Compte rendu

Ouvrage recensé :

Fréquences d'utilisation des mots en français écrit contemporain. Jean Baudot, 1992, Les Presses de l'Université de Montréal, 432 p.

par Hubert Séguin

Revue québécoise de linguistique, vol. 22, n° 2, 1993, p. 179-181.

Revue québécoise de linguistique, vol. 22, n° 2, 1993, © RQL (UQAM), Montréal

FRÉQUENCES D'UTILISATION DES MOTS EN FRANÇAIS ÉCRIT CONTEMPORAIN

Jean Baudot, 1992, Les Presses de l'Université de Montréal, 432 pages

Hubert Séguin

Au seuil de l'âge d'or du bilinguisme et du biculturalisme au pays, peu après la création vers 1964 du *Bureau des langues de la Fonction publique du Canada* – dont la vocation était d'élaborer des cours et des tests pour nos deux langues *officielles* –, l'équipe du testing, alors dirigée par le regretté Gérard W. Charbonneau secondé par l'auteur de ces lignes, a entrepris un gigantesque travail descriptif du français pour se doter de critères objectifs de sélection et de répartition des éléments à mesurer, suivant le courant de pensée d'alors. Venait juste d'être élaboré aux États-Unis le *Computational Analysis of Present-Day American English* de Francis & Kucera qui offrait un *modèle* d'enquête sur la fréquence d'emploi des mots usuels, en déterminant les bases d'un corpus (*écrit*) qui reflétât une langue fonctionnelle, c'est-à-dire bien représentée dans ses différents styles ou domaines d'utilisation: journalistique, administratif, scientifique, littéraire. L'occasion était fort opportune d'appliquer le modèle anglais (américain) au français de ces mêmes années soixante, corrigeant ainsi quelques lacunes forcément laissées par l'*Élaboration du français élémentaire* (1959) – peu après nommé *fondamental* – portant sur le français (*parlé*) en France, sans mentionner les enquêtes antérieures de Hanmon (1924), de Vander Beke (1929), etc., toutes consacrées au français *normatif*, c'est-à-dire écrit par les bons auteurs.

Cette entreprise de la Fonction publique, qui fut menée à bien et complétée vers 1968, grâce à une équipe dynamique et avec l'intelligente collaboration de l'informaticien-linguiste Jean Baudot de l'Université de Montréal, n'a jamais été livrée au public par les instances officielles qui l'avaient financièrement généreusement nourrie et, comme beaucoup d'autres projets reliés à une conjoncture politique, elle allait sombrer dans l'oubli, une fois les enthousiasmes initiaux refroidis. Il convient donc ici de rendre hommage à la remarquable persévérance de l'informaticien-linguiste qui, après maintenant le quart de siècle qui s'est écoulé depuis l'établissement du corpus, rend enfin publics les résultats uniques de cette importante enquête de vocabulaire.

Les *Fréquences d'utilisation des mots en français écrit contemporain* est l'oeuvre d'un spécialiste «qui connaît bien les aspects théoriques du problème» (Maurice Gross, Préface). Cet ouvrage nous donne, après les introductions d'usage et *quelques remarques sur la présentation et la lemmatisation des mots de classe ouverte et la catégorisation des mots grammaticaux* ainsi que *quelques statistiques*, une *liste alphabétique* des vocables (mots de textes, lemmatisés), une *liste en ordre de fréquences* de ces mêmes vocables, un *lexique*, une *liste statistique* et une *liste des sources du corpus*. Les quelque 22 000 vocables, relevés dans les quelque 800 échantillons de textes qui constituent le corpus, sont présentés en deux grandes listes – alphabétique et fréquentielle – de 185 pages chacune, indiquant, pour chaque entrée, sa fréquence d'emploi (ou d'*occurrence* dans les textes choisis), sa catégorie grammaticale et, occasionnellement, une référence au *lexique*; celui-ci donne le détail des formes de quelque 150 entrées complexes (emplois simples/locutionnels, par exemple: 338 **abord** n = 17 **abord** n, 321 **d'abord**, loc adv; ou mots à variantes en genre et nombre, par exemple: 147 **auquel** pr rel = 77 **auquel**, 39 **auxquels**, 31 **auxquelles**; ou mots à variantes de forme, par exemple: 68 373 **de** prép = 52 060 **de**, 16 313 **d'**). Les vedettes (ou entrées) et les catégories grammaticales sont établies, avec une certaine servitude, d'après les critères adoptés par la maison Robert. Le lexique est suivi d'une *liste statistique* dont les données numériques concernent la fréquence d'emploi des mots: ce tableau de chiffres indique par exemple que le mot le plus fréquent (68 373 occurrences) représente à lui seul près de 7% du total du corpus (1 040 150 occurrences), que les 6 premiers mots (les deux verbes-auxiliaires **être** et **avoir**, les deux prépositions **de** et **à**, les deux articles **le** (le,l',la,les) et **un** (un,une,des), en représentent un bon 20%, que moins de 50 mots comptent pour la moitié d'un texte, et moins de 3 000, pour les 90%; à l'inverse, les mots d'une seule occurrence, dont l'apparition est purement aléatoire, déterminée par le choix des échantillons, représentent un peu plus d'un quart du vocabulaire (6 000 vocables sur les 21 700 du corpus) et seulement la moitié de un pourcent d'un texte d'un million de mots.

Toutes ces valeurs statistiques sur le vocabulaire du français dit contemporain dépendent bien sûr de la valeur représentative du corpus. Ce dernier, tel que détaillé dans la *Liste des sources* en fin de volume, comprend en bonne majorité (66%) des textes de journaux et de revues des années 66-67, période où fut effectuée l'enquête, avec cependant quelques textes – surtout littéraires – des décennies précédentes, qui font, disons-le, un peu tache d'huile sur la contemporanéité des échantillons du corpus (1954: *Les mandarins* de Simone de Beauvoir; 1943: *Adagio* de Félix

Leclerc; 1934: *Clochemerle* de Gabriel Chevalier; 1925: *Le désert de l'amour* de François Mauriac; 1910: *Esclave ou reine* de Delly; et quelques-uns sont sans date). Il n'en reste pas moins heureusement que 90% des textes (726 sur 803) sont parus vers la même époque de cette «révolution tranquille» commencée au tout début des années soixante après celles dites de l'«obscurantisme».

Les échantillons de textes ont une longueur moyenne de 1 200 mots et participent de 15 genres d'écrits (divisés en sous-genres). La répartition nationale comprend 62 % de textes publiés en France, 37%, au Canada et 1% (9 textes), ailleurs. Quant aux types de publication, 42% des textes viennent de revues et de magazines, 25%, de livres et de manuels, 24%, de journaux, 7%, de bulletins et de rapports et 2%, de brochures et de circulaires.

«La structure du corpus peut être critiquable mais elle constitue, selon nous, un échantillon représentatif du français écrit contemporain. Des textes scientifiques y côtoient des romans, des textes administratifs, des textes légaux, des extraits de journaux et de périodiques, des récits, etc. (...). Quant à la fiabilité de la représentativité du vocabulaire, elle ne peut s'appliquer qu'aux mots dont la fréquence d'occurrences dépasse un certain seuil statistique.» (Cf. pp. 14-15). Cette remarque sur la «fiabilité de la représentativité du vocabulaire» est tout à fait juste, nous semble-t-il, pour les quelque 4 000 mots «d'usage courant», c'est-à-dire pour ceux qui, dans un texte d'un million de mots, «représentatif» de la diversité de la langue écrite, apparaissent, disons, au moins 20 fois – pour s'aligner sur l'enquête du *Français élémentaire*. Ces quelques milliers de mots de la langue usuelle de ce siècle nous semblent maintenant solidement entérinés par cette excellente enquête de vocabulaire – qui ajoute à l'objectivité des précédentes. Quant au vocabulaire «in» du français «contemporain», celui d'après mai 68, il serait à souhaiter qu'une nouvelle enquête, suivant l'heureux modèle de celle-ci, nous dise s'il faut désormais «faxer» ou «télécopier», ce qui est arrivé à nos puces et à nos souris et où nous en sommes avec nos directeurs et nos professeures.

Hubert Séguin
Université d'Ottawa