. B. (2002). *Regularity in Semantic Change*. Cambridge University Press, Cambridge.

Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism*, 15(2):147–163.

Trudgill, P. (2000). *Sociolinguistics: An Introduction to Language and Society*. Penguin Books, London.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4(3):273–290.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Twitter (2015). Evaluating language identification performance. https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-performance.

Twitter (2019). Twitter announces first quarter 2019 results. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Earnings-Release.pdf.

Uban, A., Ciobanu, A. M., and Dinu, L. P. (2019). Studying laws of semantic divergence across languages using cognate sets. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 161–166, Florence, Italy. Association for Computational Linguistics.

Underwood, G. N. (1968). Semantic confusion: Evidence from the Linguistic Atlas of the Upper Midwest. *Journal of English Linguistics*, 2(1):86–95.

Van Dijk, C., Van Wonderen, E., Koutamanis, E., Kootstra, G. J., Dijkstra, T., and Unsworth, S. (2021). Cross-linguistic influence in simultaneous and early sequential bilingual children: A meta-analysis. *Journal of Child Language*, pages 1–33.

Van Hout, R. and Muysken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, 6(1):39–62.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates.

Veltman, C. (1988). Modelling the language shift process of Hispanic immigrants. *The International Migration Review*, 22(4):545–562.

Vicente, M., Batista, F., and Carvalho, J. P. (2019). Gender detection of Twitter users based on multiple information sources. In Kóczy, L. T., Medina-Moreno, J., and Ramírez-Poussa, E., editors, *Interactions Between Computational Intelligence and Mathematics Part 2*, volume 794, pages 39–54. Springer, Cham.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2016). EN-ES-CS: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association.

Vincent, N. (2019). Analyse du traitement des anglicismes dans des guides de français québécois pour touristes. *Circula*, (9):124–147.

Vlassenroot, E., Chambers, S., Lieber, S., Michel, A., Geeraert, F., Pranger, J., Birkholz, J., and Mechant, P. (2021). Web-archiving and social media: An exploratory analysis. *International Journal of Digital Humanities*, 2:107–128.

Vulić, I., Ponti, E. M., Litschko, R., Glavaš, G., and Korhonen, A. (2020). Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Waite, P. (2021). Confederation. *The Canadian Encyclopedia*.

Wald, B. (1974). Bilingualism. *Annual Review of Anthropology*, 3:301–321.

Walker, J. A. (2015). *Canadian English: A Sociolinguistic Perspective*. Routledge, New York.

Wei, L. (2012). Conceptual and methodological issues in bilingualism and multilingualism research. In Bhatia, T. K. and Ritchie, W. C., editors, *The Handbook of Bilingualism and Multilingualism*, pages 26–51. Wiley-Blackwell, Malden.

Weiner, E. J. and Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, 19(1):29–58.

Weinreich, U. (1953). *Languages in Contact: Findings and Problems*. Mouton, The Hague.

Weinreich, U., Labov, W., and Herzog, M. I. (1968). Empirical foundations for a theory of language change. In Lehmann, W. P. and Malkiel, Y., editors, *Directions for Historical Linguistics*, pages 95–188. University of Texas Press, Austin.

Wenger, E. (2000). Communities of practice and social learning systems. *Organization*, 7(2):225–246.

White, P. (1991). Geographical aspects of minority language situations in Italy. In Williams, C., editor, *Linguistic Minorities, Society and Territory*, pages 44–65. Multilingual Matters, Clevedon.

Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv:1909.10430*.

Williams, J. and Dagli, C. (2017). Twitter language identification of similar languages and dialects without ground truth. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 73–83, Valencia, Spain. Association for Computational Linguistics.

Williams, M. L., Burnap, P., and Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Wojcik, S. and Hughes, A. (2019). *Sizing up Twitter Users*. Pew Research Center, Washington, D.C.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wolfram, W. (1993). Identifying and interpreting variables. In Preston, D. R., editor, *American Dialect Research*, pages 193–221. John Benjamins, Amsterdam.

Wood-Doughty, Z., Mahajan, P., and Dredze, M. (2018). Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 56–61, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Woods, H. B. (1999). *The Ottawa Survey of Canadian English*. Number 4 in Strathy Language Unit Occasional Papers. Queen's University, Kingston, ON.

Wu, T., Wen, S., Xiang, Y., and Zhou, W. (2018). Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security*, 76:265–284.

Xu, Y. and Kemp, C. (2015). A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, pages 2703–2708, Austin, Texas. Cognitive Science Society.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates.

Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 673–681, New York, NY, USA. Association for Computing Machinery.

Yardi, S., Romero, D. M., Schoenebeck, G., and boyd, d. (2010). Detecting spam in a Twitter network. *First Monday*, 15(1).

Yip, V. (2013). Simultaneous language acquisition. In Grosjean, F. and Li, P., editors, *The Psycholinguistics of Bilingualism*, pages 119–144. Wiley-Blackwell, Malden.

Yuan, J. and Liberman, M. (2011). Automatic detection of "g-dropping" in American English using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 490–493, Waikoloa, HI, USA. IEEE.

Yuen, A. L. (1994). Gallicisms: An analysis leading towards a prototype gallicisms checker. Master's thesis, University of Ottawa, Ottawa.

Zalizniak, A. A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak, T., and Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3):633–669.

Zenner, E., Ruette, T., and Devriendt, E. (2017). The borrowability of English swearwords: An exploration of Belgian Dutch and Netherlandic Dutch tweets. In Beers Fägersten, K. and Stapleton, K., editors, *Advances in Swearing Research: New Languages and New Contexts*, pages 107–136. John Benjamins, Amsterdam.

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27, Santiago, Chile. IEEE.

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.

Zubiaga, A. (2018). A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8):974–984.

Zubiaga, A., Pichel, J. R., Aranberri, N., Ezeiza, A., and Fresno, V. (2016). TweetLID: A benchmark for tweet language identification. *Language Resources and Evaluation*, 50:729–766.

Appendices

Appendix A

Test set for semantic shift detection

The table below presents the complete test set for the detection of contact-induced semantic shifts in Quebec English. Words are provided together with their POS tags.

Shifting words		Stable words	
chalet	N	bridesmaid	N
circulation	N	awareness	N
coordinate	N	damp	A
deceive	V	length	N
dossier	N	withdraw	V
local	N	butternut	N
pass	V	hassle	N
resume	V	chestnut	N
trio	N	deed	N
affirmation	N	moot	A
exposition	N	foam	N
manifestation	N	landline	N
population	N	clench	V
souvenir	N	tail	N
terrace	N	blatant	A
animator	N	footstep	N
deputy	N	cram	V
militant	N	breakthrough	N
nomination	N	diehard	A
portable	A	earring	N
remark	V	corn	N
definitively	R	acknowledgement	N
deception	N	darling	N
availability	N	arise	V
prudent	A	handbook	N
hesitate	V	bedding	N
grave	A	inmate	N
reparation	N	bead	N
exchange	V	feather	N
proposition	N	errand	N
occasion	N	coastline	N
ambiance	N	congressman	N
formidable	A	wart	N
entourage	N	upright	R
permit	N	begin	V
formation	N	ought	V
merit	V	balk	V
exploration	N	helm	N
boutique	N	campfire	N
laureate	N	hunch	N

Appendix B

Top 50 semantic shift candidates

The table below presents the top 50 semantic shift candidates output by the best performing model from the evaluation in Chapter 11.

1	pour	26	exposition
2	plateau	27	s
3	nt	28	corona
4	den	29	encore
5	rapport	30	trustee
6	sous	31	coupe
7	ont	32	2
8	en	33	dispatch
9	mb	34	dire
10	aux	35	appraisal
11	saison	36	vie
12	tout	37	premier
13	svp	38	overdose
14	vers	39	petite
15	pour	40	fort
16	bec	41	mtg
17	de	42	plus
18	pa	43	vu
19	trough	44	nest
20	gorge	45	staging
21	detached	46	basin
22	le	47	br
23	parfait	48	ce
24	still	49	lever
25	#venom	50	bologna

TABLE B.1: Top 50 semantic shift candidates

Appendix C

Sample clusters of tweets

This appendix presents additional examples of the cluster-based analysis.

	portraits in honour of Janet Werner's upcoming	**exposition**	at the museum . Starting November 10th , every
(1)	Collection " Come to admire Laura Granata's	**exposition**	at #CLDV
	Such a beautiful	**exposition**	!!! #mbam #art #montrealmuseum
	space will be turned into a citizens' area with	**exposition**	space and multipurpose room . #CJAD #polmtl
(2)	is WINDSOR STATION - It's now part of the	**exposition**	events space in Montreal . Its located next to the
	media students are showcasing their work at an	**exposition**	hall in Trois-Rivière .
		exposition	d'aquarelles , exhibition of my watercolor works
(3)		**Exposition**	du World Press Photo 2016 #photo #feedly
		Exposition	en cours - Galerie d'art Stewart Hall

Sample clusters for *exposition*, which is typically used in English with senses including 'opening section in fiction' and 'a comprehensive explanation'. Cluster 1 illustrates the contact-related sense 'art exhibition' (cf. Fr. *exposition*). Cluster 2 shows the effect of this usage on a specific collocational pattern (*exposition space / hall*), indicating further diffusion of the semantic shift. Cluster 3 contains occurrences of the French homograph *exposition* attested in codeswitched tweets.

	#ducks ?? With big expectations come biggest	**deceptions**	... #nhlhockey #Game7Curse #game7
(1)	Great expectations , few	**deceptions**	and stunning debuts make a unique
	fantastic example of endurance and overcoming	**deceptions**	! Thank you so much !
	The Coffee	**Deception**	: 13 Little Known Facts About Coffee
(2)	The grand	**deception**	: Looking for love , validation & peace outside of
	Kavanaugh's testimony : The immaculate	**deception**	.
	The new song	**Deception**	Bay , from Milk & Bone's second album , is out !
(3)	From their second album , out now :	**Deception**	Bay . I wonder if they are familiar with the doggy
		Deception	Bay is a masterpiece

Sample clusters for *deception*, which in English refers to the action of deceiving (misleading) someone. Cluster 1 reflects the contact-related sense 'disappointment' (cf. Fr. *déception*). Cluster 2 is a case where no determination was made by the annotators as the contexts were deemed insufficiently specific to disambiguate the possible senses. Cluster 3 exemplifies the use of the target word as a proper noun, here referring to the song "Deception Bay" by the Montreal band Milk & Bone.

	Pouring coffee beans in the water tank ... I	**definitively**	need coffee !!!
(1)	again some developers after all these years ! I	**definitively**	want to come back to Montréal next year for the
	thank you very much ♡ I'm touched , I would	**definitively**	love to work with you one day !
	This is	**definitively**	a job that should've been replaced by a small script
(2)	65% of everything in school is	**definitively**	a waste of time . Useless subjects and more > : l
	party that would bring us decades back . A party	**definitively**	far from the interests of Quebeckers and
	In 2018 ? Most	**definitively**	!
(3)		**Definitively**	!
		Definitively	Yay !!!

Sample clusters for *definitively*, whose conventional meaning in English is 'conclusively, indisputably'. Clusters 1 and 2 indicate different contexts in which it is used with the more general contact-related sense 'definitely, certainly' (cf. Quebec French *définitivement* 'definitely'). Cluster 3 shows a further generalization of that use, including as an emphatic interjection ('yes!').

.

Appendix D

Sociolinguistic protocol

This appendix presents the materials used for the sociolinguistic interviews.

PAC Protocol - Reading task 1

PAC Wordlist 1

1. start
2. pause
3. err
4. peril
5. poor
6. steer
7. scarce
8. sorry
9. fail
10. leaven
11. bury
12. fall
13. brewed
14. Mary
15. side
16. four
17. bode
18. bard
19. plant
20. room
21. foul
22. stairs
23. meat
24. dole
25. berth
26. pore
27. fair
28. bed
29. short
30. look
31. calm
32. fierce
33. gourd
34. bored
35. paws
36. here
37. for
38. foil
39. next
40. bid
41. foal
42. more
43. feel
44. sue
45. caught
46. row
47. weight
48. barred
49. heaven
50. pant
51. shepherd
52. story
53. pit
54. sport
55. pearl
56. berry
57. board
58. pat
59. paw
60. file
61. word
62. agreed
63. cook
64. purr
65. greed
66. brood
67. say
68. bad
69. weary
70. pet
71. moor
72. full
73. merry
74. knot

75. ants
76. knows
77. rose
78. far
79. put
80. fill
81. pour
82. beard
83. stir
84. spirit
85. afterwards
86. dance
87. earth
88. horse
89. fool
90. hurry
91. fir
92. leopard
93. soot
94. sighed
95. fore
96. vexed
97. pert
98. sigh
99. meet
100. jury
101. there
102. putt
103. furl
104. rows
105. pot
106. wait
107. bowed
108. farther
109. fell
110. hoarse
111. master
112. aunts
113. fur
114. pose
115. sea
116. bird
117. war
118. mate
119. bard
120. bead
121. doll
122. marry
123. nose
124. naught
125. bared
126. cot
127. father
128. choice
129. lava

PAC Wordlist 2

1.	wet	33.	heart
2.	bedding	34.	rack
3.	seal	35.	betting
4.	chutney	36.	thick
5.	little	37.	tuck
6.	earthy	38.	fan
7.	kidney	39.	meddle
8.	sinner	40.	anyhow
9.	supper	41.	loch
10.	grace	42.	badge
11.	bigger	43.	carter
12.	rung	44.	leisure
13.	bell	45.	middle
14.	sack	46.	batch
15.	lock	47.	written
16.	lab	48.	metal
17.	belly	49.	garter
18.	rum	50.	which
19.	decree	51.	graze
20.	run	52.	bishop
21.	lap	53.	fad
22.	van	54.	behave
23.	singer	55.	stronger
24.	bicker	56.	pat
25.	rubber	57.	simmer
26.	zeal	58.	sag
27.	degree	59.	duck
28.	lack	60.	berry
29.	bet	61.	bat
30.	witch	62.	this
31.	yet	63.	ridden
32.	worthy	64.	fat

PAC Protocol - Reading task 2

A Christmas interview © PAC 2021

If television personalities are anything like the rest of us, all they really want to do in Christmas week is snap at their families, criticize their friends and make their neighbours' children cry by glaring at them over the garden fence. Yet society expects them to be as jovial and beaming as they are for the other fifty-one weeks of the year. If anything, more so.

Take the Reverend Peter Smith, the TV vicar who sends out press releases in which he describes himself as “the man who has captured the spirit of the age”. Before our 9 a.m. meeting at his media office on Crawshaw Avenue, South London, he faced, he says, a real dilemma. Should he make an effort to behave like a Christian, throw his door open, offer me a cup of tea or should he just play it cool, study his fingernails in a manner that shows bored indifference and get rid of me as quickly as possible? In the end, he did neither.

“As a matter of fact, John”, he says in a loud Estuary English twang, “St Francis said, ‘At all times preach the gospel and speak whenever you have to’. But hey, he didn't mean ‘Be on your best behaviour and be happy all the time’. I could have been extra-polite to you, but the real me would have come out as I was talking. You cannot disguise what you are.”

“And what are you then, Peter?”

“Well, I'm a Christian, John. I've been one since I was 14. And I know for sure that Christianity will be judged more on what you do rather than what you have to say about it.” In many ways, Peter Smith looks exactly how you'd expect a high-profile television personality to look: tall, handsome, clean-cut and evenly sun-tanned. He doesn't wear a dog-collar. In fact, when doing his various religious programmes on Sunday mornings, he has been known to wear a black leather jacket instead, in casual mode. Today, the look is more business-like: metal-rimmed glasses, a grey suit, a blue open-neck shirt, and fashionable black shoes with large buckles. Smith is 44 but he looks a mere 24.

During the whole interview, Peter Smith stressed the need to be on the side of the poor and the needy. He also talked about his forthcoming trip to China and the masses waiting for his message there. I ventured a few questions relating to the charity trust he founded some ten years ago and which, it is generally agreed, employs eight hundred staff and runs schools, hospitals and hostels around the world. I did mention criticisms in the press of the way charitable organizations are run these days but tried not to sound hostile. He just sighed in answer to my remarks and said: “I'm only human, John. God knows I do my best and often fail. But it's no skin off my nose if our enemies sneer at some of the good work we do. Truth will out.”

PAC Protocol - Semi-structured interview

Information sheet

Date of recording _____
PAC Identifier _____
Age at date of recording _____
Place of birth _____
Current place of residence (village, town, etc.) _____
_____

Previous places of residence

place	number of years	age

Occupation _____
Other previous occupations _____
_____
_____

Education (specify until what age and what type of education)

place	type of education	age

Languages spoken

		language			
level of proficiency	basic				
	intermediate				
	fluent				
frequency of use	rarely				
	monthly				
	daily				

Informant's father, year of birth _____
Place of origin _____
Occupation _____
Education _____
Languages or local dialects spoken _____

Informant's mother, year of birth _____
Place of origin _____
Occupation _____
Education _____
Languages or local dialects spoken _____

Informant's spouse/partner, year of birth _____
Place of origin _____
Occupation _____
Education _____
Languages or local dialects spoken _____

Number of children, age and education _____
_____
_____
_____

People who played an important role during the informant's acquisition of the English language (grandparents, childminder, etc.) _____
_____
_____

Ethnic group _____

Type of accommodation of the informant (house, flat, in a residential area, housing estate, block of flats, etc.) _____

Integration into the area, relationships within the neighbourhood _____
_____
_____

Cultural and leisure activities, travels _____
_____
_____

Additional information _____
_____
_____

Information sheet on the recording

Interviewer's name (formal conversation) _____
Interviewer's name (informal conversation) _____
Length of recording _____
Place and setting of the recording _____
Location _____
Speakers _____
Ties between the interviewer and the informants
Professional _____
Friendly _____
Family _____
Other _____

Order of the situations in the recording (e.g.: formal, wordlists, text, informal) _____
_____

Main topics discussed _____
_____
_____

Quality of the recording _____

Remarks on the recording (interventions from other people, long telephone interruptions etc.)

_____
_____
_____

PAC Protocol - Thematic questionnaire

QUESTIONS RELATED TO THE CITY

1. Do you feel that you're a true Montrealer?
   [If an undeveloped yes/no response is given, continue with: What do you think being a true Montrealer means? If the response is still incomplete, continue with: When people talk about "true Montrealers", what does it mean for you?]
2. What is it like to live in your part of the city? What are the advantages and disadvantages?
3. If you had to live in another part of Montreal, or another part of the surrounding area, where would you choose to live?
4. Is there another city you would prefer to live in in Quebec or in Canada?

****

QUESTIONS RELATED TO WORK

For those in active employment:

1. Could you tell us about the things you regularly do in your work?
2. Could you explain to us what you like or what you don't like about your work?
3. If you were completely free to change your hours of work, when would you choose to work, and why?
4. Do you think you have a good work-life balance? Could you give us your reasons?
5. Would you like to change your job/the work you do in the next three years, and if so, why?

For those who are retired:

1. Could you tell us about the last job you had?
2. Could you tell us what you liked, or didn't like, about the job?
3. If you had to work again and were completely free to change your hours of work, when would you choose to work, and why?
4. When you worked, do you think you had a good work-life balance? Could you give us your reasons?
5. Did you change professions/or the type of work you did during your working life, and why?

For those who are unemployed:

1. Could you explain to us what your last job was?
2. What did you like, or what didn't you like about the job?
3. If you were completely free to choose your hours of work, when would you want to work, and why?
4. What do you see as an ideal work-life balance? Could you give us your reasons?
5. What sort of job would you like to have, and could you give reasons for this?

For teenagers and young people (who have possibly never worked, or only done short-term or part-time work):

1. Have you ever had a job, even if it was only part-time, and what did it consist of?
2. Could you explain to us what you liked or didn't like about the job?
3. If you had to work and were completely free to choose your working hours, when would you choose to work?
4. What do you see as an ideal work-life balance? Could you give us your reasons?
5. What sort of job would you like to find, and could you give reasons for this?

****

QUESTIONS RELATED TO LANGUAGE AND IDENTITY

1. Do you consider yourself a Canadian, a Quebecer, a Montrealer, or a West Islander/other? If so, in which order? Why?
2. Can you make the distinction between yourself and American speakers? What about other Canadian people? Do you speak differently from Ontarians for instance?
3. Would you say you speak a type of English that is typical of Montreal, or what people sometimes call “Montreal English”?
4. If you think that “Montreal English” exists, what would you say its main characteristics are?
5. What are the main features of Canadian and/or Quebec English for you?
6. Are there any differences in the way you speak when you are at work, when you are with friends, and when you are with your family?
7. Do you think there are any movies, TV shows, podcasts etc. that accurately reflect the way people speak English in Montreal? If so, which ones? If not, why do you think that is the case?
8. Would you say that the Montreal accent compares favourably to other accents of Canadian English?
9. What is it like living in an officially Francophone province?
10. Do you speak French fluently?
11. Do you think French influences the way you speak English? In what way?
12. What would you say it means to be bilingual for someone living in Montreal? Would you describe yourself as bilingual?
13. If you were walking around Montreal and needed to ask someone you don’t know for directions, which language would you use? Are there any situations where you would make a different decision, such as being in a specific neighbourhood or overhearing the person use [the other language]?

****

QUESTIONS RELATED TO TWITTER AND SOCIAL NETWORKS

1. Would you say that you are an active user of social media such as Twitter? [*If the answer is affirmative but for another social network, adapt the remaining questions accordingly.*] How important is Twitter to you?
2. How would you describe the way you use Twitter? What is it that draws you to it? [*Possible additional prompts: Do you actively interact with others, for example by getting involved in discussions, or do you tend to read other peoples’ tweets to pass the time? Do you tend to follow public figures, or people who you know in real life, such as your friends and colleagues? Do you use Twitter with a variety of these purposes?*]
3. In which language do you tweet most often? Do you tweet in other languages as well? If so, under what circumstances? / If not, why not?
4. Would you say that your choice of languages on Twitter is similar to the way you use them in real life?
5. Are there any characteristics of language use that you think are typical of Twitter?
6. Do you think you would be able to determine if someone is Canadian or American based on their tweets? How about determining if someone is from Montreal? If so, what would give them away?

PAC Protocol - Semantic perception test

Instructions

I will show you a series of tweets.

First of all, I would like you to read each tweet out loud, and then rate it based on how natural it sounds to you. You will use a scale from 1 to 6, where 1 means ‘very unnatural, awkward, you would never say something like that’, and 6 means ‘completely natural, just like something you might say’. Once you’ve read the tweet and understood it, try not to think too much about the rating you will choose, just go with your initial instinct.

You will also see that each tweet contains one word in bold. Once you have rated the tweet, I would like you to think of a word that you could replace it with, without changing the meaning of the tweet.

Also, if there is anything about the tweet that you would like to point out – for example, if there are words you feel are out of place, if you have trouble interpreting the tweet, if you think that the author of the tweet is very young, or old, or may not be an English speaker, anything at all that you find interesting or worth pointing out – please let me know as we go along. That would be really helpful.

We will begin with an example, just so you can see what this is going to be like, and so you can ask any questions.

#	tweet	awkward			…		natural
0	No greater disappointment than when your **dep** hasn't gotten their cheese curd delivery this morning.	1	2	3	4	5	6
1	@ Let me know your phone number & **availabilities** and I'll call you.	1	2	3	4	5	6
2	So heart warming to read your bio. You're a fantastic example of endurance and overcoming **deceptions**! Thank you so much! 🙌🏻😊	1	2	3	4	5	6
3	I believe this. I'm not there yet but my **entourage** is all approaching c in their early 30s and I can see/feel how they have a much stronger sense of purpose, direction and self. Gives me something to look forward to and I feel blessed to have people like this around me.	1	2	3	4	5	6
4	Had the best time at @ ! Loved **exchanging** with all the students and speakers! # #	1	2	3	4	5	6
5	@ **Formidable** piece of journalism! Thank you sir for the effort put to assemble together all the elements involved in this affair. https://	1	2	3	4	5	6
6	@ hey man, I am looking to change my phone and I'm **hesitating** between the Iphone 8+ and the XR, is the XR worth it or no? Thanks a lot!!	1	2	3	4	5	6
7	Looking for a developer? Our DemoDay is coming! ✨ It's the **occasion** to meet our team, invest in great talent or find a co-founde RSVP 👉 https:// # # # 🚀	1	2	3	4	5	6
8	We're glad you like it! It is not part of our plans for the moment. Than you for the **proposition**.	1	2	3	4	5	6
9	Cyclists are slowing down **circulation** and can also be very dangerous by sliding in between cars to get ahead when light is red.	1	2	3	4	5	6
10	@ I need your new work **coordinates**. I have a referral for you 😊	1	2	3	4	5	6
11	Wow - Amsterdam never **deceives**! No more energy in the body but the heart full of love! I'd like to express my... https://	1	2	3	4	5	6
12	First obligation of a municipal administration - consult on major **dossiers** that affect the population. I've always found that is the mos effective way to build consensus and make decisions that have publi support. #	1	2	3	4	5	6
13	@ The only prerequisite was that we have to be bilingua the company I work at gave us a quick **formation** and that was it!	1	2	3	4	5	6
14	We are currently in DGP in Dresden! Make sure to **pass** by our booth C32 in Hall 2 to get more information about our tremoflo C-100 and test your lungs in less than 2 minutes!	1	2	3	4	5	6
15	Nothing will change with drunk or stoned drivers, or drivers texting, until they lose their **permit** for a minimum of 6 months for the first offense, and for good if they are stupid to try that stunt a 2nd time. Education? Really? At taxpayers expense. No! Hit them hard.	1	2	3	4	5	6

#	tweet	awkward			…		natural
16	@ my bae boy and me. This picture just **resumes** how wild we were both of us from Friday to Sunday.	1	2	3	4	5	6
17	@ I rarely order fries anymore, let alone a **trio**. If I did, however, I usually go fries first. (But only a couple.)	1	2	3	4	5	6
18	Spending so much time at **locals** only to be able to play four or five games, lose two sets and then just watch people play kind of sucks.	1	2	3	4	5	6
19	The first H&M Home **boutique** will open shortly in Carrefour Laval. https://	1	2	3	4	5	6
20	nice!!! i'll **definitively** check them out, thank you!	1	2	3	4	5	6
21	Automated analysis of large chunks of data is great until the work turns into an **exploration** of all possible indexing errors and combinations there of.	1	2	3	4	5	6
22	I doesn't look very **grave**, for now, but I know tests and antibiotics ca be expensive.	1	2	3	4	5	6
23	Tragic. Where are our women scientists and innovators? I know they exist. But lets celebrate all the **laureates**, including our 17 women!	1	2	3	4	5	6
24	Sincere congratulations . Through your dedication and wonderful talent you **merit** every success	1	2	3	4	5	6
25	I'm all for slowing down and being **prudent** in difficult driving conditions, but if I can jog faster than you drive, then maybe you should just stay in #	1	2	3	4	5	6
26	Eve of # : **reparation** of gear, guitars and bonfire # # #	1	2	3	4	5	6
27	is bringing you some of its best basketball matchups. W invite you all to come watch and enjoy the **ambiance** with us! #WOR	1	2	3	4	5	6
28	Ms. , Spiritual **Animator**, has been busy with the annual #PoppyDrive honouring our fallen soldiers. Support veterans! #LestWeForget	1	2	3	4	5	6
29	This definitely resonates with me. I get **chalet** lifestyle all week long just over an hour away from the city. Especially great in the summer where 5pm means cooling off at the lake. #qualityoflife	1	2	3	4	5	6
30	The Subject Effect a @macmtl workshop explores the limits of painting portraits in honour of Janet Werner's upcoming **exposition** a the museum. Starting November 10th, every Sunday until January 5t 2020	1	2	3	4	5	6
31	Almost 500 000 ppl showed up at the Montreal walk for the environment (manifestation). Not only is this walk the biggest for environment in Quebec's history. This walk is the biggest **manifestation** for this week. And 52 more towns in the province have manifestations. 🌎🌱🔥	1	2	3	4	5	6
32	is an agressive conservative **militant**. He attacks anything liberal; good and bad. The weakness in his narrative though is that he never says anything positive or substantial about . Because there is none.	1	2	3	4	5	6

#	tweet	awkward			…		natural
33	Impressed by this new award winning #huawei Matebook small **portable**, feature packed with a really nice screen. Worth a look if yo are in the market for a new one. @ Las Vegas Convention … https://	1	2	3	4	5	6
34	I really wish my high school would've let us do this, that's seriously dope and a wonderful **souvenir**!! 😱 https://	1	2	3	4	5	6
35	The weather is still perfect for a lunch on the **terrace** #i❤ny à Greenwich Village https://	1	2	3	4	5	6
36	It's the Canadian version of "POC are dangerous" narrative? Your **affirmations** don't align with statistics and we don't like to be lied to by politicians	1	2	3	4	5	6
37	monarchy doesn't affect our lives today. sadly, we as quebecers still have to listen to the independence movement even when 80% of our elected **deputies** are federalist. the independence movement still harms growth and opportunity for quebecers today.	1	2	3	4	5	6
38	What do you think about the **nomination** of Shea Weber as the Captain of the @CanadiensMTL	1	2	3	4	5	6
39	@ says that she doesn’t care what religious symbols police wear. She wants the **population** to work and integrate becaus that’s the best way to contribute to our society. #	1	2	3	4	5	6
40	THIS IS SO COOL (I just **remarked** it)	1	2	3	4	5	6

PAC Protocol - File naming conventions
(last update: September 2017)

The following guidelines ensure the anonymous treatment of the speakers, and provide a uniform method for naming files containing all data gathered and developed within the PAC-LVTI framework across corpora. All names consist of eight positions to designate the corpus and the speaker the document belongs to.

Position 1: country (3 characters)
aus = Australia
can = Canada
eng = England
ind = India
ire = Ireland
nor = Northern Ireland
sco = Scotland
usa = United States
wal = Wales
nzl = New Zealand
sin = Singapore

Position 2: region (2 characters)
wm = West Midlands
gm = Greater Manchester
ay = Ayrshire
la = Lancashire
ns = New South Wales
ot = Otago
ca = Canterbury
qu = Québec
on = Ontario
de = Delhi
ga = Galway
du = Dublin (comté)
do = Donegal
ma = Massachussetts
mi = Michigan
mo = Missouri
ca = California
ct = Connecticut
etc.

Position 3: town/city (2 characters)
bi = Birmingham
sy = Sydney
wc = White Cliffs
de = Deniliquin
ot = Ottawa
mo = Montreal
du = Dunedin
ch = Christchurch
nd = New Delhi

ga = Galway
li = Limerick
co = Cork
ra = Rangiora
si = Singapore
bo = Boston
sl = Saint Louis
sb = Santa Barbara
etc.
IF NOT RELEVANT = 00

Position 4: number of survey (1 character)
a = 1st survey
b = 2nd survey
c = 3rd survey
etc.

Positions 5: speaker's initials (two characters) + a number (starting from 1) to distinguish speakers with the same initials (in alphabetical order of first names)
es1= Elizabeth Smith
js1 = Jason Smith
js2 = Jennifer Smith
js3 = John Smith
etc.

Position 6: task of the protocol corresponding to file (1 character)
i = informal conversation
f = formal interview
t = text
v = vowels (wordlist 1)
c = consonants (wordlist 2)
r = revised version of the PAC text
x = extra task 1
y = extra task 2
z = extra task 3

AS A REMINDER: DO NOT add a "g" or a "w" at the end of your file name to indicate the format (textgrid vs. wav) as it creates incompatibilities with some software such as DOLMEN or SPPAS!

Appendix E

Auditory analysis for a subset of speakers

This appendix presents a detailed analysis of the segmental features exhibited by a subset of interviewed speakers; they correspond to the central cluster of participants who seem to be driving the use of contact-induced semantic shifts, at least in this sample (see Chapter 14).

The presented analysis includes the full range of stimuli from the first part of the reading task. It comprises two word lists, which respectively address vocalic and consonantal features. In the table below, each stimulus is presented alongside phonemic features which would be expected in the Canadian English context, as well as realizations produced by individual speakers. This is based on an auditory analysis conducted by two annotators, with uncertainties resolved through reconciliation. Individual realizations are further associated with potential sources of influence using the following symbols:

- * = realizations that can be classified as reading errors, i.e. which represent phonological divergence with respect to reference descriptions of target lexical items;
- † = hesitation;
- § = vowel realization potentially influenced by French;
- §§ = consonant realization potentially influenced by French;
- ¶ = likely lexical conditioning given the patterns observed for other realizations by the same speaker.

If multiple realizations by the same speaker are noted for a single stimulus, this indicates corrections by the speaker.

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
Wordlist 1									
start	/stɑːɹt/ - /stɑɹt/	[ɑː]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑː]	[ɑː]
pause	/pɑːz/ - /pɑz/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
err	/ɝː/	[ɝː]	[iː] *	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[iː] *	[ɝː]
peril	/ˈpɛɹəɬ/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
poor	/pʊɹ/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
steer	/stɪɹ/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
scarce	/skɛɹs/	[ɑː] *	[ɛ]	[ɑː] *	[ɛ]	[ɛ]	[ɛ]	[ɑː] *	[ɑː] *
sorry	/ˈsɔːɹi/ - /ˈsɑːɹi/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
fail	/feɬ/	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]
leaven	/ɬɛvən/	[iː] *	[iː] *	[iː] *	[iː] *	[ɛ]	[iː] * [ɛ]	[iː] *	[iː] *
bury	/ˈbɛɹi/	[ɜ] *	[ju] *	[ɜ] *	[ɛ]	[ʌ] *	[ɛ]	[u] *	[ʌ] *
fall	/fɑːɬ/ - /fɑɬ/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
brewed	/bɹuːd/	[əuː]	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]
Mary	/ˈmɛɹi/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
side	/saɪd/	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]
four	/fɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
bode	/boʊd/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
bard	/bɑːɹd/ - /bɑɹd/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
plant	/płænt/ to [ɛ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
room	/ɹuːm/ - /ɹʊm/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[uː]	[uː]
foul	/faʊɬ/	[aʊ]	[aʊ]	[aʊ]	[aʊ]	[aʊ]	[aʊ]	[aʊ]	[aʊ]
stairs	/stɛɹz/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
meat	/miːt/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
dole	/doʊɬ/	[oʊ]	[ɔː]	[oʊ]	[oʊ]	[ɔː]	[ɔː]	[oʊ]	[oʊ]
berth	/bɝːθ/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]
pore	/pɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
fair	/fɛɹ/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
bed	/bɛd/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
short	/ʃɔːɹt/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
look	/ɬʊk/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
calm	/kɑːm/ - /kɑm/	[ɑ]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
fierce	/fɪɹs/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
gourd	/gʊɹd/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
bored	/bɔːɹd/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
paws	/pɑːz/ - /pɑz/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
here	/hɪɹ/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
for	/fɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
foil	/fɔɪɬ/	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]
next	/nɛkst/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
bid	/bɪd/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
foal	/foʊɬ/	[oʊ]	[aʊ] *	[ɔː] * †	[ɔː] *	[ɔː] * †	[ɔː] *	[oʊ]	[oʊ]
more	/mɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
feel	/fiːɬ/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
sue	/suː/	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]
caught	/kɑːt/ - /kɑt/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
row	/ɹoʊ/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
weight	/weɪt/	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]
barred	/bɑːɹd/ - /bɑɹd/	[ɛ]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɛ]	[ɑː]
heaven	/ˈhɛvən/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
pant	/pænt/ to [ɛ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]
shepherd	/ˈʃɛpɚd/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ˈʃɛfɚd] *
story	/ˈstɔːɹi/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
pit	/pɪt/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
sport	/spɔːɹt/	[ɔː] §§	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
pearl	/pɝːɬ/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]
b*y	/ˈbɛɹi/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
board	/bɔːɹd/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
pat	/pæt/	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[a] §	[æ]
paw	/pɑː/ - /pɑ/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[a] §	[ɑː]
file	/faɪɬ/	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]
word	/wɝːd/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
agreed	/əˈgɹiːd/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
cook	/kʊk/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
purr	/pɝː/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]
greed	/gɹiːd/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
brood	/bɹuːd/	[uː]	[oʊ] *	[uː]	[uː]	[uː]	[uː]	[uː]	[uː]
say	/seɪ/	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]
bad	/bæd/	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[a] §	[æ]
weary	/ˈwɪɹi/	[ɛ] *	[ɪ]	[ɛ] *	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
pet	/pɛt/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
moor	/mʊɹ/	[ʊ]	[ɔ]	[ɔ] [ʊ]	[ɔ]	[ɔ]	[ʊ]	[ɔ]	[ɔ]
full	/fʊɬ/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
m*y	/ˈmɛɹi/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
knot	/nɑt/ - /nɑːt/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
ants	/ænts/ to [ɛ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]
knows	/noʊz/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
rose	/ɹoʊz/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
far	/fɑɹ/ - /fɑːɹ/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
put	/pʊt/	[ʌ] *	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
fill	/fɪɬ/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
pour	/pɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
beard	/bɪɹd/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
stir	/stɝ:/	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[i] §	[ɝ:]
spirit	/ˈspɪɹɪt/	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]	[ɪ]
afterwards	/ˈæftə-wə-dz/	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]
dance	/dæns/ to [ɛ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]
earth	/ɝ:θ/	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]
horse	/hɔ:ɹs/	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]
fool	/fu:ɬ/	[u:]	[u:]	[u:]	[u:]	[u:]	[u:]	[u:]	[u:]
hurry	/ˈhɜɹi/ - /ˈhɜ:ɹi/	[ɜ]	[ɜ]	[ɜ]	[ɜ]	[ɜ]	[ɜ]	[ɜ]	[ɜ]
fir	/fɝ:/	[ɝ:]	[ɝ:]	[ɝ:] [ɪ]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]
leopard	/ˈɬɛpə-d/	[ɛə] *	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
soot	/sʊt/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
sighed	/saɪd/	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]
fore	/fɔ:ɹ/	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]	[ɔ:]
vexed	/vɛkst/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
pert	/pɝ:t/	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]
sigh	/saɪ/	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]	[aɪ]
meet	/mi:t/	[i:]	[i:]	[i:]	[i:]	[i:]	[i:]	[i:]	[i:]
jury	/ˈdʒʊɹi/	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]	[ʊ]
there	/ðɛɹ/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
putt	/pʌt/	[ʌ]	[ʊ] *	[ʊ] *	[ʌ]	[ʌ]	[ʌ]	[ʊ] *	[ʌ]
furl	/fɝ:ɬ/	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]	[ɝ:]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
rows	/ɹoʊz/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
pot	/pɑt/ - /pɑːt/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
wait	/weɪt/	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]
bowed	/baʊd/	[aʊ]	[oʊ] *	[aʊ]	[ɔː] *	[aʊ]	[aʊ]	[aʊ]	[aʊ]
farther	/ˈfɑɹðɚ/ - /ˈfɑːɹðɚ/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɝː] *	[ɑ]	[ɑ]	[ɑ]
fell	/fɛɬ/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[ɛ]
hoarse	/hɔːɹs/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
master	/ˈmæstɚ/	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]	[æ]
aunts	/ɑnts/ - /ænts/	[æ]	[æ]	[ɑ] ¶	[ɑ] ¶	[æ]	[ɑ] ¶	[æ]	[æ]
fur	/fɝː/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]
pose	/poʊz/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[əʊ]	[oʊ]
sea	/siː/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
bird	/bɝːd/	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]	[ɝː]
war	/wɔːɹ/	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]	[ɔː]
mate	/meɪt/	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]	[eɪ]
bard	/bɑɹd/ - /bɑːɹd/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
bead	/biːd/	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]	[iː]
doll	/dɑɬ/ - /dɑːɬ/	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]	[ɑː]
marry	/ˈmæɹi/	[ɛ]	[ɛ]	[ɛ]	[ɛ]	[æ]	[ɛ]	[ɛ]	[ɛ]
nose	/noʊz/	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]	[oʊ]
naught	/nɑt/ - /nɑːt/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
bared	/ˈbɛɹd/	[ɛ]	[ɑː] *	[ɑː] *	[ɛ]	[ɛ]	[ɑː] [ɛ]	[ɛ]	[ɛ]
cot	/kɑt/ - /kɑːt/	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]	[ɑ]
father	/ˈfɑðɚ/ - /ˈfɑːðɚ/	[ɑ]	[ɑː]	[ɑ]	[ɑː]	[ɑː]	[ɑ]	[ɑː]	[æ]
choice	/tʃɔɪs/	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]	[ɔɪ]
lava	/ˈɬævə/ - /ˈɬɑːvə/ - /ˈɬavə/	[æ]	[ɑː]	[ɑː]	[æ]	[æ]	[ɑː]	[ɑː]	[ɑː]
Wordlist 2									
wet	/wɛt/	[wɛt]	[wɛt]	[wɛt]	[wɛt]	[wɛt]	[wɛt]	[wɛt]	[wɛt]
bedding	/ˈbɛdɪŋ/ - /ˈbɛɾɪŋ/	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛdɪŋ]	[ˈbɛɾɪŋ]
seal	/siːɬ/	[siːɬ]	[siːɬ]	[siːɬ]	[siːɬ]	[siːɬ]	[siːɬ]	[siːɬ]	[siːɬ]
chutney	/ˈʧʌtni/	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]	[ˈʧʌtni]
little	/ˈɬɪɾɬ/	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]	[ˈɬɪɾɬ]
earthy	/ˈɝːθi/	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]	[ˈɝːθi]
kidney	/ˈkɪdni/	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]	[ˈkɪdni]
sinner	/ˈsɪnɚ/	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]	[ˈsɪnɚ]
supper	/ˈsʌpɚ/	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]	[ˈsʌpɚ]
grace	/ˈgɹeɪs/	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]	[ˈgɹeɪs]
bigger	/ˈbɪgɚ/	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]	[ˈbɪgɚ]
rung	/ɹʌŋ/	[ɹʌŋg]	[ɹʌŋ]	[ɹʌŋ]	[ɹʌŋ]	[ɹʌŋ]	[ɹʌŋ]	[ɹʌŋ]	[ɹʌŋ]
bell	/bɛɬ/	[bɛɬ]	[bɛɬ]	[bɛɬ]	[bɛɬ]	[bɛɬ]	[bɛɬ]	[bɛɬ]	[bɛɬ]
sack	/sæk/	[sæk]	[sæk]	[sæk]	[sæk]	[sæk]	[sæk]	[sæk] §§	[sæk]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
lock	/ɬɑk/ - /ɬɑːk/	[ɬɑk]	[ɬɑk]	[ɬɑk]	[ɬɑk]	[ɬɑk]	[ɬɑk]	[ɬɑk]	[ɬɑk]
lab	/ɬæb/	[ɬæb]	[ɬæb]	[ɬæb]	[ɬæb]	[ɬæb]	[ɬæb]	[ɬæb]	[ɬæb]
belly	/ˈbɛɬi/	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]	[ˈbɛɬi]
rum	/ɹʌm/	[ɹʌm]	[ɹʌm]	[ɹʌm]	[ɹʌm]	[ɹʌm]	[ɹʌm]	[ɹʌm]	[ɹʌm]
decree	/diˈkɹiː/	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]	[diˈkɹiː]
run	/ɹʌn/	[ɹʌn]	[ɹʌn]	[ɹʌn]	[ɹʌn]	[ɹʌn]	[ɹʌn]	[ɹʌn]	[ɹʌn]
lap	/ɬæp/	[ɬæp]	[ɬæp]	[ɬæp]	[ɬæp]	[ɬæp]	[ɬæp]	[ɬæp]	[ɬæp]
van	/væn/ to [ɛ]	[væn]	[væn]	[væn]	[væn]	[væn]	[væn]	[væn]	[væn]
singer	/ˈsɪŋɚ/	[ˈsɪŋgɚ]	[ˈsɪŋɚ]	[ˈsɪŋɚ]	[ˈsɪŋgɚ]	[ˈsɪŋɚ]	[ˈsɪŋɚ]	[ˈsɪŋɚ]	[ˈsɪŋɚ]
bicker	/ˈbɪkɚ/	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]	[ˈbɪkɚ]
rubber	/ˈɹʌbɚ/	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]	[ˈɹʌbɚ]
zeal	/ziːɬ/	[ziːɬ]	[ziːɬ]	[ziːɬ]	[ziːɬ]	[ziːɬ]	[ziːɬ]	[ziːɬ]	[ziːɬ]
degree	/diˈgɹiː/	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]	[diˈgɹiː]
lack	/ɬæk/	[ɬæk]	[ɬæk]	[ɬæk]	[ɬæk]	[ɬæk]	[ɬæk]	[ɬæk]	[ɬæk]
bet	/bɛt/	[bɛt]	[bɛt]	[bɛt]	[bɛt]	[bɛt]	[bɛt]	[bɛt]	[bɛt]
witch	/wɪʧ/	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]
yet	/jɛt/	[jɛt]	[jɛt]	[jɛt]	[jɛt]	[jɛt]	[jɛt]	[jɛt]	[jɛt]
worthy	/ˈwɝːði/	[ˈwɝːθi] *	[ˈwɝːði]	[ˈwɝːði]	[ˈwɝːði]	[ˈwɝːði]	[ˈwɝːði]	[ˈwɝːði]	[ˈwɝːði]
heart	/hɑɹt/ - /hɑːɹt/	[hɑɹt]	[hɑɹt]	[hɑɹt]	[hɑɹt]	[hɑɹt]	[hɑɹt]	[hɑɹt]	[hɑɹt]
rack	/ɹæk/	[ɹæk]	[ɹæk]	[ɹæk]	[ɹæk]	[ɹæk]	[ɹæk]	[ɹæk]	[ɹæk]
betting	/ˈbɛɾɪŋ/	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛtɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]	[ˈbɛɾɪŋ]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
thick	/θɪk/	[θɪk]	[θɪk]	[θɪk]	[θɪk]	[θɪk]	[θɪk]	[θɪk]	[θɪk]
tuck	/tʌk/	[tʌk]	[tʌk]	[tʌk]	[tʌk]	[tʌk]	[tʌk]	[tʌk]	[tʌk]
fan	/fæn/ à [ɛ]	[fæn]	[fæn]	[fæn]	[fæn]	[fæn]	[fæn]	[fæn]	[fæn]
meddle	/ˈmɛdɬ/ - /ˈmɛɾɬ/	[ˈmiːɾɬ] *	[ˈmɛdɬi] *	[ˈmɛɾɬ]	[ˈmɛdɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]
anyhow	/ˈɛnihaʊ/	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]	[ˈɛnihaʊ]
loch	/ɬɑk/ - /ɬɑːk/	[lɑʧ] *	[ɬɑk]	[ɬɑːk]	[ɬɑk]	[ɬɑk]	[ɬɑk]	[lɑʧ] *	[ɬɑːk]
badge	/bædʒ/	[bædʒ]	[bædʒ]	[bædʒ]	[bædʒ]	[bædʒ]	[bædʒ]	[bædʒ]	[bædʒ]
carter	/ˈkɑɹɚ/ - /ˈkɑːɹɚ/	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]	[ˈkɑɹɚ]
leisure	/ˈliːʒɚ/	[ˈliːʒɚ]	[ˈliːʒɚ]	[ˈliːʒɚ]	[ˈliːʒɚ]	[ˈlɛʒɚ] *	[ˈliːʒɚ]	[ˈliːʒɚ]	[ˈliːʒɚ]
middle	/ˈmɪdɬ/ /ˈmɪɾɬ/	[ˈmɪɾɬ]	[ˈmɪɾɬ]	[ˈmɪɾɬ]	[ˈmɪdɬ]	[ˈmɪɾɬ]	[ˈmɪɾɬ]	[ˈmɪɾɬ]	[ˈmɪɾɬ]
batch	/bæʧ/	[bæʧ]	[bæʧ]	[bæʧ]	[bæʧ]	[bæʧ]	[bæʧ]	[bæʧ]	[bæʧ]
written	/ˈɹɪʔn/ - /ˈɹɪɾn/	[ˈɹɪtn]	[ˈɹɪɾn]	[ˈɹɪʔn]	[ˈɹɪtn]	[ˈɹɪʔn]	[ˈɹɪʔn]	[ˈɹɪtn]	[ˈɹɪʔn]
metal	/ˈmɛɾɬ/	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]	[ˈmɛɾɬ]
garter	/ˈgɑɹɚ/ - /ˈgɑːɹɚ/	[ˈgɑɹɚ]	[ˈgɑɹdnɚ] *	[ˈgɑɹɚ]	[ˈgɑɹɚ]	[ˈgɑɹɚ]	[ˈgɑɹtɚ]	[ˈgɑɹɚ]	[ˈgɑɹɚ]
which	/ʍɪʧ/ - /wɪʧ/	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]	[wɪʧ]
graze	/gɹeɪz/	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]	[gɹeɪz]
bishop	/ˈbɪʃəp/	[ˈbaɪʃəp] *	[ˈbɪʃəp]	[ˈbɪʃəp]	[ˈbɪʃəp]	[ˈbɪʃəp]	[ˈbɪʃəp]	[ˈbɪʃəp]	[ˈbɪʃəp]
fad	/fæd/	[fæd]	[fæd]	[fæd]	[fæd]	[fæd]	[fæd]	[fad] §	[fæd]
behave	/bɪˈheɪv/	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]	[bɪˈheɪv]
stronger	/ˈstɹɑːŋgɚ/ - /ˈstɹɑŋgɚ/	[ˈstɹɑŋgɚ]	[ˈstɹɑŋgɚ]	[ˈstɹɑŋgɚ]	[ˈstɹɑŋgɚ]	[ˈstɹɑŋgɚ]	[ˈstɹɑŋgɚ]	[ˈʃtɹɑŋgɚ]	[ˈstɹɑŋgɚ]
pat	/pæt/	[pæt]	[pæt]	[pæt]	[pæt]	[pæt]	[pæt]	[pæt] §§	[pæt]

Potential influence on realization: * reading error † hesitation § French-influenced vowel §§ French-influenced consonant ¶ lexical conditioning

Stimulus	Reference	AJ1	CT1	DC1	MT1	PS1	QV1	RL1	SK1
simmer	/ˈsɪmɚ/	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]	[ˈsɪmɚ]
sag	/sæg/	[sæg]	[sæg]	[sæg]	[sæg]	[sæg]	[sæg]	[sæg] §§	[sæg]
duck	/dʌk/	[dʌk]	[dʌk]	[dʌk]	[dʌk]	[dʌk]	[dʌk]	[dʌk] §§	[dʌk]
b*y	/ˈbɛɹi/	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]	[ˈbɛɹi]
bat	/bæt/	[bæt]	[bæt]	[bæt]	[bæt]	[bæt]	[bæt]	[bæt]	[bæt]
this	/ðɪs/	[ðɪs]	[ðɪs]	[ðɪs]	[ðɪs]	[ðɪs]	[ðɪs]	[ðɪs]	[ðɪs]
ridden	/ˈɹɪʔn/ - /ˈɹɪɾn/	[ˈɹɪdn]	[ˈɹɪɾn]	[ˈɹɪɾn]	[ˈɹɪdn]	[ˈɹɪɾn]	[ˈɹɪɾn]	[ˈɹɪdn]	[ˈɹɪɾn]
fat	/fæt/	[fæt]	[fæt]	[fæt]	[fæt]	[fæt]	[fæt]	[fæt]	[fæt]