Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Étienne BRUNET, Ce qui compte. Écrits choisis, tome II. Méthodes statistiques, *Champion*, 2011, 373 pages, ISBN 9782745322258.

Lu par **Gérald PURNELLE**

CIPL, Université de Liège, Belgique

*Choix de 17 articles d’Étienne Brunet, de 1970 à 2009, qui abordent les questions théoriques et méthodologiques dans le domaine de la statistique lexicale et de la lexicométrie. À la fois histoire de la discipline et guide méthodologique, le volume constitue l’ouvrage de réflexion d’un des meilleurs spécialistes de la discipline.*

Dans la suite directe de *Comptes d’auteurs, Études statistiques, de Rabelais à Gracq*, volume dans lequel Damon Mayaffre avait tiré de la bibliographie d’Étienne Brunet un choix d’études appliquées et de monographies consacrées à divers auteurs, ce deuxième tome de ses *Écrits choisis* rassemble dix-sept contributions de portée théorique ou méthodologique — un choix judicieux, opéré « avec l’approbation du maître », et qui se signale par sa richesse et sa diversité.

Certes, Étienne Brunet n’appartient pas tout à fait à la génération des tout premiers pionniers de la statistique linguistique, mais il a rapidement rejoint ceux qu’il considère comme ses maîtres, à commencer par le père fondateur de la discipline en France, Charles Muller. Littéraire venu à l’informatique puis à la statistique à la fin des années 60, Étienne Brunet fait depuis longtemps figure de référence dans le domaine, respecté et écouté, auquel les deux volumes forment à la fois « hommage et témoignage » (Céline Poudat dans l’avant-propos).

On trouve en tête d’ouvrage deux véritables documents : le tout premier article d’Étienne Brunet, intitulé « Programme » et daté de 1970, où était décrit un programme d’ordinateur calculant la fréquence théorique et l’écart réduit pour chaque mot d’un corpus constitué de seize œuvres de Giraudoux ; et l’introduction de la thèse qu’Étienne Brunet a soutenue en 1976 et qui était consacrée au même auteur. Par-delà le côté anecdotique et historique de ces deux textes, on y observe d’emblée la convergence immédiate et définitive de trois compétences en un seul homme : le littéraire, l’informaticien et le statisticien, que l’on sent en plus d’un endroit passer avec un plaisir intellectuel intact de la formule statistique au clavier et du code au graphique. À deux reprises, l’auteur justifie le choix de conjoindre les trois compétences en un seul homme, plutôt que de confier la programmation à un informaticien. (Et lui-même continue à illustrer ce point de vue, à travers le développement permanent de son logiciel d’exploration et de statistique textuelle Hyperbase, dont une version figure sur un dvd joint à l’ouvrage.)

Mais ces premiers textes montrent en outre les grandes qualités pédagogiques dont Étienne Brunet fera toujours preuve et qui se retrouve dans chaque chapitre. Dans chaque article il explique, détaille, illustre, se met à la portée du lecteur, quel qu’il soit. C’est donc aussi à un grand pédagogue que cet ouvrage rend hommage.

Classés dans l’ordre chronologique, ces dix-sept textes esquissent par l’exemple une histoire de la discipline, de l’informatisation des textes littéraires (constitution des grandes banques données, le TLF, Frantext) à leur analyse statistique, vite baptisée lexicométrie.

C’est aussi une histoire, partielle mais instructive, des débats, voire des polémiques, théoriques et méthodologiques, qui ont agité théoriciens et praticiens de l’analyse statistique textuelle. Chacune de ses interventions montre combien Étienne Brunet fut et reste animé par une réflexion permanente sur la discipline, une vision globale de son évolution et de son avenir. Constamment il revient sur les méthodes de ses collègues pour les discuter, les comparer, les amender parfois, les utiliser et les implémenter dans son propre logiciel. Ses prises de position sont toujours courtoises et ouvertes, mais fermes.

Citons à cet égard l’article très technique (sur le plan mathématique) où il prône l’utilisation de la loi normale plutôt que la loi hypergéométrique dans le traitement des grands corpus (chapitre 4), ou celui où le « schéma d’urne » est discuté, pour avoir été la source d’un débat qui remettait en cause le rôle (statistique) du hasard dans la constitution du vocabulaire d’un texte.

La dimension à la fois historique et réflexive de l’ouvrage en fait presque l’équivalent d’un manuel ou d’une introduction théorique ou méthodologique, même si le degré de difficulté varie d’un texte à l’autre. Au point qu’il aurait peut-être été intéressant de classer ces textes dans un ordre « pédagogique » plutôt que chronologique, en allant des questions générales et des textes introduisant la discipline jusqu’aux plus pointus. Mais l’ordre chronologique réserve un joli effet de bouclage, l’avant-dernier chapitre, « Plaidoyer pour la statistique linguistique », renvoyant comme une mise à jour au 2e chapitre, qui « dessine les contours d’une statistique linguistique » (Céline Poudat).

Quant aux méthodes et objets auxquels Étienne Brunet s’est intéressé tout au long de sa carrière, on sait qu’il est l’homme des amples corpus et de la statistique lexicale. Celle-ci mobilise comme objets et concepts le mot, la fréquence, le vocabulaire, la spécificité lexicale et la distance intertextuelle (chapitres 12 et 17).

Les articles rassemblés illustrent combien la banque de données Frantext constitue le corpus de référence qu’inlassablement il explore, découpe, et prend pour référence depuis de nombreuses années. Soit il étudie toute la littérature à partir de Frantext, notamment dans son découpage chronologique (chapitre 3), soit il se concentre sur des corpus exhaustifs d’auteurs, en comparant les sous-corpus dont ils sont constitués ou en confrontant l’auteur au corpus de référence. À tous égards, les dimensions qui sont observées sont régulièrement le genre, la diachronie, la forme (vers ou prose).

Mais Brunet ne s’en est pas tenu au vocabulaire du texte (plus précisément, l’ensemble des « graphies » qui le constituent, chacune étant considérée indépendamment des autres, dans de simples dénombrements) comme substance de ses études et méthodes. Signalons sa contribution à la question de la lemmatisation : faut-il lemmatiser les textes (ce qui est coûteux ou source d’erreur) ? Le gain de précision statistique est-il assuré ? Le chapitre 9 (« Qui lemmatise dilemme attise ») apporte un éclairage surprenant : le volume des corpus mobilisés a pour conséquence que les tests statistiques sont peu sensibles à l’opposition lemmatisation / no-lemmatisation. Mais en d’autres cas le lemme est un réel apport, auquel l’auteur ne renonce pas.

Le lemme n’est pas la seule information linguistique attachée à la forme du mot : les critères morphosyntaxiques, que les méthodes modernes d’étiquetage peuvent produire, apportent un enrichissement à la statistique textuelle.

Une troisième voie, particulièrement riche, par laquelle Étienne Brunet dépasse le mot, tend à appréhender la linéarité du texte en passant du mot comme simple occurrence dans le texte (que l’on peut dénombrer, pondérer, etc.) à la séquentialité de celui-ci : cooccurrences, rafales, séquences et enrichissement du vocabulaire (chapitres 6, 13, 15)

Enfin, on relèvera combien Étienne Brunet apprécie l’exercice complexe de la comparaison des méthodes (aux chapitres 11, 12 et14, ou dans le chapitre 17, qui est un hommage à Charles Muller).

Terminons par le meilleur : le style clair et naturel d’Étienne Brunet s’enrichit d’un humour subtil et sans complexe : clins d’yeux et jeux de mots émaillent plus d’un texte. À cet égard, laissons la parole à l’informaticien qui double le littéraire : « La machine doit s’irriter très fort qu’on la compare une fois de plus à une personne et qu’on lui prête des sentiments » (p. 31).