

# Mesurer le mètre ? Poésie et statistique.

Michel Bernard

*Université de la Sorbonne-Nouvelle (Paris III)*

*13, rue de Santeuil*

*F 75231 Paris (France)*

*Michel.bernard@univ-paris3.fr*

Il peut paraître paradoxal de vouloir appliquer les méthodes statistiques au genre littéraire par excellence, la poésie, ou, pour reprendre les termes pascaliens, d'appliquer l'esprit de géométrie à l'esprit de finesse. Mais, à y regarder de plus près, les caractéristiques du langage poétique, fondé sur le rythme et les récurrences phoniques, ont depuis longtemps été reliées à la science des nombres. La langue française porte la trace de cet appariement, qui remonte au moins au *quadrivium* des études médiévales : on parle de "mètre", de "métrique", le "pied" est une ancienne mesure de longueur, un vers est qualifié de "nombreux" quand il est fortement rythmé. Ainsi, Rollin, au XVII<sup>e</sup> siècle, peut écrire : "Il y a dans l'homme un goût naturel qui le rend sensible au nombre et à la cadence ; et, pour introduire dans les langues cette espèce d'harmonie et de concert, il n'a fallu que consulter la nature"<sup>1</sup>. Dès 1910, du reste, le poète et critique russe Andréï Biély (fils du mathématicien Bougaïev) publiait le recueil *Symbolisme*, dans lequel les premiers tenants de ce que l'on appellera plus tard le "formalisme" s'employaient à analyser de manière statistique la morphologie du vers russe. Cette tradition critique, illustrée par Jakobson et Tomachevski, renforcée par l'apparition de l'informatique, a donné lieu à de très nombreuses études. Je me propose ici d'en présenter un tableau synthétique, fondé sur l'examen d'une centaine de travaux récents<sup>2</sup>.

## 1. Définition du champ

Commençons par une anecdote révélatrice. Quand Markov a développé sa théorie sur les chaînes d'événements aléatoires dépendants, il a eu l'idée curieuse de l'appliquer en premier lieu à la suite des lettres composant le texte d'*Eugène Onéguine* de Pouchkine. Si cette expérience a un intérêt pour les mathématiques, elle n'apporte bien sûr rien à la compréhension des vers de Pouchkine. Cette utilisation du texte poétique comme champ d'expérimentation des méthodes de la statistique indique une des limites de ce champ interdisciplinaire : il faut, pour que la statistique soit pertinente en littérature, qu'elle apporte une information utile, sur laquelle la critique littéraire pourra bâtir une interprétation nouvelle ou confirmer des intuitions plus anciennes. J'exclurai donc de cette revue les travaux où l'œuvre poétique ne figure que comme pierre de touche d'une méthode statistique (je pense par exemple aux études de Benzécri sur divers textes littéraires). Pour des raisons semblables, je passerai sous silence les travaux purement linguistiques, qui s'attachent moins à étudier l'œuvre dans sa singularité qu'à y découvrir les règles prosodiques générales de la langue (c'est souvent le cas des exemples littéraires utilisés par Jakobson). Enfin, pour réduire encore un domaine pléthorique, j'omettrai également les travaux sur le lexique des textes poétiques pour me concentrer sur ce qui fait la spécificité du langage poétique : son rythme, sa *métrique*.

---

1 . *Traité des études*, III, 3.

2 . Il m'est impossible de reproduire ici ne fût-ce qu'une partie de cette bibliographie. Je me contenterai donc d'indiquer les noms des chercheurs auxquels je fais référence. Les lecteurs qui voudraient obtenir des indications bibliographiques et des références plus précises pourront les trouver sur le Web, à l'adresse suivante : <<http://www.cavi.univ-paris3.fr/phalese/hubert1.htm>>.

## 2. Que compter ?

C'est en effet le caractère répétitif de son tissu phonique qui caractérise la langue poétique. Pour que la statistique trouve à s'y employer, il faut pouvoir opérer des décomptes sur des populations bien repérables. Mais c'est bien là que commencent les difficultés : comme dans la plupart des sciences humaines, aucun des objets conceptuels manipulés par la poétique n'a de définition véritablement stable. Les poéticiens ne savent toujours pas si les vers français sont constitués de pieds ou de syllabes, ce qu'est une strophe (Cornulier), en quoi consiste le rythme. Le vers hésite lui-même entre ses définitions phonique et typographique (Byers), et, pour finir, le genre poétique, depuis l'apparition du poème en prose et du vers libre, a des frontières extrêmement mouvantes (Peter). Cela suscite par exemple des remarques de ce type : " Tant que des critères objectifs n'ont pas été adoptés par un consensus des chercheurs, toute scansion dépendra forcément – dans une certaine mesure du moins – de la sensibilité du lecteur, à la différence de la scansion antique, qui ne prêtait pas à discussion. " (Volkoff). Comment, dès lors, opérer des recensements, base de toute statistique (Kurz) ? Toutes les études de métrique statistique se posent préalablement la question des définitions de départ, justifiant longuement leurs choix avant de proposer des résultats et des analyses.

Tous les éléments de la versification ont par ailleurs été étudiés avec ces méthodes : le rythme des langues à accent tonique (Bailey, Tarlinskaja, Grotjahn, Braun, Krasnoperova, Vasjutockin, Baevskij et Osipova, Youmans, Kiparsky, Jandova, Kursite, Luque-Moreno, Jackson, Monroe, Shapir, Zlatoustova et Xitina, Kayumova, Logan, Tordeur, Schmiel, Altmann, Job, Laferrière, Foley, Schutter, Greenblatt, Beaver, Evrard, Purnelle, Grinda, Laan, Crawford, Jouad, Matjas, Barber, Volkoff), les syllabes (Biggs, Cornulier, Beaudouin et Yvon, Angoujard, Grotjahn, Creed, Stephens, Gérard, Paoli), les rimes (Lilly, Campa, Juillard, McMillan, Eekman, Berca, Jones, Schourup, Ross, Blain, Barber), les motifs phoniques (Chisholm, Bailey, Malherbe, Gaudard, Ikeshita, Lord, Bowden, Voronin et Ponomareva, Chociay, Robey, Hidley, Greenberg, Schutter, Cohen, Melhem, Von Grunebaum, Beltran, Blain, Duggan, Beauchemin, Cooper et Pearsall), les strophes (Cornulier, Chung, Bluhme), la ponctuation et la longueur des phrases (Garrette, Flachmann, Water et O'Connell, Michaelson et Morton et Wake, Ross, Kylousek), la longueur des mots (Best, Crockett, Tichy, Lee et Ross), les répétitions de mots (Heinemann, Wittig), les enjambements (Bluhme, Higbie). Dans chaque cas, les chercheurs partent d'un relevé (qu'ils ne publient pas toujours) qui tend à substituer au texte poétique un schéma numérique sur lequel il est possible de pratiquer des statistiques.

## 3. Les corpus

Toutes les formes poétiques ont été étudiées statistiquement, poésie en anglais, en russe, en français, mais aussi en grec (Bowden, Higbie, Tichy, Schmiel, Greenberg, Michaelson et Morton et Wake, Laan, Martindale), en latin (Grotjahn, Job, Evrard, Balard, Luque-Moreno, Stephens Byers, Tordeur, Gérard, Altmann), en allemand (Chisholm, Galt, Schutter), en néerlandais (Schutter), en espagnol (Grigor'ev, Jandova, Beltran), en portugais (Chociay), en italien (Boldrini, Robey, Barber), en suédois (Green), en lithuanien (Kursite), en islandais (Best), en roumain (Marcus, Berca), en gallois (Crawford), en arabe (Monroe, Cohen, Paoli), en persan (Vine), en sanscrit (Wujastyk), en berbère (Jouad), en coréen (Tarlinskaja et Myong), poésies modernes et antiques, syllabiques et accentuelles. Cette universalité montre bien que le phénomène poétique est, en soi, susceptible d'analyses statistiques. C'est un des domaines où la stylométrie trouve le plus naturellement à s'employer.

En revanche, on est souvent surpris par la faible taille des corpus étudiés, qui rendent les statistiques peu significatives. Si l'on comprend qu'une étude portant sur 186 000 phonèmes (Bailey) puisse donner des résultats probants, on s'interroge sur la valeur de statistiques opérant sur 166 strophes (Cornulier), 1 000 vers (Jones) ou 1 565 inversions (Hafou). La raison de ces faibles effectifs tient d'abord à la difficulté à constituer automatiquement des transcriptions phonétiques de textes poétiques. Dans les dernières années, des expériences intéressantes (Beaudouin et Yvon) ont

montré la possibilité d'obtenir des codages fiables, ce qui devrait permettre des dépouillements beaucoup plus importants.

#### 4. Méthodes statistiques employées

Les défauts de ces statistiques, souvent pratiquées sur des populations trop faibles et mal définies, sont encore aggravés par la pauvreté de l'outillage mathématique dont s'équipent la plupart des chercheurs. Dans un grand nombre de cas, l'intitulé "étude statistique" ne recouvre au mieux qu'un relevé accompagné de quelques pourcentages. Rares sont les chercheurs qui songent à utiliser un test de khi 2 ou un écart réduit pour étayer la validité de leurs chiffres. On ne s'étonnera donc pas de trouver par exemple, dans une étude sous-titrée "Prolégomènes à une étude statistique des rimes" (Campa), des affirmations aussi imprécises que celle-ci : "D'une manière générale, 11,5 % environ des finales d'*Alcools* contiennent la tonique [□□]". Pire encore, certains se contentent de règles de trois (baptisées "fréquences relatives") pour comparer les fréquences dans deux corpus inégaux, ou tirent des conclusions de pourcentages infimes, manifestement en dessous des seuils de signification. Cette situation amène certains chercheurs (Weiss, Mineralov) à distinguer les applications purement documentaires de l'informatique, les plus courantes, et une véritable mise en œuvre des outils mathématiques, extrêmement rare. Même les règles de base de la statistique descriptive ne semblent pas toujours connues.

Il ne faut cependant pas généraliser ce constat. Plusieurs études (russes, en particulier) utilisent des outils statistiques sophistiqués, dans le cadre de recherches souvent interdisciplinaires. On a pu employer par exemple le calcul de l'entropie (Bailey, Marcus), la loi hypergéométrique (Best, Grotjahn), le coefficient de corrélation de Pearson (Tordeur, Grotjahn, Beauchemin), l'analyse régressive (Grotjahn), les chaînes de Markov (Petruszewycz, Grotjahn), l'analyse de la variance (Greenblatt, Garrette), les écarts réduits (Garrette, Beauchemin) ou les indices de dispersion (Beauchemin), de manière à établir avec beaucoup de sûreté des résultats originaux et éclairants.

#### 5. Apport de l'informatique

"With the advent of the computer age it is now more feasible to do statistical analysis, [...] Whereas numbers are no substitute for theory, with the aid of the computer it becomes possible to base theoretical claims on analyses of large bodies of data."<sup>3</sup> : cette affirmation reste valide aujourd'hui, et l'on voit se profiler, grâce notamment aux progrès des analyseurs syntaxiques et des phonétiseurs, la possibilité de traiter de vastes corpus poétiques. Il faut être conscient néanmoins que le traitement automatique de la langue poétique représente une gageure pour des programmes initialement destinés à traiter du texte documentaire. En particulier, la polysémie propre à ce type de discours tient en échec beaucoup d'algorithmes. Il suffira d'ailleurs de noter que les critiques spécialisés divergent sur la manière de scander certains vers pour laisser entendre les difficultés que peut éprouver un automate pour effectuer la même opération.

Il faut également souligner que l'usage d'une machine n'incite pas toujours les chercheurs à affiner leurs traitements statistiques. L'ordinateur ne sert le plus souvent qu'à établir des comptages, présentés sous leur forme brute, et à mettre en page tableaux et graphiques. Les logiciels utilisés, qu'ils aient été écrits sur mesure pour les besoins de la recherche (Bailey, Breydo, Baevskij et Osipova, Hidley, Wujastyk, Lee et Ross, Beaudouin et Yvon), ou qu'il s'agisse de logiciels standards (Heinemann, Robey, Vanderbok, Newman, Ross, Hawthorne), ne fournissent pas toujours des résultats statistiques évolués. Mis à part des ensembles logiciels bien conçus du point de vue statistique comme le *Statistical Package for the Social Sciences*, la plupart des logiciels utilisés sont prévus pour la lexicométrie et ne comportent pas toujours d'outils statistiques perfectionnés. Il serait souhaitable à cet égard que les poéticiens travaillent plus souvent avec des spécialistes d'informatique et de statistiques pour mettre au point des procédures fiables. Le risque de l'usage de l'informatique réside dans la trompeuse sécurité qu'il donne au chercheur, qui peut s'estimer délié

---

3 . Bjorklund (B.) : "Metre as a Human Problem", *Semiotica*, 1992, n° 88, p. 338.

par la machine de toute obligation de vérification et de validation. Bailey mettait en garde, dès 1971, contre cette méconnaissance de la statistique : “ [...] most studies of textual problems either have failed to exploit the full power of mathematical models or have run roughshod over the inherent limits of these techniques. ”<sup>4</sup>

## 6. Apports de la statistique dans le domaine littéraire

Au-delà de ces questions de méthode, c’est bien sûr l’utilité de la démarche statistique pour l’étude de la poésie qu’il faut interroger. En 1894, le latiniste Paul Lejay affirmait que “ Par la statistique seulement, grammaticale et lexicographique, on peut introduire dans la littérature un peu de rigueur et de certitude. “ Cet enthousiasme positiviste se justifie-t-il encore aujourd’hui ? Il est malheureusement indéniable que l’usage de la statistique n’a pas pu créer dans les études métriques le consensus auquel aspirent certains chercheurs. La raison tient à la constitution des données, qui suppose généralement des théories *a priori* sur le vers. Dès lors, les traitements informatiques et statistiques ne peuvent fournir que des résultats prédéterminés par les choix initiaux, ce qui explique que la plupart des travaux se bornent à vérifier statistiquement des intuitions déjà largement répandues dans la communauté littéraire. S’il ne faut pas dédaigner cette incitation à la rigueur expérimentale dans une discipline où l’on s’est trop longtemps contenté de jugements subjectifs, il ne fait aucun doute que la méthode statistique ne pourra s’imposer que lorsqu’elle permettra de découvrir des phénomènes métriques à la fois insoupçonnés et fondamentaux, sur lesquels il serait possible de bâtir une nouvelle poétique.

Pour cela, certaines conditions doivent être remplies. Les chercheurs en littérature, tout d’abord, devraient recevoir ou acquérir une formation de base en statistique, dont leurs collègues des autres sciences humaines bénéficient depuis longtemps. Ce savoir de base devrait leur permettre, à tout le moins, de pouvoir dialoguer utilement avec les statisticiens et les informaticiens, dans le cadre de recherches interdisciplinaires rendues aujourd’hui indispensables par la rapide évolution des techniques et des savoirs. De telles équipes sont à même non seulement d’appliquer à la recherche littéraire des méthodes statistiques complexes et récentes<sup>5</sup> mais aussi d’en tirer un parti pour l’avancement de la connaissance des phénomènes métriques. En particulier, il serait souhaitable de disposer de programmes qui puissent analyser les textes sans aucun codage préalable, ce qui permettrait un étalonnage standardisé de textes divers.

## 7. Apports des recherches littéraires à la statistique ?

À l’inverse, l’interdisciplinarité ainsi entendue rendrait impossible chez les statisticiens le “ complexe de Markov “ évoqué plus haut, à savoir la tentation d’instrumentaliser la littérature, et leur fournirait cependant un champ d’investigation extrêmement complexe, aussi stimulant que ceux que lui proposent aujourd’hui la biologie ou la sociologie. La principale gageure est ici de pouvoir manipuler des concepts flous et cependant couramment utilisés par la critique littéraire, d’opérer des statistiques qui créeraient leur objet plutôt que de s’appliquer, comme d’habitude, à une réalité préalablement catégorisée. Cette forme d’exploration de données (“ data mining “), indispensable dans le domaine de la métrique, permettrait certainement aux statisticiens de forger des outils réutilisables dans tous les domaines où l’on invoque des “ logiques floues “.

### ABSTRACT

*This paper is a survey of contemporary studies on poetical metrics using statistics. This field seems adapted to a statistical approach but the difficulty of defining stable objects and the weak average statistical culture of the researchers constitute serious obstacles. But interdisciplinary teams could reach interesting results by working on such “fuzzy logics”.*

---

4 . Bailey (Richard) : “ Statistics and the Sounds of Poetry “, *Poetics*, 1971, 1, p. 16.

5 . Les travaux actuels sur l’application des réseaux neuronaux aux textes littéraires sont un bon exemple de ces collaborations.