Le viol de l'urne

Étienne Brunet

▶ To cite this version:

Le viol de l'urne

Etienne Brunet

-I-

Le débat qui va nous occuper pendant quelques instants a été engagé il y a un an à Pise, au dernier Congrès de l'ALLC. Plusieurs des participants, qui se retrouvent ici, se souviennent peut-être d'une certaine table ronde fort animée où, sous l'arbitrage de notre souriant confrère belge Martin, Paul Bratley avait fait le procès de tous les travaux de statistique linguistique, en avançant que le schéma d'urne y était indûment appliqué et que personne n'avait la moindre idée des probabilités réelles qui gouvernent la distribution des mots. Bratley ne conteste pas la valeur des lois de la statistique classique, loi binomiale, loi normale, loi de Poisson. II ne parle pas de la loi hypergéométrique, peu soucieux de surenchérir et d'épurer le schéma d'urne. Car c'est l'application du schéma d'urne qu'il conteste radicalement, comme tout à fait inadéquat au domaine des mots. Le schéma d'urne suppose des tirages indépendants. Or, les mots dans la chaîne du discours sont interdépendants. Ainsi l'article appelle un substantif subséquent et le mot CHAT une fois tiré exclut un second tirage immédiat du même mot. Le modèle est donc faux dans son principe. Et, qui pis est, ses résultats le condamnent : doutant de ce que Muller appelle ses « réussites », Bratley est plus sensible aux échecs, particulièrement éclatants dans mon *Vocabulaire français de 1789 à nos jours*. Et Bratley se plaît à relever le nombre considérable d'écarts réduits qui dépassent le seuil habituel de 5% et dont la valeur absolue est au-delà de 2. On pourrait objecter que l'hypothèse nulle elle-même n'interdit pas les écarts significatifs. Le seuil habituel de 5% signifie que 95% des cas doivent se situer en deçà de la valeur 2 de l'écart réduit. Et il est tout à fait conforme au schéma d'urne que dans 5% des cas le seuil soit franchi, c'est-à-dire 1 fois sur 20. Mais il faut reconnaître que cette proportion est largement dépassée dans l'exemple qui nous occupe et que les écarts significatifs l'emportent en nombre. II nous faut donc expliquer cette anomalie qui fait de la règle l'exception et de l'exception la règle.

Ceux qui ont lu Muller attentivement connaissent la réponse. En réalité ce qu'on appelle l'hypothèse nulle désigne la part attribuée au hasard. Cette part est considérable quand les effectifs sont faibles, et elle est faible quand les effectifs sont considérables. Supposons qu'un programmeur plaisantin, voulant sans doute éprouver la sagacité du linguiste qui l'emploie, ait transformé dans les données de ce dernier toutes les occurrences de PETIT en autant d'occurrences de GRAND. Comme chacun sait, un linguiste qui compte les mots n'en comprend plus le sens et la supercherie passera inaperçue à la lecture du texte ainsi transformé. Mais échappera-t-elle à la vigilance de la statistique ? Tout dépend de la longueur du texte. Si le texte n'a qu'un millier de mots l'absence de PETIT et la surabondance de GRAND ne seront pas décelables, l'hypothèse nulle faisant écran. Si le texte au contraire porte sur un million de mots et si l'on se réfère à la norme approximative du *Trésor de la langue française*, le point aveugle où le hasard empêche toute conclusion va se rétrécir comme l'iris au soleil, laissant à découvert la supercherie : l'écart réduit, avec des valeurs énormes, va servir de clignotant et attirer l'attention du chercheur sur l'excès des GRANDS et l'écrasement des PETITS. Ainsi suivant l'étendue de l'enquête le même test peut être muet ou éloquent, alors même que les écarts sont proportionnellement semblables. Voici, en décuplant à chaque fois l'étendue de texte envisagée, comment réagirait l'écart réduit à l'anéantissement des PETITS :

étendue du texte	1 000	10 000	100 000	1 000 000	10 000 000
fréquence théorique	1,45	14,53	145,34	1453,39	14533,92
écart réduit	(-1,21)	- 3,81	- 12,06	- 38,40	- 130,17

On voit donc que la valeur de l'écart réduit croît avec l'étendue du corpus. Plus précisément – toutes choses étant semblables par ailleurs – à un rapport r d'étendue de deux corpus correspond un rapport $\sqrt{r}$ des écarts réduits (du moins dans la formule simplifiée). Et qu'on ne dise pas que les variations de l'écart réduit viennent de quelque biais mathématique, de quelque distorsion issue d'un vice de la formule. L'expérience, si on avait la patience de la tenter avec des boules, donnerait raison au calcul. Là-dessus, je renvoie à mon collègue et ami Dubrocard qui a pris la peine de construire une urne électronique, d'y jeter les 25 000 mots de son corpus de Juvénal, d'y simuler parfaitement le tirage aléatoire (en recourant à des nombres aléatoires fournis par la machine), pour récupérer finalement des résultats exactement superposables à ceux de la loi binomiale (et donc de la loi normale qui en est l'approximation). Je ne sache pas que dans beaucoup de laboratoires on ait poussé le scrupule aussi loin, jusqu'à éprouver des lois en qui tout le monde a confiance et qui remontent à Pascal.

-II-

Mais peut-on avoir confiance dans le verdict de l'urne, lorsqu'il s'agit des mots ? Oui ou non, les phénomènes de discours (ou de langue) obéissent-ils à une loi aléatoire, à un schéma d'urne ? Ici force est de répondre non. Et sur cette lancée on ira jusqu'au bout de l'aveu : aucun phénomène humain n'obéit strictement au hasard et il ne serait pas besoin de longs tourments pour que nous reconnaissions aussi que même aucun phénomène naturel ne relève absolument du hasard. En réalité le mot hasard n'a guère plus de consistance que la *virtus dormitiva* de jadis, c'est le voile qui recouvre notre ignorance. Pierre Guiraud – dont la mort vient de nous atteindre – a bien dit que les événements individuels échappaient à la prévision statistique et que Durand se mariera dans l'année, si bon lui semble, quoi qu'en disent les statistiques. Mais il pense qu'au niveau global et à une certaine échelle, tous les faits humains, si libres qu'ils puissent paraître individuellement, « ressortissent collectivement à un déterminisme statistique précis ». Or à une échelle plus haute encore, on se rend compte que le modèle statistique, s'il n'explose pas, du moins se fendille, et dans les très grands nombres, le déterminisme arrive toujours à se glisser dans une fente de l'urne et à limiter le jeu du hasard. Quand on franchit la barre du million d'observations, on entre dans un univers qui ressemble à celui de la relativité et où bien des formules qu'on croyait éprouvées cessent de fonctionner (il en est ainsi par exemple de la « loi » de Zipf ou de la distribution de Waring-Herdan). Ainsi, dans les trop petits corpus, l'urne est débonnaire et ne s'étonne d'aucun écart. Dans les très grands au contraire, elle devient pointilleuse et voit la fraude partout. Faut-il s'en indigner ou s'en accommoder ? Faut-il rejeter un instrument de mesure qui n'est pas stable, un « étalon élastique » qui se recroqueville ou se dilate selon la taille des corpus ? L'écart réduit est une balance variable et paradoxale qui est d'autant plus précise qu'elle pèse des objets plus lourds. Le chercheur doit le savoir et « relativiser » les valeurs qu'il obtient selon l'échelle où se place l'observation. Et cela se fait sans fausse honte par le choix du seuil, qui appartient en dernier ressort au chercheur. Quand on évolue dans de très grands corpus, il importe de choisir un seuil plus sévère (par exemple une valeur *5* pour l'écart réduit au lieu de *2*), afin de n'être pas écrasé par la masse des résultats

« significatifs ». Il y a probablement le même nombre de choses intéressantes à dire dans un petit et dans un gros corpus mais si l'on se sert du même filtre, du même seuil, la mesure statistique risque d'en repérer trop peu dans le premier cas et trop dans le second. Au chercheur de choisir la maille du filet.

Mais pourquoi donc s'obstiner à pêcher ainsi dans la mer des données ? Le schéma d'urne est-il le seul modèle agréé ? Et sur ce point le réquisitoire de Bratley prend la forme de soupçons douaniers : à tous ceux qui utilisent le schéma d'urne, il demande d'apporter la preuve que leurs données sont agréées et bien conformes au modèle et il regrette que trop de chercheurs escamotent cette formalité initiale.

Avouons d'emblée que cette négligence est très générale et que dans les disciplines les plus diverses la distribution normale est toujours postulée mais rarement démontrée. Et cela tient au fait que cette preuve, comme celle de l'innocence, est souvent difficile à établir. Bien des tests ont été proposés, qui permettent de vérifier la normalité des données. Citons deux des plus puissants : celui de Kolmogorov-Smirnov[1] et le test plus connu du Chi2. C'est sur ce dernier que s'appuie Bratley pour montrer que la distribution du mot AME dans les 15 tranches chronologiques du corpus ne suit pas la loi normale. On pourrait chicaner sur le choix du mot AME qui n'est pas fait au hasard : de tout le vocabulaire français, le mot AME est l'un des plus irréguliers. Il suffirait de choisir un mot mieux réparti pour que le test laisse la liberté de conclure à la normalité des données. Dès lors appliquera-t-on le test à tous les mots ? Fera-t-on deux lots en rangeant d'un côté les mots où le test est favorable à la normalité et de l'autre ceux pour lesquels il l'exclut ? Cela risque de conduire à l'impasse : on ne pourra rien faire du premier lot, par l'impossibilité d'y rejeter l'hypothèse nulle, à cause de la faiblesse des écarts. Et on ne pourra rien faire non plus du second, parce que l'importance des écarts y disqualifie le schéma d'urne et empêche toute conclusion probabiliste. Mais surtout le test du Chi2 ne donne pas les mêmes résultats selon qu'on l'applique à un mot rare ou fréquent, à un corpus de faible ou de grande dimension. Comme l'écart réduit, le Chi2, pour des écarts proportionnellement identiques, croît avec l'accroissement de l'étendue (du corpus) ou de la fréquence (du mot considéré). Et c'est pourquoi dans des corpus modestes où la statistique a d'abord été

1 Ce test est longuement développé dans Sidney Siegel, *Nonparametric Statistics for the behavioral sciences*, pp. 47-52.

appliquée, le test du Chi2, peu exigeant en de pareilles conditions, a conduit à accepter le postulat de la normalité des données linguistiques. En réalité le test n'étant pas indépendant de la taille des échantillons, sa valeur devient très relative. Et la normalité des données, pour être prouvée, doit faire appel à d'autres considérations, théoriques et expérimentales, linguistiques et mathématiques.

-III-

1- Voyons d'abord comment répondre du point de vue théorique à l'objection majeure selon laquelle le discours est fait d'éléments liés, non d'événements indépendants. L'objection vient des linguistes aussi bien que des mathématiciens. Les premiers constatent que la statistique isole des mots qui dans le discours ne prennent leur sens et leur valeur que dans l'enchaînement à d'autres mots, dans un rapport paradigmatique ou syntagmatique à d'autres unités. Les seconds constatent aussi le fait syntaxique qui brouille les probabilités et les empêche d'être fixes : le tirage d'un premier élément LE (article) rend plus probable au coup suivant celui de CHAT (ou de quelque autre substantif)[2]. Il convient tout d'abord de reconnaître les faits : la statistique lexicale en effet n'étudie pas les mots dans leur déroulement discursif, mais seulement des unités préalablement détachées. Son objet n'est pas un texte brut, mais un texte déstructuré, mis à plat, non pas une voiture en état de marche, mais une voiture en pièces détachées. Et ce sont ces pièces détachées qu'on met dans l'urne. Cette opération réductrice est-elle légitime ? C'est au linguiste d'en juger. Observons toutefois que c'est un usage constant de la science de ne considérer d'une réalité – ici le discours – qu'un point de vue particulier *à la fois* – ici les éléments lexicaux. Une étude par exemple sur des éléments chimiques du corps humain est légitime, même si elle n'épuise pas le sujet du fonctionnement physiologique. Quant à l'indépendance des tirages, elle n'est guère perturbée par le fait syntaxique. Si l'on met dans l'urne un million de mots, chacun a des attaches directes avec une dizaine d'autres, ses voisins immédiats dans la chaîne du discours, mais à l'égard du million qui reste, chaque mot reste indifférent. La syntaxe laisse le premier et le dernier mot d'un texte, ou même d'une page, tout à fait étrangers l'un à l'autre. La syntaxe a

2 Évitons ici de confondre deux combinatoires : celle où l'ordre compte et celle où les combinaisons sont acceptées dans le désordre. C'est évidemment la seconde qui est utilisée dans la statistique lexicale, puisqu'on jette les mots *en vrac* dans l'urne et qu'on ne se soucie pas de les retrouver dans l'ordre.

seulement rendu les boules de l'urne un peu poisseuses, chacune ayant tendance à se coller aux boules qui la touchent immédiatement.

2- Mais les perturbations stylistiques et thématiques sont beaucoup plus redoutables. Un texte a un sujet, une intention, il exerce des choix cohérents et systématiques dans la réalité et dans le vocabulaire. Et cette difficulté a été dès l'origine abordée par Charles Muller, qui recommande d'en mesurer l'ampleur et d'en limiter les effets en constituant des corpus homogènes, en établissant des lexiques de situation. Avant d'accepter une norme, avant de choisir un corpus, avant de risquer une comparaison, il faut avoir pris la mesure de ces faits qui échappent au jeu du hasard et que Muller enveloppe sous le terme de « spécialisation lexicale ». Mais l'objection peut tout aussi bien se retourner en faveur de la statistique lexicale. Car c'est précisément la spécificité d'un texte, d'un auteur, d'un état de langue, d'un genre littéraire qu'on cherche à définir et à mesurer, le schéma d'urne et le modèle probabiliste fournissant la référence d'où procède la mesure. Si je veux vérifier qu'une ligne est droite ou non, qu'une surface est plane ou non, je me sers d'une règle. Si la surface a des creux et des bosses, ou si la ligne a des sinuosités, je ne vais pas casser la règle, sous le prétexte qu'elle ne convient pas aux données, que la nature est rebelle aux figures idéales et que la « prévisibilité » de la règle est toujours démentie par des faits. Il y a beaucoup à dire sur la notion de prévisibilité : la règle dont je me sers ne permet pas de prévoir si la ligne que je suis va tourner à droite ou à gauche, pas plus que le thermomètre ne me permet de savoir quelle température il fera demain. En matière lexicale la règle statistique ne permet, elle aussi, que la mesure. II ne s'agit que de décrire, nullement d'expliquer, moins encore de prévoir.

3- Ainsi se justifie l'emploi du modèle probabiliste. Reste à savoir si ce n'est pas un pis-aller et si d'autres modèles ne seraient pas supérieurs. Si toutes les lignes de la nature étaient courbes, ne serait-il pas plus sage d'inventer des règles courbes et de renoncer aux droites ? Hélas, alors qu'il n'y a qu'un modèle de droite, il y a mille figures de courbes et on risque bien de construire une tour de Babel si l'on s'ingénie à fabriquer des instruments qui suivent au coup par coup la réalité multiforme. C'est ce qu'on a vu maintes fois dans le domaine qui nous occupe : combien n'a-t-on pas inventé d'indices, de formules, de rapports, de coefficients de toutes sortes ? Fondées sur l'approximation, non sur le raisonnement, ces formules épousent trop étroitement les données dont elles sont issues et se refusent aux autres. Si l'on veut éviter l'empirisme et le bricolage, force est de recourir à un modèle universel qui s'impose pareillement à toutes les données et à toutes les disciplines. Le modèle probabiliste joue cette fonction de régulation, et en attendant que naisse un nouvel Einstein, il faudra bien s'en contenter.

-IV-

En attendant cet heureux mais improbable événement, le modèle classique, le seul dont nous disposons présentement, est-il aussi mal adapté aux données que Bratley le laisse entendre ? Écartons le test, trop relatif, du Chi2 et la tentation de lui faire dire ce qu'on veut, selon la taille des données auxquelles on l'applique. Y a-t-il d'autres critères qui permettent de justifier – expérimentalement – la loi normale ? On a une chance d'avoir affaire à une population normalement distribuée quand le profil de la distribution est celui d'une courbe en cloche ou courbe gaussienne et que son dessin est symétrique autour d'un axe où se confondent la moyenne, le mode et la médiane. Il faut encore que la dispersion soit caractéristique d'une distribution gaussienne, ce qui impose le calcul de l'écart type expérimental. Et l'on doit constater qu'à une distance de un écart type de part et d'autre de la moyenne, 68% des observations se trouvent regroupées, et 95% si la distance est portée à deux écarts types. Afin de rendre la démonstration plus probante, je prendrai exprès les cas les plus défavorables, où le Chi2 invite à conclure à la non-normalité, c'est-à-dire le cas des mots fréquents, le cas des grands corpus, et, pire encore, le cas des mots fréquents dans les grands corpus. Dans cette situation-là, obéissant à la poussée des grands nombres, le Chi2 ne peut pas ne pas être élevé. Et pourtant la distribution, on va le voir, n'a rien qui puisse invalider a priori le schéma d'urne et la loi normale. Et pour permettre le contrôle nous choisirons de préférence des données déjà publiées et complètes.

Le tableau 1 reproduit les observations qu'on a enregistrées dans cinq corpus différents. Dans chaque cas, la distribution des mots ou signes étudiés a été mesurée dans des tranches égales (comptées en mots ou en pages) ou dans des sous-ensembles rendus comparables par le recours aux fréquences relatives. Certes il arrive que le modèle soit défaillant, comme en témoigne le mot ESPRIT dont la distribution est presque aussi irrégulière que celle de l'AME et dont la symétrie est prise en défaut, que l'on considère la répartition du mot dans les 335 textes du corpus ou parmi les 259 écrivains de ce même corpus (voir figure 2). C'est qu'il s'agit d'un mot thématique sur lequel s'exerce pleinement l'effet de la spécialisation lexicale et qui a pris part aux querelles idéologiques des deux derniers siècles. Il n'est d'ailleurs pas impossible d'expliquer la dissymétrie de la figure 2, qui montre à gauche, du côté négatif, une pente plus douce et des textes plus nombreux, et à droite, parmi les excédents, une chute brutale et un espace étroit. Cela donne à penser que les écarts linguistiques viennent du *trop* et non du *trop peu***.**

Tableau 1

mot ou signe	étendue du corpus	nombre de tranches	nombre d'occurrences	moyenne par tranche	médiane	écart type expér.	% entre -1 s et +1 s	% -2s +2s
	N	n	f	m	M	s	%	%
ROUSSEAU (Emile)		tranches de 3000 mots						
virgule	288541	96	13089	136,344	135	18,877	72%	96%
point	288541	96	7608	79,25	78,50	12,65	75%	94%
PROUST		tranches de 20 pages						
temps	1267069	159	1637	10,296	10	5,065	75%	95%
femme	1267069	159	1404	8,824	8	5,400	68%	97%
jour	1267069	159	1358	8,541	8	4,023	70%	96%
temps +jour +femme	1267069	159	4399	27,660	27	8,062	70%	95%
ZOLA		tranches de 50 pages						
abord	2874755	145	1040	6,97	7	3,44	70%	95%
CORPUS XIX-XX		97 sous-ensembles (fréquences relatives)						
de	70273552	97	3015363	0,0409	0,042	0,00531	75%	93%
la	70273552	97	1870137	0,0266	0,025	0,00404	72%	94%
et	70273552	97	1745247	0,0258	0,025	0,00375	78%	93%
de + la + et	70273552	97	6630747	0,0934	0,092	0,00859	67%	96%
sous-CORPUS XIX-XX		335 textes (fréquences relatives × 1 million)						
esprit	20890865	335	11492	547,40	365	583,57	88%	96%

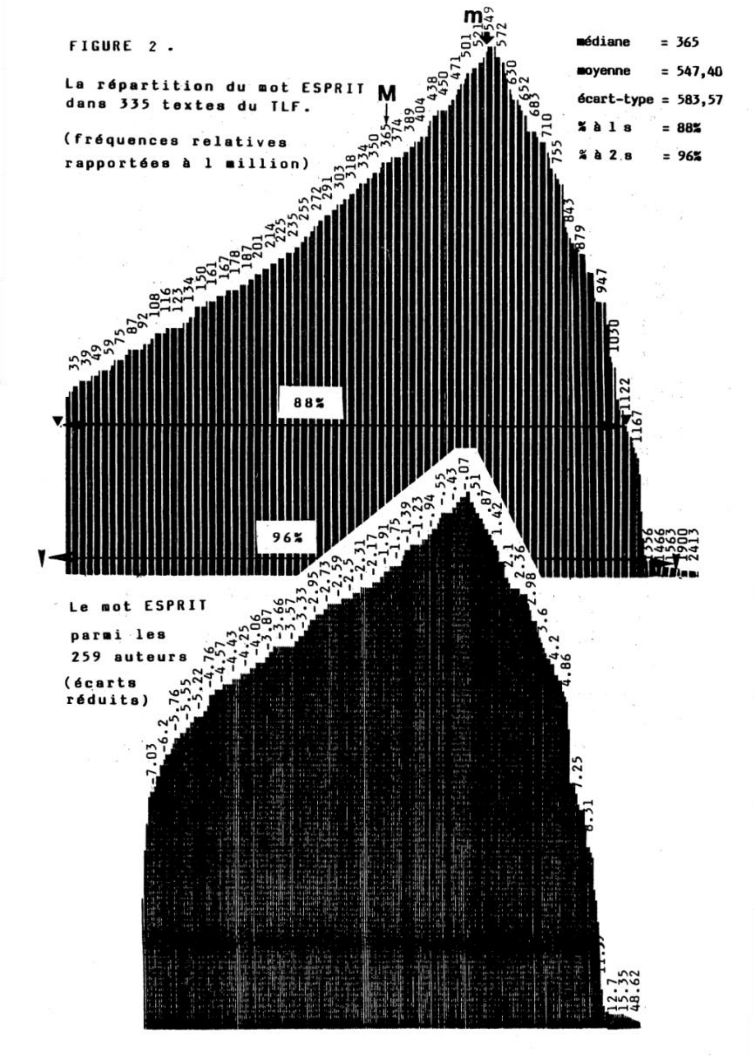

Figure 2

Un texte privilégie dans le discours une petite part du lexique, et rejette tout le reste dans l'ombre, dans la zone des déficits. Du côté positif, les écarts sont plus rares mais plus violents, du côté négatif, ils sont plus fréquents mais plus tièdes. Cependant la normalité et la symétrie ne sont pas toujours violées, comme on le voit dans le tableau 1 et dans la figure 3.

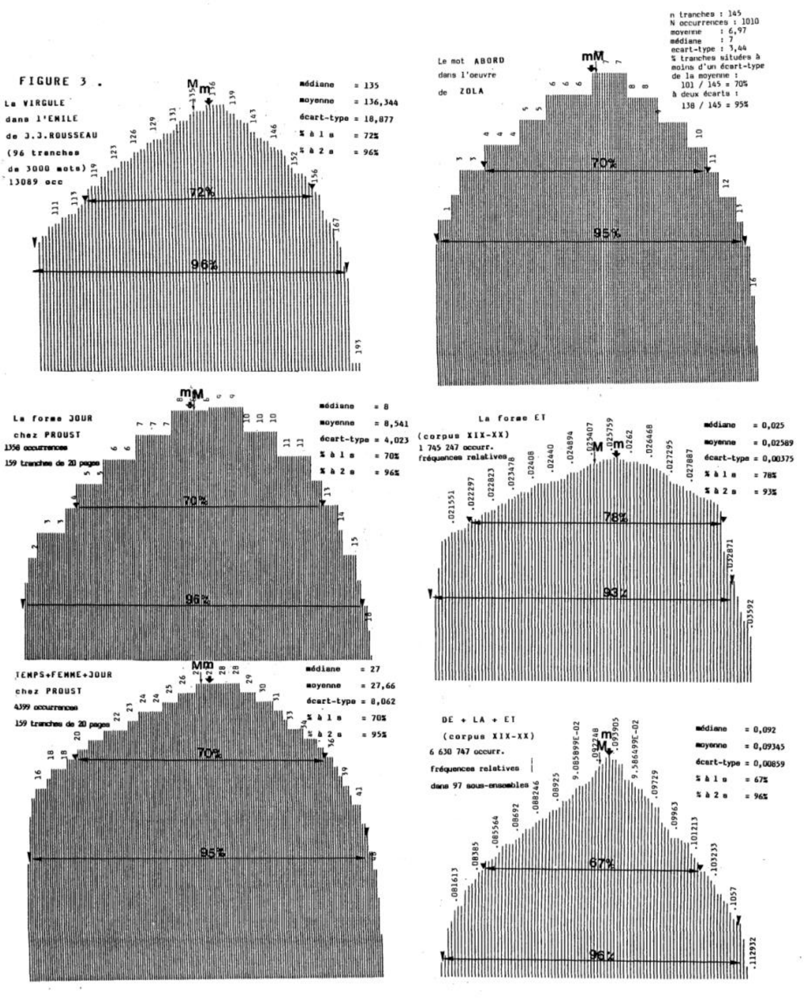

Figure 3

Qu'il s'agisse de la virgule ou du point dans l'*Émile* de J.-J. Rousseau, de la forme ABORD dans l'œuvre de Zola, des trois substantifs **(**TEMPS, FEMME et JOUR**)** qui arrivent en tête de *À la recherche du temps perdu***,** ou des trois formes les plus fréquentes **(**DE, LA et ET) du grand corpus du *Trésor de la langue française*[3], on s'écarte assez peu des caractéristiques d'une distribution normale : proximité, sinon coïncidence, de la moyenne et de la médiane, « fourchettes » acceptables à une distance de 1 et 2 écarts types, courbe en cloche à peu près régulière. Si les distorsions apparaissent parfois dans les courbes individuelles, elles tendent à se résorber dès lors qu'on opère des regroupements (DE + LA + ET dans un cas, TEMPS + FEMME + JOUR dans l'autre). Les échantillons peuvent être imparfaits pris isolément, mais leur cumul tend vers la distribution normale.

Rappelons que tous nos exemples ont été choisis dans les grands nombres et qu'en de telles occasions le test du Chi2, trop sensible à l'effet de taille, eût amené à rejeter le schéma d'urne. Naturellement on ne peut prolonger à l'infini ces expériences. Il suffit de les avoir menées dans des corpus variés et sur des mots différents pour être raisonnablement assuré que le schéma d'urne est le moins mauvais modèle qu'on puisse appliquer aux donnés littéraires et linguistiques.

-V-

1- Bratley recommande pourtant l'emploi de ce qu'on appelle la « loi faible ». En rejetant la loi forte des grands nombres, c'est-à-dire le schéma d'urne, qu'on l'appelle binomial, hypergéométrique ou normal, il a le souci du moins de ne pas laisser les linguistes totalement démunis. Et il leur propose deux modèles agréés, celui de Markov et celui de Chebyshev. Leur mérite commun est, au dire de Bratley, de pouvoir servir en toute occasion, puisque ces deux modèles ne présupposent aucun postulat sur la forme de la distribution. Peu importe que la population soit ou ne soit pas normalement répartie, les deux tests en question gardent leur valeur intacte. Voici d'abord la formule de l'inégalité de Markov :

$$prob\ (x > t) \leq m/t\ , t > 0$$

ce qui s'énonce comme suit : la probabilité pour qu'un mot ait plus de t occurrences est égale ou inférieure au rapport de la fréquence attendue (m) par la fréquence observée (t).

3 On trouvera les données dans notre *Index-Concordance* de l'*Émile*, Slatkine, tome 1, p. 572, dans notre *Vocabulaire de Proust*, Slatkine, t. 2 et 3, p. 617, et dans notre *Vocabulaire français de 1789 à nos jours*, Slatkine, t. 1, p. 8.

Appliquons le calcul au cas de l'AME dans la tranche chronologique 1833-1841 du corpus du TLF, *m* valant 3 001 et *t* 5 738 :

$$p\ (x\ > 5\,738) \leq 3\,001/5\,738$$
$$\leq 0{,}52$$

Ainsi un écart apparemment important qui va du simple au double, et qui porte sur des milliers d'observations, ne semble pas émouvoir le test de Markov qui lui accorde 1 chance sur 2. De tous les écarts observés dans le corpus du XIX-XX^{e}, le plus monstrueux concerne le mot EMPEREUR dans la tranche 1816-1832 qui retient presque la moitié des occurrences du mot quand il devrait n'en contenir que 7%. Alors que l'écart réduit atteint une valeur extrême (+140), le test de Markov est encore loin du seuil significatif (p = m/t = 647/4097 = 0,16). Ainsi des 70 000 mots contenus dans ce grand corpus, aucun – *strictement* aucun – ne serait significatif, même au seuil de 10%, si l'on se fiait à l'inégalité de Markov, ce bel instrument très pur, qui ne distille que le silence éternel.

2- L'inégalité de Chebyshev semble plus intéressante. Mais elle est plus exigeante puisqu'on doit lui fournir non seulement la fréquence théorique *m*, mais aussi l'écart type *s*. Bratley la formule ainsi :

$$prob\ (x - m > ks) \leq 1/k2\ , k\ > 0$$

Si l'on choisit un seuil de 1% (1/*k2* vaut alors 1/100 et *k* = 10), le mot devient significatif si l'écart observé entre fréquence réelle et fréquence observée dépasse dix fois l'écart type (au seuil de 5%, il faudra atteindre 4,5 fois l'écart type). Bratley applique la formule Chebyschev au mot AME dans la tranche 1833-1841; là où l'écart réduit est très élevé, il n'arrive pourtant pas au seuil de 5%. Et même si l'on choisit le cas-limite du mot EMPEREUR**,** là où nous avons rencontré l'écart record (*z* = +140), le seuil de 5% n'est pas atteint. La conclusion est donc aussi désespérante que la précédente. Pas plus que la formule de Markov, celle de Chebyshev ne permet d'extraire *un seul* mot significatif parmi les 70 000 du corpus. Ainsi, si la loi forte ne permet pas toujours d'éviter ce qu'on appelle les erreurs de première espèce (c'est-à-dire le rejet de l'hypothèse nulle quand elle est vraie), la loi faible, dans le domaine qui nous occupe, tombe systématiquement dans l'erreur de seconde espèce (c'est-à-dire l'acceptation de l'hypothèse nulle, quand celle-ci est fausse).

3- On s'explique d'ailleurs aisément la défaillance de l'inégalité de Chebyshev, dont la portée s'affaiblit lorsque croît la variance. Or précisément les écarts les plus considérables élèvent la variance, si bien que le test ne peut plus les déclarer significatifs. L'inégalité de Chebyshev souffre d'une seconde faiblesse qui tient à la nécessité de calculer un écart type sans disposer de sous-ensembles égaux. La plupart du temps, les textes que l'on compare sont d'étendue inégale. Et si nous avons pu faire les calculs qui précèdent en découpant des tranches tantôt de 3 000 mots, tantôt de 20 ou 50 pages, cela nécessite des efforts qu'on ne saurait répéter pour chaque mot. Bratley propose alors de faire des calculs de variance à partir des fréquences relatives – et nous avons procédé ainsi pour certains de nos exemples. Mais nous doutons que le procédé soit légitime. Les fréquences relatives et les pourcentages sont une transformation dangereuse des données, dont la taille est ignorée et dont les variations aléatoires ne sont pas prises en compte. Les deux séries 1, 4, 2, 3 et 1000, 4 000, 2 000, 3 000 auront la même variance si l'on raisonne sur les fréquences relatives alors que les deux distributions sont loin d'être équivalentes dans la réalité, la première étant banale et la seconde fort improbable.

Les deux formules enfin sont incapables de retenir le vocabulaire négatif, et celle de Markov par définition : puisqu'une probabilité se situe toujours entre 0 et 1, le quotient *m/t* qui la mesure doit être inférieur à 1, ce qui ne peut se faire que lorsque la fréquence observée (*t*) est supérieure à la fréquence attendue (*m*). Quant à l'inégalité de Chebyshev, si la chose n'est pas tout à fait impossible, elle ne concernerait au maximum que quelques dizaines de mots grammaticaux très fréquents – mais les écarts pour ces mots là n'ont jamais l'importance requise – et tous les mots sémantiques sont hors d'atteinte.

II n'y a donc aucun profit tirer des formules anciennes exhumées par Bratley, quelque respect qu'on leur doive. L'inégalité de Chebyshev a beau être un bijou pur et dur, résistant à tout, inaltérable et universel, il a beau être vénéré par les mathématiciens comme la pierre philosophale, sa rentabilité dans le domaine linguistique est nulle et son emploi coûteux et précaire quand les textes sont de longueur inégale. C'est comme si on proposait aux fermiers du Middle West une merveille de la technique : un soc de charrue, inusable et incassable, en diamant pur, mais pas plus long qu'une allumette !

Conclusion

1- II y a tout de même quelque profit à tirer de l'avertissement de Bratley. En matière de linguistique quantitative, les mathématiciens sont les sorciers et les linguistes, les apprentis. Les premiers pratiquent le doute systématique, les seconds s'abandonnent volontiers à la confiance naïve. Il était opportun de rappeler que le schéma d'urne est une figure idéale, sans cesse démentie par la réalité du discours. Il était sage aussi d'enseigner les vertus des méthodes non paramétriques. Mais sur ce point, les linguistes n'ont pas attendu les conseils de Bratley. Dans tous les travaux de statistique appliquée au discours, un usage constant est fait du coefficient de Spearman qui est établi sur des rangs et ne doit rien au schéma d'urne. L'étude de Corneille par Muller est fondée en grande partie sur ce coefficient. Ajoutons aussi que bien des méthodes récentes échappent partiellement au schéma d'urne et qu'en particulier les analyses factorielles se situent dans un univers descriptif où le linguiste se soucie peu de probabilités. II reste enfin que la meilleure garantie vient de la convergence des méthodes, paramétriques ou non. Nous en donnerons pour finir une illustration dans la figure 4 qui reprend l'exemple du mot ABORD chez Zola. La suite des 22 textes du corpus est reproduite dans l'ordre chronologique, de *Thérèse Raquin* au *Docteur Pascal*. La distribution de la forme ABORD y est représentée de trois façons différentes : selon la fréquence absolue du mot dans les 200 premières pages de chaque texte, selon les fréquences relatives, selon les écarts réduits. Si la première courbe s'écarte un peu parce qu'elle repose sur des textes tronqués, les deux autres sont rigoureusement parallèles et la querelle qui oppose les deux méthodes perd beaucoup de sa force.

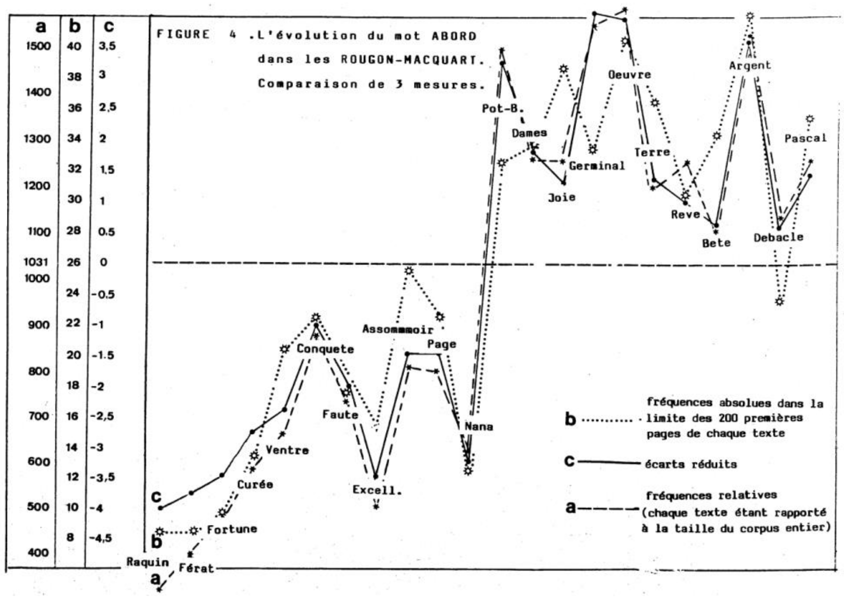

Figure 4

2- Je terminerai mon plaidoyer sur cette note d'apaisement, de compromis, presque d'excuse. S'il fallait reprendre les différents points de notre défense contre l'accusation de viol, on y trouverait la cohérence habituelle aux plaidoyers de cette sorte :

1- La loi normale n'a pas été violée.

2- De toute façon tout le monde la viole.

3- De toute façon elle n'est pas la seule.

Je demande donc un non-lieu : il n'y a pas eu viol mais accomplissement naturel. Car la statistique raffole des grands nombres comme les femmes faciles adorent les grandes fortunes. Et c'est en linguistique qu'elle rencontre le champ le plus favorable, comme le disait Guiraud, dans une formule souvent citée : « La linguistique est la science statistique type ; les statisticiens le savent bien ; la plupart des linguistes l'ignorent encore ». Il semble pourtant que les choses ont évolué depuis Guiraud et si quelques linguistes ont été convaincus, certains mathématiciens par contre ont cessé de l'être. Je ne sais si la statistique linguistique a gagné au change.