

CHAPITRE 10

Les corpus oraux montréalais à l'ère numérique

Des bobines magnétiques à la diffusion
sur la plateforme du Fonds de données
linguistiques du Québec

Wim Remysen et Hugo Saint-Amant Lamy

[...] en ce moment, je suis très angoissée pour l'avenir de nos bandes sonores. Nous n'avons pas encore trouvé le médium parfait pour en assurer une bonne conservation. Les données originales, ce n'est pas la transcription, c'est l'enregistrement. (Pierrette Thibault, citée dans Daveluy et Laforest, 1994, 19-20)

Maintenant que les corpus existent...¹

Cinquante ans après leur genèse, force est de constater que les corpus socio-linguistiques montréalais de 1971, 1984 et 1995 demeurent de puissants catalyseurs pour l'analyse et la description du français parlé au Québec, notamment dans la perspective variationniste. Il n'est pas exagéré d'affirmer que les sociolinguistes qui ont pris l'initiative de constituer le premier de ces corpus – Gillian Sankoff, David Sankoff et Henrietta Cedergren – font figure

1. Ce titre est un clin d'œil au chapitre que Gillian Sankoff a consacré, en 2018, à la recherche sur le changement linguistique, intitulé « Before there were corpora » (avant que les corpus existent).

de véritables pionnières². Non seulement leur entreprise a inspiré quantité d'autres chercheuses par la suite, mais de nombreux travaux réalisés à partir des données recueillies ont aussi grandement contribué à une meilleure connaissance de plusieurs phénomènes caractéristiques du français québécois³.

Dans un contexte encore fortement marqué par des préjugés négatifs envers cette variété de français, de tels travaux avaient, en plus de leur importance scientifique, une pertinence sociale certaine⁴. L'élaboration de corpus subséquents en 1984 et 1995, où certains locuteurs de 1971 sont interviewés à nouveau, a permis d'ajouter une dimension diachronique précieuse au riche portrait sociolinguistique du projet original (Sankoff, 2018; Blondeau, 2020). Cette richesse a suscité l'envie et l'intérêt de sociolinguistes partout dans le monde. Aujourd'hui, Montréal 1971 et ses suites n'ont aucunement perdu de leur intérêt, bien au contraire. Bien plus qu'une simple pièce d'anthologie, la désormais célèbre trilogie ouvre la possibilité d'étudier le changement qui affecte le français parlé à Montréal, notamment par des comparaisons avec des corpus constitués à une date plus récente, et permet de prendre le pouls des représentations de la langue qui existaient à l'époque. La granularité sociale et la profondeur longitudinale inégalées des trois corpus en font en outre un étalon de choix à l'aune duquel il est possible d'interpréter globalement la variation en français québécois.

Cela dit, au fil du temps, la préservation des données recueillies et leur accessibilité sont devenues des enjeux majeurs. Ainsi, la numérisation des corpus, que souhaitaient déjà les auteures de la trilogie montréalaise⁵, est relativement récente et leur mutualisation ainsi que leur dépôt n'avaient jamais été réalisés jusqu'à la mise sur pied récente du Fonds de données linguistiques du Québec (FDLQ), en 2020. C'est de cette initiative qu'il sera question dans ce chapitre, que nous dédions aux auteures qui nous ont pré-

2. Nous utilisons régulièrement dans ce chapitre le féminin générique pour souligner l'apport des femmes à la constitution des premiers corpus québécois ainsi que des corpus subséquents.

3. En fait foi l'impressionnante bibliographie des recherches menées à partir de leur corpus (ainsi que des deux autres qui ont suivi). Cette bibliographie, réalisée par Hélène Blondeau et Marty Laforest en 2015, est disponible sur le site du Fonds de données linguistiques du Québec (<https://fdlq.recherche.usherbrooke.ca/bibliographie-travaux-grands-corpus-sociolinguistiques-mtl.html>).

4. Il peut être instructif de relire la partie programmatique de l'article de Sankoff *et al.* (1976) pour s'en convaincre.

5. Voir à ce sujet les propos recueillis par Michelle Daveluy et Marty Laforest auprès des initiatrices des trois corpus (et publiés dans Daveluy et Laforest, 1994).

cédés. Soucieux d'établir des ponts entre le passé et l'avenir, et désireux de donner la parole à nos prédécesseurs, nous ponctuerons notre texte de citations de leur cru, ce qui nous permettra d'illustrer comment ce nouveau Fonds répond à des préoccupations exprimées depuis l'émergence de la linguistique de corpus au Québec dans les années 1960 relativement à la pérennisation des données.

Le Fonds de données linguistiques du Québec et la place accordée aux corpus montréalais

Je me souviens, quand Pierrette [Thibault] était étudiante, avoir eu des conversations avec elle et Diane [Vincent] sur un projet de création d'une banque de données ou d'un centre de recherches, quelque chose qui donnerait une continuité à ce travail, une institutionnalisation. (Gillian Sankoff, citée dans Daveluy et Laforest, 1994, 14)

Le Fonds de données linguistiques du Québec est une plateforme numérique documentaire destinée à mutualiser des corpus oraux et écrits constitués depuis les années 1960 en vue de décrire et d'étudier le français du Québec. Il contient des collections diverses de documents jugés représentatifs du français tel qu'il est utilisé au Québec, en tenant compte de la variabilité et de l'évolution de ses usages. Les différents corpus réunis sur la plateforme témoignent ainsi de la langue ayant cours dans diverses régions du Québec, dans des contextes d'utilisation et auprès de milieux variés ainsi qu'à différentes époques de son histoire. Certains corpus témoignent par ailleurs des rapports normatifs et identitaires particuliers que les Québécois ont entretenus ou entretiennent avec leur langue. Ces corpus métalinguistiques se prêtent bien à l'analyse des discours qui ont circulé au Québec concernant la qualité de la langue et l'acceptabilité des particularismes du français québécois.

S'inscrivant dans le domaine des humanités numériques, le Fonds est développé par une équipe de linguistes-chercheurs associés au Centre de recherche interuniversitaire sur le français en usage au Québec (CRIFUQ), en étroite collaboration avec des analystes-informaticiens du Service des technologies de l'information de l'Université de Sherbrooke (STI), et il est financé par le ministère de la Langue française du Québec. Il poursuit essentiellement trois grands objectifs: 1) assurer la pérennité des corpus linguistiques

constitués au Québec (dont plusieurs ont été réalisés avant l'ère numérique), 2) permettre la consultation et l'interrogation de leurs contenus (souvent peu accessibles aux chercheurs), 3) valoriser les pratiques linguistiques que les corpus donnent à entendre. D'outils utilisés ponctuellement pour des recherches spécifiques, les corpus intégrés au Fonds sont ainsi appelés à devenir des ressources d'usage courant, facilement accessibles et partagées par la communauté des chercheurs. Le projet contribue à la patrimonialisation des corpus concernés : outre qu'il permet la découverte de ces collectes de données, il assure une reconnaissance de leur valeur scientifique et culturelle. Et la langue préservée dans ces corpus accède à ce même rang : la plateforme préserve des traces tangibles et les inscrit dans la conscience linguistique collective. De ce point de vue, le Fonds partage certaines caractéristiques avec d'autres plateformes spécialisées permettant d'héberger des corpus, comme Cocoon et Ortolang (deux initiatives soutenues par l'infrastructure Huma-Num) en France. En même temps, il s'en démarque à plus d'un titre, notamment pour ce qui est de la consultation et de la visualisation des données. La plateforme FDLQ permet notamment d'interroger, grâce à un moteur de recherche unique, l'ensemble des corpus qu'elle diffuse, et de peaufiner les résultats d'une requête par l'activation de filtres (<https://fdlq.crifuq.usherbrooke.ca/>). Des recherches transversales de ce type ne sont pas disponibles sur les autres plateformes, où la consultation se fait corpus par corpus. Le moteur de recherche permet de trouver des mots simples, des suites de mots, des collocations ainsi que des lemmes⁶, et de formuler des requêtes incluant des métacaractères (jokers) ou des expressions régulières (regex)⁷. La plateforme FDLQ se démarque également par les différentes possibilités de navigation qu'elle offre à l'utilisateur, qui peut, par exemple, parcourir la liste des corpus et des documents contenus dans le Fonds, et mieux comprendre ainsi la composition exacte des différents contenus qu'il consulte. Il peut aussi naviguer dans le lexique du Fonds (c'est-à-dire dans la liste de l'ensemble des mots-formes⁸ attestés, triés par ordre alphabétique ou par fréquence) ou encore dans les lexiques d'un corpus ou d'un document précis. Les entrées de tous

6. Les corpus versés au Fonds sont lemmatisés avec SpaCy (lemmatisation automatique).

7. Les métacaractères correspondent à un ou plusieurs éléments non spécifiés dans une requête (ex. « pi?e » renvoie les résultats *pire*, *pige*, *pile*, etc. alors que « pi*e » renvoie aussi *piastre* et *pierre*). Les expressions régulières, de leur côté, permettent de chercher des éléments répondant à différentes conditions logiques. Par exemple, l'expression régulière « \w+z » renvoie tous les mots (\w+) se terminant par la lettre Z.

8. La lemmatisation des corpus, réalisée seulement récemment, permettra la constitution d'autres lexiques, comprenant cette fois-ci les mots attestés dans les corpus sous forme de lemmes.

ces lexiques se présentent sous forme d'hyperliens grâce auxquels on repérera par la suite les mots en contexte. Le site de présentation du Fonds apporte des informations complémentaires sur le Fonds, son équipe et ses partenaires, et contient des fiches de présentation pour l'ensemble des corpus qu'il héberge. Ce site comprend aussi un guide de consultation détaillé qui explique les principales fonctionnalités de la plateforme (<https://fdlq.recherche.usher-brooke.ca/guide/Content/>).

Les corpus montréalais occupent une place de choix parmi l'ensemble des données réunies sur la plateforme FDLQ. À l'heure actuelle, le Fonds inclut 10 corpus montréalais, dont certains ont déjà été versés intégralement (ou partiellement) alors que d'autres sont en cours de traitement. Il s'agit plus particulièrement de Montréal 1971 (Sankoff *et al.*, 1976), Montréal 1984 (Thibault et Vincent, 1990), Montréal 1995 (Vincent, Laforest et Martel, 1995), PFC-Montréal (un sous-ensemble de PFC-Québec; Côté, 2014; Côté et Saint-Amant Lamy, 2023), le corpus variationniste Hochelaga-Maisonneuve 2012 et le corpus variationniste Saint-Michel-Montréal-Nord 2013 (à propos de ces deux corpus, constitués dans le cadre du projet *Le français à la mesure d'un continent* (sous la direction de France Martineau) et versés au corpus FRAN, voir Gadet et Martineau, 2017; Martineau et Séguin, 2016), le corpus Ahuntsic-Cartierville (constitué par Davy Bigot) et Mont(REA)L 2016 (Rea, 2020). Ces corpus sont présentés de manière plus détaillée dans le tableau 10.1 ci-dessous, où ils apparaissent dans l'ordre chronologique de leur constitution. À cette liste devraient s'ajouter quelques autres corpus qui permettraient de compléter le volet montréalais du Fonds, dont le corpus Bibeau-Dugas (constitué au début des années 1960; Dugas, 1986), le corpus Centre-Sud de Claire Lefebvre (avec des entrevues menées de 1976 à 1978; Doran, Drapeau et Lefebvre, 1982), le corpus Friesner du français montréalais et du contact français-anglais-espagnol (réalisé en 2007-2008; Friesner, 2009) et le corpus réalisé auprès de Montréalais d'origine haïtienne (par Mireille Tremblay), auxquels s'ajoute la série d'entrevues menées en français auprès d'Anglo-Montréalais en 1993-1994 (par Gillian Sankoff et Pierrette Thibault, voir Thibault et Sankoff, 1993 et Sankoff *et al.*, 1997).

L'état physique sous lequel se présentent les enregistrements varie considérablement d'un corpus à l'autre. Les plus anciennes entrevues ont été sauvegardées sur des bandes magnétiques, un support d'enregistrement couramment utilisé dans les années 1960 et 1970, mais dont la consultation nécessite des machines de moins en moins disponibles dans les centres de soutien audiovisuel, ou encore sur des cassettes audio pouvant être lues sur

TABLEAU 10.1

Les corpus montréalais versés dans le Fonds de données linguistiques du Québec

Corpus	Auteurs et auteurs	Année de constitution	Nombre de participantes et de participants	Nombre et durée des entrevues	Support et format d'enregistrement
Montréal 1971	Gillian Sankoff David Sankoff Henrietta Cedergren	1971	120 personnes (60 femmes, 60 hommes)	120 entrevues (± 135 heures)	analogique (rubans magnétiques)
Montréal 1984	Pierrette Thibault Diane Vincent	1984	72 personnes (39 femmes, 33 hommes)	72 entrevues (± 100 heures)	analogique (cassettes)
Montréal 1995	Diane Vincent Marty Laforest Guylaine Martel	1995	14 personnes (5 femmes, 9 hommes)	14 entrevues (± 26 heures) auxquelles s'ajoutent des enregistrements écologiques (± 45 heures)	numérique (disque magnéto-optique, fichiers ATRAC)
PCF-Montréal (sous-corpus de PFC-Québec)	Marie-Hélène Côté	2010-2017	43 personnes (22 femmes, 21 hommes)	31 conversations libres (± 18 heures) et 43 entretiens guidés (± 14 heures) auxquels s'ajoute la lecture de deux listes de mots et d'un court texte	numérique (mémoire flash, fichiers WAV)
HoMa variationniste (Hochelaga-Maisonneuve)	Hélène Blondeau France Martineau Mireille Tremblay Yves Frenette	2012	48 personnes (24 femmes, 24 hommes)	46 entrevues (± 82 heures)	numérique (fichiers WAV)
MoMu variationniste (Saint-Michel-Montréal-Nord)	Hélène Blondeau France Martineau Mireille Tremblay	2013	20 personnes (15 femmes, 5 hommes)	20 entrevues (± 26 heures)	numérique (fichiers WAV)
MoMu écologique (Saint-Michel-Montréal-Nord)	Hélène Blondeau France Martineau Mireille Tremblay	2013	2 personnes (2 femmes)	1 entrevue (± 1 heure)	numérique (fichiers WAV)
HoMa écologique (Hochelaga-Maisonneuve)	France Martineau	2014	6 personnes (1 femme, 5 hommes)	1 entrevue (± 1 heure)	numérique (fichiers WAV)
Ahuntsic-Cartierville	Davy Bigot	2015	63 personnes (39 femmes, 24 hommes)	63 entrevues (± 42 heures)	numérique (fichiers MP3)
Mont(REA)L 2016	Béatrice Rea	2016	53 personnes (27 femmes, 26 hommes)	36 entrevues (± 60 heures)	Numérique (fichiers WAV)

des lecteurs de cassettes portatifs. Certaines bandes avaient déjà été recopiées sur des cassettes (c'est le cas notamment de Montréal 1971). Le corpus de Montréal 1995 marque une étape charnière, en ce sens que les entrevues ont été pour la première fois enregistrées sur un support numérique, notamment sur des disques magnéto-optiques, une technologie apparue vers le milieu des années 1980, mais somme toute assez rapidement disparue. Les enregistrements plus récents, réalisés à partir des années 2000, ont été faits directement sur un support numérique et non plus analogique. Nous verrons plus loin que le traitement des corpus en vue de leur diffusion sur la plateforme FDLQ tient nécessairement compte de l'hétérogénéité qui caractérise les formats d'enregistrement.

L'ensemble des corpus montréalais réunis au FDLQ représente aujourd'hui déjà près de 550 heures d'enregistrements. Cette impressionnante somme de données témoigne bien des efforts collectifs considérables déployés par les sociolinguistes pour documenter le français parlé dans la métropole (et ailleurs au Québec). Au travers de leurs choix relatifs aux informatrices et informateurs avec lesquels elles se sont entretenues, on comprend que la plupart des auteures des corpus versés au FDLQ conceptualisent le français montréalais comme étant la variété de français parlée par les francophones natifs originaires de Montréal (ou, à tout le moins, arrivés dès le jeune âge, généralement avant l'entrée à l'école), et vivant dans la ville (ou dans un quartier en particulier) au moment de l'entrevue. Cela dit, les critères de sélection précis ont pu varier d'un corpus à l'autre. En 1971, par exemple, l'équipe de Sankoff-Cedergren a recruté dans les quartiers majoritairement francophones de la ville des « individus de 15 ans et plus qui étaient à la fois francophones [de langue maternelle française] et Montréalais d'origine », donc tout individu « né à Montréal ou [...] arrivé à Montréal avant le début de l'école primaire, c'est-à-dire avant l'âge de six ans » (Sankoff *et al.*, 1976, 91-92). En 1984 et 1995, lorsque plusieurs informateurs de 1971 ont été retrouvés et réinterviewés, certains d'entre eux avaient quitté Montréal, si bien que les corpus subséquents incluent des entrevues menées auprès de personnes vivant en périphérie de Montréal (Rive-Nord ou Rive-Sud), voire (dans quelques cas plus rares) ailleurs dans la province⁹. Cette approche a permis d'évaluer l'impact de la mobilité géographique sur les usages linguistiques d'un individu (Daveluy, 2005). L'élargissement progressif de la conception même du

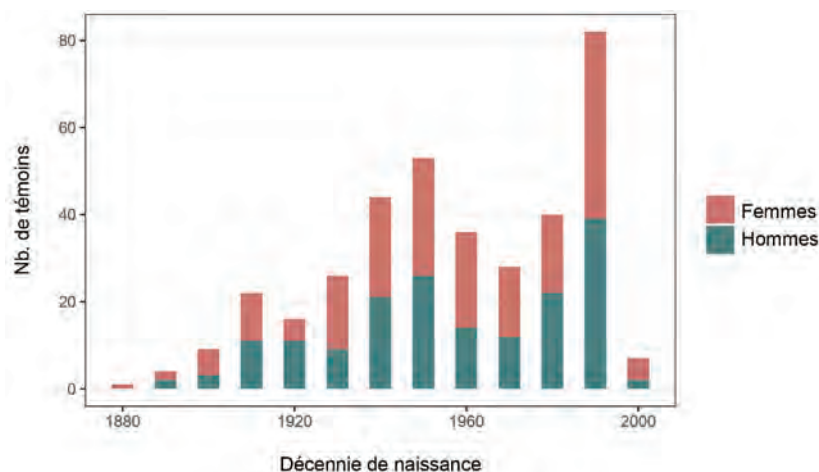
9. Selon Thibault et Vincent (1990, 22), 35 % des personnes interviewées à nouveau en 1984 vivaient à l'extérieur de l'île de Montréal au moment de leur deuxième participation.

français montréalais explique aussi que certaines collectes plus récentes incluent des entrevues menées auprès de Montréalaises et de Montréalais dont le français n'est pas la première langue apprise. On comprend mieux ainsi les effets des contacts linguistiques et de la diversité socioculturelle et linguistique croissante de la population montréalaise depuis les années 1970 sur le français parlé dans la métropole (Blondeau, 2014).

Si l'on tient compte des années de naissance des locutrices et locuteurs interrogés depuis les années 1970, le Fonds de données linguistiques du Québec permet de couvrir des pratiques linguistiques de personnes nées sur une période de plus d'un siècle (voir figure 10.1). En effet, environ 120 ans séparent la plus vieille locutrice, née en 1885 dans l'ancienne municipalité de Saint-Jean-Baptiste (maintenant intégrée à l'arrondissement du Plateau-Mont-Royal), des personnes les plus récemment rencontrées, nées dans les années 2000. Une telle richesse permettra de compléter considérablement, dans les prochaines années, notre compréhension de l'évolution du français montréalais, et de vérifier certaines hypothèses relatives à la théorie du changement linguistique. L'étude de cas présentée dans la dernière section de ce chapitre illustre concrètement l'apport du Fonds à cet égard.

FIGURE 10.1

**Nombre de locuteurs par décennie de naissance
dans les corpus montréalais**



La mutualisation et la diffusion des corpus oraux sur la plateforme FDLQ soulèvent d'importantes questions relatives aux droits d'auteur, à la propriété intellectuelle et à la confidentialité des données. Deux principes fondamentaux encadrent l'action du FDLQ à ce propos : la reconnaissance du travail des concepteurs des corpus, et la protection de la vie privée des personnes interrogées, toute l'entreprise de la linguistique de corpus reposant sur ces dernières. Chaque corpus intégré au Fonds fait l'objet d'une licence d'utilisation signée par les auteurs (ou par la ou les personnes qui les représentent) et autorisant l'intégration du corpus concerné au Fonds. Cette entente balise essentiellement la diffusion qui peut être faite des données, garantit l'anonymisation des entrevues, protège l'intégralité du corpus et assure le respect de la propriété intellectuelle et des droits moraux¹⁰. Plus particulièrement, l'entente-type prévoit que les corpus ne soient jamais rendus accessibles dans leur totalité sur la plateforme et que les résultats de recherche se limitent à un contexte de 60 mots avant et après le mot-pivot ou les mots-pivots (une limite de 12 secondes s'applique également aux extraits audio), à moins qu'une telle diffusion soit autorisée explicitement par l'auteur. Chaque contenu extrait d'un corpus est systématiquement accompagné d'une référence bibliographique qui signale l'auteur et le nom du corpus. De toute évidence, le contenu des ententes varie d'un corpus à l'autre, suivant le consentement obtenu auprès des personnes participantes et suivant le souhait exprimé par les auteurs. Dans certains cas, en effet, il arrive que les entrevues puissent être diffusées de manière intégrale (c'est le cas notamment du corpus PFC-Québec), alors que, dans d'autres, l'accès est plus restrictif.

Le traitement des corpus : des enregistrements à la préparation des données

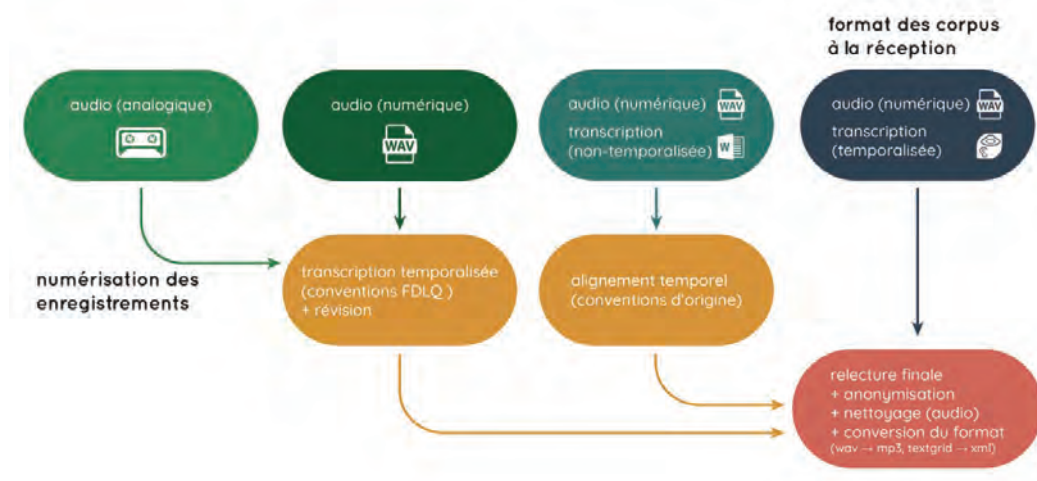
[N]ous avons passé des heures et des heures à discuter sur les conventions de transcription en se disant que c'était un outil qu'on faisait une fois pour toutes, mais ce n'est plus vrai. La transcription n'est aujourd'hui qu'un point de départ, un canevas de base. (Pierrette Thibault, citée dans Daveluy et Laforest, 1994, 20)

10. Il va de soi que chaque auteur de corpus garde l'entière responsabilité de ses droits sur son œuvre. Au Canada, il est d'ailleurs impossible de céder les droits moraux, catégorie de droits qui « permet à l'auteur d'une œuvre d'en revendiquer la création » (source : <https://educaloi.qc.ca/capsules/le-droit-dauteur-pour-la-protection-de-la-creation/>).

Les problèmes et les enjeux, nombreux pour ce qui est du traitement des corpus, concernent toutes les étapes menant à la mise à disposition des données, incluant leur anonymisation. De façon générale, le déroulement des opérations est le même pour tous les corpus : numérisation des fichiers sonores, transcription temporalisée des contenus et préparation à la diffusion (voir la figure 10.2 pour une représentation schématique de la chaîne de traitement). Dans certains cas, par exemple celui du corpus Mont(REA)L 2016, seuls les fichiers sonores sont disponibles. L'équipe du FDLQ s'occupe alors de l'entièreté du processus, à commencer par la numérisation, si les données sont stockées sur un médium analogique. Dans d'autres, une partie du traitement a déjà été effectuée. Par exemple, dans le cas des corpus Montréal 1971, Montréal 1984 et Montréal 1995, il existe déjà une transcription, mais celle-ci n'est pas temporalisée, c'est-à-dire il n'y a pas d'association structurée entre le son et le texte. L'alignement temporel des transcriptions existantes est alors assuré par l'équipe du FDLQ. Enfin, certains corpus comme PFC-Montréal sont transmis sous une forme presque complète. Il ne reste alors qu'à faire quelques ajustements en vue d'une intégration harmonieuse à la base de données.

FIGURE 10.2

Le traitement des corpus oraux en vue de leur diffusion sur la plateforme



Afin d'assurer la cohérence du Fonds, un certain nombre de principes de transcription sont respectés. Ceux-ci varient légèrement selon que les corpus sont déjà transcrits, ou non, au moment de leur réception. S'ils sont transcrits, la préoccupation principale est le respect des orientations éditoriales établies par les auteurs, ce qui comprend la préservation des conventions de transcription et le maintien des marquages extralinguistiques existants (codages, marquages, etc.). L'intervention du FDLQ dans la transcription se limite alors à la correction de coquilles et d'autres erreurs évidentes. Lorsque les transcriptions n'existent pas déjà, elles sont réalisées par l'équipe du FDLQ à l'aide du logiciel Praat (Boersma et Weenink, 2023), au format TextGrid. Les principes de base guidant cette opération sont la lisibilité des données et leur relative transparence pour un public composé en grande partie de non-spécialistes. Les conventions employées sont fondées sur celles du projet Phonologie du français contemporain (PFC; Durand, Laks et Lyche, 2002), et plus particulièrement celles développées pour son volet québécois (Côté, 2014; Côté et Saint-Amant Lamy, 2023).

Afin d'assurer un équilibre entre la représentation fidèle du contenu du signal sonore et la lisibilité par l'usager, les conventions PFC préconisent une transcription en orthographe standard de l'ensemble des éléments lexicaux effectivement produits par les locuteurs (et seulement de ceux-là). Par exemple, la séquence [ʒāvwajpɔsɔa:bɔs] serait transcrite «j'envoie pas ça à la boss». Le *ne* de négation n'est pas transcrit puisqu'il est complètement absent du signal, alors que l'article *la* est inclus car une trace de sa présence persiste dans l'allongement du [a]. Suivant les conventions PFC-Québec, la variation de nature phonétique/phonologique n'est pas représentée, mais la variation morphologique l'est. Ainsi, le mot *pas* est transcrit de la même façon, qu'il soit prononcé [pa], [pɔ] ou [pɔ̃]. Toutefois, la distinction entre «envoie» et «envoie» est faite afin de marquer l'opposition entre deux radicaux existants du verbe *envoyer*. Pour faciliter davantage encore la consultation par le public, les transcriptions faites par l'équipe du FDLQ ne comprennent pas de codages ou de marquages idiosyncrasiques¹¹. De plus, dans un souci de contrôle de la qualité, ces transcriptions font toutes l'objet

11. Par exemple, le recours aux codes et marquages dans la transcription (fictive) suivante permet de relever des phénomènes absents de la transcription orthographique standard, comme la liaison, l'intonation et l'alternance codique: «je suis T à veille de dire comme ↑ [angl] that's enough drama for me». Bien qu'ils communiquent des renseignements intéressants pour les chercheurs, ces éléments graphiques additionnels alourdissent la transcription et rendent son accès par le grand public plus difficile.

d'une révision complète. Les conventions adoptées pour les différents corpus, qu'ils soient déjà transcrits ou non, sont constamment mises à jour dans des guides de transcription détaillés.

Un des principaux intérêts de la plateforme du FDLQ pour l'exploitation des corpus oraux est la possibilité d'écouter, dans son contexte de production, un mot ou une expression tapée dans le moteur de recherche. Cette fonctionnalité nécessite un balisage temporel serré, obtenu par une segmentation des transcriptions en intervalles courts, n'excédant généralement pas 7 ou 8 secondes. Les séquences de silence dépassant 150 millisecondes sont également identifiées et isolées. Enfin, les passages où les voix de plusieurs locuteurs se chevauchent sont clairement délimités afin de bien représenter la temporalité des interactions et de mettre en évidence les séquences où l'environnement acoustique est plus chargé. La segmentation du signal sonore et sa transcription sont réalisées dans des fichiers TextGrid, sous Praat. Lorsque l'entièreté du traitement est effectuée par l'équipe du FDLQ, la transcription et la segmentation sont simultanées. Les corpus comprenant des transcriptions non temporalisées, de leur côté, font l'objet d'une procédure d'alignement particulière en trois étapes : le traitement préalable des transcriptions originales, le transfert du contenu textuel et l'ajustement des rendus finaux. Lors de la phase de prétraitement, le contenu des fichiers originaux est transféré dans un fichier de format TXT où chaque ligne correspond à un tour de parole. Les longs tours de parole dépassant souvent la durée maximale visée pour les intervalles, ils sont donc divisés en séquences plus courtes par l'insertion de ruptures après les points et les virgules. Dans Praat, un membre de l'équipe d'alignement insère ensuite des balises temporelles au long du fichier sonore, balises correspondant à chaque saut de ligne présent dans le fichier TXT. Le contenu textuel peut ensuite être transféré automatiquement du TXT au TextGrid, en associant chaque ligne à un intervalle. Une fois la transcription importée dans Praat, elle est révisée de façon à obtenir le même format que les transcriptions effectuées en totalité par l'équipe du FDLQ (ajustement des frontières d'intervalles, segmentation des intervalles trop longs, isolement des silences, ajouts d'éléments manquants, etc.).

Tous les fichiers, peu importe les particularités associées à leur traitement, font l'objet d'une procédure de finalisation en vue de leur publication. Cette procédure comporte une relecture complète de la transcription, l'anonymisation des informations sensibles, l'archivage des différentes versions des fichiers et le nettoyage des fichiers sonores. La tâche de relecture a pour

but d'assurer que les fichiers définitifs soient de la plus haute qualité. Elle comprend la correction des coquilles restantes, les ajustements nécessaires à l'alignement temporel ainsi que la résolution de problèmes divers soulevés dans les phases précédentes du traitement¹². La personne responsable de la finalisation de chaque fichier doit s'assurer de la protection des renseignements personnels des locuteurs qui y interviennent. La procédure d'anonymisation prévoit l'élimination de tout élément permettant une identification directe (nom, prénom) ou indirecte (adresse, date de naissance, employeur, noms des amis/parents/enfants, etc.). Dans le fichier sonore, cela se fait en remplaçant le passage sensible par un bruit périodique de la même durée et suivant la courbe intonative¹³. Dans le texte, l'élément retiré est remplacé par un terme générique précédé d'un N majuscule. Comme l'illustre bien l'exemple ci-dessous (tableau 10.2), la méthode peut alourdir le texte et rendre plus difficile son interprétation. Pour pallier ces inconvénients et pour conserver la dimension humaine de l'interaction, il pourrait être pertinent de réfléchir à une forme de pseudonymisation qui permettrait minimalement de remplacer les anthroponymes par des équivalents appropriés et vraisemblables, leur choix étant basé sur les données disponibles relativement à la fréquence des prénoms selon l'année de naissance du locuteur¹⁴.

TABLEAU 10.2

Anonymisation des entrevues (exemple fictif)

Version non anonymisée	Ma fille Claudine, elle est gérante au Steinberg puis elle a engagé Mario.
Version anonymisée	Ma fille Nprénom, elle est gérante au Nentreprise puis elle a engagé Nprénom.
Version pseudonymisée	Ma fille {Christiane}, elle est gérante au Nentreprise puis elle a engagé {Alain}.

12. Généralement, la personne qui s'occupe de cette tâche est la troisième à travailler sur le fichier. Elle suit celle ayant effectué la première transcription et celle ayant procédé à la révision (dans le cas des fichiers entièrement traités par le FDLQ) ou à l'alignement (dans le cas des transcriptions initialement non temporalisées). Les différentes personnes intervenant sur un fichier peuvent partager à même le fichier TextGrid leurs questions et leurs incertitudes par rapport à la transcription.

13. L'opération est effectuée avec un script développé par Daniel Hirst (2013).

14. Les données pour la période 1890-1980 sont tirées de Duchesne (1997) et celles pour la période 1980-2020 de la banque de prénoms de Retraite Québec (<https://www.retraite-quebec.gouv.qc.ca/fr/services-en-ligne-outils/banque-de-prenoms>).

Des considérations liées à la gestion de l'espace et aux fonctionnalités de lecture mènent à convertir les fichiers audio (en général au format WAV) au format MP3. Les fichiers sont également nettoyés pour assurer leur facilité d'écoute et leur uniformité. Le nettoyage est effectué dans le logiciel de restauration audio iZotope RX Advanced et comprend, entre autres, une correction du ronflement de secteur, une diminution de la réverbération et du bruit ainsi qu'une normalisation du volume sonore. Les fichiers de transcription sont convertis au format XML. Après ces dernières interventions, les fichiers sont prêts pour une diffusion sur la plateforme FDLQ, étape qui est assurée par l'outil de gestion décrit dans la section suivante.

La gestion des corpus : des métadonnées à la publication des entrevues

La constitution de mégacorpus comprend un ensemble d'activités de gestion de données qui demandent un investissement et des efforts considérables pour assurer la sauvegarde des renseignements généraux et les transcriptions. Il est certain qu'une politique d'accessibilité des données pour d'autres chercheurs contribue à assurer un rendement satisfaisant, compte tenu du travail fourni. (Thibault et Vincent, 1990, 29)

La gestion des contenus diffusés sur le FDLQ constitue un autre défi important dans la réalisation du projet, ne serait-ce qu'en raison de l'hétérogénéité des corpus, du nombre élevé de documents concernés ou encore de la confidentialité de certaines données et des droits d'auteur à respecter. Cette gestion est assurée par une interface spécialement conçue pour les besoins du projet, baptisée Comet. Il s'agit d'un outil collaboratif qui se présente sous la forme d'une application Web accessible à l'aide d'un navigateur. Comet permet plus particulièrement de consulter, d'éditer et de publier les différents corpus du Fonds (et les documents qui les constituent), mais aussi de gérer l'ensemble de leurs métadonnées. Les interfaces Comet et FDLQ communiquent entre elles par un intergiciel à messages. Lorsqu'un document est publié par un administrateur, Comet envoie un message à la plateforme FDLQ qui constitue son canal de diffusion. L'outil est également doté d'un moteur de recherche qui permet de faire des requêtes textuelles dans les corpus.

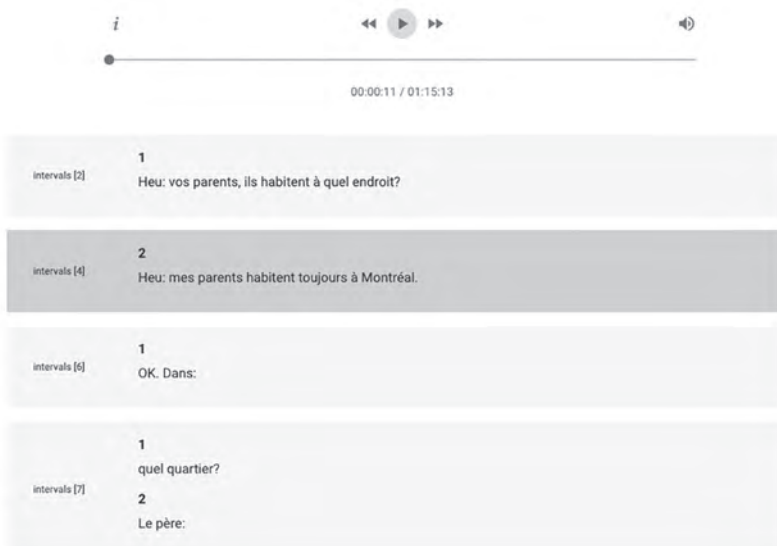
L'interface Comet permet de gérer conjointement les documents sources (fichiers audio) et les transcriptions (fichiers textuels) de toutes les entrevues

faisant partie des corpus oraux. Lorsqu'une entrevue est jugée prête en vertu du protocole de traitement présenté dans la section précédente, les différents fichiers ainsi que les métadonnées qui s'y rattachent sont importés. Cette importation peut se faire entrevue par entrevue ou encore en lot, lorsqu'on veut importer un corpus complet (ou un sous-corpus composé de plusieurs entrevues). Tous les fichiers sont stockés sur un serveur sécurisé, hébergé à l'Université de Sherbrooke. La transcription peut être visualisée de différentes manières – sous forme d'intervalle (voir figure 10.3a), ou sous forme de dialogue, avec des tours de parole qui alternent (voir figure 10.3b) – grâce à une feuille de style CSS qui est adaptée aux particularités de chaque corpus.

FIGURE 10.3

Visualisation sous forme d'intervalle (a) ou sous forme de dialogue (b) de la transcription de l'entrevue menée auprès de Bertrand P. (Corpus Montréal 1984)

a)



b)



L'alignement temporel permet par ailleurs d'écouter l'entrevue en suivant la transcription en temps réel. Cette fonctionnalité est très utile dans la mesure où elle permet de retourner facilement dans l'enregistrement original lorsqu'on veut écouter la prononciation d'un mot ou encore lorsqu'on veut valider un élément de la transcription. Lorsqu'un collaborateur détecte une erreur ou une coquille, il est facile de la corriger à l'aide de l'éditeur XML intégré à l'interface Comet et de publier la version corrigée. Il est également possible d'ajouter des pièces jointes à un corpus ou à une entrevue. Cette fonctionnalité peut être utile pour conserver des documents qui témoignent des différentes étapes dans la constitution du corpus, comme le protocole d'édition, ou pour archiver d'autres versions du fichier audio.

La gestion des métadonnées est au cœur des fonctionnalités de Comet. Dans la mesure où le projet adhère aux principes définis par l'initiative FAIR (<go-fair.org>) – qui encourage à rendre les données scientifiques faciles à trouver, accessibles, interopérables et réutilisables –, les contenus diffusés sur la plateforme doivent en effet être dotés de données de référence riches, structurées et pérennes. L'outil permet d'associer des métadonnées autant à des corpus qu'aux documents qui les constituent, qu'il s'agisse de métadonnées descriptives (l'auteur, la date et le titre d'un document, par exemple, ou encore les métadonnées de nature sociologique et démographique relatives

aux locuteurs enregistrés¹⁵), structurelles (comme l'appartenance du document à un corpus précis) ou administratives (informations concernant les droits d'accès ou les droits d'auteur, par exemple, ou concernant l'état d'un document dans le processus d'édition).

Comet est accessible seulement aux personnes autorisées par le responsable du projet. Contrairement à la plateforme FDLQ, qui ne requiert pas d'authentification (pour la simple raison que l'affichage des données se fait suivant les limites autorisées par les ententes signées), l'interface Comet est dotée de plusieurs niveaux d'autorisation qui permettent de contrôler minutieusement l'accès aux données. Les principaux niveaux d'accès concernent la lecture (consultation) et l'exploration (requêtes textuelles) des documents (accès minimal), l'édition (ajout d'un corpus ou d'un document, ajout de métadonnées, intervention au besoin dans la transcription, etc.), l'approbation et la diffusion. Seuls les administrateurs du projet ont la possibilité de publier un document ou un corpus sur la plateforme FDLQ et de gérer les accès accordés.

L'exploitation des données : l'exemple des mots en /ar#/

Le bien durable a été obtenu en raison d'une problématique justifiant la captation d'un ensemble de données qui pourront être observées sous différents angles, au-delà de ce qui avait été prévu par les concepteurs du projet. Si l'entreprise est coûteuse, elle se rentabilise par la quantité d'observables qu'elle offre aux chercheurs et par l'étalement dans le temps et dans l'espace de son utilisation. (Vincent, 2009 : 86)

Pour les linguistes, les bases documentaires de grande taille, comme le FDLQ, présentent l'avantage indéniable de permettre l'étude de phénomènes moins fréquents, difficiles à cerner de façon satisfaisante dans des ensembles plus limités. Afin d'illustrer les nouvelles possibilités de recherche ainsi créées, nous présentons ici l'analyse d'un phénomène de ce type : la neutralisation des voyelles basses antérieure et postérieure devant un /r/ final. Nous limitons cette analyse aux corpus montréalais du Fonds.

15. Les données sociodémographiques (âge, scolarité, profession, etc.) peuvent déjà être consultées dans l'interface Comet par les chercheurs qui ont les autorisations requises, mais ne sont pas encore disponibles sur la plateforme du Fonds. L'ajout est prévu dans une version ultérieure qui sera diffusée depuis le site Web du projet.

Si l'opposition /a/ - /ɒ/ se maintient bien en français québécois (ex. *patte-pâte, mal-mâle*), on observe une tendance à favoriser la postérieure devant une rhotique finale (ex. *égare* [egaʁ] → [egɒʁ], *égard* [egɒʁ]; voir Côté, 2012, 2016). Côté évoque comme facteur contributif à ce phénomène l'important déséquilibre statistique qui existe entre l'ensemble, très limité, des mots où l'on trouve traditionnellement la voyelle antérieure (catégorie /aʁ#/; ex. *gare, guitare, prépare*) et celui, beaucoup plus important, où l'on trouve la voyelle postérieure (catégorie /ɒʁ#/; ex. *phare, départ, part*). Or, ce déséquilibre représente un obstacle de taille pour l'exploration du phénomène. Par exemple, dans une étude des voyelles prérhotiques produites en parole spontanée, Saint-Amant Lamy (2022) ne recense que 42 occurrences de mots appartenant à la catégorie /aʁ#, contre 623 occurrences pour /ɒʁ# (sur un total de plus de 13 000 voyelles prérhotiques). Dans ces conditions, les possibilités d'analyses sont évidemment très limitées.

Une solution possible au problème de la rareté consiste à amener délibérément les locuteurs à produire les variables étudiées, par exemple dans le cadre d'une tâche de lecture. C'est l'approche adoptée par Côté (2012, 2016), qui met en lumière une distinction générationnelle claire dans les données de PFC-Trois-Rivières: alors que les locuteurs plus âgés (nés entre 1931 et 1958) favorisent nettement [a] dans les mots de la catégorie /aʁ#, les plus jeunes (nés entre 1986 et 1989) adoptent à divers degrés la variante [ɒ]¹⁶. Toujours à partir d'un exercice de lecture, Michaud et Remysen (2025), observent aussi une utilisation nettement plus fréquente de [ɒ] pour ces mots à Sherbrooke et à Victoriaville chez les jeunes de moins de 30 ans, nés depuis 1990. La tâche de lecture a toutefois le désavantage de favoriser l'hypercorrection, associée à la variante antérieure. Ainsi, deux locutrices de l'échantillon de Côté (2012, 2016) ont dû être exclues en raison de leur propension à antérioriser toutes les voyelles basses, y compris dans les mots de la catégorie /ɒʁ# (ex. *homard* [omaʁ]).

En agrégeant de grandes quantités de données orales issues d'entrevues semi-dirigées ou de conversations spontanées, le FDLQ permet de régler en bonne partie le problème de la rareté et d'éviter celui de l'hypercorrection en contexte de lecture. Nous avons ainsi effectué des recherches dans le Fonds pour chacun des mots de la classe /aʁ#/ identifiés dans Dumas (1986, 1987)

16. Quatre jeunes sur cinq favorisent [ɒ] dans *guitare* et *gare*, deux d'entre eux optent aussi pour cette voyelle dans *bulgare*. Le jeune locuteur JG la produit également dans *prépare*, ce qui suggère que la neutralisation est totale chez lui.

et Côté (2012, 2016), soit *bagarre*, *bulgare*, *cigare*, *cithare*, *déclare*, *démarre*, *gare*, *guitare*, *tare* et tous les verbes se terminant par le morphème *pare* (ex. *prépare*, *sépare*, *compare*, *répare*, *accapare*). Ces requêtes ont permis d'identifier 237 occurrences dans les corpus Montréal 1971, Montréal 1984, Montréal 1995, PFC-Montréal et Mont(REA)L 2016, qui totalisent ensemble 4,24 millions de mots. Le nombre de cas reste modeste, mais il s'avère suffisant pour une analyse générale des facteurs sociaux et linguistiques qui conditionnent la variation.

Les 237 voyelles ont été classées de façon impressionniste par accord inter-juge¹⁷. Seules deux catégories ont été distinguées : [a] et [ɒ]. L'exercice s'est avéré plutôt aisé, puisque les deux voyelles s'opposent non seulement par le timbre, mais souvent aussi par la durée et la diphtongaison¹⁸. Trois facteurs temporels sont disponibles pour l'analyse : l'année de naissance des locuteurs, leur âge au moment de l'entrevue et l'année de l'enregistrement. Deux autres facteurs externes ont été étudiés, soit le sexe et la catégorie professionnelle. Le classement de la catégorie professionnelle est tiré de Thibault et Vincent (1990), mais rapporté sur une échelle de 1 à 3, soit : 1-cols bleus, 2-cols blancs, 3-cadres et professionnels. Les facteurs internes retenus pour l'analyse sont la catégorie grammaticale (nom, verbe) et la fréquence du lexème dans l'ensemble des corpus. La catégorie grammaticale est retenue car elle est associée à deux réalités morphologiques favorisant potentiellement le maintien du [a] dans la catégorie verbale : la domination du morphème *-pare*, qui pourrait assurer une certaine robustesse paradigmatique, et l'existence de nombreuses formes dérivées où la neutralisation vers [ɒ] est moins productive (ex. *séparer*, *réparais*, *comparez*, *préparerai*). La prise en compte de la fréquence permet de vérifier si un phénomène de diffusion lexicale est à l'œuvre.

Avant de procéder à la modélisation du phénomène de neutralisation, une question doit être posée quant à la nature de celui-ci : s'agit-il d'un cas de gradation selon l'âge ou d'un changement en cours ? Les études antérieures, basées sur des données synchroniques, ne permettent pas d'apporter une réponse définitive. La présence dans le FDLQ de corpus montréalais élaborés à différentes époques rend toutefois la chose possible. Une série de régressions logistiques a été effectuée afin de déterminer la proportion de la

17. Dans les cas de désaccord entre les deux auteurs, une locutrice de français québécois ayant une formation en phonologie (Marie-Pier Picard) a été sollicitée pour trancher.

18. La voyelle [a] est la seule voyelle ne pouvant être diphtonguée en syllabe finale fermée par /ʁ/.

variance [aʁ#] - [ɔʁ#] expliquée par chacun des facteurs temporels (voir tableau 10.3). On constate que l'année de naissance est de loin le facteur le plus explicatif, ce qui indique qu'un changement est en cours¹⁹. L'âge au moment de l'entrevue et l'année de l'enregistrement ont donc été écartés de l'analyse qui suit.

TABEAU 10.3

Portion de la variance expliquée par chacun des facteurs temporels

Facteur	R ² de Nagelkerke
Année de naissance	0,239
Âge lors de l'entrevue	0,126
Année de l'enregistrement	0,066

Les statistiques descriptives concernant les facteurs catégoriques sont présentées à la figure 10.4 et celles pour les facteurs continus, à la figure 10.5. Afin de dégager les tendances générales dans les données, nous avons d'abord testé la relation statistique existant entre chaque facteur et le taux d'emploi de la voyelle postérieure, notamment grâce à des tests t et des mesures de corrélation. La corrélation entre l'année de naissance et le taux de réalisations [ɔʁ#] a été mesurée à partir des moyennes de chaque locuteur. Dans le cas de la fréquence lexicale, la corrélation a été mesurée à partir des moyennes pour chaque lexème. Les tests suggèrent que les variables jouant un rôle significatif sont la catégorie professionnelle, la classe grammaticale et l'année de naissance. Des comparaisons par paire pour la catégorie professionnelle montrent qu'il n'y a pas de distinction significative entre les cols bleus et les cols blancs, mais que ces deux groupes se distinguent de la catégorie des cadres et professionnels ($p < 0,001$). Aucune tendance claire n'émerge pour le sexe et la fréquence lexicale. En ce qui a trait à la classe grammaticale, il semble que les noms soient nettement plus propices à la neutralisation que les verbes.

Nous avons par la suite élaboré un modèle de régression logistique binaire (fonction *glm* dans R) afin de mesurer de façon intégrée le rôle des différents facteurs de variation. La catégorie vocalique (0 = [aʁ#], 1 = [ɔʁ#]) a été retenue comme variable dépendante, et l'année de naissance du locuteur, son sexe

19. Dans un cas de gradation selon l'âge, le facteur le plus explicatif serait l'âge au moment de l'entrevue.

FIGURE 10.4

Taux de réalisation du type [ɒɪ#] en fonction du sexe, de la catégorie professionnelle et de la classe grammaticale

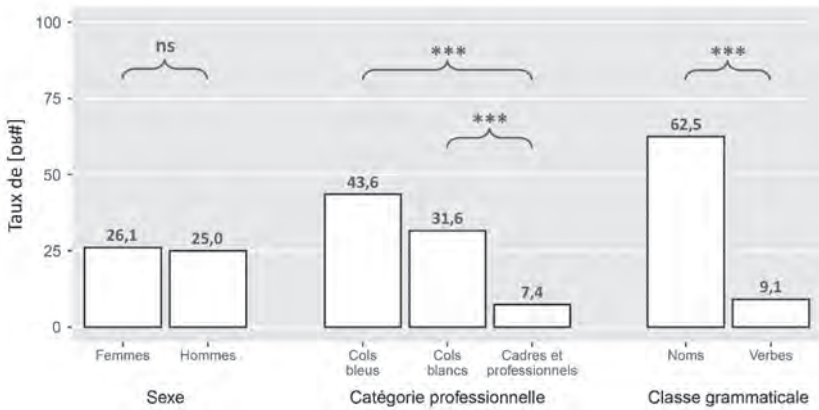
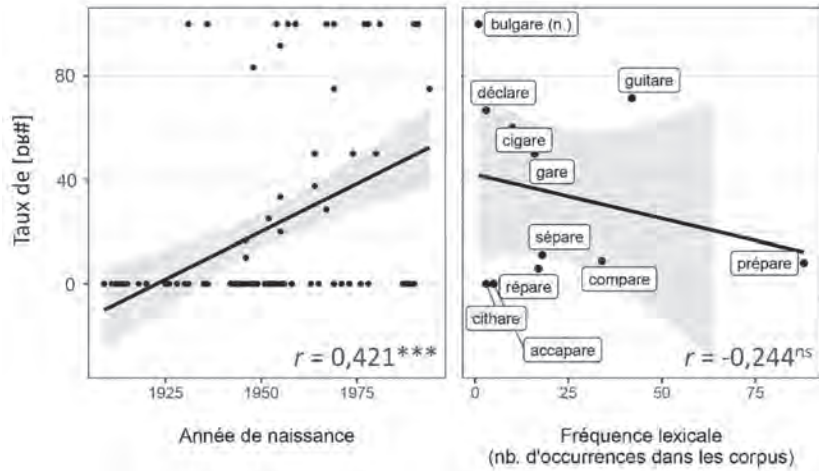


FIGURE 10.5

Taux de réalisation du type [ɒɪ#] en fonction de l'année de naissance et de la fréquence lexicale



(F = 0, H = 1), sa catégorie professionnelle, la fréquence lexicale du mot et la catégorie grammaticale de celui-ci (nom = 0, verbe = 1), comme variables indépendantes. Le modèle de régression logistique s'avère significatif²⁰ et permet d'expliquer 55,9 % de la variance (R^2 de Nagelkerke) et de classer correctement 87,3 % des observations. Lorsque les facteurs sont pris en compte ensemble, on relève un effet significatif de l'année de naissance, de la catégorie professionnelle et de la classe grammaticale (voir tableau 10.4). Pour l'année de naissance, les chances que la variante postérieure soit produite augmentent d'environ 5 % par année. Elles diminuent toutefois d'environ les deux tiers pour chaque échelon dans la catégorisation professionnelle. Enfin, les chances que la variante [ɒʁ#] soit réalisée sont environ 20 fois plus élevées dans les noms que dans les verbes. Le sexe et la fréquence lexicale n'ont pas d'effet significatif sur le choix de la voyelle.

TABLEAU 10.4

Tableau récapitulatif de la régression logistique binaire

Variable	B	Erreur type	p	Rapport des cotes (RdC)	IC 95 % pour RdC	
					Borne inf.	Borne sup.
Année	0,049	0,012	<0,001	1,051	1,027	1,077
Catégorie professionnelle	-1,094	0,328	<0,001	0,335	0,170	0,624
Classe grammaticale (verbe)	-3,055	0,517	<0,001	0,047	0,016	0,123
Fréquence lexicale	-0,002	0,008	0,793	0,998	0,982	1,014
Sexe (homme)	-0,904	0,483	0,061	0,405	0,151	1,024

Ces résultats permettent de reconstituer sommairement l'évolution de la neutralisation de /a/ et /ɒ/ devant un /ʁ/ final (voir figure 10.6). Il semble que le processus débute avec quelques locuteurs nés dans le deuxième quart du 20^e siècle et appartenant aux catégories des cols bleus et des cols blancs²¹. Les cadres et les professionnels suivent avec un décalage d'une génération. Ce schéma pourrait suggérer que le changement se situait initialement sous

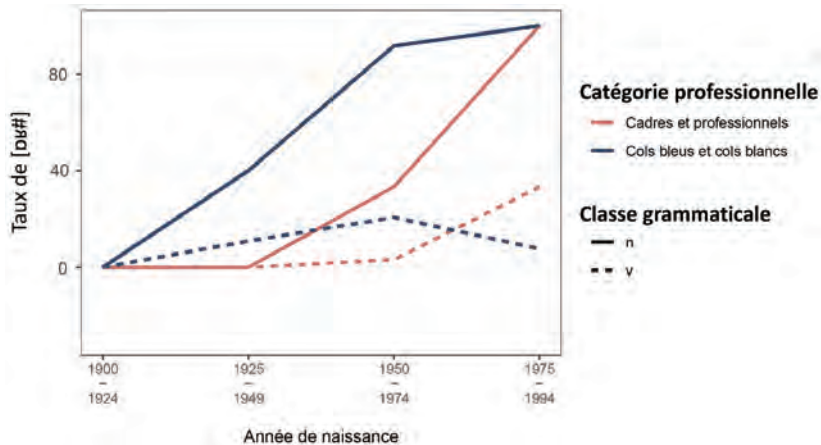
20. $\chi^2(5, N = 237) = 22,337, p < 0,001$

21. Les deux catégories ont été combinées à la figure 10.6 puisqu'aucune différence significative dans l'usage des variantes [aʁ#] et [ɒʁ#] n'a pu être identifiée avec les tests t.

le seuil de la conscience. Le changement s'observe essentiellement dans les noms (ex. *gare, guitare, cigare*), la réalisation [aʁ#] demeurant largement dominante dans les verbes (ex. *prépare, sépare*). Il est possible que l'alternance du radical en /aʁ#/ avec des formes dérivées, où le [a] se maintient de façon plus robuste, ait contribué à freiner la neutralisation dans cette dernière catégorie. Toutefois, chez les jeunes cadres et professionnels, la variante [ɔʁ#] semble gagner du terrain au sein de la classe verbale, avec un tiers des occurrences étudiées. Cette tendance est à confirmer (elle est ici basée sur un petit nombre d'observations), mais elle pourrait indiquer qu'un changement dans le statut de la variante postérieure est en cours. La neutralisation étant complétée (et le conditionnement social éliminé) pour la catégorie nominale, il est probable que la pression systémique en faveur de la forme [ɔʁ#] soit devenue suffisamment forte pour contrebalancer les facteurs ayant jusque-là permis le maintien de [aʁ#] dans le paradigme *prépare-sépare-répare*.

FIGURE 10.6

Taux de réalisation du type [ɔʁ#] en fonction de la classe grammaticale et de la catégorie professionnelle



* * *

Alors, nous avons voulu faire quelque chose comme ça, sur le français de Montréal, une grosse étude sociolinguistique. C'était des folies de jeunesse! (Henrietta Cedergren, citée dans Daveluy et Laforest, 1994, 12)

Outre qu'elle pérennise des collectes de données dont la valeur patrimoniale n'est plus à démontrer, la plateforme FDLQ permet de mutualiser des données qui, autrement, auraient pu rester fragmentées et dispersées. Grâce aux « folies de jeunesse » des sociolinguistes québécoises de la première heure et aux progrès récents des humanités numériques, nous sommes en mesure de rassembler des données en nombre considérable et de reconstituer certaines étapes dans l'évolution récente du français au Québec. Le FDLQ ouvre ainsi la porte à de nouvelles avancées en matière d'analyse, par la taille des corpus rassemblés ainsi que par la profondeur chronologique couverte par les données réunies. La plateforme favorise aussi les analyses comparatives, puisqu'elle permet de rapprocher et de croiser plus facilement des données recueillies dans différentes régions du Québec.

Nous avons illustré les nombreuses possibilités d'analyse offertes par le FDLQ à partir de l'exemple d'un phénomène de prononciation en particulier, la neutralisation des voyelles basses en position prérhotique. S'il reste beaucoup à découvrir sur ce phénomène, en particulier sur ses conditions d'apparition, il apparaît nettement qu'une meilleure compréhension repose, d'abord et avant tout, sur une meilleure description du phénomène. La rareté des occurrences des mots appartenant à la classe /aʁ#/ dans la parole spontanée avait jusqu'ici rendu cette tâche difficile. Un outil comme le FDLQ peut alors se révéler une aide précieuse pour la description et l'analyse de ce genre de phénomènes peu fréquents.

Références

- BLONDEAU, Hélène (2014), La nature métropolitaine du Montréal « d'aujourd'hui » et le français « d'ici », dans Wim REMYSEN (dir.), *Les français d'ici: du discours d'autorité à la description des normes et des usages*, Québec, Presses de l'Université Laval, 205-239.
- BLONDEAU, Hélène (2020), Du cheminement sociolinguistique des individus: l'apport des études longitudinales à l'étude de la variation et du changement en français parlé, dans Wim REMYSEN et Sandrine TAILLEUR (dir.), *L'individu et sa langue: hommages à France Martineau*, Québec, Presses de l'Université Laval, 13-46.

- BLONDEAU, Hélène et Marty LAFORÉST (2015), *Bibliographie des travaux fondés sur les données des grands corpus sociolinguistiques de Montréal*, <https://fdlq.recherche.usherbrooke.ca/bibliographie-travaux-grands-corpus-sociolinguistiques-mtl.html>.
- BOERSMA, Paul et David WEENINK (2023), *Praat, doing phonetics by computer*, version 6.3.10., <http://www.praat.org/>.
- CÔTÉ, Marie-Hélène (2012), Laurentian French (Quebec): extra vowels, missing schwas and surprising liaison consonants, dans Randall GESS, Chantal LYCHE et Trudel MEISENBURG (dir.), *Phonological variation in French: illustrations from three continents*, Amsterdam, John Benjamins, 235-274.
- CÔTÉ, Marie-Hélène (2014), Le projet PFC et la géophonologie du français laurentien, dans Jacques DURAND, Gjert KRISTOFFERSEN et Bernard LAKS (dir.), *La phonologie du français: normes, périphéries, modélisation*, Nanterre, Presses universitaires de Paris Ouest, 175-198.
- CÔTÉ, Marie-Hélène (2016), Variation in Canada: Trois-Rivières in Quebec, dans Sylvain DETEY, Jacques DURAND, Bernard LAKS et Chantal LYCHE (dir.), *Varieties of spoken French*, Oxford, Oxford University Press, 449-462.
- CÔTÉ, Marie-Hélène et Hugo SAINT-Amant LAMY (2023), The Phonologie du français contemporain project in Quebec: methodological and dialectometric considerations, dans Elissa PUSTKA, Carmen QUIJADA VAN DEN BERGHE et Verena WEILAND (dir.), *Corpus dialectology: from methods to theory (French, Italian, Spanish)*, Amsterdam/Philadelphie, John Benjamins, 60-83.
- DAVELUY, Michelle (2005), *Les langues étendards: allégeances langagières en français parlé à Montréal*, Québec, Éditions Nota Bene.
- DAVELUY, Michelle et Marty LAFORÉST (1994), Entretiens: *Tout ça a commencé en 1970...*, *Culture*, 14(2), 11-26.
- DORAN, Michèle, Lyne DRAPEAU et Claire LEFEBVRE (1982), Le projet Centre-Sud: corpus et méthodologie, dans Claire LEFEBVRE (dir.), *La syntaxe comparée du français standard et populaire: approches formelles et fonctionnelles*, tome 2, Québec, Office québécois de la langue française, 401-420.
- DUCHESNE, Louis (1997), *Les prénoms, des plus rares aux plus courants au Québec*, Saint-Laurent, Trécaré.
- DUGAS, André (1986), *Le corpus Bibeau-Dugas*, document interne, Université du Québec à Montréal.
- DUMAS, Denis (1986), Le statut des « deux a » en français québécois, *Revue québécoise de linguistique*, 5(2), 167-196.
- DUMAS, Denis (1987), *Nos façons de parler: les prononciations en français québécois*, Québec, Presses de l'Université Laval.
- DURAND, Jacques, Bernard LAKS et Chantal LYCHE (2002), La phonologie du français contemporain: usages, variétés et structure, dans Claus D. PUSCH et Wolfgang RAIBLE (dir.), *Romanistische Korpuslinguistik: Korpora und gesprochene Sprache*, Tübingen, Gunter Narr, 93-106.
- FRIESNER, Michael (2009), *The Social and Linguistic Predictors of the Outcomes of Borrowing in the Speech Community of Montréal*, thèse de doctorat, University of Pennsylvania.

- GADET, Françoise et France MARTINEAU (2017), Le maillage du français en Amérique du Nord, dans un cadre de francophonie, dans Beatrice BAGOLA et Ingrid NEUMANN-HOLZSCHUCH (dir.), *L'Amérique francophone: carrefour culturel et linguistique*, Bern, Peter Lang, 11-40.
- HIRST, Daniel (2013), Anonymising long sounds for prosodic research, dans Brigitte BIGI et Daniel HIRST (dir.), *Tools and Resources for the Analysis of Speech Prosody*, Aix-en-Provence, Laboratoire Parole et Langage, 36-37.
- MARTINEAU, France et Marie-Claude SÉGUIN (2016), Le Corpus FRAN: réseaux et maillages en Amérique française, *Corpus*, 15, 55-87.
- MICHAUD, Josiane et Wim REMYSEN (2025), L'analyse géophonologique du français québécois et la diffusion de quelques traits de prononciation dans le Centre du Québec, dans Luc BARONIAN et Sandrine TAILLEUR (dir.), *Les français d'ici en mouvement: hier et maintenant*, Québec, Presses de l'Université Laval, 323-346.
- REA, Beatrice (2020), « Je m'ai fait mal quand j'ai tombé »: a real- and apparent-time study of auxiliary alternation in intransitive and pronominal verbs in spoken Montréal French (1971-2016), thèse de doctorat, University of Oxford.
- SAINT-AMANT, Lamy (2022), *Initiation des changements phoniques: le cas de l'écosystème rhotique du français laurentien*, thèse de doctorat, Université de Lausanne.
- SANKOFF, David, Gillian SANKOFF, Suzanne LABERGE et Marjorie TOPHAM (1976), Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale, *Cahier de linguistique*, 6, 85-125.
- SANKOFF, Gillian (2018), Before there were corpora: the evolution of the Montreal French Project as a longitudinal study, dans Suzanne EVANS WAGNER et Isabelle BUCHSTALLER (dir.), *Panel studies of variation and change*, New York, Routledge, 21-51.
- SANKOFF, Gillian, Pierrette THIBAUT, Naomi NAGY, Hélène BLONDEAU, Marie-Odile FONOLLOSA et Lucie GAGNON (1997), Variation in the use of discourse markers in a language contact situation, *Language Variation and Change*, 9(2), 191-218.
- THIBAUT, Pierrette et Gillian SANKOFF (1993), Diverses facettes de l'insécurité linguistique: vers une analyse comparative des attitudes et du français parlé par des Franco- et des Anglo-Montréalais, *Cahiers de l'Institut de linguistique de Louvain*, 19(34), 209-2019.
- THIBAUT, Pierrette et Diane VINCENT (1990), *Un corpus de français parlé. Montréal 84: historique, méthodes et perspectives de recherche*, Québec, Département de langues et linguistique, Université Laval.
- VINCENT, Diane (2009), Corpus, banques de données, collections d'exemples: réflexions et expériences, *Cahiers de linguistique*, 33(2), 81-96.
- VINCENT, Diane, Marty LAFOREST et Guylaine MARTEL (1995), Le corpus de Montréal 1995: adaptation de la méthode d'enquête sociolinguistique pour l'analyse conversationnelle, *Dialangua*, 6, 29-45.