Loi hypergéométrique et loi normale. Comparaison dans les grands corpus.

Étienne Brunet

Loi hypergéométrique et loi normale. Comparaison dans les grands corpus

Etienne Brunet

L'idée de cet exposé m'est venue d'un lapsus qui n'était pas seulement de la langue mais aussi du cerveau et qui, il y a quelques années, m'avait fait confondre la loi multinomiale et la loi hypergéométrique. Cela se passait ici même devant le public du laboratoire de Saint-Cloud, réuni au séminaire de G. Th. Guilbaud. Cette bévue, heureusement redressée par le maître, m'avait laissé troublé pendant le reste de mon exposé, qui d'ailleurs ne reposait nullement sur ces deux lois.

J'ai conscience que je risque des bévues plus graves encore aujourd'hui puisque j'ai entrepris de parler de la loi hypergéométrique devant un public hypercritique et hyperaverti, devant des mathématiciens et des praticiens qui la connaissent bien mieux que moi et la mettent en œuvre bien plus souvent. Quand un littéraire de mon espèce s'aventure dans le champ mathématique des possibles et des probables, il a toutes les chances - et c'est la probabilité qu'il peut le mieux mesurer - d'attirer sur lui les erreurs, comme l'aimant la limaille ou le goudron les plumes de volaille. Je demande donc aux spécialistes de prendre patience, assurés qu'ils sont de se divertir lorsque mes raisonnements mathématiques, comme les discours apologétiques de Sganarelle, se casseront le nez - ce qui ne saurait tarder.

1°) Quoique la méthode hypergéométrique soit mieux connue ici que partout ailleurs, il est sans doute nécessaire d'en montrer d'abord le mécanisme. Soit un corpus de 12 mots emprunté à la première phrase de *La Recherche du temps perdu* : « Je n'avais pas le temps de me dire : "je m'endors" ». Que Proust me pardonne de couper un tronçon dans sa prose et de l'appeler hideusement corpus. Dans ce tronçon déjà fort menu, nous allons détacher un sous-ensemble que nous nommons texte,

et qui sera constitué du discours direct, c'est-à-dire des 3 derniers mots. En choisissant d'autres mots, on aurait pu isoler un autre texte. Le problème est de savoir combien de textes différents peuvent être extraits, combien de combinaisons de 3 éléments peuvent être réalisées quand on dispose de 12 éléments au total. On peut trouver la solution avec les mains, si l'on manipule les 12 cubes d'un jeu de construction, ou les 12 figures d'un jeu de cartes. On peut la trouver avec le raisonnement, en isolant les 2 premiers mots et en dénombrant toutes les possibilités d'association avec un troisième, puis en isolant le 1er et le 3ème, etc. Si les mains et le cerveau ont correctement fonctionné, on doit trouver 220. Car il y a 220 façons d'extraire 3 mots sur un ensemble de 12, autrement dit 220 façons de jouer au tiercé dans le désordre, c'est-à-dire sans que l'ordre des 3 éléments extraits soit pris en compte. Ce résultat est donné plus commodément - et en économisant la fatigue du raisonnement par des tables ou par un calcul, soit qu'on consulte le triangle de Pascal, 12e ligne, 3e colonne, soit qu'on fasse le calcul grâce à la formule usuelle (avec les symboles T (= corpus) et t (= texte) qu'on utilise à Saint-Cloud) :

$$\left(\frac{T}{t}\right) = \frac{T\,!}{t\,!\,(T-t)!} = \frac{12\,!}{3\,!\,9\,!} = \frac{10\,\times 11\,\times 12}{1\,\times 2\,\times 3} = 220$$

Mettons en réserve ce nombre 220 qui va servir de dénominateur, de pondération, au calcul qui va suivre.

Je m'intéresse en effet au mot JE que je rencontre 1 fois dans le sous-corpus et 2 fois dans le corpus et je voudrais savoir quelle est la probabilité d'obtenir cette répartition. Il faut donc calculer combien de combinaisons de 3 éléments sur 12 contiennent un JE et un seul. Pour cela trichons un peu et mettons l'œil et la main dans l'urne où nous ferons 2 tas : l'un contiendra les 2 JE, l'autre les 10 autres mots. Pour avoir une combinaison gagnante (i.e. qui contienne un JE), il faut puiser 1 mot dans le premier tas et 2 autres dans le second. Dans le 1er tas les possibilités sont faibles : puisqu'il n'y a que deux éléments, il y a deux combinaisons, ce qui est confirmé par la formule générale :

$$\left(\frac{2}{1}\right) = \frac{2\,!}{1\,!\,(2-1)!} = \frac{2}{1\,\times 1} = 2$$

Dans le 2e tas la même formule nous donne :

$$\left(\frac{10}{2}\right) = \frac{10\,!}{2\,!\,(10-2)!} = \frac{9\,\times 10}{2} = 45$$

Le nombre de combinaisons favorables est le produit de ces deux résultats indépendants, soit $2 \times 45 = 90$ sur 220 possibilités, soit une probabilité de $\frac{90}{220} = 0{,}41$. La formule d'ensemble s'écrit ainsi avec les paramètres T (étendue du corpus), t (étendue du sous-corpus), f (fréquence du mot dans le corpus), k (fréquence du même mot dans le sous-corpus) :

$$\frac{\binom{f}{k}\binom{T-f}{t-k}}{\binom{T}{t}} = \frac{\frac{f\,!}{k\,!\,(f-k)!}\,\frac{(T-f)\,!}{(t-k)\,!\,(T-f-t+k)!}}{\frac{T\,!}{t\,!\,(T-t)!}} = \frac{f\,!\;(T-f)!\,t\,!\,(T-t)!}{k\,!\;(f-k)!\,(t-k)!\,(T-f-t+k)\,!\,T\,!}$$

Pour rencontrer les 2 occurrences de JE dans le sous-ensemble, il faut que la main gauche saisisse les 2 éléments du 1er tas et que la droite en prenne 1 sur les 10 du 2ème tas, soit 10 combinaisons sur 220. C'est ce que donne aussi la formule pour $k = 2$.

Enfin, pour n'avoir aucune occurrence de JE, il faut puiser 3 éléments parmi les 10 du second paquet, soit $\frac{10\,!}{3\,!\,(10-3)\,!} = 120$.

Il n'y a pas d'autres possibilités que ces trois résultats, comme le prouve leur sommation à 1 :

pour $k = 0$ $p0 = \frac{120}{220} = 0{,}545$

pour $k = 1$ $p1 = \frac{90}{220} = 0{,}41$

pour $k = 2$ $p2 = \frac{10}{220} = 0{,}045$

$$p0 + p1 + p2 = \frac{120 + 90 + 10}{220} = \frac{220}{220} = 1$$

Le raisonnement et la formule concordent parfaitement. Aucune contrainte d'aucune sorte dans le modèle, sinon que k doit être inférieur ou égal à f (la fréquence d'un mot dans une partie ne pouvant être supérieure à celle du même mot dans l'ensemble, ce qui est l'évidence). Aucun autre modèle n'apporte à l'esprit autant de satisfaction. Le moins éloigné est le modèle binominal.

2°) Ce dernier convient lui aussi à la mesure des probabilités discrètes. Comme notre exemple est de faible dimension, un développement restreint de la formule suffira. La probabilité pour que *k* soit égal à 0, 1 et 2 est donnée par les termes suivants :

q^2, $2pq$ et p^2 (si l'on imagine deux tirages ($f = 2$) en posant $p = t/T = 3/12$)

dont la somme vaut 1[1]. Je renvoie aux ouvrages de Muller le lecteur qui serait peu familiarisé avec cette loi, dont les résultats sont ici comparables à ceux du modèle géométrique :

Pour $k = 0$, on obtient 0,5625 (contre 0,545)

pour $k = 1$, on obtient 0,375 (contre 0,409)

pour $k = 2$, on obtient 0,0625 (contre 0,045)

La différence vient uniquement du fait que le tirage est exhaustif dans un cas et non-exhaustif dans l'autre. La loi binomiale suppose en effet constante la composition de l'urne après chaque tirage, et si au second tirage, on mesure *p* non plus par le rapport constant $\frac{3}{12}$, mais par le quotient réajusté $\frac{3}{11}$ ou $\frac{2}{11}$ selon que le coup précédent a donné tel ou tel résultat (i.e. selon que la probabilité *p* ou *q* a été activée), on retrouve alors très exactement les résultats de la loi hypergéométrique. Bien entendu la probabilité *q* doit être ajustée de la même façon.

3°) Par contre, ni la loi normale ni la loi de Poisson ne trouvent ici leur justification. Dans notre exemple où *T* et *f* sont très faibles, l'application de la loi normale serait désastreuse. Rappelons que le calcul a besoin de deux paramètres : la moyenne *m* et l'écart-type σ. La moyenne est assimilée à la fréquence théorique, i.e. *fp*, *p* étant le rapport du sous-ensemble à l'ensemble, soit = 0,25. D'où $m = fp = 2 \times 0{,}25 = 0{,}50$. L'écart-type n'est pas expérimental mais théorique et s'exprime par la formule :

$$\sigma = \sqrt{npq} = \sqrt{mq}$$
$$= \sqrt{0{,}50 \times 0{,}75} = 0{,}61$$

[1] Rappelons que chaque terme de la formule générale s'écrit $\binom{f}{k} p^k q^{f-k}$ et que, *k* variant de 0 à *f*, on compte *f* +1 probabilités.

Ce recours à l'écart-type théorique n'est autorisé que si on a affaire à une distribution normale et symétrique, ce qui n'est pas le cas ici.

Le résultat le montre bien : les paramètres *m* et σ permettent de calculer un écart réduit :

$z = \frac{k-m}{\sigma}$ où *k* est la fréquence observée, par exemple 1.

z vaut donc $\frac{1-0,5}{0,61} = 0,82$.

Reste à convertir cet écart réduit en probabilité, ce qu'on peut faire à l'aide des tables, ou par une approximation qu'on réalise aisément sur les calculatrices de poche. Ici la probabilité est de 0,20, ce qui est très différent de la réalité 0,41. Le calcul donne 0,20 également pour *k* = 0 et 0,007 pour *k* = 2. La sommation des 3 évènements possibles ne dépasse pas 0,41. Reste donc une probabilité de 0,59 pour des évènements impossibles, plus probables, comme *k* = 0,5 où p vaut 0,5, ou *k* = 0,25 ou 0,75 où *p* vaut 0,141. La distorsion vient en partie du fait que la loi normale fait intervenir une intégrale et que la variable est supposée continue, alors que les fréquences sont des variables discrètes qui ne peuvent avoir que des valeurs entières, positives ou nulles, et dont les probabilités sont cumulatives. Quand je calcule la probabilité que le mot JE apparaisse *k* fois ou moins, ou bien *k* fois ou plus, il faut faire la somme des probabilités de 0 à *k*, ou bien de *k* à *f*, respectivement. Il faut monter l'escalier jusqu'à la marche *k* ou le descendre à partir de la marche *k*. On peut simuler cette fonction discrète si l'on utilise la formule de la loi normale, sans avoir recours aux tables, en cumulant les valeurs obtenues de 0 à *k*, ou de *k* à *f*. C'est ce que nous avons fait pour notre exemple où les paramètres *m* = *fp* et $\sigma = \sqrt{fpq}$ ont été empruntés à la loi binomiale et transférés à la loi normale.

Car pour une taille *f* croissante, la distribution binomiale tend vers la distribution normale et cela d'autant plus rapidement que *p* se trouve plus proche de 0,5.

$$\frac{f}{k}\, p^k\, q^{f-k} = \frac{1}{\sqrt{2\pi fqp}}\, e^{-\frac{1}{2}\left(\frac{k-fp}{\sqrt{fqp}}\right)^2}$$

Les conditions ne sont évidemment pas remplies ici, puisque *f* vaut 2 seulement. Et les résultats, quoique meilleurs que les précédents, restent éloignés de la réalité. Ils sont également éloignés de la logique. Il n'y a en effet que trois résultats possibles : *k* = 0, 1 ou 2. Or la sommation des trois probabilités obtenues (0,467 ; 0,467 ; et 0,032) n'est pas égale à 1

comme on le souhaiterait. C'est que la courbe de Gauss ne connaît pas de limites et qu'allant de $-\infty$ à $+\infty$, elle ne s'arrête ni à 0, ni à 2, qui sont ici les limites absolues de notre exemple. Si l'on veut compléter les vides et obtenir la sommation à 1 de toutes les probabilités théoriques, il faut imaginer des sous-fréquences négatives, ou supérieures à la fréquence *f*. Le modèle propose ainsi une probabilité de 0,032 pour $k = -1$, ce qui est absurde.

Dans l'exemple limite que nous avons choisi, la loi de Poisson serait peut-être moins mal venue, quoiqu'elle s'appuie sur le seul paramètre *m*. Mais elle a besoin elle-aussi d'un nombre suffisant de tirages (or ici *f* vaut 2) et son application est recommandée lorsque *p* est faible, ce qui n'est pas le cas non plus.

La conclusion est claire : dans le domaine des fréquences (qu'il s'agisse de mots, de phonèmes ou de quelque autre unité linguistique), un seul modèle est universellement exact : le modèle hypergéométrique et un seul autre acceptable : le modèle binomial. Les autres modèles (loi normale ou loi de Poisson) ne conviennent pas à tous les cas et notre exemple vient de les prendre en défaut. Ajoutons que les deux premières sont transparentes et qu'avec un peu d'effort, même un esprit littéraire peut en démonter le mécanisme. En particulier le raisonnement hypergéométrique évite avec élégance certaine impasse où tombent la loi normale et la loi binomiale, quand on utilise le schéma d'urne. Il y a en effet deux manières de concevoir le tirage des mots dans l'urne. Dans notre exemple nous nous intéressions à la fréquence *f* (2) dans une urne où la probabilité *p* était de $\frac{3}{12} = 0{,}25$ (t=3, T=12), ce qui conduisait à une fréquence théorique :

$$m = \frac{2 \times 3}{12} = 0{,}5$$

à un écart-type : $\sigma = \sqrt{mq} = \sqrt{0{,}5 \times 0{,}75} = 0{,}6123$

et à un écart réduit : $\varepsilon = \frac{k-m}{\sigma} = 0{,}8165$ (pour $k = 1$).

Mais on obtient un autre résultat si on suit un raisonnement différent, tout aussi légitime, où l'on institue une probabilité $p = 2/12$, qui représente la proportion des JE dans le corpus, et où l'on s'intéresse au texte *t* (=3). Si la fréquence attendue est bien la même $m = \frac{3\times2}{12} = 0{,}5$ la probabilité q étant différente ($q = \frac{10}{12}$ et non plus $\frac{9}{12}$).modifie la valeur de l'écart-type et conséquemment de l'écart réduit, lequel devient alors :

$z = 0{,}7746$ au lieu de $0{,}8165$. Or la loi hypergéométrique donne miraculeusement le même résultat que l'on procède d'une façon ou d'une autre. Inversons en effet les paramètres *t* et *f*, ce qui correspond à une mise en place différente du schéma d'urne :

sachant le nombre total de combinaisons à 2 éléments $(\frac{T}{f})$ je mesure maintenant parmi celles-ci le nombre de cas où un des deux éléments appartient au texte t (pour $k = 1$), soit $\left(\frac{t}{k}\right) \times (\frac{T-t}{f-k})$. On obtient comme précédemment $p = \frac{27}{66} = 0{,}41$.

Admirable est cette indifférence de la loi hypergéométrique aux tripotages des urnes, au point que le recours de l'esprit à cette représentation traditionnelle devient inutile. La loi hypergéométrique s'impose ainsi de droit divin, et ses ministres peuvent, comme le demandait certain tract de mai 68, uriner dans l'urne. Ou plus exactement le schéma d'urne n'est véritablement respecté que dans la loi hypergéométrique, qui est la seule à vider l'urne au moment du dépouillement et à constater honnêtement combien de bulletins ont telle forme (appartiennent à *t*) et telle couleur (appartiennent à *f*), alors que les autres lois s'empêtrent à démêler, comme la vieille scholastique, si la forme précède la couleur ou la couleur la forme.

Mais il est temps de suspendre notre exercice d'école maternelle et de ranger notre jeu de construction, pour aborder la réalité linguistique à plus grande échelle. Et d'un saut je propose qu'on passe de l'infiniment petit à l'infiniment grand, et qu'à la première phrase de *La Recherche du temps perdu* on ajoute toutes les autres, et même toute la littérature du XX^{e} et du XIX^{e} siècles qui a été dépouillée à Nancy. L'expérience des sciences physiques nous a montré qu'en changeant d'échelle, certaines lois sont devenues fragiles, qu'on pensait établies sur la brique et le roc, et qu'au contraire des approximations faites de bric et de broc peuvent alors être utiles.

1°) Je proposerai d'abord un exemple intermédiaire qui appartient aux grands nombres par certains de ses paramètres et aux petits nombres par certains autres. Il concerne le mot ABETI dont on recense 55 occurrences dans le corpus du *Trésor de la Langue Française*, et qui est absent dans la première tranche. Les paramètres du mot, *f* et *k*, ont donc respectivement pour valeur 55 et 0, et les paramètres du texte, *T* et *t*, 70 273 552 et 5 857 336. La loi hypergéométrique nous donne une probabilité de 0,00834, soit moins d'une chance sur cent pour que

l'absence du mot dans la première période puisse être attribuée au hasard. Or la loi binomiale propose exactement le même résultat, par des voies nettement plus rapides, puisqu'il suffit d'élever la probabilité q à la puissance *f*.

$$q^f = (1 - \frac{t}{T})^{55} = 0{,}00834$$

En toute rigueur la probabilité q n'est pas constante mais la zone de variation est extrêmement faible entre la probabilité de départ :

$$1 - \frac{5\ 857\ 336}{70\ 273\ 552} = 0{,}916649496$$

et celle d'arrivée :

$$1 - \frac{5\ 857\ 336}{(70\ 273\ 552 - 54)} = 0{,}916649432$$

Le même parallélisme est constaté entre la loi hypergéométrique et la binomiale, pour toutes les valeurs de *k*, conne on peut le voir sur le graphique 1 où les points coïncident parfaitement. Comme entre ces deux lois la seule différence vient de là, on peut en conclure que l'exhaustivité du tirage ne joue aucun rôle dans les conditions où nous nous trouvons. Voilà donc levée une première hypothèque qui pesait aussi sur la loi normale.

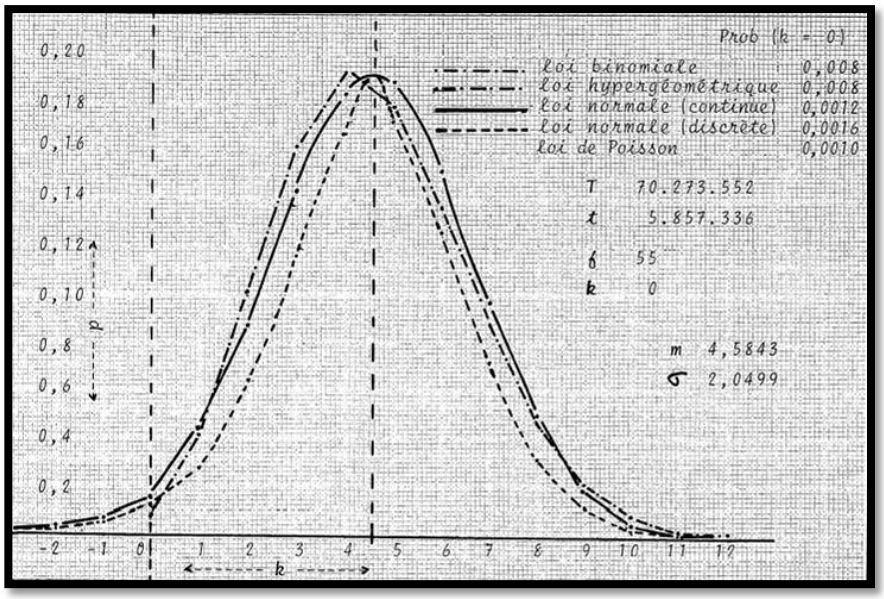

Figure 1 : Le vocable « abêti » dans la 1ère tranche du TLF

2°) Peut-on lever les autres hypothèques ? Observons les courbes superposées dans le graphique 1. L'étalon de référence est constitué par le modèle hypergéométrique. L'évidence visuelle y découvre la forme caractéristique de la courbe en cloche, ce qui suggère l'idée d'une distribution normale. Si l'on construit la courbe normale avec les paramètres m (4,5843) et σ (2,0499), on obtient la cloche intérieure (en pointillés sur la figure 1) qui s'étale plus largement aux extrémités gauche et droite et se rétrécit dans sa position centrale.

D'où vient cette distorsion ? Du fait que dans la loi normale la variable est supposée continue et qu'on ne mesure pas la probabilité qu'elle soit égale à k mais qu'elle soit inférieure à k. Ainsi pour $k = 4$, loi hypergéométrique et loi binomiale proposent la valeur 0,194 alors que la loi normale n'atteint que 0,168. En fait cette probabilité 0,168 correspond à l'intégrale comprise entre les valeurs 3 et 4 de la variable. Généralement les tables de la loi normale contiennent plutôt les probabilités pour que x soit inférieur (ou supérieur) à telle ou telle valeur. Il s'agit d'une probabilité cumulée. Dans le cas présent pour $k < 4$ nous obtenons : $F(k < 4) = 0{,}388$, alors que le modèle hypergéométrique fournit la probabilité 0,509 pour le cumul des probabilités obtenues lorsque $k = 0, 1, 2, 3$ et 4. De l'autre côté de la moyenne (ou du mode), la loi normale sous-estime de la même façon les probabilités. Pour $k > 5$ on trouve la valeur $p = 0{,}42$ contre $(1\text{-}0{,}509) = 0{,}491$ dans le modèle exact. Il s'ensuit un phénomène étrange ; si l'on ajoute l'une à l'autre les 2 probabilités pour $k < 4$ et $k > 5$ on ne rejoint pas l'unité. Il manque en effet l'intervalle compris entre 4 et 5 dans lequel se situe le mode ou sommet de la courbe et dont la mesure est la suivante : $1 - (p(k < 4 + p(k > 5)) = 1 - (0{,}388 + 0{,}420) = 1 - 0{,}808 = 0{,}192$. Il nous arrive ainsi la mésaventure de l'apprenti géomètre à qui l'on demandait de compter les arbres et qui avait relevé les intervalles. La conséquence est que la loi normale fournit des probabilités plus faibles que le modèle exact et qu'elle exagère les écarts. Ce biais qu'on peut appeler le biais de la continuité s'exerce dans les deux sens, des deux côtés de la courbe. Il est constant et symétrique et comme tel il n'est pas dangereux. Il suffit d'adopter un seuil plus sévère, 0,01 au lieu de 0,02, et les conclusions seront aussi certaines que celles de la loi hypergéométrique. Mais il n'est pas difficile de corriger ce biais, ce qu'on peut faire de deux ou trois façons :

- On pourrait monter l'échelle en mettant le pied entre les barreaux, en partant de 0,5 au lieu de 0. Ainsi, si je mesure la probabilité que *k* se situe entre 3,5 et 4,5, je calcule à peu près la probabilité d'obtenir $k = 4$. Car la surface du trapèze qui a pour bases 3,5 et 4,5 est égale à l'aire d'un rectangle de même largeur et de longueur 4. Avec cette correction, dite de continuité, la probabilité pour obtenir $k = 4$ passe à 0,175 et celle qui mesure $k \leq 4$ s'élève à 0,483.

- Le même résultat est obtenu si l'on applique directement la formule originelle de la loi normale en faisant le calcul pour chacun des évènements $k = 0,1, 2... f$, comme nous l'avons expliqué précédemment. Par cette méthode qui fait l'approximation de la loi binomiale par la loi normale, on obtient la courbe en cloche en traits pleins de notre graphique. Cette fois nous nous rapprochons sensiblement du diagramme hypergéométrique.

- Les deux corrections se recouvrant exactement, on préférera la première qui est plus facile à mettre en œuvre et qui l'emporte aussi en simplicité sur un troisième procédé où l'écart réduit se calcule ainsi :

$$z = 2\left(\sqrt{(k+1)q} - \sqrt{(f-k)p}\right).$$

3°) Voilà donc levées les deux premières hypothèques. Quoique de type exhaustif et discontinu, les distributions lexicales des grands corpus peuvent être soumises à la loi normale. Reste la dernière hypothèque qui concerne la symétrie. Notre exemple nous suggère que la symétrie n'est pas parfaite : l'extrémité gauche est tronquée, au moment où *k* prend son départ (qui ne peut être que 0), alors que la loi normale poursuit imperturbablement son chemin de part et d'autre vers l'infini[2]. La troncature s'exerce aussi à droite, *k* ne pouvant dépasser *f*. Mais en général la distorsion est moins grave de ce côté, sauf lorsque *p* est proche de 1, c'est-à-dire lorsque le sous-ensemble *t* tend à se rapprocher de l'ensemble *T*. À l'endroit de la troncature, l'effet de la symétrie imposée par la loi normale contredit celui de la continuité. La continuité (quand elle n'est pas corrigée) de la loi normale affaiblit les probabilités ; au contraire la symétrie prolongée au-delà des valeurs 0 et *f* ajoute aux très faibles probabilités une plus-value injustifiée. Et c'est précisément ce qui se produit dans l'exemple du mot ABETI. Son absence dans la première tranche, dont la probabilité réelle est de 0,008, paraît moins surprenante à la loi normale qui estime à 0,012 les chances de cette observation.

[2] En réalité, si le mode coïncide avec un nombre entier ou se situe à mi-chemin entre 2 entiers, la symétrie est parfaite, au moins jusqu'au point de troncature.

La conclusion qu'on peut tirer de cet exemple particulier est que les grands corpus autorisent un emploi prudent de la loi normale. Mais nous avons choisi exprès les conditions les plus défavorables avec des paramètres *k* et *f* très faibles, compte tenu de la taille de notre corpus. La fréquence 55 dans un corpus de 70 millions de mots est le fait d'un vocable rare. Dans les corpus ordinaires, qui sont au moins cent fois plus faibles, cette fréquence correspondrait tout au plus aux hapax et échapperait à toute partition et à tout calcul de ce genre. En fait notre exploitation du corpus de Nancy n'a jamais dépassé ni même atteint ce cas-limite et en règle générale nous avons posé la barre dix fois plus haut, à $f = 500$, ce qui circonscrit utilement le foisonnement de la recherche des spécificités lexicales, puisque à ce niveau, il reste encore près de 7 000 vocables différents à exploiter.

1°) Le verbe ABIMER qui a 575 occurrences, se situe aux abords de cette nouvelle frontière. Cette fois-ci, nous envisageons la septième tranche (de 1870 à 1880) du corpus. *k* peut y recevoir un nombre suffisant de valeurs de chaque côté du mode (33), pour que les effets de la discontinuité et de la dissymétrie soient effacés. Si nous superposons la distribution de la loi normale, et celle de la loi hypergéométrique, la coïncidence est presque parfaite. Et pour apprécier la probabilité de la fréquence effectivement rencontrée ($k = 38$), on peut recourir à l'un ou à l'autre procédé sans changer le résultat ($p = 0{,}209$ et $0{,}207$ respectivement). Même en l'absence de correction de la loi normale, le résultat est satisfaisant ($p = 0{,}19$). À plus forte raison lorsqu'on envisage des mots moins rares, les résultats tendent à se confondre. C'est le cas du substantif ABIME représenté dans la figure 2. Les courbes y ont été établies non sur les probabilités mais sur l'abscisse de ces probabilités dans la représentation standardisée, c'est-à-dire sur la valeur de l'écart réduit[3]. L'avantage est qu'on réduit ainsi l'échelle des variations et que de -10 à +10 on peut parcourir une gamme très étendue de probabilités (jusqu'à 10^{-24}). Il y a superposition pure et simple dans le cas du mot ABIME et rien n'y subsiste des légers écarts qu'on observait dans l'exemple du vocable ABETI. D'ailleurs, même dans ce dernier cas, les différences qui paraissent les plus nettes (par exemple dans la tranche de 1875, deux points séparent les valeurs de *z*) sont insignifiantes quand on les exprime en termes de probabilité ($p < 10^{-10}$).

[3] Bien sûr la loi hypergéométrique ne donne pas directement un écart réduit mais une formule simple permet la conversion de toute probabilité en écart réduit et la transformation inverse (qui est plus usuelle).

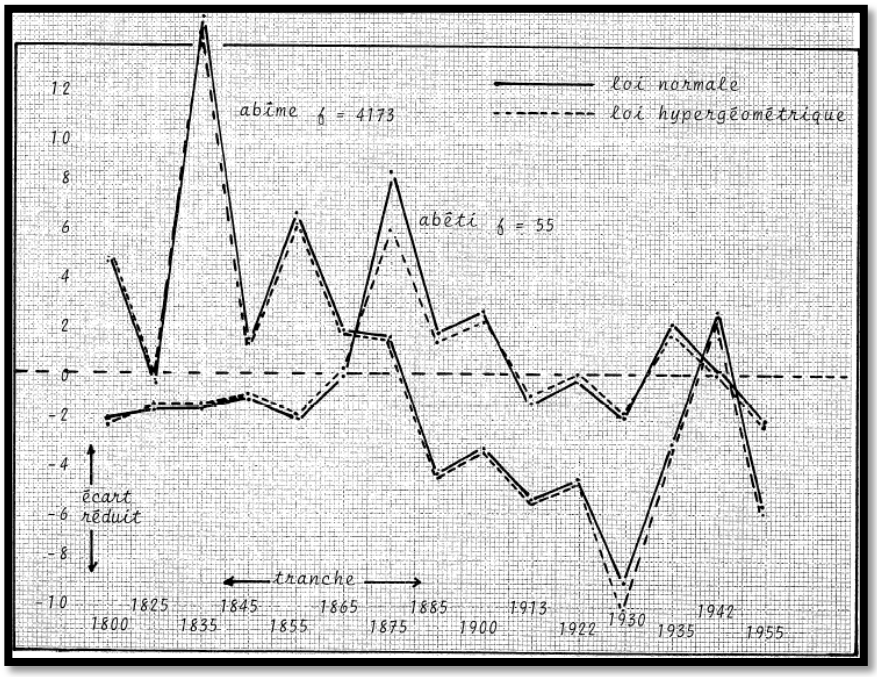

Figure 2 : Courbes des vocables « abêti » et « abîme » du TLF

2°) Le tableau ci-dessous montre que les différences s'amenuisent encore lorsqu'on considère les mots fréquents et qu'elles ne portent plus guère alors que sur la 2ème ou 3ème décimale de l'écart réduit.

Tableau 3 : Différence (valeur absolue) des écarts réduits, calculés selon la loi hypergéométrique et selon la loi normale

	rangs	abêti f = 55	abblu-tion f = 100	aberra-tion f = 225	abîmer 470	abîme 4284	abord 24597	après 69730	au 374026	à 1396290
zone des déficits	1	0,158	0,041	0,362	0,470	0,034	0,305	-	-	-
	2	0,114	0,070	0,042	0,392	0,272	0,249	0,103	0,062	-
	3	-0,039	-0,181	0,035	0,091	0,280	0,270	0,047	0,055	-
	4	-0,055	-0,227	-0,006	0,057	0,157	0,239	0,033	0,041	-
	5	-0,179	-0,259	-0,058	0,020	0,144	0,125	0,004	0,017	-
	6	-0,182	-0,229	-0,076	-0,020	0,052	0,132	-0,004	0,016	0,041
	7	-0,268	-0,268	-0,129	-0,061	0,052	-0,006	-0,007	0,007	0,047
	8	-0,285	-0,248	-0,184	-0,061	-0,038	-0,016	-0,009	-0,001	0,047
zone des excédents	9	-0,346	-0,254	-0,096	-0,080	-0,041	-0,013	-0,007	-0,007	0,046
	10	-0,090	-0,256	-0,103	-0,084	-0,045	-0,058	-0,008	-0,006	0,050
	11	-0,078	-0,027	-0,098	-0,126	-0,052	-0,123	-0,015	-0,013	0,049
	12	-0,317	-0,151	-0,141	-0,154	-0,082	-0,155	-0,088	-0,016	-
	13	-0,395	-0,198	-0,175	-0,171	-0,194	-0,257	-0,157	-0,062	-
	14	-0,489	-0,440	-0,452	-0,232	-0,403	-0,282	-0,189	-0,088	-
	15	-2,579	-1,014	-1,028	-0,718	-	-	-0,215	-	-

Précisons que dans chaque série les éléments ont été classés dans l'ordre croissant des écarts réduits (du plus grand des déficits au plus grand des excédents). Les valeurs du tableau représentent le correctif qu'il faudrait donner (en valeur absolue) à l'écart réduit de la loi normale, si l'on voulait ajuster la probabilité à celle de la loi hypergéométrique. On voit que la correction est généralement une diminution, qui reste faible relativement aux valeurs de l'écart réduit, et cela d'autant plus que la fréquence du mot augmente. Nous avons déjà donné l'explication de la tendance de la loi normale à surestimer les écarts et à sous-estimer les probabilités correspondantes. Si l'on avait ajouté aux données la valeur 0,50 pour ajuster la continuité de la loi normale au caractère discret des fréquences, le correctif aurait été moins fréquemment négatif. Ce qui est nouveau, c'est le renversement de tendance à mesure que la fréquence augmente. Dans les hautes fréquences, la loi normale surestime les probabilités qui se situent loin de la zone centrale de la courbe, de part et d'autre du mode. Cela est dû au tirage non exhaustif, dont l'effet cesse d'être tout à fait négligeable lorsque *f* augmente et se rapproche de *T*. On le comprendra à l'aide d'un exemple simple où la loi binomiale est utilisée successivement de manière exhaustive et non-exhaustive. Donnons à *T*, *t* et *f* les valeurs respectives 20, 5 et 5. Comme on tire 5 éléments, le binôme est au 5e degré et se développe ainsi :

Tableau 4 : Tirages exhaustif et non exhaustif

n	prob (k=n)	tirage 1 non exhaust.	tirage 2. exhaustif (avec modification des valeurs de p et q à chaque épreuve)		écart entre 1 et 2
0	q^5	0,237305	$\frac{15}{20}\cdot\frac{14}{19}\cdot\frac{13}{18}\cdot\frac{12}{17}\cdot\frac{11}{16}$	0,193692	+ 0,043613
1	$5pq^4$	0,395508	$5\cdot\frac{5}{20}\cdot\frac{15}{19}\cdot\frac{14}{18}\cdot\frac{13}{17}\cdot\frac{12}{16}$	0,440209	- 0,044701
2	$10p^2q^3$	0,263672	$10\cdot\frac{5}{20}\cdot\frac{4}{19}\cdot\frac{15}{18}\cdot\frac{14}{17}\cdot\frac{13}{16}$	0,293473	- 0,029801
3	$10p^3q^2$	0,087891	$10\cdot\frac{5}{20}\cdot\frac{4}{19}\cdot\frac{3}{18}\cdot\frac{15}{17}\cdot\frac{14}{16}$	0,067724	+ 0,020167
4	$5p^4q$	0,014648	$5\cdot\frac{5}{20}\cdot\frac{4}{19}\cdot\frac{3}{18}\cdot\frac{2}{17}\cdot\frac{15}{16}$	0,004837	+ 0,009811
5	p^5	0,000977	$\frac{5}{20}\cdot\frac{4}{19}\cdot\frac{3}{18}\cdot\frac{2}{17}\cdot\frac{1}{16}$	0,000064	+ 0,000913
		1,000000		1,000000	0,000000

Le tirage non exhaustif augmente les probabilités éloignées du mode ici ($\frac{5\times5}{20} = 1{,}25$), au détriment des valeurs de *k* plus proches. Or, dans le

cumul des probabilités qu'on réalise pour calculer Prob ($k \leq n$) ou Prob ($k \geq n$), on ne rencontre pas les valeurs centrales (sauf si *n* est égal au mode). En partant de *n* en direction de 0 ou de *f*, on ajoute les unes aux autres des probabilités un peu gonflées. Si donc les effets de troncature s'effacent dans les hautes fréquences, où il devient quasi impossible que *k* atteigne 0 ou *f*, la dissymétrie ne disparaît pas pour autant car les modifications de p et de q dans un tirage exhaustif ne sont pas équivalentes (sauf si $p = q = 0{,}50$). L'effet du tirage exhaustif est plus sensible à gauche, dans la zone des déficits, là où la probabilité forte (q) est soumise à l'exponentiation. Ainsi pour $T = 100$, $t = 25$, $f = 50$, la loi binomiale rendue exhaustive[4] donne $\text{Prob}(k \leq 9) = \text{prob}(k \geq 16) = 0{,}082577$ (car 9 et 16 sont symétriques de part et d'autre du mode 12,5), alors que le modèle binomial non exhaustif propose respectivement les valeurs 0,163684 et 0,163083, qui non seulement sont supérieures à la valeur exacte, mais diffèrent entre elles. C'est pourquoi dans le tableau 3, la non-exhaustivité, plus sensible dans la zone des déficits, y combat plus efficacement l'effet de la continuité. D'où la diagonale du graphique, où se marque le renversement dissymétrique de la tendance. Mais il faut répéter que ces distorsions sont extrêmement faibles dans les grands corpus et qu'elles ne mettent en cause ni les classements ni les conclusions qu'on peut établir à partir des résultats de la loi normale. Les courbes qu'on dessine avec la loi normale sont superposables à celles du modèle géométrique, et il faudrait une loupe pour distinguer deux points qui ont pour ordonnées 3,861 et 3,867, ces valeurs étant fournies respectivement par les lois hypergéométrique et normale pour l'article AU dans la 6e tranche.

Ajoutons qu'avec un corpus de 70 millions d'unités la marge de sécurité est considérable. Une expérimentation le prouve qui consiste, pour f et k constants, à réduire t et T d'un facteur 10, 100, 1000 et davantage. La loi normale qui n'envisage que le rapport t/T fournit évidemment les mêmes résultats. Mais ceux de la loi hypergéométrique bougent à peine :

[4] La loi hypergéométrique produit exactement le même résultat.

Tableau n° 5 : Réduction du corpus. Résultats dans la 1ère tranche

	f	k	p	1/10	réduction de T et t 1/100	1/1000
abêti	55	0	0,8340 E - 2	0,8338 E -2	0,8324 E-2	0,8180 E-2
aberration	225	5	0,1168 E - 3	0,1166 E -2	0,1147 E-3	0,9727 E-4
abîmer	470	20	0,2055 E - 5	0,2041 E -5	0,1908 E-5	0,9313 E-6
abîme	4824	437	0,7859 E - 6	0,7336 E -6	0,3521 E-6	--
abord	24597	1673	0,1768 E -18	0,5087 E-19	0,4625 E-27	--

On constate qu'avec un corpus dix fois moindre la loi hypergéométrique aurait encore à peu près le même effet et coïnciderait le plus souvent avec la loi normale.

3°) La question se pose alors de l'utilité du modèle hypergéométrique dans les grands corpus. On peut penser que ce modèle mérite bien son préfixe, et que la loi normale mérite bien aussi son nom. Puisque les deux modèles se rejoignent dans les grands nombres, il est normal de préférer l'outil le plus simple, le plus rapide et le moins coûteux, plutôt qu'un instrument hypersensible dont la haute précision est payée trop cher. Nous avons pu calculer que le coût de la loi hypergéométrique est en effet 100 fois supérieur à celui de la loi normale et le tableau ci-dessous donne le rapport de coût pour différentes fréquences, pour différentes précisions et pour différentes machines.

Tableau n°6 : Comparaison des temps de calcul du modèle hypergéométrique et de la loi normale

	En millièmes de seconde pour une série de 15 résultats sur ordinateur Iris 50				En secondes pour 1 résultat sur calculatrice HP 97	
	loi normale	hypergéométrique 10^{-8}	10^{-16}	10^{-48}	loi normale	hypergéométrique 10^{-8}
abêti	10	562	773	747	1	35/46
abblution	9	603	856	797	1	40/52
aberration	10	806	921	1089	1	45/98
abîmer	10	727	1075	1202	1	49/125
abîme	9	742	1168	1796	1	40/175
abord	9	788	1375	3019	1	60/470
TOTAL	57	4228	6168	8650	6	617
rapport de coût hyper/ normal		74	108	152		103

Ce n'est pas le lieu ici de développer la technique de calcul hypergéométrique. Rappelons seulement qu'il fait intervenir des factorielles, qui sont impraticables lorsqu'il s'agit de grands nombres, et

qu'il faut convertir en logarithmes. Mais le logarithme de T! lorsque T vaut 70 millions est lui-même un nombre de 10 chiffres. On est donc obligé de faire tous les calculs en double précision, ce qui ralentit l'ordinateur. Le calcul de la factorielle et de son logarithme est bien entendu indirect et fait appel à l'approximation de Stirling. Là-dessus je renvoie le lecteur à l'article de Pierre Lafon dans *Travaux de Lexicométrie et de lexicologie politique*, n°2, nov. 77, p. 100. Je me contenterai d'indiquer le maillon manquant dans la chaîne des calculs qu'on y trouve détaillés : $\log n! = n \log n - n + \frac{\log 2\pi n}{2} + \frac{1}{12n}$.

Or si l'on veut circonscrire la probabilité qu'un évènement ait lieu *k* fois, il faut appliquer à 9 termes différents ce calcul complexe qui comporte 11 opérations élémentaires. À partir de là, pour trouver les probabilités inférieures ou supérieures à *k*, il faut procéder par récurrence à des itérations dont le nombre est parfois fort élevé s'il s'agit d'une haute fréquence. Ainsi pour le mot APRES nous avons compté une moyenne de 500 itérations, ce qui est d'un coût prohibitif[5]. Ajoutons que la précision extrême qui est requise rend malaisée l'utilisation des calculatrices de poche. Même les plus perfectionnées, comme le modèle Hewlett-Packard 97, ne disposent pas d'un nombre suffisant de chiffres significatifs et avec 10 chiffres seulement tous les résultats sont faux. Même l'ordinateur renâcle devant certaines difficultés : lorsque la probabilité devenait extrêmement faible (de l'ordre de 10^{-35}), la fonction exponentielle a cessé de fonctionner sur l'Iris 50 dont nous nous servions. Et c'est ce qui explique les lacunes de notre tableau 3. À ce dépassement de capacité du côté de la précision, s'ajoute un dépassement du côté de la mémoire, lorsque pour les fréquences faibles on constitue une table de résultats préétablis comme le recommande Pierre Lafon. Dans un très grand corpus, les fréquences faibles montent au moins jusqu'à 100. Le tableau exigera donc $\frac{100\times103}{2}$ = 5 150 colonnes et 15 lignes soit 77 250 éléments et 7 fois plus si nous distinguons – comme nous l'avons fait – les 7 genres littéraires dans chacune des 15 périodes. Il existe peu d'ordinateurs au monde qui puissent procurer au chercheur, pour un même tableau, plus de 2 millions d'octets de mémoire ($77\,250 \times 7 \times 4 = 2\,163\,000$).

[5] Voici pour chacune des 15 tranches chronologiques le nombre d'itérations nécessaires : 893, 295, 342, 406, 545, 821, 864, 905, 787, 610, 894, 761, 451, 597 et 99.

Faut-il insister davantage ? Dans les très grands corpus, non seulement la loi hypergéométrique est inutile, mais elle est ruineuse et impraticable[6]. On ne peut songer à l'appliquer à chacune des 105 sous-fréquences de chacune des 200 000 formes recensées dans le corpus du TLF. Pour des raisons budgétaires autant que scientifiques, la loi hypergéométrique doit être réservée aux petits corpus ou au contrôle partiel des grands. Ses promoteurs en conviendront parfaitement. Était-il nécessaire alors de pourfendre les moulins à vent ? Sans doute un peu, car si des tendances divergent actuellement en matière de lexicométrie ou de statistique linguistique (j'utilise les deux appellations pour ne pas prendre parti), il est nécessaire de se rendre compte que les mathématiques sont parfaitement neutres dans le débat. L'utilisation de la loi normale ou de la loi hypergéométrique n'est pas un choix d'école, elle est dictée par la taille du corpus ou la taille du budget. Point n'est besoin, pour justifier la loi normale ou la binomiale, d'imaginer une langue extérieure au discours et l'englobant, une population dont le discours ne serait qu'un échantillon. Quand on dispose d'un très grand corpus, point n'est besoin d'extrapoler et de s'aventurer aux frontières du vocabulaire et du lexique, aux confins du discours et de la langue. Bien au contraire, il est plus facile de s'enfermer dans un grand corpus que dans un petit. On y découvre tellement plus de paysages. On peut donc revendiquer pour la loi normale et ceux qui s'en servent le droit à la modestie et à la rigueur. S'en servir n'est pas outrepasser son objet mais y circuler librement et sans grande dépense. Il y a bien sûr des règles de circulation, et quelque danger quand l'espace est exigu. Peut-être alors vaut-il mieux marcher à pied en mesurant ses pas. Mais quand de grands espaces et de grands corpus s'ouvrent à l'appétit du chercheur, il serait sot de se priver d'un moyen si économique et si rapide de locomotion.

[6] Il n'est pas certain que la pratique en soit très aisée, là où elle est possible. Les résultats qu'elle traduit en termes de probabilités sous forme de virgule flottante sont souvent moins lisibles aux littéraires que le déchiffrage d'une écart réduit transcrit en virgule fixe.

DISCUSSION

Ont pris part à la discussion : Frédéric BON, Etienne BRUNET, Pierre LAFON, Bernard QUEMADA, Maurice TOURNIER.

P. Lafon conteste les affirmations d'E. Brunet sur l'économie réalisée. Il maintient que pour les textes traités à Saint-Cloud (50 000 à 250 000 occurrences) le calcul hypergéométrique est préférable.

F. Bon pense qu'il y a un problème de modèle et un problème d'approximation, un choix de modèle et un choix d'approximation. Il lui semble que l'hypothèse générale des travaux de lexicométrie selon lesquels la norme est une norme de corpus et non pas une norme de langue impose le modèle hypergéométrique. Il y a ensuite de la technique calculatoire, et à ce stade ce modèle peut être approximé par une loi : binomiale, de Poisson,etc...

E. Brunet est d'accord avec cela, le modèle qui convient est bien le modèle hypergéométrique. Mais il pense qu'on doit avoir recours dans les grands corpus à la loi normale. Les calculs de P. Lafon passent aussi par des approximations. Si on a moins de 200 000 occurrences, il faut, selon lui, passer par le modèle hypergéométrique. Pour sa part, il traite des corpus de plus de 700 000 occurrences, mais persiste à imaginer le modèle comme un modèle de corpus.

M. Tournier croit que derrière les problèmes de modèles ce qui est en jeu, c'est la conception qu'on se fait de l'étude du discours. Le raisonnement de Herdan qui consiste à dire qu'un corpus est représentatif de la langue, que la fréquence d'un mot peut être considérée comme un de ses attributs au même titre que les autres, lui paraît faux.

M. Tournier : « Méfions-nous des dictionnaires de fréquences, ce sont toujours des dictionnaires de corpus. Nous prônons une statistique du texte et pas une statistique de langue ».

B. Quémada : La langue française, c'est le corpus des corpus. M. Brunet travaille dans un corpus plus vaste qui va vers des aspects plus globaux ; nos amis de Saint-Cloud travaillent sur des corpus plus réduits qui se présentent avec d'autres ambitions, plus fines au niveau des interprétations.