Les mystères des vocables

Étienne ÉVRARD

COSSETTE (André) : 1994, ***La richesse lexicale et sa mesure*** (Paris–Genève, Champion-Slatkine, «Travaux de linguistique quantitative», n° 53), 194 pp. ISBN 2–85203–354–2.

La notion de richesse lexicale, ou richesse du vocabulaire, est souvent utilisée dans l'analyse littéraire; malheureusement elle est fort vague. André Cossette (A.C. dans la suite) consacre la 1re partie de son livre à tenter de la débrouiller et examine de quelle manière des techniques quantitatives pourraient la préciser. Il observe que les facteurs qui entrent dans la composition de la richesse lexicale sont nombreux; il insiste particulièrement sur la distinction entre une richesse thématique (certains thèmes disposant d'un lexique plus abondant que d'autres) et une richesse stylistique, mais même cette distinction, théoriquement claire, se perd dans la complexité des faits. Quant aux indices quantitatifs proposés, ils reposent sur le fait que le nombre d'occurrences et celui des vocables augmentent au fur et à mesure, mais selon une fonction difficile à définir : non seulement le vocabulaire croît plus lentement que le nombre d'occurrences, mais il croît de plus en plus lentement. Dès lors, la plupart des chercheurs qui se sont préoccupés du problème ont cherché à définir un indice capable d'indépendance par rapport à la longueur des textes. A.C. cite rapidement, selon l'ordre chronologique, la caractéristique K de Yule, l'indice R de Guiraud, celui de Herdan, l'utilisation de la loi binomiale par Ch. Muller et par moi[1], et enfin les indices de Ét. Brunet et de D. Dugast. Comme pour la notion de richesse lexicale, A.C. insiste sur le fait que l'évaluation des

[1] A.C., p. 19, n. 58 et p. 111, n. 196 et 197, note que cette application a été proposée simultanément et indépendamment par Ch. Muller et par moi, mais il néglige de donner la référence du 1er article que j'ai écrit sur le sujet («Note sur le calcul d'une distribution du vocabulaire» dans *Bull. Amis de l'Université de Liège*, 36 (1964), 33ss.

✉ Université de Liège; Laboratoire d'Analyse Statistique des Langues Anciennes; 32, place du 20-Août; B–4000 Liège (Belgique).
Fax : + 32 41 23 25 45 et + 32 41 66 57 02 E-mail : u0013a1@vm1.ulg.ac.be

Extrait de la Revue Informatique et Statistique dans les Sciences humaines XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.

indices obtenus par application des formules reste entachée d'imprécision et qu'un effort est à faire en ce domaine.

C'est à cet objet qu'est consacrée la 2e partie. A.C. soumet à sa critique les diverses vérification proposées quant à la valeur des indices, pour leur sensibilité tant à la longueur des textes qu'à la richesse du vocabulaire (ce second aspect étant souvent négligé.) De nombreux moyens de vérifications ont été employés par les inventeurs même des indices, mais c'est surtout la méthode présentée par Nathan Ménard qui retient l'attention d'A.C. Cette méthode, il la trouve intéressante mais croit qu'il y a lieu de la compléter, tant parce que certains indices n'y ont pas été soumis que parce que les échantillons de Ménard ne sont pas assez longs. A.C. a donc utilisé deux corpus, un de 45000 mots, que Ch. Muller a constitué au moyen de bulletins météorologiques du journal *Le Monde* et un autre de 23000 mots tirés d'une pièce de théâtre d'un auteur québécois, M. Dubé. A.C. détaille abondamment les procédures qu'il a utilisées pour la lemmatisation des textes et pour la constitution d'échantillons de diverses longueurs qui soient aléatoires pour éviter les accidents dus à des hétérogénéités.

La troisième partie, qui est la plus longue (près de 100 pages) utilise ces échantillons pour vérifier les diverses formules et procédures. Je me bornerai à noter rapidement les conclusions obtenues de la sorte. La caractéristique K de Yule semble assez insensible à la longueur des textes; en revanche, l'effet sur K des additions de vocables ou d'occurrences est différent selon qu'elles concernent des vocables à forte ou à faible concentration. Il faut donc conclure que « la valeur de K n'est liée qu'occasionnellement et accessoirement à la richesse du vocabulaire »[2].

Le jugement porté sur l'indice V_m de Yule-Herdan n'est pas très favorable : cet indice ne cesse d'être sensible à la longueur du texte qu'à partir de 25000 occurrences environ et il réagit beaucoup plus à un petit nombre de vocables très fréquents qu'à un grand nombre de vocables de faible fréquence.

L'indice R de Pierre Guiraud mérite encore moins de confiance. On peut montrer en effet qu'il a une tendance à augmenter jusqu'à une certaine longueur de texte, puis qu'il baisse, le point d'inflexion étant variable et lié à la richesse lexicale.

Herdan avait proposé un nouvel indice, qu'il appelle C et qui s'inspire d'une loi de la biologie connue sous le nom de loi de la croissance relative. Si ingénieux que soient les commentaires de Herdan, il est clair que cet indice est lui aussi sensible à la longueur et est donc inapte à la comparaison de textes de longueurs différentes.

Le chapitre le plus long du livre est consacré à l'évaluation du vocabulaire théorique d'un texte à partir du vocabulaire d'un autre texte plus long, grâce à l'application de la loi binomiale. On peut alors comparer le vocabulaire théorique au vocabulaire observé dans le texte le plus court. A.C., à la suite de Ch. Muller lui-même, note que le modèle théorique le plus adéquat serait la loi hypergéométrique, puisque les tirages de mots

---

[2] À la p. 72, la formule (7) a été tronquée : il faut lire

$$K = 10000\frac{S_2}{S_1^2} - \frac{1}{S_1} \qquad \text{ou} \qquad K = 10000\left(\frac{\sum f_i^2 x_i}{N^2} - \frac{1}{N}\right).$$

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.

devraient être exhaustifs. Mais on peut montrer que, dans les conditions où on se trouve, la loi binomiale en donne une approximation tout à fait satisfaisante : les calculs produits par A.C. le montrent à suffisance. De plus la loi hypergéométrique est d'une application extrêmement coûteuse, par exemple en temps d'ordinateur, et la dépense n'en paraît pas justifiée ici. A.C. se fait aussi l'écho des discussions entre Muller et Brunet concernant le fait que la loi binomiale néglige la spécification lexicale. Cela vient de ce qu'elle se fonde nécessairement sur l'hypothèse nulle. De toute manière, on peut dire que les écarts entre les nombres théoriques produits par la loi binomiale et le vocabulaire réel d'un texte sont instructifs à la fois quant au vocabulaire de spécialisation et quant à la richesse lexicale. A.C. s'étend ensuite assez longuement sur le rôle de l'écart réduit pour apprécier la signification de cet écart et il compare en détail l'écart type observé à l'écart type théorique, qui peuvent servir tous deux au calcul de l'écart réduit. Au total, la loi binomiale fournit un outil intéressant, même si, dans les résultats qu'il produit, il reste trop d'incertitude dans la distinction de la spécialisation et de la richesse lexicale.

L'avant-dernier indice étudié est W d'Étienne Brunet. La formule de cet indice contient un élément a dont la valeur se fixe par des essais et peut d'ailleurs être optimisée. Cette optimisation permet d'ajuster la formule à ce que Brunet appelle les différents niveaux de composition : phrases, chapitres, textes complets, ensemble de textes. L'idée est intéressante, mais, comme le note A.C., elle reste sommaire et imprécise. Par ailleurs, elle repose sur un critère stylistique et méconnaît le facteur thématique. Enfin, il faut noter que les niveaux de composition sont liés à des types de longueurs; dans ces conditions, il sera assez difficile, s'il y a variation de l'indice, de déterminer ce qui vient de la sensibilité à la longueur et ce qui vient du niveau de composition. Les vérifications expérimentales montrent que l'indice de Brunet est significativement sensible à la longueur jusqu'à un certain niveau (*circa* 45000 mots), mais qu'au-delà il atteint peut-être une certaine stabilité.

Quant à l'indice Uber de Dugast, les expériences sur échantillons ne lui sont pas favorables. Par ailleurs, si quelques idées de Dugast paraissent intéressantes, il semble, si l'on s'en rapporte à ce qu'en dit A.C., qu'il n'ait pas toujours la précision souhaitable; en outre, on le voit proposer trois sinon quatre manières de calculer un vocabulaire théorique, ce qui est beaucoup.

Une synthèse finale bien documentée et très ferme arrive à la conclusion que la loi binomiale est « la seule qui permette une évaluation juste de la richesse du vocabulaire à travers les variations de la longueur de texte (p. 167).

Vient ensuite une analyse de la richesse lexicale dans la pièce *Bilan*, qui avait fourni les échantillons. Cette analyse remarquable montre comment les données et les indices statistiques sont de nature à permettre une compréhension en profondeur d'un texte littéraire. Après l'aridité inévitable de développements techniques, elle révèle à quel point les méthodes quantitatives peuvent donner vie à une analyse littéraire et découvrir les ressorts d'une action dramatique.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.

Voilà un livre auquel on recourra souvent : les informations qu'on y trouve sont de grande exactitude; les tests et vérifications, impitoyables mais parfaitement justifiés; les appréciations se montrent fermes mais pondérées[3].

[3] On me permettra peut-être de rappeler que, dans mon article « Deux programmes d'ordinateur pour l'étude quantitative du vocabulaire » *Revue-Org. intern. pour l'ét. des L. anc. par ordinateur*, 1967, n° 3, p. 81–95 et plus encore dans « Richesse et mode d'enrichissement d'un vocabulaire » dans *L'ordinateur et les recherches littéraires et linguistiques* (Actes de la XI[e] conf. intern. de l'ALLC, 1984), Champion-Slatkine, 1985, pp. 147–152, j'ai proposé d'utiliser l'entropie comme mesure de la richesse et de la distribution du vocabulaire, et j'ai montré que, pour une même œuvre, les entropies correspondant aux longueurs successives forment une courbe dont une bonne approximation est donnée par une parabole de la forme $y = ax^2 + bx$, où le paramètre b correspond à la tendance à la diversification, tandis que a (négatif) est un indice d'essoufflement. Cette méthode, au lieu de chercher à oblitérer les effets de la longueur du texte, thématise au contraire ce facteur et en calcule les effets. Par ailleurs, comme les entropies successives ont un minimum et un maximum (0 s'il y a autant de vocables que d'occurrences et le log de la longueur du texte si toutes les occurrences relèvent du même vocable), on peut plus facilement les évaluer que des indices qui n'ont pas de limites supérieures.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.