

Article

« L'élaboration du vocabulaire fondamental du québécois parlé »

Normand Beauchemin

Revue québécoise de linguistique, vol. 11, n° 2, 1982, p. 113-125.

Pour citer cet article, utiliser l'adresse suivante :

<http://id.erudit.org/iderudit/602490ar>

Note : les règles d'écriture des références bibliographiques peuvent varier selon les différents domaines du savoir.

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter à l'URI <http://www.erudit.org/apropos/utilisation.html>

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche. Érudit offre des services d'édition numérique de documents scientifiques depuis 1998.

Pour communiquer avec les responsables d'Érudit : erudit@umontreal.ca

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL DU QUÉBÉCOIS PARLÉ*

Normand Beauchemin

L'idée même d'élaborer le vocabulaire fondamental d'une langue à des fins pédagogiques n'était déjà pas très nouvelle quand Gougenheim, Michéa, Rivenc et Sauvageot (1964) réalisèrent ce travail extraordinaire à l'époque du *français fondamental*. On trouve dans l'introduction et les cinq premiers chapitres de leur volume¹ une sorte d'histoire critique excellente des vocabulaires simplifiés et des vocabulaires plus ou moins fondamentaux élaborés en différentes langues jusqu'à 1960. Sans croire qu'ils ont tout dit sur le sujet, nous estimons que leur approche est suffisamment documentée, éclairante et solide pour servir de base à nos propres travaux sur le québécois parlé.

Avant d'aborder les questions importantes de la définition de notre corpus, de notre méthode de dépouillement et de comptage et des résultats que nous espérons, arrêtons-nous un moment pour expliciter le contenu des

* Communication présentée au XVI^e Congrès international de linguistique et philologie romanes (section II), Palma de Mallorca.

1. On peut également voir l'essentiel de ces notions à l'article "Vocabulaire", dans Pottier (1973).

deux expressions principales du titre de cet article: 1) vocabulaire fondamental 2) québécois parlé.

Vocabulaire: Nous entendons ici l'ensemble des unités lexicales qui apparaissent dans un texte ou un ensemble de textes, constitués en objet d'observation appelé corpus. À toutes fins utiles, ce "vocabulaire" ne sera toujours qu'une partie effectivement réalisée en discours de l'ensemble des unités du lexique de la langue considérée.

Fondamental: Les unités lexicales observées dans le corpus et qui seront communes à tous les textes ou à la grande majorité des échantillons de discours observés seront appelés *vocabulaire fondamental*. Pour reprendre la formule de Ch. Müller nous devrions avoir là "le vocabulaire que très peu d'individus ignorent" (1968, p. 137), si, bien entendu, nos échantillons sont suffisamment étendus et représentatifs.

Il ne s'agit donc ni d'un *trésor* (où l'on tente de collectionner toutes les unités lexicales différentes utilisées par les membres d'une communauté linguistique), ni d'un *vocabulaire passif* (où l'on ajouterait aux données de l'observation directe du corpus le vocabulaire que d'autres sortes d'études pourraient révéler comme connu passivement par les locuteurs de la communauté linguistique), ni d'un *vocabulaire disponible*, c'est-à-dire des mots qu'un individu connaît et qu'il utilise parfois mais que de fait il n'a pas utilisés dans le texte qu'on a dépouillé, ni surtout ce que ce qu'on est convenu d'appeler *les vocabulaires techniques* et dont on n'a aucune définition vraiment exhaustive.

Québécois parlé: Faute de moyens, nous ne comptons pas, du moins pour l'instant, dépasser les limites géographique et politique du Québec dans

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL DU QUÉBÉCOIS PARLÉ

le choix de nos textes. Ajoutons aussi qu'il s'agit strictement, pour nous, du parler francophone du Québec: nous ne sommes pas sans savoir qu'il y a un million et demi d'habitants du Québec pour qui la langue maternelle est autre que ce que nous appelons, du moins entre nous francophones du Québec, le *québécois*, mais pour l'instant seule nous intéresse la langue des francophones nés au Québec de parents québécois. Nous sommes bien conscients de n'avoir utilisé jusqu'ici que des critères externes pour cerner cette réalité linguistique appelée le *québécois*, que certains définissent comme une variété régionale du français et que d'autres, aidés en cela par l'étiquette même, considèrent comme une langue parente, proche parente même, du français commun mais assez différente toutefois pour n'être pas le *français commun*. Nous n'avons pas l'intention de trancher la question dans l'immédiat: il nous manque, entre autres, une bonne description de toutes ses structures internes, pour mesurer la distance linguistique qui le sépare du français commun ou l'y unit encore assez pour empêcher cette distinction. Nous pourrions, bien sûr, parler du *franco-québécois* pour souligner cette parenté et réserver le terme de *québécois* pour définir "l'ensemble de tous les traits linguistiques spécifiques du français parlé au Québec" (voir à ce sujet, Verreault (1979, p. 144)) ce qui suggère, évidemment, que ces traits sont connus, et qu'ils déterminent une espèce *québécois* distincte d'un genre appelé *français*.

Notre projet porte sur le "parlé": mais de quelle parole s'agit-il? Comme linguiste croyant et pratiquant, j'avais considéré cette notion comme l'une des nombreuses évidences qui nous mettent à l'abri de nous-mêmes sans que nous les remettions vraiment en question: pour moi, était parlé ce qui était dit, c'est-à-dire prononcé à l'aide de l'appareil phonatoire,

et cela s'opposait à ce qui est écrit, c'est-à-dire mis sur papier, le plus souvent, et fait pour être lu. Mais des sociologues sont montés à l'assaut de mes évidences et des sémioticiens déguisés en linguistes de la parole ou en analystes du discours m'ont obligé à y voir quelques brèches. Les linguistes de la maison Larousse vont même dans le même sens: "Ce divorce [...entre la langue écrite et littéraire et la langue parlée] n'est pourtant pas inéluctable: il arrive fréquemment [...] que la langue écrite garde assez de contact avec l'usage oral pour que tout locuteur puisse rester lecteur des ouvrages de son siècle et du précédent. C'est le cas en France, où l'écrit et le parlé au niveau du XX^e siècle ne peuvent être donnés pour "deux langues" que par une métaphore, passée dans l'usage par des linguistes, dont il ne faut pas être dupe. Cela dit, on doit reconnaître qu'il existe un écart sensible aujourd'hui entre le français écrit tenu (...pas forcément littéraire) et le français parlé..." (*Le Grand Larousse de la langue française* (1976), t. V, p. 3983).

Je sais aussi depuis toujours qu'à la limite des auteurs écrivent comme on parle, et que des locuteurs se font accuser, en langue populaire, de "parler comme un grand livre". Pour moi, en théorie on peut bien ne pas trouver de coupure nette dans cette sorte de continuum; mais la pratique nous force à en poser une. D'autre part, ce qui est dit du français hexagonal est au moins à vérifier une fois avant qu'on ne le dise du québécois, ne serait-ce que par acquit de conscience.

Le corpus: Espérons qu'il apparaîtra maintenant plus clairement pourquoi nous avons insisté un peu sur la définition des termes du titre de cet exposé. L'ensemble des textes déjà étudiés et ceux qui restent à é-

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL EN QUÉBÉCOIS PARLÉ

tudier doivent être nécessairement de québécois parlé, ou considérés comme tels, selon des critères qui, pour l'instant, ne peuvent être qu'externes. Nous avons déjà constitué un ensemble de 300 000 mots de parole, prononcés en réponse libre à des questions ouvertes, lors d'une enquête sociolinguistique, dont le travail sur le terrain s'est déroulé entre 1972 et 1974 dans la région du sud-est du Québec². La transcription de 100 de ces enquêtes (environ 325 000 mots) a déjà été publiée sous le titre *Echantillons de textes libres* (Beauchemin et Martel (1975-1978) et Beauchemin, Martel et Théoret (1980-1981)). Il s'agit d'une langue parlée de conversation ordinaire, à un niveau plutôt familier. Les questions ont été les mêmes dans tous les cas pour amorcer la conversation sur des lieux communs de la vie québécoise: la langue, le temps de l'enfance, le temps des fêtes, les "sucres", le mariage, la maladie, etc. Nous comptons bien, au cours des mois (ou des années?) qui viennent, élargir ce corpus en y ajoutant d'autres textes de langue parlée qui nous permettent de voir la variété de vocabulaire commun qu'apporteraient par exemple des contes populaires enregistrés sur le vif, des monologues de quelques chansonniers populaires, quelques textes radiophoniques³ mais considérés par tous comme étant de la véritable langue parlée, et peut-être enfin, avec quelque réticence, quelques pièces de théâtre très populaires reconnues par les non spécialistes comme étant en langage ordinaire de tous les jours.

Notre conception large de "langue parlée", n'a pas de scrupules ma-

2. Le questionnaire et les principales conditions de l'enquête sont donnés dans Beauchemin (1972) et Beauchemin (1977).

3. Par exemple, du genre de ceux qu'on trouve dans Pagé (1976).

jeurs à englober ainsi dans un même corpus des éléments qui, selon plusieurs théories, sont fort divers. Ce n'est pas par inconscience que nous ignorons la diversité des intentions de locuteurs, la diversité des auditeurs, la diversité des moyens et des règles de chaque genre, la diversité des fonctions du langage dans un même texte et a fortiori dans plusieurs, la diversité des régions québécoises et des traces qu'elles laissent dans le vocabulaire, etc. La diversité dans tous ces cas devrait venir appuyer encore ce que nous trouverons de vocabulaire fondamental (ou commun à tous les membres de la communauté).

Disons un mot de la taille de ce corpus. Nous envisageons de ne pas dépasser le demi million de mots pour des raisons d'ordre pratique et empirique. L'accessibilité des textes qui répondent à nos critères nous permettra un choix plus confortable si nous n'avons à trouver que 100 000 mots de contes, monologues et textes radiophoniques déjà disponibles sous forme imprimée. Même chose pour le genre de pièces de théâtre qui pourraient satisfaire à nos critères. Enfin nous avons déjà un peu plus de 300 000 mots de nos propres enquêtes, et dont la moitié est déjà lemmatisée et sur disque d'ordinateur. Ce sera un échantillon bien maigre si on le compare aux 70 millions de mots de langue littéraire du *Trésor de la langue française* de Nancy. Mais notre corpus pourra sainement être comparé à celui du *Français fondamental* (312 135 mots) pour le français parlé, ou à celui du *Frequency Dictionary of the French Words* de Juilland (500 000 mots) pour le français écrit. À notre connaissance, il n'y a pas de théorie qui puisse infirmer ou confirmer la justesse de la taille d'un échantillon dans ce genre de travail. Tout au plus sait-on empiriquement que les comparaisons, pour être éclairantes, doivent se faire sur

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL EN QUÉBÉCOIS PARLÉ

des échantillons de taille à peu près semblable.

Méthode de dépouillement et de comptage: Les modèles statistiques du lexique et du vocabulaire d'une langue sont plus puissants à notre connaissance, quand on fonctionne sur des listes de fréquence de vocables (c'est-à-dire d'entrées de dictionnaire sous lesquelles on trouverait les mots du texte). Nous avons donc besoin, entre autres, de chaque vocable, de sa fréquence et de sa dispersion dans le corpus pour décider, en fin de compte, de son appartenance au *vocabulaire fondamental*. Cet exercice de découpage des unités d'un texte et de leur lemmatisation pose plusieurs problèmes pratiques déjà exposés souvent⁴ et sur lesquels nous ne nous arrêterons pas ici, du moins pour ce qui touche les postulats linguistiques qu'il présuppose. À cause de l'originalité et de l'efficacité des méthodes que nous avons mises au point, nous croyons toutefois utile d'en parler ici à titre d'information et pour fins de discussion.

Profitant d'abord de l'expérimentation de notre collègue et associé de recherches Pierre Martel qui, minutieusement, méthodiquement et surtout très patiemment, avait dépouillé à la main un premier échantillon d'environ 50 000 occurrences, nous avons adapté à nos besoins, en les développant, un certain nombre de programmes informatiques créés à Toulouse pour le Professeur Roche qui travaillait à des textes littéraires portugais. Nous avons exposé ailleurs cette méthode et les résultats partiels déjà obtenus (Beauchemin, à paraître). En janvier 1979, nous avons donc

4. Voir Muller (1977), p. 11-40 en particulier. Sur nos choix de norme de dépouillement, voir l'introduction de notre *Vocabulaire fondamental du québécois parlé* (1979), p. iv et suivantes.

l'expérience de près de 200 000 occurrences lemmatisées, indexées et comparées de différentes façons. Nous y avons puisé les 3 000 formes les plus fréquentes, qui représentent environ 80% des occurrences de nos textes. Les 3 000 formes ont été organisées en dictionnaire (fichier informatisé séquentiel indexé) où chacune est une entrée donnant accès à un article contenant le vocable, la classe grammaticale, et quelques commentaires dans les cas d'homographie ainsi qu'un certain nombre de renseignements nécessaires à leur traitement informatique.

À l'aide d'un système d'ordinateur relativement simple⁵ et des plus courants, et d'une batterie de programmes ad hoc (écrits en macro-assembleur et stockés en binaire sur disquette, après compilation par le même système), nous pouvons traiter nos textes de façon presque automatique: les textes ayant été entrés sur clavier ASCII, apparaissent par tranche sur écran et chacun des mots (groupe de lettres précédé et suivi d'un espace) est comparé aux 3 000 formes du dictionnaire. Si le mot s'y trouve et qu'il n'a pas d'homographes, il est automatiquement transféré dans un fichier de sortie, prêt au comptage, avec son article de dictionnaire. S'il est au dictionnaire comme homographe, toutes les possibilités apparaissent sur l'écran et le système attend une intervention humaine pour l'inscrire au fichier de sortie selon la solution retenue: il suffit d'appuyer sur une seule touche (chiffre de 1 à 9) pour indiquer ce choix et déclencher la suite du travail automatique. Si le mot n'apparaît pas au dictionnaire, le système fait apparaître sur l'écran une matrice sim-

5. L'unité centrale est un micro-ordinateur ONTEL-OP1(64K), (INTEL-8080), muni d'une mémoire auxiliaire à 2 disquettes et d'une imprimante NEC.

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL DU QUÉBÉCOIS PARLÉ

ple à remplir (classe grammaticale et vocable, s'il est différent de la forme). Une fois fournis ces renseignements, le travail automatique se continue jusqu'à la fin du texte. Dans le fichier de sortie, sur disquette, chaque unité comporte une référence au numéro du texte, à la page et à la ligne du document de départ, ce qui permet les vérifications ordinaires, les corrections nécessaires des fautes détectées à n'importe quelle étape, et finalement, l'établissement des concordanciers désirés.

Le traitement complet d'une occurrence de texte se fait en 0,23 seconde. Fabriquer des index de toutes sortes ou des concordances à partir de ces fichiers n'est toujours l'affaire que de quelques secondes de temps-machine sur un ordinateur courant (IBM 360-45 ou mieux). Nous avons élaboré tous les programmes pour le faire et nous le faisons régulièrement.

Tous ceux qui ont déjà eu à travailler à la main à la confection d'index et de concordances, ou qui ont travaillé à lemmatiser à la main des formes d'unités lexicales comprendront que l'avantage principal de la technique rapidement exposée ici est d'assurer une application rigoureuse et constante de la "norme" que nous nous sommes donnée au départ. Mentionnons enfin que le dictionnaire informatisé est très facilement perfectible par le même programme qui sert à le créer ou à le retoucher.

Résultats attendus: Que nous fournira le corpus ainsi dépouillé? Des listes de vocables et de formes, ainsi que leur fréquence pour chaque texte individuel, une liste par sous-ensemble qui nous intéressera et une liste pour l'ensemble. Dès lors, la dispersion de chaque vocable à travers les différents textes nous permettra d'établir une sorte d'indice

d'usage ou d'utilité: la fréquence (f) d'un vocable, comme on sait, peut fluctuer énormément d'un texte à l'autre, d'un locuteur à un autre, d'une situation à une autre, etc. Aussi, pour en tenir compte, il nous faudra combiner cette f de chaque mot à un indice de dispersion (D) qui exprime la régularité ou l'irrégularité des occurrences d'un mot dans les sous-ensembles du corpus⁶. C'est par le calcul de cet indice d'usage que nous pourrons établir les 2 000 ou 3 000 ou 5 000 vocables les plus communs et partant, les plus utiles pédagogiquement au moins, du québécois parlé.

Nous croyons pouvoir aussi distinguer à l'aide de ces 3 indices (Fréquence, Dispersion et Usage) le vocabulaire commun des vocabulaires thématiques, spécialisés ou régionaux qui se trouvent dans nos textes. Le vocabulaire commun sera constitué des unités fréquentes et réparties de façon régulière à travers tous les sous-ensembles. Les autres vocabulaires seront identifiés surtout par une répartition irrégulière et une fréquence moyenne ou basse. La combinaison des deux indices en une échelle unique (Usage) servant à classer tous les vocables du corpus nous servira à décider du sort des mots qui obtiennent une fréquence et une dispersion moyenne et, nous l'espérons, à éclairer les cas douteux.

La fréquence relative des formes de vocables fréquents et très polymorphiques (par exemple, *être*, *avoir*, *aller*) nous apparaît aussi du plus

6. Pour une discussion plus poussée de cet aspect, voir Juilland (1970). Rappelons ici ses formules:

Indice d'usage (U): $f \times D$

Indice de dispersion: $1 - \frac{V}{\sqrt{n - 1}}$, ou V est le coefficient de varia-

tion et n le nombre de branches du corpus.

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL DU QUÉBÉCOIS PARLÉ

haut intérêt pédagogique: l'enseignant au primaire pourrait s'en inspirer pour répartir plus efficacement ses efforts en les concentrant davantage sur les formes les plus fréquentes.

La graphie fantaisiste de plusieurs de nos auteurs qui veulent écrire "comme on parle" pourrait être aussi grandement aidée: une brève comparaison entre toutes les graphies d'une même forme fera apparaître la graphie la plus fréquente et pourra servir par là même à un début de standardisation, au moins pour les expressions ou les syntagmes figés et peut-être lexicalisés dans la pensée populaire du type: *ça fait que, à cette heure, en tous les cas, qu'est-ce que c'est que ça?* ou tout simplement pour des mots isolés comme *bean* (écrit *bine, binne, bean*), *draught* (écrit *draffe, draft, drafte*), *dump* (écrit *dompe, dump, dumpe*), etc.

Conclusion: Nous n'avons exposé ici qu'un aspect de l'ensemble de notre projet, celui de l'étude du vocabulaire commun du québécois parlé, qui, à notre avis, est un instrument de travail nécessaire au psychologue, au psychiâtre, à l'enseignant, à l'audiologiste, au publicitaire, et au linguiste, pour ne mentionner que ceux-là.

Dégager la spécificité du québécois parlé commun par rapport au français commun présente aussi un intérêt théorique certain: archaïsmes et néologismes, emprunts à l'anglais et aux dialectes de France, rejets conscients ou ignorances involontaires du vocabulaire familier quotidien de l'hexagonal, etc., pourront enfin apparaître de façon plus évidente encore par la connaissance plus précise de ce vocabulaire fondamental du québécois parlé.

Mais nous sommes bien conscients qu'un corpus de langue parlée de 500 000 mots mis en mémoire d'ordinateur, même sous une forme écrite et souvent discutable, pourra servir à plusieurs autres types d'études: standardisation de la graphie, morphologie et syntaxe des parties de discours, syntaxe de la phrase et de l'énoncé, modèles de dialogues et méthode d'enquête orale, marques linguistiques de la différence des sexes, et bien d'autres études encore, pourront utilement être conduites à partir d'un corpus aussi facilement accessible.

*Normand Beauchemin
Université de Sherbrooke*

L'ÉLABORATION DU VOCABULAIRE FONDAMENTAL DU QUÉBÉCOIS PARLÉ

RÉFÉRENCES

- BEAUCHEMIN, N. (1972) *Document de travail n° 1: Le questionnaire*, collection "Recherches sociolinguistiques dans le région de Sherbrooke", Sherbrooke, 46 p.
- BEAUCHEMIN, N. (1977) *Document de travail n° 11: Données sociologiques*, collection "Recherches sociolinguistiques dans la région de Sherbrooke", Sherbrooke, 54 p.
- BEAUCHEMIN, N. (1979) *Vocabulaire fondamental du québécois parlé*, Sherbrooke.
- BEAUCHEMIN, N. (à paraître) "Méthode de cueillette et d'analyse du vocabulaire fondamental du québécois parlé", communication présentée au Colloque sur les parlers régionaux, Québec (1979), à paraître dans les *Actes du colloque*.
- BEAUCHEMIN, N. et P. MARTEL (1975-1978):
Echantillon de textes libres n° I (1975), Sherbrooke, 236 p.
Echantillon de textes libres n° II (1975), Sherbrooke, 268 p.
Echantillon de textes libres n° III (1977), Sherbrooke, 209 p.
Echantillon de textes libres n° IV (1978), Sherbrooke, 291 p.
- BEAUCHEMIN, N., P. MARTEL et M. THÉORET (1980-1981):
Echantillon de textes libres n° V (1980), Sherbrooke, 245 p.
Echantillon de textes libres n° VI (1981), Sherbrooke, 290 p.
- GOUGENHEIM, G., R. MICHÉA, P. RIVENC et A. SAUVAGEOT (1964) *L'élaboration du français fondamental* (1^{er} degré), Paris, Didier.
- Le grand Larousse de la langue française*, t. V (1976).
- JUILLAND, A.G. (1970) *Frequency Dictionary of French Words*, La Haye, Mouton.
- MULLER, Ch. (1968) *Initiation à la statistique*, Paris, Larousse.
- MULLER, Ch. (1977) *Principes et méthodes de statistique lexicale*, Paris, Hachette.
- PAGÉ, P. (1976) *Le comique et l'humour à la radio*, Montréal, Editions La Presse.
- POTTIER, B. (1973) *Le langage*, collection "Les dictionnaires du savoir moderne", Paris, Denoël.
- VERREAULT, C. (1978) "Les adjectifs en *-able* en franco-québécois", dans L. BOISVERT, M. JUNEAU et C. POIRIER, *Travaux de linguistique québécoise*, n° 3, Les Presses de L'Université Laval.