Cahier de linguistique

La richesse lexicale individuelle comme marqueur sociolinguistique

Nathan Ménard and Laurent Santerre

Number 9, 1979

URI: https://id.erudit.org/iderudit/800081ar
DOI: https://doi.org/10.7202/800081ar

See table of contents

Publisher(s)

Les Presses de l'Université du Québec

ISSN

0315-4025 (print)
1920-1346 (digital)

Explore this journal

Cite this article

Ménard, N. & Santerre, L. (1979). La richesse lexicale individuelle comme marqueur sociolinguistique. *Cahier de linguistique*, (9), 165–188. https://doi.org/10.7202/800081ar

LA RICHESSE LEXICALE INDIVIDUELLE COMME MARQUEUR SOCIOLINGUISTIQUE

1. LES MARQUEURS SOCIOLINGUISTIQUES

On sait, depuis Labov surtout, que les traits phonétiques sont des marqueurs sociolinguistiques privilégiés dans la mesure où les auditeurs peuvent être conscients des traits soumis à variation; c'est pourquoi Labov distingue bien entre indicateurs, marqueurs et stéréotypes. On peut se demander dans quelle mesure le *vocabulaire* utilisé dans la communication orale varie selon les strates sociales établies par l'âge, le sexe, l'instruction, le métier, etc., et dans quelle mesure ces variations sont perçues par les auditeurs au point d'entrer en ligne de compte dans les jugements favorables ou défavorables qu'ils portent sur les locuteurs.

1.1 *Marqueurs phonétiques*

Les variations linguistiques peuvent être très grandes au sein d'une société, sans que les individus en soient même conscients; c'est le cas de la chute des voyelles hautes dans le français parlé actuellement à Montréal (Santerre, 1975). Ce phénomène d'une très grande fréquence dans toutes les couches de la société passe complètement inaperçu, même de ceux qui soignent leur langage, à la radio et à la télévision; pourtant il réduit considérablement le nombre de syllabes dans les énoncés. Par contre, la diphtongaison est connue de la majorité des locuteurs montréalais; les locuteurs de tel groupe social, ou en telles circonstances, peuvent se démarquer volontairement de ceux qui diphtonguent toujours ou de ceux qui ne diphtonguent jamais; ce trait peut servir de marqueur sociolinguistique (Santerre, 1977). Les variations conscientes entrent en jeu dans la stratification sociale par l'imitation des formes de prestige ou par la stigmatisation des formes

réputées populaires. Mais les variations inconscientes sont une source peut-être plus puissante des changements linguistiques, parce que les grammairiens, l'école, les locuteurs eux-mêmes ne peuvent les freiner.

D'ailleurs le jeu de l'imitation et de la stigmatisation des formes linguistiques dans une société est très difficile à cerner; les jugements des auditeurs varient considérablement selon leurs caractéristiques personnelles, et il n'est pas rare de voir des locuteurs pratiquer couramment et inconsciemment ce qu'ils stigmatisent chez les autres. Par exemple, la diphtongaison de *père, mère* à Montréal est stigmatisée par les gens instruits, en général. Pourtant, les mêmes personnes qui ne voudraient pas diphtonguer les voyelles longues antérieures ne s'entendent pas diphtonguer les voyelles postérieures, comme dans *base, côte*. Quant aux sujets peu instruits qui diphtonguent d'ordinaire, ils se corrigent quand ils disent frère [fʀeʀ] et prononcent [fʀa^{e}ʀ]. Ainsi donc, une même forme est valorisée par un groupe social et stigmatisée par un autre, dans la même ville.

Pour mieux analyser les forces ou les contraintes sociologiques qui s'exercent sur une langue, on aurait besoin de savoir dans quelle mesure les jugements subjectifs des auditeurs sur la qualité ou les convenances du langage de leurs concitoyens sont corroborés par les mesures objectives des paramètres linguistiques. Le langage des gens est-il plus évalué d'après les structures syntaxiques, ou par la prononciation, le débit, l'accentuation, l'intonation, ou par l'aisance à formuler, ou encore par les choix du lexique ? Est-on plutôt sensible à l'apparition dans le discours de certains mots qui font bien, ou par le large stock lexical qui empêche un trop grand nombre de répétitions ?

1.2 *Marqueurs lexicaux*

L'étude statistique du vocabulaire a déjà vérifié nombre d'hypothèses sur les différences individuelles et les oppositions de groupes. Mais les résultats ne sont pas toujours concluants, en ce sens qu'ils ne permettent pas de rejeter nettement le rôle du hasard. Parfois ils se révèlent même contradictoires. C'est qu'il faudrait pousser plus loin l'analyse de cette variable linguistique ou encore refaire les expériences sur des bases plus

larges ou plus finement calculées afin de trouver des explications nécessaires. Une définition plus nuancée de la *richesse lexicale* dissipera certaines confusions voire des erreurs méthodologiques qui ont pu entacher quelques calculs antérieurs. De même la prise en considération des structures du lexique, tant au point de vue quantitatif - loi de Zipf par exemple - qu'au point de vue qualitatif (catégories grammaticales[1], champs lexicaux, classes étymologiques, classes morphologiques), sera beaucoup plus éclairante pour la sociolinguistique que la manipulation de données indifférenciées. Il va sans dire que les problèmes méthodologiques sont cuisants, et cet aspect nous préoccupe, dans cet article, davantage que les corrélations comme telles, encore que, prises comme résultats provisoires, celles-ci ne manquent pas d'intérêt.

..3 *État de la question*

Il existe une longue tradition des études de la richesse lexicale individuelle dans la langue écrite. Les caractéristiques lexicales des textes selon les genres littéraires et selon les époques, de même que les caractéristiques des auteurs et même des époques des auteurs, servent à la datation et à l'authentification des oeuvres. Les travaux de stylométrie reposent en grande partie sur l'emploi privilégié en fréquence et en qualité d'une partie déterminée du lexique, de même que sur la prépondérance accordée à telles catégories grammaticales. Carroll (1969) a cherché à savoir jusqu'à quel point les lecteurs pouvaient être conscients des caractéristiques lexicales des textes en prose et comment leurs jugements subjectifs s'en trouvaient influencés.

Les recherches sur la langue parlée sont beaucoup plus rares, en français en particulier. Malécot (1976) a étudié la fréquence d'occurrences des mots dans la conversation à Paris. En français québécois, nous avons les textes de discours libres, et les premiers résultats de l'enquête publiée par Beauchemin et Martel (1972, 1975) pour la région de Sherbrooke.

---

1. Notre étude se limitera ici à la *richesse lexicale* et aux *catégories*.

À Montréal même, D. Sankoff (1975) a fait une première approche globale d'analyse sociolinguistique sur l'ensemble du corpus Sankoff-Cedergren. Même si nos expériences visent un plus grand nombre de sujets, nous nous restreignons dans cet article à un sous-ensemble de douze dossiers où l'on retrouve les caractéristiques des locuteurs montréalais. On sait comment ce corpus de français est sociologiquement balancé (D. Sankoff, *et al*, 1976). C'est un risque d'en tirer un sous-échantillon, et pour cette raison quelques-unes des conclusions d'ordre statistique n'ont ici qu'une valeur expérimentale.

2. PROBLÈMES MÉTHODOLOGIQUES

Ils se posent à la fois sur le plan linguistique et sur le plan statistique.

2.1 *Définition de la richesse lexicale individuelle*

Ce n'est pas nouveau en linguistique, mais tout le monde ne s'entend pas sur la portée et les limites d'une telle notion. Les statisticiens se contentent bien souvent d'un simple rapport entre nombre de mots différents ou vocables[2] (V) et nombre de mots du *texte* (N) ou occurrences totales[3] (type/token). Ceux qui ont abordé la question par le biais des études de style ont mis l'accent beaucoup plus sur les aspects fonctionnels et esthétiques.

Nous écartons ici volontairement les considérations sur la souplesse, la précision, le caractère imagé, expressif et recherché du vocabulaire dans le but d'en inférer une certaine richesse. Non pas que le vocabulaire ne possède ces propriétés, mais avant de les quantifier, il faut songer sérieusement aux moyens de faire échec à la subjectivité et à l'arbitraire. On peut y parvenir en adoptant certains instruments de mesure, courants en psycholinguistique notamment. Toutefois, il s'agit d'opérations coûteuses et encore délicates.

---

2. Voir à la fin une liste des symboles.

3. Par texte nous entendons tout discours clos, parlé ou écrit.

Nous préférons considérer le *vocabulaire* comme un ensemble clos, fini : c'est la somme des vocables contenus dans un corpus ou fragment de corpus - à ne pas confondre avec le *lexique* qui a un caractère virtuel et est un ensemble ouvert, même si Guiraud et d'autres lexicologues ont déjà avancé des chiffres pour le lexique individuel - 30 000 mots en moyenne pour un Français cultivé. Cette opposition n'est pas non plus celle, courante, entre le *vocabulaire actif* et le *vocabulaire passif*, en ce sens que chacun des mots du vocabulaire actif d'un locuteur a une probabilité non nulle de s'actualiser dans son discours alors que le vocabulaire passif inclut tout mot qu'il comprend ou est susceptible de comprendre, même s'il ne l'utilise jamais - sauf par psittacisme. La *richesse lexicale individuelle* - malgré l'ambiguïté de l'adjectif - est un lieu de comparaison entre deux textes ou deux groupes de textes en fonction : 1) de leur étendue respective et du nombre de vocables relevés dans chacun d'eux; 2) de l'accroissement du vocabulaire au fur et à mesure que le discours progresse.

.2 *Accroissement et lexique en jeu*

La relation entre le nombre de vocables et la longueur du texte donne une vision statique de la richesse lexicale. Nous avons pris pour chaque témoin des fragments de 2000 mots, ce qui facilite la comparaison. Toutefois, malgré les précautions normales au niveau de l'échantillonnage, certaines questions se posent :

1) Même si pour N = 2000 on constate que le locuteur *A* utilise un vocabulaire plus riche que *B*, comment peut-on affirmer qu'avec un texte un peu plus court ou un peu plus long le classement ne sera pas renversé ?

2) Même si au point de vue thématique on a pu dans l'enquête minimiser les variations dues aux sujets de conversation (les témoins répondaient pratiquement aux mêmes questions et ils n'étaient pas limités par des données scientifiques ou techniques), il est possible qu'un témoin n'ait pas voulu donner sa plein mesure ou au contraire ait essayé de se surpasser en faisant appel momentanément à toutes ses ressources lexicales pour s'exprimer. Les raisons seraient diverses : distance vis-à-vis de l'interviewer, choix du niveau de style, etc. Il s'agit là de contraintes inhérentes à

toute enquête sur la performance, mais est-ce qu'elles ne pèsent pas lourd dans toute extrapolation sur le comportement verbal d'une classe sociale par exemple ?

À la première question on peut répondre par un calcul de l'accroissement du vocabulaire. Des méthodes assez sûres permettent d'y parvenir, notamment l'application de la loi binomiale telle que l'a formulée Ch. Muller (1977). Mais déjà nous obtenons une courbe en divisant chaque texte par tranches de 250 mots et en comparant la valeur V à différents points. Un exemple frappant est la courbe des témoins 49 et 7 (voir tableau I en annexe) Le témoin 49 tient la tête jusqu'à N = 1000, mais sa moyenne d'accroissement est de 37,46 par opposition à 41,50 pour le témoin 7. Comparez également les témoins 31 et 115, 115 et 73. En cas d'ex aequo à N =2000 ou à n'importe quelle longueur de texte, la courbe de progression (valeur observée) peut intervenir comme 2^{e} critère; c'est son côté pratique. Le cas échéant, on prendra comme norme les valeurs de la *pente* et de la fonction V/N telles que les ont calculées Guiraud et Herdan (texte écrit) et D. Sankoff (corpus montréalais).

S'il faut l'aborder de front, la deuxième question restera sans réponse vraiment satisfaisante. On peut l'escamoter en faisant confiance aux opérations de normalisation, qui corrigent les accidents dans les distributions, et à la loi des grands nombres, en espérant que de tels cas constituent l'exception.

Mais en serrant la réalité de plus près, on se rend compte que dans la structure quantitative d'un vocabulaire (tableau des fréquences), les mots qui ne sont pas du tout ou presque pas répétés (hapax et mots de fréquence 2) forment un ensemble qui permet d'obtenir par extrapolation le stock lexical qu'un individu met en oeuvre dans une situation de communication, et ce, indépendamment du fait qu'il a pu en puiser beaucoup (vocabulaire riche) ou peu (vocabulaire relativement pauvre) pour construire son discours. C'est la notion de *lexique en jeu* qui intervient ici et pour le calcul duquel on a fait énormément de progrès récemment (Kalinin, 1965; Ménard, 1972; Muller, 1977; G.R. Roy, 1977). Les résultats sont d'autant plus intéressants qu'ils émanent sinon des chercheurs indépendants, du moins d'expériences diverses.

Les hypothèses à la base de ces calculs établissent un lien étroit entre les règles complexes de la redondance - qui à un certain niveau échappent complètement à la volonté du sujet parlant - et la taille du lexique. Des différentes formules proposées, c'est celle de Kalinin (1965, 2e version) que nous retenons en dépit de sa complexité. Celle de Muller serait beaucoup plus facile à appliquer, d'autant plus qu'elle s'accompagne d'une table. Mais compte tenu de la longueur de nos fragments, nos textes ne remplissent pas toutes les conditions nécessaires à ce calcul.

Signalons enfin que le *lexique en jeu* ne se confond pas avec le *vocabulaire disponible*, même si au niveau des opérations mentales (associations, rappels) il sont sans doute régis par les mêmes règles (voir à ce sujet Vikis-Freiberg, 1974, pour une expérience avec des sujets québécois).

Ces deux paramètres - accroissement du vocabulaire et lexique en jeu - donnent une version dynamique de la richesse lexicale et nous éclairent au moins sur la valeur définitive ou provisoire des classements obtenus à l'aide de V.

Les normes de dépouillement

Nous avons parlé de mots et de vocables. Mais il y a différentes façons de délimiter et de dénombrer ces unités, et à moins d'adopter des normes strictes, les résultats ne sont pas comparables. Nous avons adopté les normes suggérées par Charles Muller et ses disciples, mais avec des modifications qui tiennent compte non seulement du caractère oral de notre corpus, mais encore des unités de fonctionnement autonomes (éléments lexicalisés) dans le français québécois. À cet égard, l'équipe de Sherbrooke (Beauchemin et Martel) avait déjà affronté les mêmes difficultés et pris des décisions auxquelles nous nous sommes ralliés en grande partie.

Par exemple :

- Nous avons tenu compte du texte intégral, avec les hésitations, les phrases non terminés, et même les incohérences inhérentes à la parole spontanée.

- Nous avons réuni les variantes morphologiques sous une forme neutre, après transcription graphique normalisée. C'est la lemmatisation. De cette façon, quand nous donnons la fréquence du verbe *aller*, nous sommes sûrs d'avoir inclus *ira*, et de ne pas confondre la *porte* et je *porte* un chapeau.
- Astheure (à cette heure), pantoute (pas en tout) et autres unités de fonctionnement de ce genre ne sont pas séparés.

En ce qui concerne les catégories grammaticales, dans tous les cas où il pouvait y avoir un ambiguïté (entre adjectif et adjectif substantivé, entre adjectif verbal et adjectif), c'est l'emploi du mot en contexte qui a servi de critère déterminant. Rappelons par ailleurs qu'il est possible de traiter le texte de manière à comptabiliser les structures syntagmatiques et syntaxiques de surface. Le dénombrement des catégories grammaticales comme telles n'est nullement une limite théorique à cet égard.

2.4 *Les tests et calculs statistiques*

Nous utilisons les tests statistiques usuels et très connus en sciences humaines : formule de Spearman pour les corrélations des rangs, chi-carré lorsque les tables de contingence étaient possibles, indice Z pour la comparaison de moyennes lorsque les calculs de moyenne et d'écart type sont permis. Pour rejeter l'hypothèse nulle, le seuil de probabilité est fixé à 5 %.

Le calcul du lexique en jeu (K) exige quelques explications. Kalinin a proposé deux formules. Pour diverses raisons (cf. Roy, 1977) nous n'appliquerons que la deuxième. Il s'agit d'une intégrale. Elle exige que quantitativement le texte obéisse aux règles du discours naturel et qu'entre autres le tableau des fréquences vérifie la loi de Zipf - avec les corrections que l'on sait - pour les fréquences basses. Elle requiert la connaissance de deux données observées : V_1 (ou effectif des mots de fréquence 1) et V_2 (effectif des mots de fréquence 2).

$$K = 2\, V_2 \frac{e^{x_o}}{x_o}$$

Nous déterminons x_o en cherchant dans les tables de Jahke-Emde la valeur qui vérifie l'équation suivante :

$$\frac{V_1}{2V_2} \quad e^{-x_o} \quad = \quad \int_{x_o}^{\infty} \quad \frac{e^{-u}}{u} \quad du$$

Le deuxième terme de l'équation peut se reformuler $E_i(-x)$ pour fins de consultation des tables. Par exemple, pour le témoin 13 si x = 0,133, nous obtenons pour l'équation les valeurs

$$1,5676 \approx 1,5695$$

ce qui est une bonne approximation permettant le calcul de K qui est alors 1271.

RÉSULTATS ET INTERPRÉTATION

Les valeurs de V et de K

Pour des raisons évidentes le vocabulaire dans le discours oral est moins riche qu'à l'écrit. Par exemple, les textes les plus riches de notre corpus se comparent à *l'Étranger* de Camus, dont la prose est dans la moyenne par rapport à d'autres romans ou à des écrits didactiques.

$\overline{V}$ N tranches	500	1000	1500	2000
L'Étranger	201,0	329	436	530
Témoin 65	199,5	319	420	511

En observant les données du tableau I (en annexe), il est déjà possible de faire un classement des témoins à chacune des 8 tranches. Les valeurs de V pour 250 ≤ N ≤ 1000 sont plus fiables dans la mesure où il s'agit de moyennes. D'ailleurs, étant donné qu'on a pour chaque témoin 8 tranches de 250 mots, on peut mesurer le degré de constance ou d'homogénéité de chaque texte grâce au coefficient de variation. À ce sujet, le témoin 13, avec un écart type de 2,88 et un c.v. (coefficient de variation) de 0,0251, est le cas idéal - pratiquement pas de variation. Le témoin 4 est presque un cas problème pour notre échantillonnage : écart type 13,56 et c.v. = 0,1316.

Observons maintenant le classement selon le vocabulaire (V) aux points N = 250 et N = 2000, et selon le lexique en jeu (K) au point N = 2000 (les tranches inférieures ne remplissant pas toutes les conditions nécessaires à l'application de la formule de Kalinin).

Classement des témoins

Rang	1	2	3	4	5	6	7	8	9	10	11
Selon V											
N = 250	(115 65)	--	31	21	73	13	10	12	49	7	23
N = 2000	(73 65)	--	31	115	21	10	13	12	7	49	23
Selon K											
N = 2000	73	10	115	65	31	12	21	13	49	7	23

À côté des cas où le classement demeure stable : témoins 31, 49, 7, 23, 4, il y a des surprises, par exemple la remontée du témoin 73, le témoin 115 en perte de vitesse. La valeur K confirme la supériorité à long terme du #7 et fait voir que le #10 a en réserve bien plus de ressources qu'il n'en a fait voir. Dès lors on peut se permettre de faire les calculs de corrélatio avec un peu plus de confiance - ou de prudence - les anomalies ou contradictions éventuelles étant déjà pressenties.

3.2 *Les corrélations*

Le tableau II ne donne aucune preuve de corrélation entre richesse lexicale et instruction pour l'ensemble des témoins, ni entre instruction et lexique en jeu. Le contraire aurait été étonnant, compte tenu des conclusions de Sankoff (1975). Toutefois on se rend compte que c'est l'âge des témoins qui neutralise l'instruction dans ce domaine, et que ce soit avec l'indice t de Student-Fisher ou l'indice ρ, le vocabulaire de même que le lexique en jeu croissent avec l'âge. En isolant ce facteur par le regroupement des témoins en deux générations, on obtient sept fois sur dix des corrélations entre richesse lexicale ou lexique en jeu et instruction.

Au tableau III, la comparaison entre la performance des deux générations (qu'on a regroupées en veillant à ce que la moyenne d'instruction soit à peu près la même d'un côté comme de l'autre) donne des différences significatives dans tous les cas à l'avantage des "vieux".

De même, les témoins regroupés en classe économique "favorisée" (en fonction de la profession, du quartier, des ascendants) démontraient une richesse de vocabulaire et un lexique en jeu plus forts que leurs antagonistes. Bien sûr, il y a ici intervention de l'instruction comme 3^{e} terme à l'avantage du groupe "favorisé" d'autant plus facilement que l'âge a été neutralisé dans ce regroupement.

Les catégories

La distribution des catégories grammaticales est importante dans la mesure où l'on peut émettre l'hypothèse que, dans une langue donnée, elles ont une probabilité relativement stable, au point que les natifs deviennent sensibles à des écarts stylistiques reliés à la fréquence de ces catégories. Linguistes et stylisticiens y recourent non seulement pour comparer deux écrivains ou deux groupes d'écrivains (cf. les stylogrammes de Zemb) mais encore pour faire des rapprochements ou marquer les différences dans la structure quantitative de deux ou plusieurs langues.

Considérons les occurrences. Prise globalement, cette distribution (tableau V) montre des irrégularités qui ne sont pas dues au hasard, avec un χ^2 de 346 pour 77 degrés de liberté; la probabilité d'une distribution aléatoire est infime. Notons toutefois que ces différences ne sont pas importantes pour une longueur de texte de N = 250 par témoin (il aurait fallu χ^2 = 98,77 alors qu'il atteint à peine 45). Il est pour le moment difficile d'expliquer ces irrégularités. Nous n'avons opposé que les groupes socio-économiques. Si l'on excepte l'emploi des adjectifs, les différences constatées entre ces groupes ne sont pas significatives (tableau VII).

En faisant les mêmes calculs sur le nombre de mots différents (tableaux VI et VIII), en fonction des catégories, les différences sont plus marquées. Le groupe "favorisé" emploie une plus grande variété d'adverbes et d'adjectifs que le groupe "défavorisé". Il faudrait d'ailleurs se reporter aux

constatations déjà faites à la section 3.2. Ce même groupe avait un vocabu laire plus riche; les vocables qu'il possède en plus sont surtout des adver et des adjectifs.

Du point de vue syntaxique et stylistique les adverbes et les adjectif sont des éléments à valeur prédicative très forte et qui permettent de préc ser la pensée à peu de frais. La tradition scolaire leur accorde beaucoup d'importance, du moins dans l'expression écrite. Que le discours du groupe "favorisé" soit plus proche du discours académique, rien de plus vraisembla ble. Mais notre analyse du corpus n'est pas assez fine pour nous permettre de tirer cette conclusion.

CONCLUSIONS

La structure quantitative du vocabulaire n'est pas un ensemble de données aussi évidentes que les réalisations phonétiques et, à priori, les interlocuteurs n'en prendraient conscience qu'exceptionnellement et indirectement (indices de pauvreté ou de richesse extrême, taux de redondance, respect de conventions stylistiques). Néanmoins, dans les limites permises par notre sous-échantillon, on peut retenir certaines constatations que nous vérifierons ultérieurement dans l'ensemble du corpus :

1) À condition de la définir avec un peu de rigueur et de tenir compte de la progression en fonction de la longueur des textes, la *richesse lexical individuelle* est un facteur pertinent qui permet de regrouper les locuteurs en fonction de l'âge et, à l'intérieur des groupes d'âge, en fonction du niveau d'instruction. Elle intervient également pour distinguer les groupes socio-économiques définis en fonction de plusieurs critères.

2) La notion de *lexique en jeu* - dont nous sommes loin d'avoir exploit toutes les possibilités dans cet article - permet de faire de meilleures extrapolations sur les ressources lexicales des locuteurs et par conséquent de savoir dans quelle mesure le classement est définitif.

3) La fréquence relative des catégories grammaticales, bien qu'elle varie selon les témoins, n'intervient que faiblement pour distinguer des groupes socio-économiques. Par contre, il y a lieu de s'interroger sur l'emploi privilégié des adjectifs et des adverbes par le groupe "favorisé".

Nathan Ménard
Laurent Santerre
Département de linguistique
Université de Montréal

BIBLIOGRAPHIE

CARROLL, J.B. (1969), "Vectors of prose style", dans *Statistics and Style*, Elsevier, Dolezel and Bailey.

BEAUCHEMIN, N. et P. MARTEL (1972), *Recherches sociolinguistiques dans la région de Sherbrooke* (série de documents et d'études), Université de Sherbrooke.

VIKIS-FREIBERGS, V. (1974), *Fréquence d'usage des mots au Québec*, Montréal P.U.M.; Ménard, N. (1974), compte rendu dans *Livres et auteurs québécois*.

GUIRAUD, P. (1954), *les Caractères statistiques du vocabulaire*, Paris, P.U.F.

HERDAN, G. (1966), *The Advanced Theory of Language as Choice and Chance*, Springer-Verlag.

KALININ, V.M. (1965), "Functionals related to the poisson distribution and the statistical structure of a text", dans *Proceedings of Steklov Institute of Mathematics*, p. 79.

LABOV, W. (1976), *Sociolinguistique*, Paris, Éditions de Minuit.

MALÉCOT, A. (1976), "Fréquence d'occurrences des mots dans la conversation", dans *Revue d'acoustique*, n° 38.

MÉNARD, N. (1972), *Mesure de la richesse lexicale*, thèse dactylographiée, Université de Strasbourg (à paraître, Slatkine, Genève).

MULLER, CH. (1977), *Principes et méthodes de statistique lexicale*, Paris, Hachette.

ROY, G.-R. (1977), *Contribution à l'analyse du syntagme verbal. Étude morphosyntaxique et statistique des coverbes*, Paris, Klincksieck; Québec, P.U.L.

SANKOFF, D. (1975), "Vocabulary richness : A sociolinguistics analysis", dans *Science*, vol. 190, novembre.

SANKOFF, D. et G., *et al.* (1976), "Méthodes d'échantillonage et utilisation de l'ordinateur dans l'étude de la variation grammaticale", dans *la Sociolinguistique au Québec*, Montréal, P.U.Q., les Cahiers de l'Université du Québec, n° 6, p. 85-125.

SANTERRE, L. (à paraître), "la Disparition des voyelles hautes et la coloration consonantique en français québécois", *Actes du 8e Congrès international des sciences phonétiques*, 1975, Leeds, à paraître.

SANTERRE, L. et J. MILLO (1978), "Diphtongization in Montreal French", dans *Linguistic Variation : Models and Methods*, D. Sankoff (édit.), New York, Academic Press.

ZEMB, J.-M. (1970), "la Stylométrie", dans *la Stylistique, lectures*, Guiraud et Kuents (édit.), Paris, Klincksieck, p. 214-222.

ANNEXES : TABLEAUX STATISTIQUES

Symboles utilisés

N	nombre de mots (unités de texte); occurrences.
V	nombre de vocables du texte; mots différents.
K	lexique en jeu, d'après la formule de Kalinin.
s	écart type
c.v.	coefficient de variation
z	indice de comparaison des moyennes. Différence significative si $z \geq 2$, au seuil de 5 %.
ρ	indice de Spearman - corrélation des rangs.
χ^2	chi-carré (test de Pearson).

TABLEAU I

Vocabulaire (V) et lexique en jeu (K) pour une longueur de texte (N) - écart-type (s) - coefficient de variation (c.v.)

N			250	500	750	1000	1250	1500	1750	2000	
	s	c.v.	$\overline{V}$	$\overline{V}$	$\overline{V}$	$\overline{V}$	V	V	V	V	K
Témoin											
4	13,56	0,1316	103,12	161,25	203,50	248,50	282	308	342	374	692
7	8,08	0,0749	107,87	170,50	219,50	271,50	324	363	403	438	1164
10	7,62	0,0674	113,12	182,25	237,50	295,00	341	381	421	474	1680
12	5,18	0,0460	112,62	182,75	242,00	293,50	346	388	431	465	1305
13	2,88	0,0251	114,37	182,50	243,00	289,50	342	387	427	472	1271
21	7,95	0,0675	117,87	189,50	248,50	303,50	346	388	427	475	1289
23	7,35	0,0689	106,62	164,50	206,00	254,50	278	315	342	386	885
31	5,13	0,0423	118,62	193,25	259,00	311,00	369	410	451	488	1348
49	7,57	0,0687	110,12	175,75	223,50	274,50	292	343	385	429	1169
65	9,07	0,0747	121,50	199,50	261,00	319,00	366	420	468	511	1441
73	4,95	0,0420	117,75	195,00	259,00	317,50	373	416	465	511	1812
115	6,02	0,0496	121,50	190,50	257,00	304,00	348	406	441	476	1556

TABLEAU II

Calculs de corrélations

	Longueur de texte	Coefficient de Spearman	Seuil de 5 % Interprétation
1. Richesse lexicale et instruction (sans distinction d'âge)	N = 500 2000	+ 0,486 + 0,465	non significatif non significatif
2. Richesse lexicale et instruction (les jeunes: ≤ 30 ans)	N = 250 500 1000 2000	+ 0,828 + 0,830 + 0,830 + 0,714	significatif significatif significatif non significatif
3. Richesse lexicale et instruction (les vieux: > 30 ans)	N = 250 500 1000 2000	+ 0,700 + 0,943 + 0,886 + 0,843	non significatif significatif significatif significatif
4. Coefficient de variation dans le vocabulaire et instruction (sans distinction d'âge)	N = 250	- 0,206	non significatif
5. Coefficient de variation et instruction (les jeunes: ≤ 30 ans)	N = 250	- 0,310	non significatif
6. Coefficient de variation et instruction (les vieux: > 30 ans)	N = 250	- 0,600	non significatif
7. Lexique en jeu et instruction	N = 2000	+ 0,287	non significatif
8. Lexique en jeu et instruction (les jeunes)	N = 2000	+ 0,371	non significatif
9. Lexique en jeu et instruction (les vieux)	N = 2000	+ 0,830	significatif
10. Lexique en jeu et âge	N = 2000	+ 0,797	significatif

TABLEAU III

Richesse lexicale - lexique en jeu : comparaison de moyenne entre deux groupes d'âge

Longueur de texte	Groupe d'âge	Moyenne de V	Variance	Écart réduit	Seuil de 5 % Interprétation
N = 250	≤ 30 > 30	110,6 116,9	35,42 9,73	z = 2,10	différence significative
N = 500	≤ 30 > 30	175,6 188,9	162,23 23,52	z = 2,17	différence significative
N = 1000	≤ 30 > 30	276,3 304,1	545,31 70,45	z = 2,50	différence significative
N = 2000	≤ 30 > 30	435,0 481,5	2218,67 218,92	z = 2,11	différence significative
		Moyenne de K			
N = 2000	≤ 30 > 30	1104,0 1498,0		z = 2,78	différence significative

TABLEAU IV

Richesse lexicale - lexique en jeu : comparaison entre deux groupes socio-économiques

Longueur de texte	Groupes socio-économiques	Moyenne de V	Variance	Écart réduit	Seuil de 5 % Interprétation
N = 250	Favorisé Défavorisé	118,8 110,2	7,07 20,23	z = 3,77	différence significative
N = 2000	Favorisé Défavorisé	491,6 434,4	278,64 1455,10	z = 3,24	différence significative
		Moyenne de K			
N = 2000	Favorisé Défavorisé	1486,0 1169,0		z = 2,07	différence significative

TABLEAU V

Catégories grammaticales (occurrences)

Témoins	23	49	7	31	115	12	13	65	4	10	73	21	Total
Verbes	371	358	350	342	358	382	405	354	391	352	343	413	4419
Noms communs	236	289	320	271	300	278	310	310	242	309	355	259	3479
Adverbes	277	145	136	180	157	161	161	179	179	147	151	189	2062
Noms propres	21	25	6	19	22	9	32	10	9	9	22	7	191
Pronoms	428	373	391	396	373	414	326	369	489	371	303	448	4681
Déterminants	209	267	272	225	255	231	241	246	208	255	290	228	2927
Adj. qual.	86	82	94	109	105	93	117	136	74	109	101	97	1203
Autres catégories	372	461	431	458	430	432	408	396	408	448	435	362	5041
Total													24003

TABLEAU VI

Catégories grammaticales (mots différents : vocables)

Témoins	23	49	7	31	115	12	13	65	4	10	73	21
Verbes	86	86	83	100	96	100	88	105	82	89	93	104
Noms communs	123	144	178	173	166	174	170	194	116	184	195	174
Adverbes	52	52	41	60	63	53	53	58	53	50	57	53
Noms propres	15	18	6	14	15	8	18	9	8	8	15	6
Pronoms	25	27	26	26	26	25	23	24	26	28	22	32
Déterminants	13	23	18	18	16	11	14	13	17	19	23	14
Adj. qual.	40	44	53	60	55	59	69	75	39	60	65	54
Autres catégories	32	35	33	37	39	35	37	33	33	36	41	38
Total	386	429	438	488	476	465	472	511	374	474	511	475

TABLEAU VII

Distribution des catégories grammaticales : comparaison de moyennes entre deux groupes socio-économiques (OCCURRENCES)

	Groupes socio-économiques	Moyenne	Variance	Écart réduit	Seuil de 5 % Interprétation
Verbes	Favorisé Défavorisé	360,4 373,9	535,44 456,98	z = 0,93	différence non significative
Noms communs et pronoms	Favorisé Défavorisé	671,4 692,4	66,64 561,96	z = 1,18	différence non significative
Adverbes	Favorisé Défavorisé	165,6 176,3	139,04 2000,70	z = 0,56	différence non significative
Adjectifs	Favorisé Défavorisé	113,6 90,7	153,44 109,63	z = 3,02	différence significative
Adverbes et adjectifs	Favorisé Défavorisé	279,2 267,0	603,70 1861,71	z = 0,57	différence non significative

TABLEAU VIII

Distribution des catégories grammaticales : comparaison de moyennes entre deux groupes socio-économiques (MOTS DIFFERENTS : VOCABLES)

	Groupes socio-économiques	Moyenne	Variance	Écart réduit	Seuil de 5 % Interprétation
Verbes	Favorisé Défavorisé	96,4 90,0	33,84 63,14	z = 1,47	différence non significative
Noms communs et pronoms	Favorisé Défavorisé	203,8 183,1	130,96 730,98	z = 1,66	différence non significative
Adverbes	Favorisé Défavorisé	58,2 50,6	10,96 16,24	z = 3,27	différence significative
Adjectifs	Favorisé Défavorisé	64,8 50,1	48,16 63,26	z = 3,09	différence significative
Adverbes et adjectifs	Favorisé Défavorisé	123,0 100,7	27,20 67,38	z = 5,32	différence significative