Cahier de linguistique

Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale

David Sankoff, Gillian Sankoff, Suzanne Laberge et Marjorie Topham

Numéro 6, 1976

La sociolinguistique au Québec

URI : https://id.erudit.org/iderudit/800043ar
DOI : https://doi.org/10.7202/800043ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université du Québec

ISSN

0315-4025 (imprimé)
1920-1346 (numérique)

Découvrir la revue

Citer cet article

Sankoff, D., Sankoff, G., Laberge, S. & Topham, M. (1976). Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale. *Cahier de linguistique*, (6), 85–125. https://doi.org/10.7202/800043ar

MÉTHODES D'ÉCHANTILLONNAGE ET UTILISATION DE L'ORDINATEUR DANS L'ÉTUDE DE LA VARIATION GRAMMATICALE[1]

1. LES BUTS D'UNE RECHERCHE SOCIOLINGUISTIQUE

À différents objectifs de recherche conviennent différentes méthodologies ; pour comprendre les raisons qui ont motivé les techniques que nous allons décrire, il est bon de savoir au départ quels sont les buts d'une recherche sociolinguistique tels, que nous les concevons, et comment ces buts diffèrent de ceux du linguiste conventionnel, d'une part, et de ceux du sociologue du langage, d'autre part.

Le linguiste vise à construire une grammaire qui va exposer le modèle de compétence linguistique d'un individu, ou d'un locuteur hypothétique d'un certain dialecte. Son but final est de représenter

---

1. Nous tenons à remercier plusieurs amis qui nous ont aidé dans différentes étapes de cette enquête. Jacques Brazeau et Marc Leduc du Centre de sondage de l'Université de Montréal qui ont servi d'aviseurs pour l'échantillonnage ; Melvin Rothman, C.R., qui nous a donné des conseils concernant le contrat visant à protéger nos informateurs ; Jean Baudot du Centre de calcul de l'Université de Montréal, qui a autorisé le prêt des poinçonneuses ; Henrietta Cedergren, la codirectrice (avec Gillian Sankoff) de l'enquête sur le français montréalais ; et Martine Landriault qui a soigneusement revisé et corrigé le manuscrit. Les fonds qui nous ont permis d'effectuer le test préliminaire en 1970 venaient du ministère de l'Education, Québec ; le Conseil des arts a subventionné l'enquête actuelle. Cet article a été publié sous le titre : "Sample survey methods and computer assisted analysis in the study of grammatical variation", dans R. Darnell, édit., *Canadian Languages in their Social Context*, Edmonton (Canada), Linguistic Research, Inc., 1973, p. 7-63.

dans leur ensemble tous les jugements de grammaticalité que son informateur (ou son intuition personnelle) lui fournit, de telle sorte que sa grammaire soit pleinement "générative" dans le sens bien connu du terme.

Le sociologue du langage, pour sa part, délaisse généralement les détails grammaticaux de la description linguistique pour s'intéresser davantage à la découverte et à la compréhension des rapports entre la langue et les diverses forces ou institutions sociales. Des sujets tels que les rapports entre une population et une langue dominante, le prestige ou le déshonneur attaché à deux codes alternatifs ; la conversation ou la perte d'une langue occasionnée par des politiques sociales ; les implications sociales de la standardisation linguistique ; les attitudes face à la langue ; voilà autant de sujets appartenant à la sociologie du langage.

D'une certaine manière, le sociolinguiste se distingue de ces deux approches ; il tente de comprendre les structures linguistiques imbriquées dans la matrice culturelle et sociale. Les détails de phonologie ou de syntaxe sont examinés à l'intérieur de leurs contextes d'usage. Ainsi le sociolinguiste diffère, d'une part, du linguiste parce qu'il vise moins à constituer une grammaire complète d'une certaine langue ou dialecte (reposant souvent, en partie du moins, sur les descriptions données par les linguistes), qu'à cerner la nature et l'étendue de la diversité grammaticale à l'intérieur d'une communauté linguistique. Il cherche à voir comment cette diversité reflète la stratification sociale, la dispersion géographique, le changement linguistique, la variation stylistique, les diverses fonctions de certaines formes linguistiques et la dynamique des interactions entre les individus.

D'autre part, le sociolinguiste diffère du sociologue en ce sens que le premier s'intéresse principalement aux détails de la structure linguistique du comportement verbal, alors que le second s'appuie

généralement sur quelques distinctions linguistiques très apparentes (ex. : entre le français et l'anglais).

Les objectifs des sociolinguistes ne peuvent être atteints en prenant pour base uniquement les techniques et les perspectives méthodologiques tirées soit de la linguistique, soit de la sociologie. La technique de la linguistique traditionnelle utilise un seul informateur et se fonde sur le langage de ce dernier, sur ses intuitions se rapportant à la grammaticalité, ou même sur les deux à la fois. Cette technique est inadéquate pour deux raisons : premièrement, un seul informateur ne peut représenter la diversité systématique qui existe à l'intérieur d'une communauté linguistique ; cependant, nous devons tenir compte de cette variabilité car elle possède une structure et un sens pour les membres de cette communauté. Deuxièmement, le langage (généralement soutenu) et les jugements de grammaticalité obtenus sont tous deux sujets à une autocorrection consciente ou inconsciente qui tend à éliminer l'emploi des formes familières et non standard. À l'autre extrême, les techniques sociologiques généralement utilisées conviennent encore moins pour une recherche sociolinguistique. Des cas où l'individu choisi doit répondre à un questionnaire, ou doit être interviewé sur son langage et ses opinions, en sont des exemples frappants. Analyser la structure d'un certain aspect de la compétence linguistique à travers des interviews exige de l'interviewer un plus grand soin et une plus grande expérience. Comme l'information obtenue par les linguistes, les réponses aux enquêtes traitant de : qui parle telle variété non standard X ? quand parle-t-on telle variété ? et qu'est-ce que l'on en pense ? sont sujettes à l'influence de normes conscientes et inconscientes que tout locuteur possède à propos de la langue.

Pour résoudre ces problèmes, les sociolinguistes ont besoin à la fois de données linguistiques précises utilisées par les linguistes et de techniques d'échantillonnage employées par les sociologues.

En combinant ces deux aspects, ils veulent s'assurer que le langage de tous les segments d'une communauté soit adéquatement représenté. Pour éluder les difficultés posées par les normes linguistiques, les sociolinguistes recherchent le langage exprimé spontanément (plutôt que suscité par le linguiste), et visent ainsi à saisir les fonctions propres à la communication des formes linguistiques produites par les locuteurs.

Bien que les sociolinguistes aient aussi des buts comparatifs plus vastes (ex. : la distribution dans différentes langues de formes linguistiques ou de règles grammaticales particulières en rapport avec des situations de communication), notre étude du français parlé à Montréal s'insère dans le champ des analyses de la sociolinguistique descriptive portant sur le comportement linguistique de communautés définies, comme cela a été mentionné plus haut. Notre but précis était d'obtenir des données sur la nature, l'étendue et la fonction de la diversité linguistique à l'intérieur du français parlé par les Montréalais, afin d'éclaircir la situation d'une population souffrant d'une forte aliénation linguistique. L'opinion populaire distingue deux variétés de français : un "bon" français, se rapprochant d'un français standard international et un "mauvais" français associé à "non-instruit" ou à "classe ouvrière" ou même associé à français "canadien" ou "québécois". C'est sur ce dernier point que l'opinion s'appuie, même si elle reconnaît d'autres distinctions (comme celle du langage de la campagne et de la ville). Cette aliénation est constamment renforcée par les annonces publicitaires qui affirment "bien parler, c'est se respecter", et qui exhortent les gens à "bien parler pour être bien compris". Notre objectif était de démontrer le caractère marginal du "français international" sur la scène québécoise et de rendre manifeste qu'il n'est pas nécessaire de "bien parler" pour être "bien compris" ! Nous voulions contribuer à une meilleure compréhension du français parlé au Québec en considérant ses aspects

propres non comme des erreurs ou aberrations ou encore en termes de mélange non structuré d'anomalies grammaticales, mais en tant qu'éléments d'un système cohérent partagé par tous les membres de la communauté. Nous espérions montrer que la diversité linguistique existante fait partie d'une structure sociolinguistique complexe mais systématique.

L'utilisation de techniques d'échantillonnage minutieuses, la difficulté d'obtenir de bons enregistrements et les efforts considérables pour manipuler d'une façon très méthodique un vaste corpus de données dont il est question dans le présent rapport, ne constituaient pas un exercice inutile. Ces éléments s'avéraient nécessaires à une explication exhaustive et de portée sociale, d'un système sociolinguistique urbain complexe. La communauté linguistique que nous avons étudiée est située dans la ville de Montréal qui est à la fois le lieu d'une concentration maximale de francophones et le centre de communication le plus important pour la langue française en Amérique du Nord. Notre but spécifique était de décrire le système sociolinguistique des francophones qui sont montréalais d'origine.

Les méthodes exposées dans ce rapport poursuivent le développement de techniques utilisées dans des études sociolinguistiques antérieures effectuées dans des communautés linguistiques urbaines. Ces techniques ont été initialement employées par Labov (1966) et discutées ultérieurement par plusieurs auteurs ; Shuy *et al.*(1968) en fournissent l'exposé le plus détaillé. Le présent article rapporte les étapes de notre travail avec l'espoir que la connaissance des problèmes que nous avons rencontrés et les solutions que nous avons mises à l'essai pourront être utiles lors d'études futures.

2. PROCÉDÉS D'ÉCHANTILLONNAGE

Pour en arriver à une description aussi complète que possible de la structure du français tel qu'il est parlé par les Montréalais, il nous a semblé essentiel de cerner trois dimensions ou sources de variation :

(1) la variation conditionnée par l'environnement linguistique immédiat, comme dans le cas d'une règle phonologique d'omission qui se réalise différemment dans un environnement vocalique et dans un environnement consonantique ;

(2) les variations intra-individuelles, dues à l'autocorrection ou à l'influence plus indirecte de diverses situations et facteurs sociaux comme les formalités d'un événement, la volonté particulière du locuteur de se mettre dans une certaine perspective vis-à-vis de son interlocuteur, etc.

(3) la variation interindividuelle, attribuable à des facteurs tels que l'éducation, l'influence du groupe de pairs, la présence ambiante d'un dialecte régional, etc.

Il était clair qu'il serait extrêmement difficile, sinon impossible, d'examiner tous ces aspects d'un seul coup. D'une part, pour découvrir l'étendue et la structure de la variation interindividuelle, nous avions besoin d'un échantillon d'individus couvrant l'échelle des caractéristiques sociales selon les dimensions qui, possiblement, auraient une influence directe ou indirecte sur leur langage. Il était important d'obtenir un échantillon pris au hasard afin d'éliminer tout biais systématique dans la sélection des sujets. Nous avions décidé que les sujets appartenant à l'échantillon devaient être enregistrés dans des circonstances semblables, contrôlant ainsi, du moins dans une certaine mesure, l'influence de la situation. Nous admettions cependant qu'il serait virtuellement impossible de saisir toute l'étendue des styles ou variétés de cet ensemble de personnes, notamment parce que les enregistrements devaient déjà être assez longs pour assurer un effectif suffisant de variables linguistiques dans un nombre convenable de contextes linguistiques. Nous avons choisi des sujets interviewés à l'intérieur d'une seule situation qui permettait un cadre passablement informel, celui du foyer (cf. § 4.1). D'autre part, la tâche de découvrir l'étendue des répertoires individuels devait être

réalisée par la méthode d'étude de cas type, c'est-à-dire en entrant en contact avec des individus ou groupes d'individus qui accepteraient d'être enregistrés dans des situations de communication linguistique variées et échelonnées sur une période de plusieurs semaines ou de plusieurs mois. Nous espérions que la familiarité établie avec le chercheur au cours de la période d'étude aurait tendance à amener la disparition de l'autocorrection causée par la présence de ce dernier ; ceci semble d'ailleurs s'être vérifié au cours d'une étude de cas actuellement terminée. Dans cet article nous ne développerons pas davantage cette partie de l'étude qui n'a été qu'amorcée. Ce rapport ne décrit donc que les techniques élaborées pour l'étude des dimensions (1) et (3) mentionnées aux pages précédentes.

L'objectif visé par les procédés d'échantillonnage était de représenter l'étendue de la variation linguistique interindividuelle qui existe dans le français des Montréalais d'origine, et ce, avec un nombre minimum d'informateurs. Nous voulions enregistrer chaque individu pendant environ 45 minutes à 1 heure, et restreindre de ce fait le corpus à un maximum approximatif de 100 heures de conversation ; nous pensions que ce serait le maximum que nous pouvions espérer traiter de façon adéquate. Nous ne pouvions raisonnablement enregistrer plus d'une centaine de répondants. Étant donné la régularité des données linguistiques en général (cf. Labov, 1969), nous avons estimé qu'un échantillon de cette taille était suffisamment large pour représenter adéquatement la variation existant dans la communauté et pour examiner si celle-ci était ordonnée de façon stratifiée. Nous nous sommes alors trouvés devant un autre problème, celui de garantir la représentation de toute la diversité possible des comportements linguistiques tout en s'assurant que le choix final des répondants soit fait au hasard.

Nous avons déterminé que l'échantillon se composerait de tous les individus de 15 ans et plus qui étaient à la fois francophones et Montréalais d'origine. Ce dernier critère définissait celui qui était

né à Montréal ou était arrivé à Montréal avant le début de l'école primaire, c'est-à-dire avant l'âge de six ans. Bien qu'aucune statistique de cette population spécifique ne soit disponible, nous avons trouvé que les personnes s'identifiant ethniquement comme *Français* constituaient 64 % de la population totale de Montréal selon le recensement de 1961. Ceux-ci incluent naturellement un très grand nombre de personnes qui ne sont pas nées à Montréal, entre autres ceux venus de la campagne. L'échantillonnage se faisait à partir d'un choix d'adresses prises au hasard dans un annuaire. Nous ne pouvions fixer, à priori, de limites au nombre d'adresses qu'il nous fallait éventuellement sélectionner pour assurer la quantité suffisante d'adresses remplissant les critères d'échantillonnage, pour les raisons suivantes :

(1) Il n'y avait aucune façon d'identifier les adresses convenant à notre critère de "Montréalais d'origine" (même la connaissance du nom du résident ne pouvait apporter de précision en ce sens) ;

(2) nous ne pouvions faire que des suppositions quant au pourcentage de Montréalais d'origine par rapport à la population totale ou à la population francophone. Il a donc été décidé de prendre le nombre exact d'adresses requis pour l'échantillon. Pour les cas d'adresses non satisfaisantes, les interviewers ont eu comme instructions de poursuivre leur recherche en suivant l'ordre croissant des numéros de la rue (sans toutefois la traverser), et ce jusqu'à ce qu'ils trouvent une maison et un sujet correspondant aux critères de l'échantillonnage.

Pour épargner aux interviewers de vaines recherches de répondants, nous avons cherché des adresses uniquement dans les secteurs de recensement qui possédaient une concentration passablement élevée de personnes s'identifiant comme francophones ; on opta pour un pourcentage minimum de 64 %. Le recensement de 1961 indiquait que sur l'ensemble des 326 secteurs de l'île de Montréal, 193 logeaient au moins 64 % de

francophones. Pour différentes raisons cinq d'entre eux furent éliminés (ex. : certains contenaient uniquement des édifices publics) ; ceci ramena à 188 le nombre de secteurs de recensement à partir desquels nous projetions de tirer notre échantillon. La population totale de ces secteurs de recensement en 1961 était de 995 905 et la population totale de francophones de 829 189.

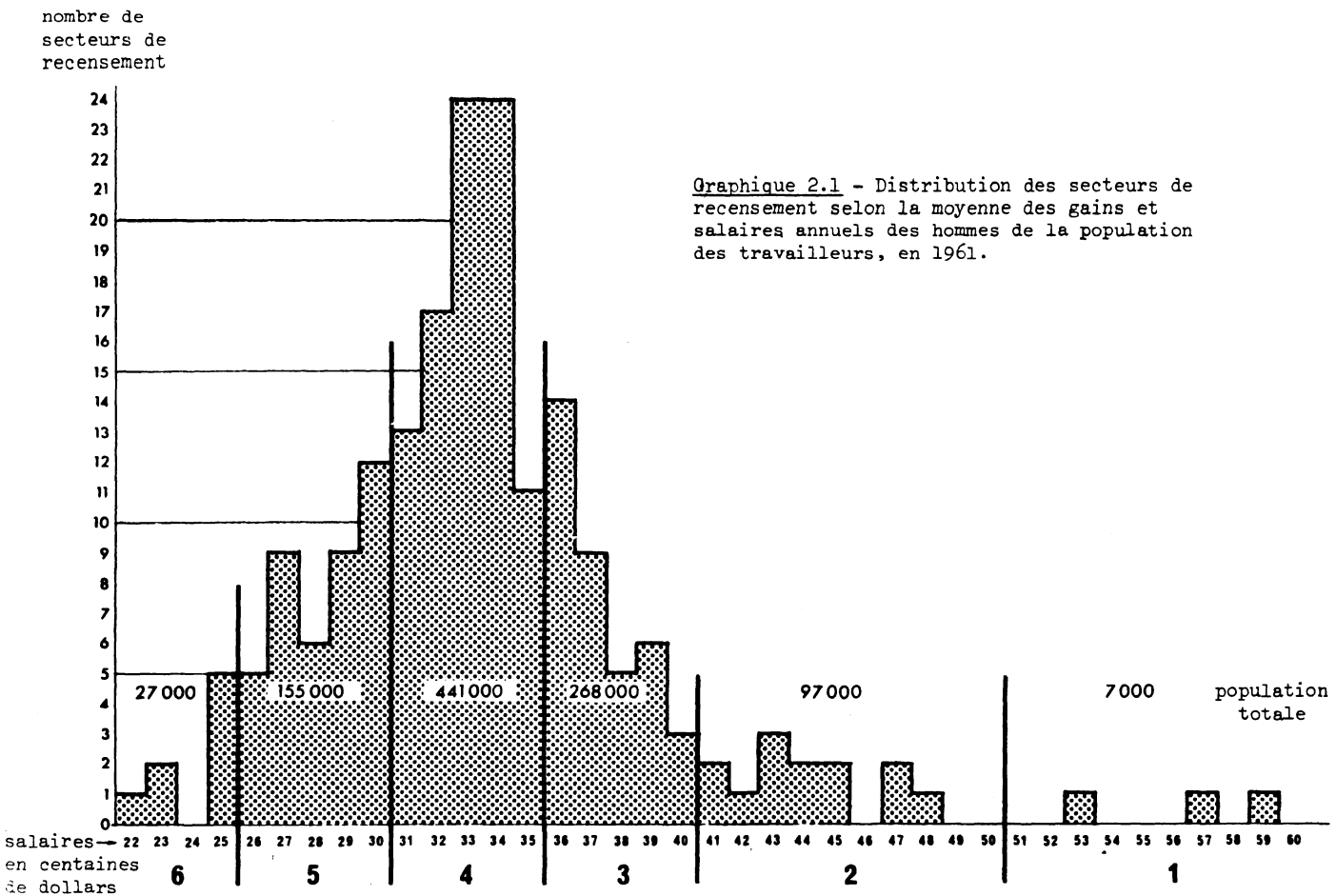
L'examen des données du recensement des 188 secteurs nous amena à penser qu'il serait possible d'obtenir une bonne distribution de répondants selon leur instruction, leur occupation et leur salaire simplement en stratifiant l'échantillon en termes de salaire moyen pour chacun des secteurs de recensement. Après avoir établi la liste des 188 secteurs de recensement selon la moyenne des gains et salaires annuels de la population mâle faisant partie de la main d'oeuvre, nous avons divisé la liste en six catégories : $5 100 et plus ; $4 100 — $5 099 ; $3 600 — $4 099 ; $3 100 — $3 599 ; $2 600 — $3 099 ; et $2 200 — $2 599. Selon le bulletin du recensement, "les secteurs de recensement sont conçus de façon qu'ils aient à peu près la même étendue et la même population et que chacun soit suffisamment homogène quant à la situation économique et les conditions de vie" (Recensement du Canada 1961, Bulletin CT-4 , p. 3). La distribution de la population dans les 188 secteurs de recensement, selon la moyenne des gains et salaires annuels des hommes de la population des travailleurs., est représentée par le graphique 2.1. La majorité de la population se trouve dans les catégories 3, 4 et 5, c'est-à-dire que leurs salaires annuels, en 1961, se situaient entre $2 600 et $4 099. Étant donné que nous voulions être sûrs de représenter les extrêmes de la population quant au revenu et autres caractéristiques qui lui sont reliées, nous avons résolu de choisir 20 individus dans chacune des six catégories, sans tenir compte de l'importance de leur population, ce qui donnait un total de 120 pour l'échantillon.

Afin d'assurer une dispersion aussi bien géographique qu'économique chez les répondants, et ce, au cas où il y ait, à l'intérieur de la ville, des différences linguistiques régionales ne comportant pas de relation avec la strate économique, nous avons dressé une carte qui indiquait les six catégories de l'échantillon. Un quota de 20 adresses par catégorie fut alors partagé géographiquement, lorsqu'il y avait une certaine dispersion entre les secteurs de recensement d'une catégorie ; cela est exposé au tableau I. Pour choisir les adresses à l'intérieur des quotas de 5, 10 ou 20 répondants pour chaque sous-section de l'échantillon, nous avons procédé comme suit : un programme fournissait des numéros pris au hasard, spécifiant la page, la colonne et la position dans la colonne d'adresses de l'*Annuaire des rues de Montréal*[2].

Chaque adresse ainsi obtenue était ensuite pointée sur la carte et retenue seulement si elle se trouvait à l'intérieur des 188 secteurs de recensement que nous avions délimités. Quand les quotas pour une sous-section précise étaient atteints, toute adresse ultérieure fournie pour cette sous-section était rejetée. Les catégories 2 à 5 furent remplies bien avant les catégories 1 et 6 ; puis le programme fut révisé afin qu'il ne fournisse que les pages et les numéros des colonnes qui renvoyaient aux adresses appartenant à ces deux secteurs relativement petits.

---

2. Annuaire Lovell de 1970. Pour s'assurer que les adresses plus longues n'aient pas plus de chances de "sortir" que les courtes, nous avons considéré qu'une adresse était choisie lorsque le point indiqué par le programme tombait sur la première ligne de l'adresse.



Graphique 2.1 - Distribution des secteurs de recensement selon la moyenne des gains et salaires annuels des hommes de la population des travailleurs, en 1961.

TABLEAU I

Sous-divisions géographiques de l'échantillon

Catégorie de l'échantillon	Aire géographique	Nombre de secteurs de recensement	Population	Quota de répondants
1	Outremont	3	7 166	20
2.1	Côte des Neiges ; nord	8	53 160	10
2.2	St-Léonard ; est	5	43 734	10
3.1	nord	13	102 308	5
3.2	Verdun, Ville St-Pierre	4	23 649	5
3.3	est, Pointe-aux-Trembles	6	69 761	5
3.4	Rosemont	12	72 415	5
4.1	Lachine, Verdun	16	81 974	5
4.2	St-Michel, Anjou ; Rivière des Prairies	7	87 145	5
4.3	Plateau Mont-Royal ; centre	65	272 138	10
5.1	St-Henri ; le port de	10	39 816	10
5.2	Centre [Montréal	31	115 555	10
6	Centre-sud	8	27 084	20
Total		188	995 905	120

Une fois les adresses obtenues, des quotas étaient aussi fixés à l'intérieur de chaque subdivision en fonction de l'âge et du sexe de façon à couvrir l'échelle des âges des deux sexes pour toutes les catégories de l'échantillon. Ces quotas sont présentés dans le tableau II.

TABLEAU II

Quotas établis pour le nombre de répondants selon le sexe, l'âge et la sous-section de l'échantillon

Catégorie de l'échantillon	Âge								Total	
	15-19		20-34		35-54		55-			
	H	F	H	F	H	F	H	F	H	F
1	3	2	2	3	2	3	3	2	10	10
2.1	1	2	1	1	2	1	1	1	5	5
2.2	1	1	2	1	1	1	1	2	5	5
3.1	1	1	1	-	-	1	-	1	2	3
3.2	-	1	1	-	1	1	1	-	3	2
3.3	-	1	1	1	-	1	1	-	2	3
3.4	1	-	1	-	-	1	1	1	3	2
4.1	1	-	-	1	1	-	1	1	3	2
4.2	-	1	1	1	1	-	-	1	2	3
4.3	2	1	1	1	1	2	1	1	5	5
5.1	1	2	1	1	2	1	1	1	5	5
5.2	1	1	1	2	1	1	2	1	5	5
6	3	2	2	3	3	2	2	3	10	10
Total	15	15	15	15	15	15	15	15	60	60

La tâche de remplir les quotas selon l'âge et le sexe dans chaque sous-section était entièrement laissée à l'interviewer qui devait s'assurer, à la porte de la maison, s'il y avait une personne qui accepterait de passer l'entrevue, et si cette personne correspondait aux catégories d'âge et de sexe qui n'étaient pas déjà complètes pour

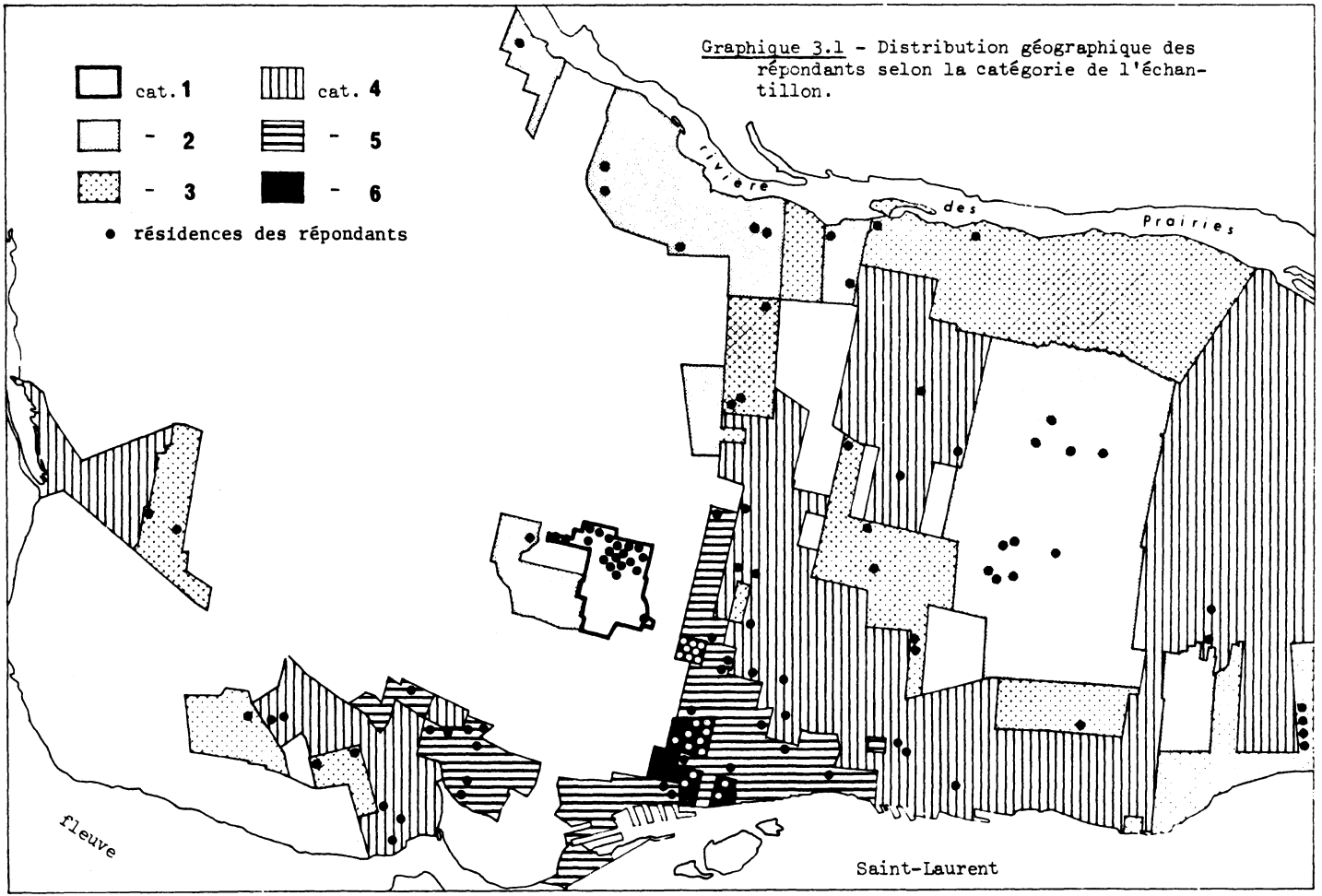
cette sous-section. Il est important de souligner que les quotas de départ furent établis pour les besoins de l'échantillonnage et non pour des fins d'analyse. En ce sens, les 120 répondants localisés par les procédés d'échantillonnage devaient être étudiés selon leurs caractéristiques propres, découvertes lors de l'entrevue, plutôt que par la catégorie dans laquelle ils se trouvaient catalogués, selon leur lieu de résidence. La section 3 décrit en détail les caractéristiques sociales réelles des gens choisis comme répondants.

3. ANALYSE DE L'ÉCHANTILLON

Dans la section 2 (cf. tableau II), les procédés d'échantillonnage fixaient les quotas en termes d'âge et de sexe pour chacune des 13 unités de l'échantillon afin d'assurer que les 30 individus de chacun des quatre groupes d'âges, de même que les 60 hommes et 60 femmes, soient distribués équitablement dans l'échelle économique et géographique. Dans cette section, nous allons examiner en détail les caractéristiques de l'échantillon que nous avons obtenu en suivant les procédés et quotas formulés dans la section 2 ; on pourra alors vérifier si notre échantillon représente réellement l'étendue de la variation selon les aspects désirés. Nous allons d'abord regarder les caractéristiques explicitement impliquées dans notre technique d'échantillonnage, ensuite nous considérerons les autres traits que nous souhaitions être en variation à l'intérieur de notre échantillon.

3.1 *Distribution géographique*

Le graphique 3.1 montre la dispersion géographique des 120 personnes interviewées en indiquant la distribution de la population échantillon à travers les régions à prédominance francophone (64 % ou plus) de Montréal. Les secteurs ayant moins de 64 % de francophones constituent les espaces vides de la carte ; ils incluent une grande partie de l'ouest de la ville aussi bien que plusieurs petits îlots encerclés par des secteurs à majorité francophone. Les deux plus



Graphique 3.1 - Distribution géographique des répondants selon la catégorie de l'échantillon.

fortes concentrations de points sur notre carte se trouvent dans deux régions dont le taux de population est relativement faible : la zone de la classe supérieure d'Outremont (1), au centre de l'île dans la partie nord du Mont-Royal, et le secteur centre-sud (6) de la ville où vivent les Montréalais les plus pauvres. Les deux niveaux économiques suivants qui composent, avec ce groupe économiquement le plus bas, la moitié inférieure de l'univers de notre échantillon se regroupent aussi dans le centre de la ville. Les individus choisis dans la catégorie 5 sont parsemés à travers ce secteur, de même que ceux de la catégorie 4.

Pour les catégories 2 et 3, nous remarquons que la catégorie 3, représentant la classe moyenne, constitue le groupe le plus éparpillé géographiquement, c'est pourquoi nous l'avons divisé en quatre régions représentatives (voir section 2). Les 20 individus de la catégorie 3 sont assez bien répartis à travers ces quatres régions. Les quatre individus qui ne sont pas conformément localisés sur la carte du graphique 3.1 sont des résidents de l'extrême est de Pointe-aux-Trembles. La catégorie 2 est divisée en deux régions : le centre nord (incluant à la fois le secteur de Côte-des-Neiges voisin de la catégorie 1, et la rive nord de l'île de Montréal) ; et la banlieue est de Saint-Léonard et ses environs. À nouveau, les individus choisis étaient assez bien dispersés.

Le tableau III résume en termes de chiffres la distribution géographique des répondants et indique que les 120 informateurs furent tirés d'un total de 62 des 188 secteurs de recensement de l'échantillon.

3.2 *Salaire*

La stratification de notre échantillon en 6 catégories visait à fournir une dispersion à la fois géographique et économique parmi les répondants. Toutefois, comme nous ne voulions pas questionner de façon explicite les personnes sur leur salaire, nous avons utilisé plusieurs autres mesures pour nous assurer que les sujets couvraient l'échelle

TABLEAU III

Distribution des répondants par secteur de recensement

Catégorie de l'échantillon	Nombre de secteurs de recensement (s.r.)	Nombre de secteurs de recensement représentés par des répondants	Nombre de répondants	Nombre de secteurs de recensement des 20 répondants/total de s.r.
1	3	3	20	3/3
2.1	8	6	10	8/13
2.2	5	2	10	
3.1	13	5	5	13/35
3.2	4	3	5	
3.3	6	2	5	
3.4	12	3	5	
4.1	16	3	5	15/88
4.2	7	3	5	
4.3	65	9	10	
5.1	10	7	10	17/41
5.2	31	10	10	
6	8	6	20	6/8
Total	188	62	120	

des revenus. Bien que les secteurs de recensement paraissaient relativement homogènes, nous n'avions aucun moyen de garantir que les 20 individus sélectionnés n'étaient pas tous des "anormaux", en ce sens qu'il fallait que leurs caractéristiques économiques personnelles correspondent à la moyenne de leur secteur respectif. Pour le vérifier nous avons fait un tableau de la distribution de nos informateurs selon la catégorie occupationnelle du chef de ménage[3]. Nous demandions à

3. Selon le bulletin du recensement, "Un ménage : c'est la personne ou le groupe de personnes occupant un logement." (Recensement du Canada 1961, Bulletin CT-4, p. 3).

tous les répondants l'occupation précise de ce dernier. Les résultats furent compilés selon les catégories suivantes : professionnel — gérant, col blanc, col bleu, ouvrier manuel, chômeur, ménagère, étudiant. Les 18 chefs de maison retraités furent classés d'après leur dernier emploi.

Le tableau IV indique que les gérants et professionnels se concentrent dans les catégories 1 et 2, que les cols blancs se retrouvent principalement dans les catégories 3 et 4, et que les gens sans emploi se regroupent dans les catégories 5 et 6. Nous avons rencontré des cols bleus dans les catégories 2 à 6. Il en est de même pour les ouvriers manuels qui, toutefois, sont plus nombreux dans les catégories 4, 5 et 6. Cette répartition tend à confirmer l'hypothèse selon laquelle les individus choisis correspondaient à la moyenne du secteur de recensement dans lequel ils habitaient. Il apparaît que les chefs de ménage sont en nombre à peu près égal dans chacun des quatre groupes occupationnels, et que dans 14 cas le chef de ménage est sans emploi.

L'instruction est une autre mesure témoignant de l'étendue socio-économique de l'échantillon et démontrant la corrélation entre les caractéristiques sociales des individus interviewés et la catégorie de l'échantillon auquel ils appartiennent. La section inférieure du tableau V révèle que les informateurs possédant une instruction post-secondaire se rassemblent dans les catégories 1 et 2, que les gens ayant fait leur cours secondaire prédominent dans les catégories 3 et 4, et que les personnes n'ayant fréquenté que l'école primaire sont fortement concentrées dans les catégories 5 et 6. Les sujets de 15 à 19 ans furent exclus du tableau parce que la majorité est encore aux études et que leur nombre d'années d'études terminées ne correspond pas à leur place dans l'échelle sociale. Il ressort du tableau V que les plus jeunes ont complété une scolarité plus avancée. Cette observation vaut pour toutes les catégories de l'échantillon.

TABLEAU IV

Distribution des répondants selon l'occupation du chef de ménage et la catégorie de l'échantillon

Occupation	Catégorie de l'échantillon 1	2	3	4	5	6	Total
Professionnel-gérant	[17]	[6]			1		24
Col blanc	2	3	[9]	[9]	2	1	26
Col bleu		4	[6]	3	3	[6]	22
Ouvrier		4	4	[7]	[7]	[7]	29
Chômeur	1	2	1		6	4	14
Ménagère		1		1	1	1	4
Étudiant						1	1
Total	20	20	20	20	20	20	120

[ ] indique les maxima pour chaque catégorie et chaque profession.

TABLEAU V

Niveau d'instruction des répondants de 20 ans et plus selon l'âge et la catégorie de l'échantillon

Âge	Niveau d'instruction	Catégorie de l'échantillon			Total
		1 et 2	3 et 4	5 et 6	
20-34	Primaire	-	-	[3]	3
	Secondaire	3	[7]	[5]	15
	Post-secondaire	[7]	3	4	14
35-54	Primaire	1	4	[8]	13
	Secondaire	[5]	[6]	1	12
	Post-secondaire	[4]	-	-	4
55 +	Primaire	[4]	[5]	[9]	18
	Secondaire	[3]	[3]	1	7
	Post-secondaire	[3]	2	-	5
Tous les âges	Primaire	5	9	[20]	
	Secondaire	11	[16]	7	91
	Post-secondaire	[14]	5	4	

Primaire = école primaire (terminée ou non)

Secondaire = études secondaires (terminées ou non)

Post-secondaire inclut des cours techniques et spécialisés.

TABLEAU VI

Distribution des répondants selon le type de résidence et le niveau de l'échantillon

Type de résidence	Catégorie de l'échantillon						Total
	1	2	3	4	5	6	
Maison détachée	[10]	4	3				17
Maison semi-détachée	[5]	1		1	1		8
Duplex	3	[5]	[5]	4			17
Appartement	2	4	[9]		2	5	22
Logement		6	3	15	[17]	15	56
Total	20	20	20	20	20	20	120

Le type d'habitation fournit une indication supplémentaire sur le niveau des revenus des répondants. Le tableau VI fait voir que les individus de la catégorie 1 occupent surtout des maisons détachées ou semi-détachées. Ceux des catégories 2 et 3 occupent des types d'habitation variés, et la majorité de la population des catégories 4, 5 et 6 vit dans des logements.

TABLEAU VII

Distribution des répondants selon qu'ils sont ou non propriétaires de leur résidence et selon la catégorie de l'échantillon

	Catégorie de l'échantillon						Total
	1	2	3	4	5	6	
Locataires	4	11	13	17	20	16	81
Propriétaires	16	9	7	3	-	4	39
Total	20	20	20	20	20	20	120

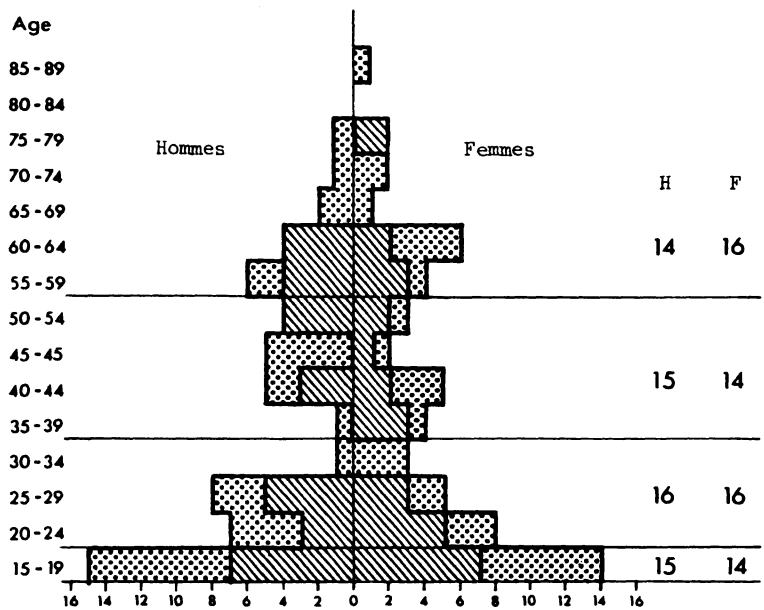
Comme on peut le constater dans le tableau VII, la proportion de gens qui possèdent leur propre maison diminue régulièrement de la catégorie 1 à la catégorie 5 ; il y a toutefois quelques propriétaires dans la catégorie 6 (deux de ces derniers ont un revenu qui provient de la location de chambres).

3.3 *Âge et sexe*

La technique d'échantillonnage non seulement veillait à la stratification en termes de niveaux (ou catégories) de revenus, mais exigeait aussi que l'échantillon contienne un nombre égal d'individus des deux sexes, divisés en quatre groupes d'âges. Examinons maintenant la distribution réelle des répondants par intervalles de cinq ans selon le sexe, l'âge et la catégorie de l'échantillon ; nous nous référons au graphique 3.2. Dans le groupe des 15 à 69 ans, la plupart des intervalles de cinq ans contiennent au moins un représentant de chaque sexe et de chacune des deux moitiés de l'échantillon (les catégories 1, 2 et 3 par rapport aux catégories 4, 5 et 6). Diverses erreurs (ex. : un sujet admet, dans le déroulement de l'entrevue, qu'il a effectivement 34 et non 35 ans) ont fait en sorte que certains totaux des huit sous-sections âge-sexe de l'échantillon présentent un quota de 14 ou de 16 plutôt que de 15 comme nous l'avions établi au départ. Nous avons néanmoins réussi à obtenir un total de 60 hommes et 60 femmes ; les quatre groupes d'âges étant formés de 30, 29, 32 et 29 membres.

3.4 *Lieu d'origine*

Un critère très important de notre échantillon exigeait que l'éventuel candidat devait être originaire de Montréal ou devait y avoir vécu au moins depuis l'âge de 6 ans. Des 120 répondants de l'échantillon, 107 sont nés dans la ville. Pour vérifier l'influence possible de dialectes régionaux sur le langage des informateurs, nous avons posé des questions sur le lieu d'origine des parents. Près de la moitié de ceux-ci (117 sur 240) étaient aussi des Montréalais.



Graphique 3.2 - Distribution des répondants selon le sexe, l'âge et la catégorie de l'échantillon.

Comme on pouvait s'y attendre, les plus jeunes étaient davantage susceptibles d'avoir des parents nés à Montréal. Cette tendance ne fut pas très forte : 34 des 117 parents nés à Montréal ont des fils ou filles entre 15 et 19 ans ; 31 parents en ont entre 20 et 34 ; 28 autres en ont entre 35 et 54 ans ; et finalement 24 en ont de 55 ans et plus. En tout, 75 des 120 répondants possèdent au moins un parent montréalais.

Les renseignements concernant les parents apportent un peu de lumière sur le modèle d'immigration à Montréal pour les 3 ou 4 générations précédentes mais ils n'aident pas à clarifier le problème de la dimension de l'univers à partir duquel notre échantillon fut tiré. Comme nous l'avons expliqué dans la section 2, il n'existait aucune donnée numérique sur la proportion de Montréalais d'origine ayant le français comme langue maternelle. Toutefois nous pouvions calculer, à partir des données du recensement de 1961, que la proportion de gens s'identifiant comme "Français" à l'intérieur des secteurs de recensement d'où nous avons extrait notre échantillon était de 83,2 %. Les données recueillies par les interviewers démontrent que la proportion de Français rencontrés est seulement un peu plus faible que cela. Les interviewers devaient compléter des formulaires indiquant à combien d'autres portes ils avaient frappé et précisant pourquoi ces adresses ne fournissaient pas de répondants. Ils devaient continuer à le faire jusqu'à ce qu'ils trouvent le répondant requis. Ces formulaires furent remplis pour 88 des 120 entrevues ; un interviewer récalcitrant fut responsable de la majorité des formulaires non rapportés. Les résultats sont fournis au tableau VII.

TABLEAU VII

Renseignements sur les personnes mises en contact avec l'interviewer mais n'ayant pas passé l'interview

Type de contact	Pas de renseignements	Quelques renseignements
Interviews terminées, formulaires remplis		88
Pas de réponse (y compris magasins, commerces, etc.)	388	
Refus	91	
Non-francophones		107
Francophones, nés à l'extérieur de Montréal ou arrivés après l'âge de 6 ans.		102
Montréalais francophones dont le quota âge-sexe a déjà été rempli*.		146

* Les interviewers devaient suivre des quotas en termes d'âge et de sexe pour chaque région de l'échantillon (voir section 2) ; ils devaient ainsi sonner à la porte de plusieurs répondants qui auraient éventuellement accepté avant de trouver une maison où il y avait quelqu'un ayant l'âge et le sexe demandés pour remplir le quota pour chaque district.

Comme nous n'avons aucun renseignement sur les caractéristiques de ceux qui n'étaient pas à la maison ou qui ont refusé de passer l'entrevue, les pourcentages furent calculés sur la base des 443 personnes entrées en communication avec l'interviewer. Les non-francophones regroupent 107 des 443 répondants contre un pourcentage de 75,8 de francophones. Vu que ce chiffre est près de la proportion réelle de francophones dans la population échantillon (83,2 %), il est possible que la proportion réelle de francophones arrivés à la ville après l'âge

de six ans, c'est-à-dire 102 des 336 francophones, reflète aussi la proportion approximative de ces personnes dans la population totale. Nous supposons alors qu'environ 70 % de la population francophone de Montréal est née à la ville ou y est arrivée à l'âge de 6 ans. Si cette approximation est acceptable[4], cela signifie que la population totale de la région de l'échantillon possédant les caractéristiques de "locuteur autochtone", telles que nous les avons définies, serait de l'ordre de 400 000 à 450 000.

3.5 *Coût des entrevues (temps des interviewers)*

Le coût des 120 entrevues n'est pas exorbitant, en dépit des chiffres du tableau VII qui montrent qu'un grand nombre de contacts n'ont pas entraîné des entrevues complètes (834 contacts infructueux pour 88 entrevues réalisées). Les interviewers étaient payés sur une base hebdomadaire ($100) plutôt qu'au nombre d'entrevues complétées. Un autre montant de $100 fut accordé pour une période d'une semaine d'entraînement. L'objectif de chaque interviewer était d'effectuer deux entrevues par jour, c'est-à-dire dix entrevues par semaine de cinq jours. Le plan initial était d'avoir deux interviewers à temps plein, qui pourraient effectuer 40 entrevues chacun, et deux interviewers à temps partiel pour en effectuer chacun 20. En réalité, cinq interviewers furent employés et chacun d'eux a effectué le nombre d'entrevues suivant : 37, 34, 26, 20, 3. Ils ont travaillé environ de 2 à 5 semaines.

On s'est aperçu qu'une semaine de 10 entreveues était trop chargée, surtout à la dernière semaine de l'enquête, parce qu'il fallait trouver les données manquantes du tableau II. Nous n'avons pas insisté pour qu'ils parviennent à l'objectif hebdomadaire, mais plutôt pour qu'ils maintiennent la qualité des entrevues. L'interviewer qui a

---

4. Cette estimation n'est évidemment pas très exacte étant donné que les interviewers continuaient à sonner aux portes jusqu'à ce qu'ils rencontrent un répondant approprié.

effectué 37 entrevues a pris près de cinq semaines pour les réaliser, les autres fonctionnant à une vitesse relativement semblable. Le salaire total payé aux interviewers, à la fois pour la période d'entraînement et la période d'entrevues proprement dite, a été de $ 2 000, soit un peu plus de seize dollars par entrevue.

3.6 *Résumé*

Dans cette section, nous avons tenté de démontrer non seulement que les répondants se distribuaient selon les critères (âge, sexe, région géographique) établis par notre technique d'échantillonnage, mais aussi qu'ils s'échelonnaient d'après des caractéristiques non fixées par l'échantillonnage telles que l'éducation, l'occupation et le type de résidence. La corrélation entre des traits typiques d'un revenu élevé (ex. : vivre dans une maison détachée, en être propriétaire, avoir un haut degré d'instruction et occuper un emploi de gérant ou de professionnel) et le fait d'habiter un secteur de recensement à revenu élevé, comme la corrélation entre des traits d'un faible revenu (ex. : des études primaires ou moins, être sans emploi ou ouvrier manuel, vivre dans un logement loué) et le fait d'habiter un secteur de recensement à faibles revenus tend à appuyer la méthode d'échantillonnage utilisée. Nous pouvons être sûrs que les répondants furent sélectionnés par une technique qui garantissait l'élément "hasard" bien que tous n'avaient pas une chance égale d'être choisis (voir section 2) — et que l'échantillon comprenait des répondants d'une gamme très étendue.

4 DONNÉES RECUEILLIES

4.1 *Réalisation et contenu des entrevues*

Tous les interviewers parlaient le français québécois comme langue maternelle ; trois d'entre eux étaient natifs de Montréal, les deux autres y habitaient depuis assez longtemps pour s'être familia-

risés avec le langage de Montréal. À l'exception d'un, ils étaient tous de sexe féminin et se situaient entre 22 et 24 ans. Tous avaient étudié la sociolinguistique et trois d'entre eux avaient participé à une étude pilote l'été précédent. Pour ce projet-ci, ils ont été soumis à une période d'entraînement d'une semaine pendant laquelle ils ont appris à utiliser le matériel technique et les règles élémentaires de l'interview.

Le but des interviews était d'obtenir des enregistrements de bonne qualité dans une conversation informelle. Les interviewers tentaient toujours de créer une atmosphère de détente entre eux et l'interviewé surtout lorsque ce dernier était quelque peu intimidé par le magnétophone. Ils encourageaient les autres membres de la famille à demeurer dans la pièce et à participer à la discussion ; le magnétophone fonctionnait pendant toute la période où l'interviewer se trouvait dans la maison. L'entrevue commençait par une série de questions sur les coordonnées de l'informateur : âge, occupation, instruction, lieu de naissance des parents, etc. (cf. appendice A*). Ensuite l'interviewer essayait d'amener le répondant à parler de divers sujets, y compris le langage, reliés au thème "Vie et coutumes au Québec". Les informateurs n'étaient pas avisés de l'intérêt particulier que nous portions à leur langage. Nous pensions que cela entraînerait une certaine conscience de soi et susciterait une autocorrection. D'ailleurs, quelques études sur les opinions et attitudes exprimées dans ces entrevues sont actuellement en cours ; toutefois nous admettons que la dissimulation du but principal de l'entrevue, bien que nécessité par la méthodologie, pose, quant à l'éthique, un sérieux dilemme auquel nous n'avons pu trouver de solution adéquate. Afin de combler cette lacune nous avons fait le

---

* On peut se procurer les appendices relatifs à cet article en s'adressant à Mme Henrietta Cedergren, Département de linguistique de l'Université du Québec à Montréal.

maximum pour protéger les répondants contre toute conséquence nuisible découlant de leur coopération (décrit ci-dessous en 4.2).

Le test préliminaire nous a permis d'établir une liste de sujets intéressants susceptibles de provoquer chez les répondants des commentaires spontanés et prolongés ; les questions étaient groupées en trois sections principales d'après les thèmes suivants : (1) vie et coutumes au Québec dans le passé (ex. : la célébration des fêtes en famille ; les souvenirs des jeux d'enfance) ; (2) vie moderne à Montréal (ex. : l'évolution du rôle de l'Église) ; et (3) opinions sur la langue. L'appendice B fournit un exemplaire du questionnaire-guide. Nous demandions aux interviewers de couvrir les trois sujets généraux ; ils pouvaient néanmoins poser des questions qu'ils jugeaient à propos tout en n'étant pas forcés de poser toutes les questions de la liste. Cependant, il était important qu'ils obtiennent les renseignements personnels des informateurs car ces données allaient nous être nécessaires pour l'analyse (les trois formulaires de l'appendice A devaient être remplis quotidiennement et ceci pour chaque répondant interviewé). À la fin de l'entrevue, on demandait aux répondants de lire un petit texte intitulé "Une soirée de hockey" (joint à l'appendice C). Ce texte contenait un certain nombre de paires minimales (ex. : *patte — pâte*) servant à mesurer les effets d'un style de lecture formelle sur la phonologie des informateurs et à fournir un contexte encore plus uniforme permettant la comparaison entre individus.

Les entrevues étaient enregistrées sur des rubans BASF de 900 pieds, bobine de 5 pouces. Nous utilisions des micros lavallières *Electrovoice* et des magnétophones Uher 4000 Report L. La durée des interviews variait entre une demi-heure et une heure et demie ; au total, nous possédons environ 100 à 120 heures de conversations enregistrées. Nous avons été frappés par le fait que les répondants discutaient très ouvertement de leurs opinions et de leurs expériences. Plusieurs d'entre eux ont tenté d'impressionner l'interviewer à divers moments

et de différentes façons (y compris par l'emploi de formes linguistiques marquées "+ polies" ou "+ formelles"), mais de telles stratégies se manifestent aussi dans la plupart des genres d'interaction. Dans le but de mesurer l'influence de l'interviewer, nous comptons aussi comparer chez les informateurs "cas type" leur langage "d'entrevue" et celui employé dans des situations plus intimes.

4.2 *Transcription*

L'un des problèmes majeurs que nous avons rencontrés fut de rendre accessible et utile au maximum cette quantité considérable de données enregistrées. Il a été décidé de faire transcrire les entrevues sur des cartes perforées IBM plutôt que de les faire dactylographier. Ainsi elles devenaient disponibles pour divers types de traitements par ordinateur[5]. Cette décision entraîna d'autres problèmes dont le plus important relève du nombre limité de symboles sur le clavier de la poinçonneuse. Ce facteur nous obligea à employer une transcription selon l'orthographe du français standard au lieu d'une transcription phonétique. Les secrétaires recevaient l'instruction de respecter la syntaxe des informateurs, l'ordre des mots, etc. ; et nous leur fournissions un ensemble de conventions que nous avions établies pour la transcription, y compris les pauses et les hésitations. Le langage des interviewers fut identifié par le chiffre 1 ; celui des informateurs par le chiffre 2, et tout autre locuteur par d'autres chiffres. Nous avons utilisé des magnétophones Uher 5000 fabriqués pour la transcription et équipés de pédales permettant l'avance et le recul ; les secrétaires portaient des écouteurs qui les isolaient, autant que possible, du bruit des poinçonneuses.

---

5. Nous sommes très reconnaissants à Martha Laferrière qui nous suggéra l'idée d'utiliser des techniques automatiques pour la manipulation des données.

Étant donné que toutes les analyses linguistiques établies sur ces transcriptions devaient ultérieurement utiliser les éditions imprimées des entrevues, et non directement les cartes perforées, nos conventions de typographie visaient à faciliter la vitesse de perforation plutôt que la clarté ou la lisibilité des cartes. Ces derniers aspects sont du ressort des routines d'édition du programme. Lors de la perforation, la transcription passe d'une carte à l'autre sans se soucier de mettre des traits d'union aux mots qui sont coupés à la fin des cartes. Ne pouvant utiliser de gommes à effacer, liquide opaque, papier correcteur, ou autres procédés du genre pour la correction rapide des erreurs, la technique adoptée consistait à mettre un astérisque immédiatement avant et immédiatement après l'extrait à effacer sur une carte. La correction d'erreurs est aussi facilitée par le fait que la carte comportant une erreur peut être enlevée et remplacée par une carte correctement poinçonnée (un procédé analogue ne peut s'appliquer à la correction d'une ligne de texte dactylographié). Une autre convention stipulait que plusieurs espaces vides consécutifs ne comptaient que comme un seul espace lors de l'impression. Ceci signifiait que si, par exemple, les mots "les jours" étaient orthographiés "la jours", cette erreur pouvait être corrigée, premièrement en reproduisant de façon automatique, un duplicata jusqu'à la lettre "l" ; deuxièmement, en poinçonnant "es" après le "l" sur la nouvelle carte ; troisièmement, en laissant libre le reste des espaces de cette même carte ; quatrièmement, en reproduisant sur une seconde carte, la carte incorrecte, et ce à partir du mot "jours", laissant libre le début de la seconde carte vierge. Les deux nouvelles cartes se trouvent ainsi correctement transcrites.

Le programme pour l'impression est le premier de trois programmes que nous utilisons à grande échelle dans le traitement des données. Ce programme prend comme entrée un bloc de cartes contenant une inter-

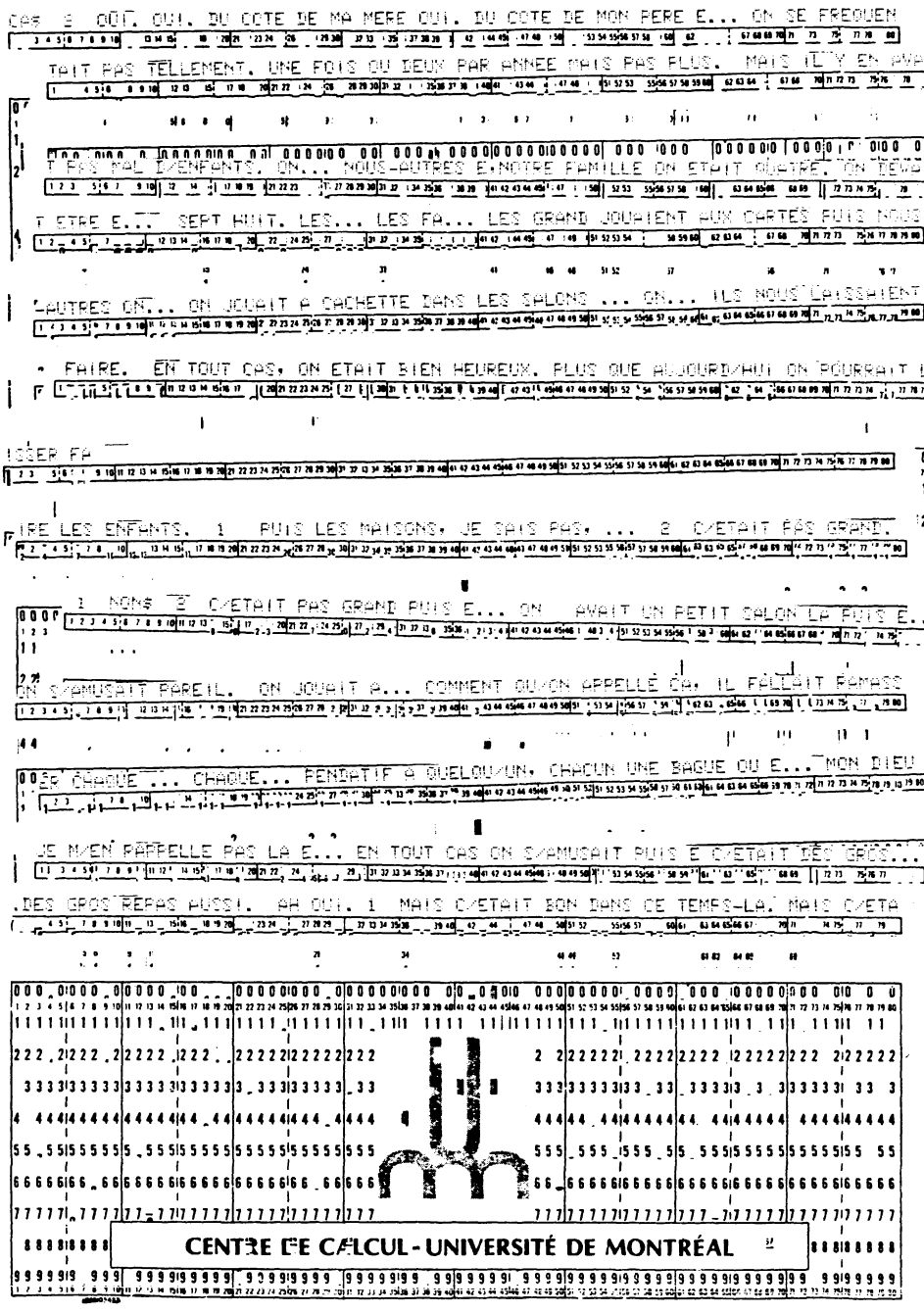


Figure 4.1 - Exemple d'une transcription sur des cartes perforées.

2 OUI. OUI. DU COTE DE MA MERE OUI. DU COTE DE MON PERE E... ON SE FREQUENTAIT PAS TELLEMENT. UNE FOIS OU DEUX PAR ANNEE MAIS PAS PLUS. MAIS IL Y EN AVAIT PAS MAL D'ENFANTS. ON... NOUS-AUTRES E,NOTRE FAMILLE ON ETAIT QUATRE. ON DEVAIT ETRE E... SEPT HUIT. LES... LES PA... LES GRAND JOUAIENT AUX CARTES PUIS NOUS-AUTRES ON... ON JOUAIT A CACHETTE DANS LES SALONS ... ON... ILS NOUS LAISSAIENT FAIRE. EN TOUT CAS, ON ETAIT BIEN HEUREUX. PLUS QUE AUJOURD'HUI ON POURRAIT LAISSER FAIRE LES ENFANTS.

1 PUIS LES MAISONS, JE SAIS PAS, ...

2 C'ETAIT PAS GRAND.

1 NON?

2 C'ETAIT PAS GRAND PUIS P... ON AVAIT UN PETIT SALON LA PUIS E... ON S'AMUSAIT PAREIL. ON JOUAIT A... COMMENT QU'ON APPELLE CA, IL FALLAIT RAMASSER CHAQUE ... CHAQUE... PENDATIF A QUELQU'UN, CHACUN UNE BAGUE OU E... MON DIEU JE M'EN RAPPELLE PAS LA P... EN TOUT CAS ON S'AMUSAIT PUIS E C'ETAIT DES GROS....DES GROS REPAS AUSSI. AH OUI.

1 MAIS C'ETAIT BON DANS CE TEMPS-LA. MAIS C'ETAIT LA PERSONNE EN TOUT CAS QUI RECEVAIT, IL ME SEMBLE... C'EST MOINS FORT AUJOURD'HUI LA JE PENSE...

2 MAIS C'EST A QUI EST-CE QUE QUI RECEVRAIT PAS JE PENSE AUJOURD'HUI. AH OUI.

1 AUJOURD'HUI EST-CE QUE VOUS VOUS ETES SEULEMENT VOTRE FAMILLE ENSEMBLE OU SI...?

2 NON. ON RECOIT ENCORE MAIS E... LA PLUS... PLUS CA VA, PLUS E... IL Y A D'ENFANTS. COMME MOI J'EN AI TROIS. MA SOEUR ELLE EN A TROIS. PUIS MON FRERE EN A UN PUIS MA... MON AUTRE SOEUR EN ATTEND UN. CA FAIT QUE LA QUAND ILS SE RENCONTRENT LA, C'EST EFFRAYANT COMME... COMMENT C'EST TANNANT. PEUT-ETRE QU'ILS VONT VENIR QU'A S'AMUSER MAIS...QUAND CA SE TIRAILLE LA, CA VIENT QUE... MAIS ENCORE JUSQU'A DATE LA E... ON EST TOUJOURS SORTI DANS LE TEMPS DES FETES. COMME AU JOUR DE L'AN LA, DEPUIS QUE JE SUIS MARIEE, JE RECOIS LES DEUX FAMILLES.

1.

2 OUI. ON EST P... ON EST A PEU PRES E... UNE VINGTAINE.

1 TOUS EN MEME TEMPS?

2 OUI. MAIS JE PENSE QUE CA VA ETRE LA DERNIERE ANNEE LA. CA FAIT TROP DE MONDE. LA... LA FAMILLE S'AGGRANDIT BIEN TROP. MAIS CHEZ... A CHAQUE ANNEE JE SAIS JAMAIS LAQUELLE FAMILLE QUE J'INVITERAI PAS. LA JE VOUDRAIS PAS E... IL Y A JAMAIS PERSONNE , POURTANT IL ... IL DEVRAIT EN AVOIR UN... QUI S'OFFRE BIEN... BIEN JE PEUX PAS SUR SON COTE A MOI PARCE QUE...

1 EST-CE QUE C'EST PARCE QUE VOUS AVEZ UNE GRANDE MAISON?

2 C'EST PARCE QUE J'AI COMMENCE DE MEME. PUIS E... JE SAIS PAS. ILS S'OFFRENT PAS D' RECEVOIR. MAIS CA ME... CA ME FAIT

Figure 4.2 - Edition imprimée qui contient la transcription des cartes (fig. 4.1)

view transcrite. Cette opération a pour but d'éditer l'entrevue en un format facilement lisible et possédant les caractéristiques suivantes :

i) des pages numérotées consécutivement ;

ii) 28 lignes à double interligne par page et 120 caractères par ligne (par opposition à 80 par ligne sur une carte perforée) ;

iii) des alinéas faits à chaque changement de locuteur ;

iv) l'élision de toutes paires d'astérisques avec l'ensemble des éléments qui s'y trouvent encadrés ;

v) la réduction à un seul espace d'espaces libres consécutifs sur une carte ;

vi) aucune coupure de mots à la fin des lignes.

Une interview moyenne demandait environ 800 cartes perforées, ce qui donnait à la sortie de l'ordinateur environ 30 pages d'édition. Des exemples de cartes perforées et d'une édition imprimée sont reproduits respectivement pages 116 et 117 (fig. 4.1 et 4.2).

Alors qu'une série de cinq ou six questions et réponses brèves peuvent s'insérer dans une ou deux cartes, la même séquence devient plus lisible dans l'édition obtenue où elle est disposée en une série de cinq ou six petits paragraphes d'une ligne. Le programme permet donc à la fois la vitesse et la facilité dans la perforation et la lisibilité de la transcription.

L'utilité de nos procédés de correction et de notre programme d'édition devient encore plus apparente dans l'étape suivante de notre traitement des données ; celle-ci est effectuée par des assistants de recherche qui ont pour tâche la correction et l'uniformisation des transcriptions. Utilisant l'édition imprimée, l'assistant de recherche écoutait à nouveau la bobine et corrigeait la transcription où cela s'avérait nécessaire. Les secrétaires reprenaient alors les cartes impliquées dans cette correction, après quoi une nouvelle édition était

imprimée. Les cartes corrigées étaient de plus conservées pour des changements ou corrections futures ; nous comptons en effet écouter les bobines, corriger et uniformiser toutes les transcriptions une fois de plus. Il a été démontré (Hays 1970) que peu importe le nombre de fois qu'un enregistrement est retranscrit, chaque transcripteur percevra certaines différences et fera quelques changements ; cependant nous pensons qu'en termes d'argent investi, un procédé de transcription à trois étapes, couvrant une période d'environ deux ans, est convenable pour obtenir un ensemble uniforme et raisonnablement correct de transcriptions.

Dès que la première transcription fut terminée, un assistant de recherche fut chargé d'effacer le nom et l'adresse des répondants sur la bobine originale, et de faire une copie de la bobine pour utilisation future lors de l'analyse. Par la suite, on a demandé une impression des cartes sans le nom du répondant. Ainsi les noms et adresses des informateurs ne font pas partie des données de base. Les renseignements personnels des répondants (âge, sexe, instruction, etc.) étaient codés et poinçonnés sur un ensemble de cartes utilisées dans l'analyse et ne comportaient pas davantage de nom ou d'adresse.

En résumé, nos données de base se présentent sous la forme suivante :

i) 120 bobines d'entrevues enregistrées (2 exemplaires) ;

ii) 64 boîtes, complètes pour la plupart, de cartes perforées contenant les transcriptions, soit un total de 100 000 cartes ;

iii) des éditions imprimées (en plusieurs exemplaires) dans un format lisible ;

iv) nous conservons présentement les transcriptions corrigées sur une bobine maîtresse (pour ordinateur). Jusqu'à maintenant, 40 entrevues et plus de 20 boîtes de cartes sont conservées sur une seule bobine au centre de calcul.

Évidemment, plusieurs problèmes linguistiques (ex. : de phonétique, de morphologie) exigent que le chercheur écoute les enregistrements. Toutefois, même pour ce genre de travail, il est bien pratique de pouvoir suivre, à l'aide d'une transcription, l'enregistrement d'une conversation. Dans le cas de problèmes syntaxiques, sémantiques et lexicaux, il est possible d'obtenir des programmes capables de relever dans la bobine maîtresse certains ensembles de formes de surface, ce que nous décrirons plus en détail dans la section 5.

4.3 *Coût de la manipulation des données*

Jusqu'à présent, les étapes mentionnées en 4.2 ne sont pas entièrement terminées ; cependant, la partie du travail actuellement accomplie est suffisante pour nous permettre d'estimer de façon raisonnable quel sera le coût total de la manipulation des données. À l'exception de quelques-unes, les entrevues ont été transcrites une première fois, et environ 60 d'entre elles ont été corrigées par un assistant de recherche. une secrétaire ayant par la suite repoinçonné les cartes comportant une erreur. Les secrétaires consacraient en moyenne deux jours à une première transcription et une journée à la correction des cartes. En moyenne, les assistants de recherche peuvent corriger l'original de deux transcriptions par jour.

Pendant les 14 mois qui ont suivi la fin de la réalisation des entrevues, nous avons donné aux secrétaires un montant de $4 500 et $4 200 aux assistants de recherche employés spécifiquement pour travailler sur le corpus. Quelques corrections ont aussi été apportées par des chercheurs qui ont utilisé les transcriptions dans l'analyse de problèmes précis. Nous croyons qu'il faudra environ de quatre à cinq mois et une somme supplémentaire de $3 000 pour terminer la première étape de corrections. La seconde étape s'étendra probablement sur une période de six mois et coûtera près de $4 000. De plus, nous avons l'intention d'employer un programmeur qui nous conseille dans

le développement futur des techniques de traitement automatique des données ; cela augmentera d'environ $1 500 notre montant des salaires.

Il est difficile d'estimer le coût du temps de l'ordinateur. Il n'y a pas eu de distinction faite entre l'utilisation de l'ordinateur du présent projet et celle des programmes déjà inscrits au compte de l'un des auteurs (D.S.) affilié au Centre de recherches mathématiques. Nous croyons qu'un projet de cette dimension exige de l'ordinateur de 2 à 3 heures pour la préparation des données (sur des séries IBM 360, ou un ordinateur de séries CDC 6000) ; toutefois, nous ne nous sommes pas limités à cette durée. À cette somme il faut ajouter $500 pour les cartes d'ordinateur et autres fournitures.

En dépit du fait que nous avons tenté de restreindre au minimum le coût de traitement des données, certains pourront juger exorbitant le total obtenu, soit près de $17 700 (ce qui n'inclut pas les $2 500 déboursés pour l'équipement d'enregistrement et les $500 pour les bobines de ruban magnétique). Ce montant se trouve justifié si l'on accepte que la perspective d'ensemble de la sociolinguistique dépend d'une analyse réunissant les aspects linguistiques et sociolinguistiques (cf. section 5), et que cette analyse est impossible sans un nombre considérable et systématique de données du type de celles que nous avons rassemblées.

5 ÉTUDES DES DONNÉES

Les analyses effectuées jusqu'à maintenant par les directeurs du projet, H.J. Cedergren et Gillian Sankoff, et par les étudiants que nous dirigions, relèvent du domaine de la phonologie et de la syntaxe. Bien qu'aucune de ces études n'utilisait l'ensemble des 120 entrevues (principalement en raison du fait que les procédés de traitement des données, décrits en 4.2, n'ont pas encore été appliqués à

tout le corpus) presque toutes ont employé des techniques quantitatives facilitées par ces procédés. Les analyses portant sur : 1) l'élision du L dans les articles et les pronoms ; 2) l'élision du QUE complétif ; 3) l'utilisation du QUE explétif ; 4) les structures interrogatives ; 5) les changements dans l'emploi et la structure sémantique des pronoms impersonnels ont employé des méthodes quantitatives, y compris la formulation des règles variables.

La plupart de ces études comportaient les étapes suivantes :

a) la définition de la variable linguistique ;

b) la recherche des différentes formes de cette variable à l'intérieur du corpus. Les premières étapes impliquaient l'audition des bobines simultanément avec la lecture des transcriptions imprimées des entrevues. Nous avons déjà commencé à utiliser le second de nos trois programmes servant à la manipulation des données (le premier consistant dans le programme d'impression décrit en 4.2). Ce dernier rédigé par M. Guy Poulin du Projet de traduction automatique de l'Université de Montréal, est essentiellement un programme de concordance. Il utilise comme "entrée" une liste de mots et cherche à l'intérieur de la bobine maîtresse toutes les occurrences de ces différents mots. Par la suite, il imprime chaque phrase dans laquelle l'un de ces mots se trouve. Par exemple, la liste d'entrée peut se composer comme suit : tu, toi, tes, ton, t', ta ; la sortie sera alors formée de toutes les phrases contenant une ou plusieurs de ces formes de la deuxième personne du singulier. Pour tout problème qui n'est pas uniquement d'ordre phonologique, cela représente une économie de temps considérable par rapport aux recherches et procédés non automatisés ; c'est d'ailleurs un aspect que nous souhaitons améliorer et utiliser davantage dans nos futurs travaux ;

c) l'élaboration d'hypothèses et la formulation de règles grammaticales traitant de la variable, de même que les facteurs sociaux et (ou) stylistiques qui l'influencent ;

d) la sélection d'un sous-échantillon approprié de répondants pour vérifier ou rejeter les hypothèses ;

e) la compilation de toutes les occurrences de la variable et de leur contexte d'émission telles qu'elles apparaissent dans le langage du sous-échantillon ;

f) l'analyse statistique de la variation ; ceci renvoie à notre troisième programme qui donne le résultat d'un calcul de probabilité maximum, dans la formulation de la règle variable (Labov 1969) du phénomène grammatical étudié (cf. Cedergren et Sankoff, 1974). Il est possible et assez facile de vérifier chez un ou plusieurs individus donnés si une règle quelconque est variable ou catégorique dans un contexte linguistique et extra-linguistique particulier. Le programme détermine la valeur des paramètres des règles variables, se servant comme entrée des données brutes de fréquence des variantes selon leurs différents contextes. La sortie du programme est la contribution respective des divers éléments de l'environnement linguistique et extra-linguistique dans la probabilité d'application de la règle ;

g) la reformulation des règles.

L'effort déployé dans la codification des données et dans les procédés visant à les rendre accessibles au maximum, et ce, dans divers formats, n'a pas été uniquement avantageux pour les chercheurs directement associés au projet. Nous avons mis les données à la disposition de plusieurs autres linguistes, moyennant la signature d'un contrat (Appendice D) qui comprend un certain nombre de clauses visant à protéger les informateurs. Jusqu'à présent, six individus et des groupes

de recherches, autres que notre groupe, intéressés à des problèmes allant de la phonétique à la sémantique, ont utilisé nos données.

6. RÉSUMÉ

L'objectif de cet article était de fournir des renseignements sur les éléments méthodologiques de la recherche Sankoff-Cedergren sur le français parlé à Montréal, notamment en ce qui a trait aux procédés impliqués dans la cueillette, la manipulation et l'analyse des données pour la partie "enquête" du projet. Des techniques à caractère plus ethnographique sont appliquées dans la collecte des données de la partie "cas types" du projet, mais nous comptons employer des méthodes semblables à celles décrites ici pour la transcription, le traitement et l'analyse des conversations enregistrées.

Aux critiques qui affirment que les méthodes quantitatives n'apportent rien d'autre qu'ennui, perte de temps et d'argent et que celles-ci n'aboutissent qu'à prouver des généralités bien connues sur les rapports entre le langage et la stratification sociale, nous répondons qu'il est vrai que les études descriptives détaillées des communautés urbaines complexes supposent une bonne part de travail laborieux et parfois ennuyeux. Cependant, on peut diminuer cet inconvénient par l'emploi maximum de techniques électroniques et automatiques dans le traitement des données. C'est là une des raisons principales qui nous a incités à faire des innovations dans ce domaine. Par ailleurs, de telles méthodes sont indispensables à la compréhension des rapports subtils et complexes entre individus et à l'explication de la façon dont ils se servent de la langue. Souvent il est impossible de cerner ces rapports autrement qu'en termes quantitatifs (pour de plus amples élaborations de ces arguments, voir H.J. Cedergren et D. Sankoff, 1974, et G. Sankoff, 1971).

Dans une situation sociolinguistique du type de celle que nous étudions, caractérisée par une aliénation linguistique marquée, il est urgent et nécessaire de comprendre et démontrer la dimension systé-

matique des variétés non standard, et ce, en termes de leurs détails linguistiques. Selon nous, cette compréhension peut davantage être obtenue par le type d'étude que nous avons choisi d'entreprendre. En présentant l'éventail des problèmes méthodologiques entraînés par une telle étude et les solutions que nous leur avons apportées, nous espérons que notre expérience sera utile à d'autres chercheurs engagés dans des projets analogues.

David Sankoff
Gillian Sankoff
Suzanne Laberge
Marjorie Topham

Université de Montréal