Répertoire des fréquences du français

Michel LENOBLE

Jean BAUDOT, *Fréquences d'utilisation des mots en français écrit contemporain*, Montréal, Presses de l'Université de Montréal, 1992, 431 pp. ISBN 2-7606-1563-4. $ 38. Distribué par Gaëtan Morin éditeur, C.P. 180, Boucherville (Québec) – CANADA J4B 5E6[1].

Jean Baudot, professeur au Département de linguistique de l'Université de Montréal, vient de publier un ouvrage répertoriant les fréquences d'utilisation des mots en français écrit contemporain. Cette recherche 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[2].

La présente publication s'inscrit aussi dans la lignée des travaux de Juilland[3], Imbs[4], Engwall[5] et Brunet[6]. 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 roumain et de l'espagnol et devaient être suivies par ceux du portugais et de l'italien. Seul le projet de dictionnaire

[1] Une version anglaise de ce compte rendu sera prochainement disponible dans la revue électronique *SCHOLAR. On line Information for Natural Language Processors*.
E-mail : scholar@cunyvm.bitnet.
Cette revue électronique est dirigée par Jo RABEN. E-mail : jqrqc@cunyvm.cuny.edu [N.D.É.].

[2] *Nusvensk frekvensordbok baserad på tidningstext*, 1971.

[3] *Frequency Dictionary of French Words*, 1970.

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

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

[6] *Le Vocabulaire français de 1789 à nos jours*, 1981

✉ Département de Littérature Comparée; Université de Montréal; C.P. 6128; Succ. « A »; Montréal (Québec); CANADA H3C 3J7.
E-mail : lenoblem@ere.umontreal.ca

de fréquences pour le portugais ne semble pas avoir abouti. Étonnamment, hormis l'étude de Brown, Baudot ne mentionne aucune de ces recherches, pas même celles portant sur des corpus francophones.

Le volume de Baudot, préfacé par Maurice Gross, présente en fait cinq listes différentes accompagnées d'une importante introduction. Ce répertoire a été constitué à partir d'un corpus de 803 échantillons de textes d'environ 1 000 à 1 500 mots chacun, pour un total de 1 040 150 occurrences lemmatisables. Les textes, dont la majorité ont été rédigés entre 1960 et 1967, se répartissent en 15 genres discursifs, exactement comme lors des études précitées d'Allén et des corpus de Brown et de LOB. D'après l'introduction, les origines nationales des échantillons se distribuent entre la France (62 %), le Canada (37 %) et les autres pays (1 %).

Les collaboratrices de Jean Baudot signent la partie introductive (pp. 13–22) consacrée à la « Présentation du répertoire » où sont décrits très clairement les divers champs des différentes listes ainsi que les principes qui ont présidé aux travaux de réalisation du répertoire et notamment à la lemmatisation. Elles nous précisent que le dictionnaire servant de « norme lexicographique » pour déterminer le découpage des entrées discrètes isolées lors de la préparation du corpus a été le *Petit Robert* (1985) et occasionnellement, elles ont eu recours au *Petit Larousse*, au *Grand Dictionnaire encyclopédique Larousse* (1982) et au *Grand Robert de la langue française* (1985). Les auteurs indiquent aussi en quelles circonstances elles se sont écartées de la nomenclature lexicographique. De plus, l'équipe de recherche n'a pas conservé dans ses relevés les noms propres de lieux, de personnes ou de marques de commerce ainsi que les expressions de langues étrangères.

Baudot rassemble en un bref passage (pp. 23–25) les quelques constatations qu'il tire des résultats statistiques de son étude.

La première liste (pp. 26–210) énumère en ordre alphabétique tous les mots lemmatisés accompagnés de leur fréquence absolue, de leur catégorie grammaticale et éventuellement d'un renvoi au lexique. La liste compte 21 684 types.

La deuxième liste (pp. 211–394) reprend les mêmes informations que celles présentées dans la première, mais triées, cette fois-ci, par ordre décroissant de fréquence.

Le lexique (pp. 395–403), troisième liste, nous donne des renseignements supplémentaires expliquant ponctuellement, là où il importait de pallier la trop grande généralité des principes de lemmatisation, la façon dont les entrées ont été lemmatisées.

Quatrièmement, la liste statistique (pp. 404–414), disposée en cinq colonnes, répertorie pour chaque fréquence d'occurrence, le nombre de mots ayant cette fréquence, le nombre cumulatif de mots ayant une fréquence égale ou supérieure, le nombre cumulatif d'occurrences ayant une fréquence égale ou supérieure et le pourcentage cumulatif du total des occurrences.

La dernière liste (pp. 415–430) mentionne les références des textes d'où ont été puisés les échantillons constituant le corpus.

*
* *

La qualité de la présentation typographique, la facilité de navigation parmi les diverses listes et la clarté des explications introductives témoignent du savoir-faire

indéniable de l'équipe de recherche et font de ce volume un outil de travail pratique. Le choix de la définition du mot et la sélection d'une norme lexicographique ont été faits uniquement en fonction de critères opératoires, ce qui permet à cette publication de présenter, en ce qui concerne ses diverses listes, une qualité constante.

Le répertoire de Baudot se veut représentatif du français écrit contemporain, ce qui peut paraître bizarre puisque la majorité des textes ont une date de publication qui remonte au moins à un quart de siècle. Il nous semble que la communauté scientifique serait en droit de s'attendre, vu le nombre de textes tout récents disponibles sur support électronique, à un corpus « plus contemporain », datant de ces dernières années.

Contrairement aux études de Engwall (1962–1968), de Hofland et Johnsson (LOB–1961), de Kucera et Francis (Brown–1961) et de Allén (1965) qui portaient sur des corpus comprenant des textes pour la plupart publiés la même année, le répertoire de Baudot recense des textes s'étalant sur plus de 55 ans. L'auteur n'indique néanmoins aucune intention de réaliser une étude diachronique comparable à celle de Brunet (1789 à aujourd'hui). Cette non-homogénéité chronologique se double d'une hétérogénéité géographique des lieux de production des textes. Rien n'indique cependant dans la préface que l'équipe de recherche voulait produire un répertoire représentatif de la francophonie. Si tel avait néanmoins été le cas, la distribution souffrirait alors d'écarts flagrants de sur- et de sous-représentation. Puisque le but de l'entreprise n'était ni de produire une étude de l'ensemble de la francophonie, ni de réaliser une étude diachronique, il eût été, en conséquence, néanmoins relativement facile d'élaguer le corpus afin de le rendre plus cohérent géographiquement et chronologiquement.

En contradiction avec ce qui est indiqué dans l'introduction (p. 14), la lecture rapide des dates de publication des sources nous donne à voir seulement 63 textes datant d'avant 1960 et non pas 77. Les calculs à cet égard sont d'autant plus aléatoires que 36 références de source ne comportent aucune date de parution et que le texte n° 198 (p. 418) manque dans l'énumération. Par ailleurs, le relevé de la distribution géographique de provenance des textes est également erroné (p. 14) puisque 9 textes y sont attribués à des pays autres que la France et le Canada. Or, un parcours rapide de la cinquième liste a permis de noter au passage entre autres 13 références à la *Libre Belgique*, 8 références à des « Marabout Flash », 3 références à des romans de Simenon, sans compter deux rapports de la *Kredietbank* émanant tous de Belgique. C'est donc une relative consolation de savoir que le gros du travail a été réalisé sur ordinateur, vu les difficultés remarquées dans les décomptes « manuels » présentés dans l'introduction.

On peut difficilement souscrire, de nos jours, à l'argumentation des auteurs (p. 14) quant à la nécessité de s'en tenir à « un corpus de taille raisonnable » (un million de mots) en partie « à cause des différents traitements informatiques à effectuer », étant donné que, par exemple, le *TLF* a géré, avec des machines plus anciennes, des corpus dont la taille était plus de 100 fois supérieure. On ne peut non plus se résoudre à accepter la dénomination « genres littéraires » (p. 14) attribuée à des publications comme des revues, des magazines, des manuels, des journaux, des brochures, des circulaires, etc. Il eût été plus heureux de retenir le terme « genres discursifs ».

Même s'il est intéressant de comparer les résultats globaux du livre de Baudot avec ceux des autres études du domaine francophone et de vérifier si les mêmes tendances et fonctionnements de l'économie des textes se retrouvent également dans les études portant

sur d'autres champs linguistiques, on ne trouve que trop peu de commentaires analytiques (pp. 23–25) des statistiques proposées dans ce volume. À défaut, les auteurs auraient peut-être pu suggérer des pistes de recherche pour les utilisateurs du répertoire en les invitant, ceci n'étant qu'exemple, à travailler à partir de données comme les premiers mots, noms communs, parmi les plus fréquents (p. 211), à savoir : *homme*, *monsieur*, *an*, *temps*, *vie*, *monde*, etc.

Il est dommage qu'à chaque entrée de la liste des sources ne soit pas ajoutée sa catégorie discursive. Par ailleurs, la liste statistique ne mentionne malheureusement pas les fréquences relatives pour les occurrences et ne fournit aucune comparaison des résultats (rapport type/occurrence) pour le même corpus lemmatisé et non-lemmatisé. Aucun résultat n'est calculé en fonction des genres discursifs différents, cette variable n'entrant, de toute évidence, jamais en ligne de compte dans la présente étude. On peut le regretter, car une fois le travail de préparation du corpus terminé, il est relativement aisé d'obtenir de telles informations, qui sont, somme toute, à portée de la « main/clavier ». Il faut en conséquence croire qu'il n'entrait pas dans l'intention des auteurs de réaliser ces types de recherche; ils indiquent par ailleurs aux lecteurs que les listes informatiques permettant néanmoins de poursuivre l'étude sont disponibles sur demande. Dès lors, on voit difficilement ce que vient apporter de neuf à la lexicométrie en général un ouvrage qui ne se démarque en rien (ni par la date du corpus, ni par l'origine des textes [par ex. aucunement spécifiques au français d'Amérique du Nord], ni par le type de résultats proposés) de ceux dans la lignée desquels il semblait vouloir s'inscrire.

Fort heureusement, cet ouvrage vise un public plus large (linguistes, psychologues, anthropologues, professionnels des industries de la langue, personnels des départements d'éducation) que celui des spécialistes de la lexico-statistique, public aux yeux duquel il devrait avoir plus que le simple mérite d'exister. Le répertoire des *Fréquences d'utilisation des mots en français écrit contemporain* est donc à utiliser en gardant à l'esprit les limites et les lacunes de sa conception et de sa réalisation. Il restera toujours, pour les plus exigeants, la possibilité de poursuivre le travail à partir des listes qui sont disponibles sur demande.