Dictionnaire de fréquence du français

Michel LENOBLE

BEAUCHEMIN (Normand), MARTEL (Pierre), THÉORET (Michel) : 1992, ***Dictionnaire de fréquence des mots du français parlé au Québec : fréquence, dispersion, usage, écart réduit*** (American University Studies, Series XIII, Linguistics, Vol. 26, New York, San Francisco, Bern, Baltimore, Frankfurt am Main, Berlin, Wien, Paris, Peter Lang Publishing), LIII + 767 pp. ISBN 0-8204-1740-8.

Les trois professeurs et chercheurs à l'Université de Sherbrooke, au Québec, viennent de publier un dictionnaire de fréquence des mots du français parlé au Québec. Cette publication est en fait largement inspirée des fameuses études effectuées à l'Université Brown, dans les années soixante, pour l'anglais américain et, parallèlement, dans les années soixante-dix, pour l'anglais britannique, d'abord à Lancaster, ensuite à l'Université d'Oslo et au Centre de calcul de Bergen (LOB). D'ailleurs, dans la même tradition, Sture Allén a réalisé un travail similaire pour la langue suédoise, en 1971[1].

La présente publication s'inscrit aussi dans la lignée des travaux de Juilland[2], Imbs[3], Engwall[4] et Brunet[5]. Les publications de Juilland appartenaient en réalité au cadre plus large d'une collection d'études semblables pour le latin et les langues romanes et avaient été précédées à l'époque par des dictionnaires de fréquence du

[1] ALLÉN (Sture) : 1971, *Nusvensk frekvensordbok baserad p tidningstext.*

[2] JUILLAND : 1970, *Frequency Dictionary of French Words.*

[3] IMBS : 1969–1971, *Dictionnaire des fréquences — Trésor de la langue française.*

[4] ENGWALL : 1974, *Fréquence et distribution du vocabulaire dans un choix de romans français.*

[5] BRUNET (Étienne) : 1981, *Le Vocabulaire français de 1789 à nos jours* (Genève, Paris : Slatkine-Champion, collection « Travaux de linguistique quantitative », 17), pp. 852, 512, 454.

✉ Michel LENOBLE; Littérature Comparée; 3465, avenue Laval; Montréal; Québec H2X 3C7 (Canada).
E-mail : lenoblem@ere.umontreal.ca

roumain et de l'espagnol, et devaient être suivies par ceux du portugais et de l'italien. Seul le projet de dictionnaire de fréquence pour le portugais ne semble pas avoir abouti.

Le présent dictionnaire vient aussi complémenter l'ouvrage de Jean Baudot[6], dont une partie du corpus sous étude (37 %) était constitué de textes en français écrit provenant du Canada[7]. Langue écrite et langue orale peuvent ainsi être aisément comparées. C'est aussi une contribution significative à la description systématique des us linguistiques du Québec[8].

Le présent volume, préfacé par Jean-Denis Gendron, propose cinq listes différentes accompagnées d'une très importante introduction. Ce dictionnaire a été compilé à partir de dix tranches de corpus de langue parlée de 100 000 occurrences chacune. Le corpus total de 1 000 000 d'occurrences, compte un peu plus de 11 000 vocables différents. Les tranches se répartissent en deux sous-corpus de 500 000 occurrences, le premier consacré à la langue orale spontanée et le second à la langue orale non-spontanée (textes écrits pour être dits). L'ensemble du corpus a été lemmatisé.

La majorité des textes retenus dans les corpus datent des années soixante-dix. Quatre des cinq tranches de corpus de langue orale spontanée proviennent des années 1972 à 1978; seule la tranche de la Ville de Québec n'est pas datée. Les cinq tranches de langue orale non-spontanée s'inscrivent dans une fourchette de dates plus large : contes (1946–1978), théâtre (1971–1979), monologues (1956–1980), textes radiophoniques (1945–1969), téléromans (1954–1978).

Les cinq tranches de langue orale spontanée couvrent la presque totalité du territoire québécois : Montréal, Québec, Estrie, Saguenay Lac-Saint-Jean. L'Estrie est surpondérée puisque deux sous-corpus viennent de cette région, alors que la Beauce, Trois-Rivières, l'Outaouais, la Gaspésie, la Basse Côte-Nord, l'Abitibi, le Nouveau Québec sont absents de l'enquête. La provenance géographique du sous-corpus de langue orale non-spontanée n'est nullement mentionnée et ne semble pas avoir été un facteur de sélection. Par contre les chercheurs se sont efforcés de ne retenir que des locuteurs « québécois » et d'éliminer les « néo-québécois ». Rien ne précise si seuls les francophones ont été pris en compte, ou si des québécois anglophones ou allophones ont pu se glisser parmi les « informateurs ».

---

[6] BAUDOT (Jean) : 1992, *Fréquences d'utilisation des mots en français écrit contemporain* (Montréal : Presses de l'Université de Montréal), 431 pp.

[7] Pour de plus amples informations concernant cette publication, consulter notre compte rendu publié sous la référence suivante : LENOBLE (Michel) : 1992 « Répertoire des fréquences du français », *Revue. Informatique et Statistique dans les Sciences humaines*, 28, pp. 321–324.

[8] Différentes études descriptives des pratiques linguistiques du Québec se sont attachées à décrire d'autres aspects particuliers de la langue orale. Voir, entre autres, BAZERGUI (N.), CONNORS (K)., LENOBLE (M.), MAJKRAK (B.) : 1990, *Acquisition du français (L2) chez les adultes à Montréal : Études morphosyntaxiques et discursives* (Montréal, Office de la langue française) collection « Langues et sociétés », 163 pp.

Les auteurs ont réservé les premières pages de leur dictionnaire à une « Présentation » (pp. IX–XLVIII) de l'ouvrage. Ils y décrivent en détails le corpus (pp. XI–XV), le traitement du corpus (pp. XV–XXV) et l'analyse statistique (pp. XXV–XLVI). Dans la section suivante, « Bibliographie » (pp. XLIX–LIII), ils fournissent toutes les informations bibliographiques concernant les ouvrages théoriques ou de référence pour la statistique et la lemmatisation, les ouvrages disponibles similaires au leur, et les références des textes composant les divers sous-corpus.

Beauchemin, Martel et Théoret apportent un soin particulier à décrire, dans la partie « Traitement du corpus », les problèmes inhérents au traitement de corpus de langue orale et de mots dont la graphie n'est pas encore fixée par la norme. Ils précisent les choix méthodologiques qu'ils ont adoptés à ce sujet. Ils abordent de façon exhaustive toutes les facettes de la lemmatisation du corpus, sans éluder la question des noms propres, des interjections, des variantes graphiques, des verbes complexes, etc. Ils exposent clairement les principes discriminants dont ils se sont servis pour régler les cas tant de types morphologique que syntaxique, tout comme les cas d'homographes, d'anglicismes et de québécismes.

Dans la section intitulée « Analyse statistique », les trois auteurs traitent de huit sujets distincts : la fréquence, l'écart-réduit, la dispersion, l'usage, les résultats généraux, le vocabulaire de structure, l'étendue du vocabulaire et les anglicismes. Pour chaque sujet, ils précisent toujours leur définition des termes, ils établissent les principes de travail, explicitent leur processus de calcul, fournissent des exemples et commentent les résultats, le tout, dans un langage concis et accessible.

La majeure partie du dictionnaire est consacrée aux « Listes des vocables » (pp. 1–766). La première liste (p. 2) présente et explicite les indices que le lecteur retrouve dans la présentation des quatre listes principales du dictionnaire.

La liste alphabétique des vocables et des formes (pp. 3–678) donne à voir, pour chaque vocable et ses formes dérivées, les fréquences par tranche de sous-corpus, ainsi que les totaux pour la langue spontanée, la langue non-spontanée et le corpus complet. La liste mentionne également les écarts-réduits partiel et total, ainsi que les indices de dispersion et d'usage, là où des fréquences absolues atteignent des seuils significatifs.

La troisième liste (pp. 679–726) énumère les vocables triés par ordre de fréquence décroissante et accompagnés de la mention de la classe grammaticale à laquelle ils appartiennent.

La quatrième liste (pp. 727–745) reprend les vocables triés selon l'ordre décroissant des indices de dispersion. Chaque vocable est présenté avec sa classe grammaticale tout comme sa fréquence absolue. Seuls les vocables ayant une fréquence de 10 ou plus sont repris dans cette liste.

La dernière liste (pp. 747–765) mentionne les vocables triés cette fois-ci selon l'ordre décroissant des indices d'usage. Chaque entrée de la liste comporte également la fréquence absolue et la classe grammaticale du vocable.

Pour clore l'ouvrage, les auteurs ont disposé une « Page guide » (p. 766) présentant tous les champs de la liste principale, à savoir le dictionnaire de fréquence des vocables

et des formes, présenté par tranche, par sous-corpus et pour le corpus complet. Ils y font un rappel synthétique des paramètres retenus pour exécuter les calculs de tous les indices, sous-totaux et totaux, tout en spécifiant la signification de chaque champ composant les entrées du dictionnaire.

*
* *

Beauchemin, Martel et Théoret nous proposent ici un ouvrage en tous points admirable et d'une grande valeur scientifique. Fruit d'un travail long et exigeant, leur dictionnaire s'avère être un atout important tant pour les chercheurs en linguistique que pour les pédagogues du français L1 et L2. Cette publication vient combler un vide de façon magistrale, en plus de donner de la réalité linguistique de la langue française orale du Québec une description scientifique d'une précision et d'une rigueur inestimables. Cette étude rend possible enfin les études comparatives entre les langues françaises parlées de la francophonie, et entre langue parlée et langue écrite. Resterait peut-être à étudier la langue des dialogues romanesques afin de voir s'ils présentent les marques statistiques d'une parenté plus forte avec la langue orale ou la langue écrite. Les résultats de la recherche des trois linguistes de l'Université de Sherbrooke constituent une contribution importante à la connaissance des comportements statistiques des gros corpus ainsi qu'à la culture statistique dont nous avons besoin pour saisir les lois de fonctionnement des textes et percevoir en quoi certaines langues, certains niveaux de langue ou certains types discursifs s'écartent de la norme.

De toute évidence, les auteurs, qui ne sont d'ailleurs pas à leur coup d'essai dans le domaine puisqu'ils ont déjà publié des travaux sur la langue de l'Estrie, font preuve de beaucoup de savoir-faire et de métier. En effet, ils sont tout à fait conscients des problèmes que pose l'établissement d'un dictionnaire de vocables lemmatisés. Aucun de leurs choix théoriques ou méthodologiques n'est passé sous silence : chaque problématique est débattue sans naïveté. Les questions les plus complexes sont abordées avec maturité et avec la sagesse de ceux qui ont une longue fréquentation et une grande intelligence des phénomènes linguistiques et de leur traitement informatique. Chacune des sections de la partie « Analyse statistique » offre plus qu'une simple énumération de formules. Les auteurs apportent tous les éléments théoriques nécessaires à la compréhension des enjeux; la qualité de leurs explications et de la justification de leurs choix est à souligner. De plus, pour chaque indice, total, moyenne ou calcul, ils poussent beaucoup plus loin que ne le font les ouvrages similaires, les pistes d'analyse et d'interprétation des résultats présentés. Leurs commentaires et réflexions sont autant d'ajouts à la connaissance des caractéristiques des types discursifs représentés par chaque tranche du corpus. La section « Présentation » témoigne de qualités intrinsèques telles que celle-ci constitue, même prise séparément du dictionnaire qu'elle accompagne, une publication de référence pour ceux qui projettent entreprise similaire. C'est aussi un texte de référence pour ceux qui enseignent la lexico-statistique et la dictionnairique.

La présentation physique de l'ouvrage, la qualité typographique, la clarté de la mise en page, la lisibilité des listes et la maniabilité du livre en rendent la consultation et l'exploitation aisée et agréable.

*
* *

La composition du corpus peut faire l'objet de débats interminables et nous savons combien les ressources financières à la disposition des chercheurs leur imposent des limitations difficiles à justifier d'un point de vue scientifique. On ne peut donc que souhaiter à l'équipe de recherche du dictionnaire de pouvoir disposer de fonds suffisants pour compléter leur corpus en y ajoutant une série de tranches importante : d'une part, pour couvrir les zones géographiques ne figurant pas encore dans leurs relevés et, d'autre part, pour rendre justice à d'autres composantes inhérentes à la langue orale : le discours spontané, le discours quotidien, l'improvisation théâtrale, pour ne citer que ceux-là. On peut cependant difficilement souscrire, d'un point de vue scientifique, à l'argumentation des auteurs (p. XIII) quant à la justification de la surpondération des textes provenant de l'Estrie, cette région « méritait cet honneur ». On ne peut aussi se résoudre, la démonstration n'ayant point été faite, à accepter le commentaire manquant un peu de retenue épistémologique (p. XXVI) que « tout autre dépouillement de textes du même genre que les nôtres confirmerait, selon nous, les données recueillies et présentées ici ». Le lecteur aurait probablement aimé également être informé de la distribution par tranche d'âge et par sexe des locuteurs/informateurs qui ont collaboré à la constitution des corpus.

Le dictionnaire propose comme troisième liste un index des vocables par ordre de fréquence décroissante présentant un changement subit de présentation que rien ne vient motiver (pp. 687–689) et en contradiction avec ce qui est indiqué dans la « Présentation » (p. XXVII), ladite liste poursuit son énumération jusqu'à la fréquence la plus basse 1 et ne s'arrête pas à la fréquence 35, ce qui eût d'ailleurs été ardu à justifier d'un point de vue lexico-statistique.

On peut regretter que les auteurs n'aient pas produits d'étude de distribution des occurrences par fréquence afin de voir dans quelle mesure la fonction de Zipf se vérifiait ici également. Dès lors que le corpus général se divise en deux parties distinctes, il eût été aisé de faire profiter le chercheur de résultats de comparaison des corpus en terme de vocabulaire absent dans une des deux composantes[9]. Il était somme toute aisé d'obtenir de telles informations, une fois le travail de préparation du corpus terminé.

---

[9] Une étude particulièrement intéressante du « vocabulaire absent » est celle qu'en fait Étienne Brunet dans le cadre de sa recherche plus large concernant le vocabulaire de Proust. Voir BRUNET (Étienne) : 1983, *Le vocabulaire de Marcel Proust avec l'index complet et synoptique de « À la recherche du temps perdu » d'après les données de l'Institut National de la langue française* (Genève : Slatkine), 3 vol., 1 918 pp.

Dommage aussi que les chercheurs sherbrookois ne donnent pas plus d'informations concernant le lemmatiseur semi-automatique dont ils font mention (p. XXVI). Références complètes, présentation des principes du programme, références de rapports techniques ou d'articles qui y seraient consacrés, informations techniques concernant la programmation, les plate-formes pouvant le supporter et disponibilité ou distribution n'auraient pas été superflus.

La bibliographie n'offre aucune référence à des publications scientifiques concernant la lemmatisation et le «Dictionnaire orthographique inverse et fréquentiel» y est attribué de façon un peu «servile» à Beaudot J., reprenant ainsi l'information évidemment erronée figurant sur la couverture et sur la deuxième de couverture du dit ouvrage. La différence d'épellation du nom de l'auteur entre celle mentionnée sur ces deux pages, et celle indiquée à la première page du tapuscrit (p. I) et dans le reste de l'ouvrage aurait dû les amener à se poser la question de savoir s'il ne fallait pas plutôt attribuer l'ouvrage au linguiste et lexicographe connu, Jean Baudot. Ont-ils donc réellement consulté l'ouvrage ?

Il est malheureux de constater qu'un ouvrage de cette qualité contienne quelques imperfections : les renvois internes en chiffres romains figurant aux pages XII et XXXIX sont erronés et doivent être remplacés respectivement par des renvois aux pages XL et XLI; les titres de colonne du deuxième tableau de la partie de droite de la page XXXII sont décalés d'un tabulateur vers la gauche; la dizaine de fautes d'orthographe et de coquilles sont inattendues et regrettables. Le peu d'effort qu'aurait exigé une dernière révision du manuscrit de l'introduction aurait certes procuré un caractère plus fini à la publication. Il est enfin amusant de constater qu'une équipe travaillant à la lemmatisation de corpus utilise le mot «espace» au masculin (p. XVIII), alors qu'il est ici clairement repris dans son acception typographique qui est, faut-il le rappeler, de genre féminin.

Le dictionnaire de fréquence des mots du français parlé est, hors de tout doute, un ouvrage de très grande valeur scientifique et témoigne d'un savoir-faire indéniable de l'équipe rédactionnelle. Il sera un outil de travail précis, sérieux et clair tant pour le public large des linguistes, psychologues, anthropologues et pédagogues, que pour les spécialistes de la lexico-statistique. Son introduction, d'une rare intelligence et d'une grande honnêteté scientifique est un modèle à suivre. Les commentaires et les pistes d'interprétation des résultats statistiques préliminaires sont une invitation à consulter les analyses que les auteurs pourraient publier par ailleurs à partir des recherches et des travaux qu'ils ont effectués pour la rédaction du dictionnaire. Il est à espérer que l'équipe de Beauchemin, Martel et Théoret dispose des moyens matériels et financiers pour poursuivre leurs travaux fondateurs et fondamentaux pour la recherche linguistique de la francophonie et du Québec et ainsi puisse donner un portrait aussi précis que possible de l'ensemble du spectre des activités langagières.