Navigation dans les rafales

Étienne BRUNET

UMR 6039 « Bases, corpus et langage », Université de Nice, Faculté des Lettres, 06204 NICE

Abstract

The chief object of lexicometry sofar has been frequencies, little attention being paid to sequences. Yet, the divisions of a corpus, on which numbers and frequencies are based are an a priori notion which should sometimes be questioned. The study of sequences makes it possible to ignore divisions and follow the progress of a word or any other linguistic object throughout a corpus without having to stop at the barriers of texts. The graphic representation of each occurrence enables the eye to spot distributions in bursts or "rafales", to appreciate the lulls and the rhythm of speech. Several tests are suggested and experimented in order to measure the corresponding distribution, among which the probability measure devised by Pierre Lafon within the framework of the hypergeometric model. The remaining task is to compare the distributions of two or several words and to measure the distances between them. In order to achieve this, the hypergeometric model is again used rather than other well-known indices such as Dunning's “likelihood ratio" or Church's "mutual information".

Résumé

La lexicométrie s’est surtout attachée jusqu’ici aux fréquences, sans trop s’occuper des séquences. Or la segmentation du corpus sur laquelle effectifs et fréquences sont fondés est un a priori qu’on doit parfois remettre en question. L’étude des séquences permet de s’en affranchir et de suivre le parcours d’un mot ou de tout objet linguistique tout au long du corpus, sans s’arrêter aux barrières des textes. La représentation graphique de chaque occurrence permet à l’œil de repérer les rafales, les bonaces et le rythme du discours. Pour mesurer la répartition constatée, plusieurs tests sont proposés et mis à l'épreuve, dont la mesure en probabilité établie par Pierre Lafon à partir du modèle hypergéométrique. Reste à comparer la répartition de deux ou plusieurs mots et à mesurer la distance qui les sépare ; là encore le modèle hypergéométrique est adopté et préféré à d'autres indices connus, comme le "rapport de vraisemblance" de Dunning ou "l'information mutuelle" de Church.

Mots-clés : topologie, séquences, cooccurrences, proxémie, calcul hypergéométrique.

1. Introduction

Les fonctions statistiques supposent habituellement une segmentation du corpus en textes séparés et sont fondées sur des fréquences ou des effectifs observés dans ces textes. Cette segmentation se justifie souvent parce que les textes rassemblés se distinguent par la date, le genre, l’auteur, le thème ou tout simplement le titre. Quand il s’agit d’œuvres distinctes, la partition du corpus semble aller de soi. Pourtant cela ne va pas parfois sans quelque arbitraire. Les nécessités du traitement posent souvent des problèmes de sélection, de jalons et de frontières et imposent des équilibres, des regroupements ou des sectionnements. Or paradoxalement la décision doit être prise avant que le traitement puisse l’éclairer. Au stade du traitement, la statistique, étant essentiellement comparative, est amenée à durcir les oppositions entre les parties de l’ensemble. Et la segmentation initiale, bonne ou mauvaise, se trouve alors artificiellement justifiée.

Il serait de meilleure méthode de n’imposer aucune segmentation au départ de l’entreprise et de laisser à l’analyse du contenu le soin d’en suggérer une. C’est ainsi que procède souvent le

dépouillement des enquêtes d'opinion : le texte libre est traité à plat, en continu, et si des balises permettent de se repérer dans le flux textuel, elles restent étrangères au traitement et permettent seulement d'interpréter les oppositions et les classements qui ont été obtenus sans leur concours. Ainsi fait le logiciel Alceste, qui ne veut connaître strictement que les relations proches que les mots ont entre eux dans le discours, indépendamment de toute appartenance, qu'il s'agisse du métier, de l'âge, du sexe, de l'origine, de la fortune ou, dans le domaine littéraire, du genre, de l'auteur, de l'époque et du titre. C'est pourquoi, même quand la segmentation a été faite a priori, il est prudent de la contrôler par le truchement d'Alceste. Il convient aussi en s'appuyant sur les balises XML de préparer des partitions virtuelles, que l'on peut ou non activer, si le logiciel utilisé en donne la possibilité[1].

Ce n'est pas le cas malheureusement de notre logiciel Hyperbase, dont les jalons, une fois définis, ne peuvent être déplacés, sauf à refaire le traitement. Sans doute avons-nous jeté un pont vers Alceste, en lui fournissant les données pour en recevoir les leçons. Sans doute avons-nous aussi proposé une fonction qui ignore les barrières établies entre les textes, et, rétrécissant la fenêtre aux dimensions de la page, ne veut considérer que les relations de voisinage observées dans ce cadre étroit. Mais on a cru devoir aller plus loin et observer les mots dans le détail de leur répartition tout au long du corpus, en négligeant les frontières des textes, et sans se soucier de constituer des effectifs, des fréquences et des sous-fréquences, chaque occurrence étant considérée individuellement, indépendamment du texte où elle se trouve. On rejoint là une perspective ouverte par Pierre Lafon dans la dernière partie de sa thèse, où il oppose les séquences aux fréquences.

2.

2.1 Représentation graphique

Graphiquement la chose est aisée à représenter, soit qu'on établisse sur un plan la suite linéaire des paragraphes du corpus, en les répartissant symboliquement à la queue leu leu, de place en place et de ligne en ligne, comme l'écriture le fait pour les mots, soit qu'on adopte tout bonnement la ligne droite pour représenter dans l'espace la séquence temporelle du discours. Mais, dans ce dernier cas, la largeur réduite de l'écran impose une unité plus large que le paragraphe. Et nous avons choisi la page, d'autant que, contrairement aux paragraphes, les pages sont de longueur constante. Et pour être moins serré, le graphique emprunte la diagonale. Le mot recherché est représenté séquentiellement de la première occurrence (en bas à gauche) à la dernière (en haut à droite), chaque point étant déterminé par la position du mot dans le corpus - c'est l'abscisse - et le numéro de l'occurrence - c'est l'ordonnée. La diagonale obtenue est plus ou moins régulière selon que l'objet représenté est plus ou moins régulièrement distribué. Quand les points se rapprochent et s'orientent vers la verticale, il s'agit d'une « rafale », c'est-à-dire d'une concentration des occurrences due à quelque cause locale, thématique ou stylistique. Quand les points s'espacent et s'inclinent à l'horizontale, cela correspond à une raréfaction momentanée de l'objet recherché.

Le nombre de pages parcourues d'une occurrence à l'autre donne la mesure de la distance. Les pages étant de longueur voisine, cette mesure est plus fiable que celle des paragraphes et presque aussi précise que celle des mots. La distance est alors convertie en pixels. Mais, le nombre de pixels étant limité sur l'écran, on a réduit à un échantillon les mots très fréquents

[1] Le logiciel *Lexico3* offre cette opportunité.

(une occurrence sur 2 ou 3 ou plus), de telle sorte qu'on ait au maximum 360 points à représenter et ainsi une lisibilité acceptable.

Le graphique 1, précisément, représente un mot dont la fréquence est naturellement élevée dans un corpus romanesque[2]. L'*amour* n'est pourtant pas équitablement partagé. Il jaillit verticalement dans la *Nouvelle Héloïse*, dans *Indiana* de Georges Sand et dans *Un Amour de Swann* et il s'étiole en un mince filet languissant de Flaubert à Verne et Zola. Cela, l'histogramme l'aurait dit aussi, mais de façon abrupte et carrée, sans la finesse et la fluidité des détails, sans les changements de rythme qu'on observe à l'intérieur même d'un texte, par exemple dans *Les Chouans*, *Madame Bovary*, *Du côté de chez Swann*, ou *Le Temps retrouvé*. Précisons que si les lignes en pointillés symbolisent le passage d'un texte à un autre, cette grille a été surajoutée au graphique, comme les méridiens sur le globe terrestre, la segmentation du corpus n'entrant nullement en ligne de compte dans le traitement.

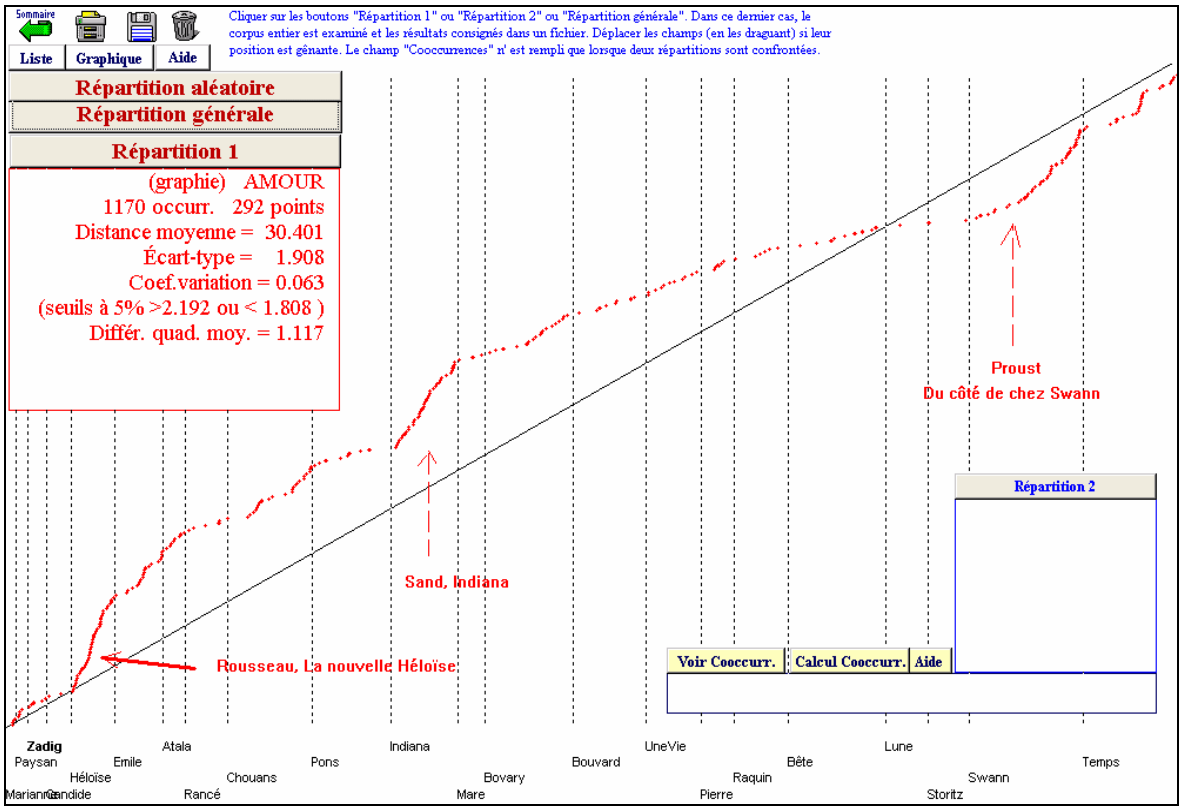


Figure 1. Représentation graphique du mot amour, *de Marivaux à Proust*

Cependant si l'œil est satisfait de la représentation, des variations de la pente et de la gradation des pleins et déliés, la raison a des raisons de s'inquiéter. Car elle ne sait trop si les sinuosités observées sont ou non le fruit du hasard. Alors que la hauteur des « bâtons » de l'histogramme donne une réponse brutale, mais claire et chiffrée, on a besoin ici d'outils supplémentaires pour apprécier en probabilité l'orientation de la courbe et confirmer

[2] En exploitant le même corpus déjà présenté aux JADT de Saint-Malo, on croit pouvoir s'autoriser à n'en donner qu'une brève description. Il s'agit d'œuvres narratives choisies dans la littérature française, de Marivaux à Proust, à raison de deux textes par auteur. La liste détaillée apparaît à la page 222 des JADT 2002. On a écarté ici les quatre derniers textes pour des questions de copyright. Restent en lice 11 écrivains : Marivaux, Rousseau, Voltaire, Chateaubriand, Balzac, Sand, Flaubert, Maupassant, Zola, Verne et Proust. C'est ce corpus-témoin qui sert d'exemple dans la diffusion de notre logiciel Hyperbase.

l'impression visuelle. On aura recours aux calculs traditionnels qui s'offrent à toute distribution : la moyenne des intervalles, l'écart-type de leur répartition et le coefficient de variation qui combine moyenne et écart-type. Mais leur témoignage est décevant car s'il reflète les irrégularités du profil, il dépend aussi de la fréquence du mot, rendant délicate la comparaison de deux distributions inégales. On a donc eu recours à d'autres mesures propres aux données sérielles.

2.2. Le test de la « différence quadratique moyenne successive »

La première est connue sous le nom de « différence quadratique moyenne successive ». Elle correspond à la formule :

$$\delta 2 = \frac{1}{f} \sum_{i=1}^{i=f} (x(i+1) - xi)^2$$

pour *i* variant de 1 à *f* (*f* = nombre d'occurrences de l'objet recherché).

Pour apprécier sa valeur, on doit la rapprocher de la variance σ2, où la sommation est celle des carrés des écarts à la moyenne[3]. Or δ2 est moins sensible que σ2 aux variations lentes, quand des intervalles de même type se suivent, soit courts, soit longs. Ce genre de distribution, qui correspond aux rafales, est marqué par un rapport δ2/σ2 plus faible que le seuil convenu. C'est le cas du mot *amour* où le test propose la valeur 1.117, qui est très inférieure à la limite basse 1.808 où l'on franchit le seuil de 5%.

La distribution inverse, où les distances sont soumises à une alternance rapide et régulière, peut rarement être observée dans les textes (sauf quand s'exerce quelque contrainte syntaxique ou prosodique). Le rapport δ2/σ2 est alors plus élevé que la limite dévolue au hasard. Les deux seuils qui fixent l'espace de l'hypothèse nulle sont calculés pour chaque fréquence[4] et la valeur du quotient δ2/σ2 est considérée comme significative lorsqu'elle échappe à la fourchette indiquée. Un exemple simple aidera à comprendre les deux cas de figure.

Une distribution en rafales est analysée dans la moitié gauche de la figure 2. Les 20 occurrences du mot s'agglutinent au début et à la fin, encadrant un passage où le mot est rare et les intervalles importants. Le rapport δ2/σ2 est alors faible, soit 0,50, en dessous de la valeur basse du seuil 1,04. Inversement la figure de droite propose une distribution alternée où les intervalles courts et longs se suivent de façon si systématique que le hasard ne peut en rendre compte. La valeur du test, 3,59, se situe au-delà de la fourchette haute 2,96. Le mécanisme se dérègle toutefois dans certaines distributions exclusives où toutes les occurrences sont agglutinées aux mêmes endroits. La figure 3 correspond à ce schéma : dans une surface de 118 pages, le mot n'apparaît que dans les 8 premières et les 10 dernières.

---

[3] Est-il utile d'en rappeler la formule $\sigma 2 = \frac{1}{f} \sum_{i=1}^{i=f} (xi - \bar{x})^2$?

[4] Selon la formule $\eta g, \eta d = 2 \pm 2 \mid \mathrm{ca} \mid \sqrt{\frac{\mathrm{N}-2}{(\mathrm{N}-1)(\mathrm{N}+1)}}$, la valeur *ca* étant lue dans les tables de la distribution normale, au seuil de 5%, soit 1,6449.

DIFFERENCE QUADRAT. MOYENNE SUCCESSIVE

fourchette pour seuil de 5% et pour N = 20 : 1,041 2,959

DISTRIBUTION PAR RAFFALES

	distance avec	écart à la	écart au	écart succes	écart au
1	1	-6	33	2	4
2	2	-5	23	1	1
3	2	-5	23	0	0
4	3	-4	14	1	1
5	1	-6	33	-2	4
6	4	-3	8	3	9
7	5	-2	3	1	1
8	14	7	53	9	81
9	15	8	68	1	1
10	12	5	28	-3	9
11	18	11	127	6	36
12	11	4	18	-7	49
13	12	5	28	1	1
14	14	7	53	2	4
15	5	-2	3	-9	81
16	6	-1	1	1	1
17	3	-4	14	-3	9
18	2	-5	23	-1	1
19	2	-5	23	0	0
20	3	-4	14	1	1
somme	135	s2=	29	d2=	15
moyenn	7				
				test =	0,50

RÉPARTITION ALTERNÉE (cycle court)

	cycle court	écart à la moyenne	carré	écart successif	carré
1	1	-6	36	8	64
2	10	3	9	9	81
3	2	-5	25	-8	64
4	13	6	36	11	121
5	1	-6	36	-12	144
6	10	3	9	9	81
7	1	-6	36	-9	81
8	8	1	1	7	49
9	3	-4	16	-5	25
10	12	5	25	9	81
11	4	-3	9	-8	64
12	11	4	16	7	49
13	5	-2	4	-6	36
14	14	7	49	9	81
15	5	-2	4	-9	81
16	13	6	36	8	64
17	3	-4	16	-10	100
18	13	6	36	10	100
19	2	-5	25	-11	121
20	9	2	4	7	49
somme	140	s2=	21	d2=	77
moyenne	7				
				test =	3,59

Figure 2. Distribution en rafales et distribution alternée

	distance exclusive	écart à la moyenne	écart au carré	écart successif	écart au carré
1	1	-5	24	0	0
2	5	-1	1	4	16
3	1	-5	24	-4	16
4	0	-6	35	-1	1
5	0	-6	35	0	0
6	0	-6	35	0	0
7	1	-5	24	1	1
8	0	-6	35	-1	1
9	100	94	8855	100	10000
10	1	-5	24	-99	9801
11	2	-4	15	1	1
12	5	-1	1	3	9
13	0	-6	35	-5	25
14	1	-5	24	1	1
15	0	-6	35	-1	1
16	0	-6	35	0	0
17	0	-6	35	0	0
18	0	-6	35	0	0
19	0	-6	35	0	0
20	1	-5	24	1	1
somm	118	s2=	468	d2=	994
moyer	5,90				
				test =	2,12

	distance	écart à la moyenne	carré	écart successif	carré
1	1	-5	24	0	0
2	5	-1	1	4	16
3	1	-5	24	-4	16
4	0	-6	35	-1	1
5	0	-6	35	0	0
6	0	-6	35	0	0
7	1	-5	24	1	1
8	50	44	1945	49	2401
9	50	44	1945	0	0
10	1	-5	24	-49	2401
11	2	-4	15	1	1
12	5	-1	1	3	9
13	0	-6	35	-5	25
14	1	-5	24	1	1
15	0	-6	35	-1	1
16	0	-6	35	0	0
17	0	-6	35	0	0
18	0	-6	35	0	0
19	0	-6	35	0	0
20	1	-5	24	1	1
somn	118	s2=	218	d2=	244
moyer	5,90				
				test =	1,12

Figure 3. Distribution exclusive

Un fossé de 100 pages se creuse au milieu et la rupture est comptée deux fois, quand on y tombe et quand on en sort. Il en résulte un accroissement intempestif de δ2, qui rend aveugle au caractère pourtant irrégulier de la distribution. La valeur 2,12 laisse croire que le hasard est seul en cause. Il suffirait de peu de chose pour que le test soit opérant : ainsi si une seule occurrence du mot se rencontrait dans les 100 pages désertiques (comme dans la moitié droite de la figure 3), le test deviendrait probant (1,12). Le cas de telles distributions déséquilibrées n'est pas exceptionnel dans la réalité. Il arrive qu'en accord avec le thème traité un même texte accapare toutes les occurrences d'un mot. Force est donc de corriger cette anomalie, ce qu'on peut faire avec une simple mesure d'écrêtement : toute valeur extrême qui dépasse un seuil convenu est ajustée à la valeur limite précisée par ce seuil. Certes ce correctif introduit dans notre programme amortit les secousses les plus brutales, mais il vaut sans doute mieux recourir à un test plus fiable et plus constant.

2.3. L'autocorrélation

Il y a peu d'amélioration à espérer d'un test proche, connu sous le nom d'autocorrélation. Les ingrédients sont les mêmes. La variance sert pareillement de pondération. Et la mesure est fondée semblablement sur l'écart entre un intervalle et le suivant (ou les suivants). Et la formule se présente comme une généralisation de la précédente, avec l'intervention d'un facteur supplémentaire : le décalage *h*, grâce auquel la mesure peut s'exercer non pas nécessairement sur le suivant immédiat, mais sur un élément décalé de *h* places. Cela permet de déceler des phénomènes de rythme, ou de phase, ou de mesurer la portée d'une contrainte ou d'une influence qu'un élément exerce sur son voisinage.

L'autocorrélation *Rh*[5] se présente ainsi : $$Rh = \frac{\sum xi \cdot x(i+h) - \bar{x}\sum xi}{\sum (xi - \bar{x})^2}$$

Appliquée au problème qui nous occupe, la formule met en œuvre un décalage de 1 et ses résultats ne diffèrent guère de ceux qu'on vient d'obtenir avec le test précédent. La spécialisation du mot *amour* y est pareillement soulignée par la valeur 0,44, très supérieure à la limite haute du seuil à 5% : 0,09. Mais aussi les mots exclusifs échappent à son emprise, comme le nom *Ralph* qui concentre ses 311 occurrences dans le seul roman *Indiana* et dont la distribution rencontre le point aveugle du test, lequel propose la valeur de -0,006, à l'intérieur de la fourchette de l'hypothèse nulle (-0,0962 < *rh* < 0,0898)[6].

2.4. Le test de Lafon

La mesure la plus appropriée semble être celle que propose Pierre Lafon, dans "*Dépouillements et statistiques en lexicométrie*"[7]. Elle repose elle aussi sur un calcul de variance un peu particulier, ainsi formulé : $$\frac{1}{f} \sum_{i=1}^{i=f} xi(xi-1)\ /2$$

On calcule la moyenne (ou espérance mathématique) et l'écart-type de cet indicateur, ce qui permet d'apprécier la probabilité du résultat obtenu, au moyen de l'écart réduit, soit :

z = (valeur observée - valeur théorique) / écart-type

La valeur théorique est calculée par la formule : $m = \frac{(T-f)(T-f-1)}{f(f+1)}$

et l'écart-type par : $\sigma = \frac{mT(T+1)(f-1)}{f(f+1)(f+2)(f+3)}$

où *f* désigne la fréquence du mot et *T* l'étendue du texte, évaluée avec l'unité de mesure choisie pour les distances xi, c'est-à-dire ici le nombre de pages du corpus.

---

[5] Rh, étant un coefficient de corrélation, évolue entre -1 et +1.

[6] Dans le cas d'une distribution exclusive, comme toutes les occurrences se rencontrent dans l'espace réduit d'un seul texte du corpus, l'intervalle est petit entre la première occurrence et la dernière (et à plus forte raison entre chacune et la suivante), mais il y a un désert immense entre la dernière et la première, car le parcours est conçu comme circulaire, la première page prenant la suite de la dernière. Il y a ainsi autant d'intervalles que d'éléments, et l'espace du corpus est balayé intégralement, quel que soit le mot considéré, rare ou fréquent, exclusif ou non.

[7] Pierre Lafon, *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, Genève-Paris, 1984 : 97-110.

Cette mesure est plus sûre que les précédentes dans les cas extrêmes où toutes les occurrences d'un mot sont concentrées dans un même texte du corpus. L'écart réduit atteint alors des valeurs très élevées, qui signalent les monopoles exclusifs. Ce sont souvent les noms propres.

En dehors de ce cas particulier, les différentes mesures accordent généralement leurs témoignages. Il faut toutefois éviter l'emploi du test de Lafon quand on procède par échantillonnage et que la fréquence trop élevée du mot considéré rend nécessaire l'extraction d'une occurrence sur 2 ou plus. Il en résulte un effet de lissage qui perturbe le calcul de la loi hypergéométrique. Bien entendu dans le cas inverse, celui des fréquences faibles, toutes les méthodes sont pareillement fragilisées et l'on évitera, dans cette situation, de leur faire une confiance absolue.

2.5. Application à des données aléatoires

Avant d'aller plus loin il convient de vérifier la validité des méthodes, en proposant au calcul un mot virtuel dont la répartition serait aléatoire.

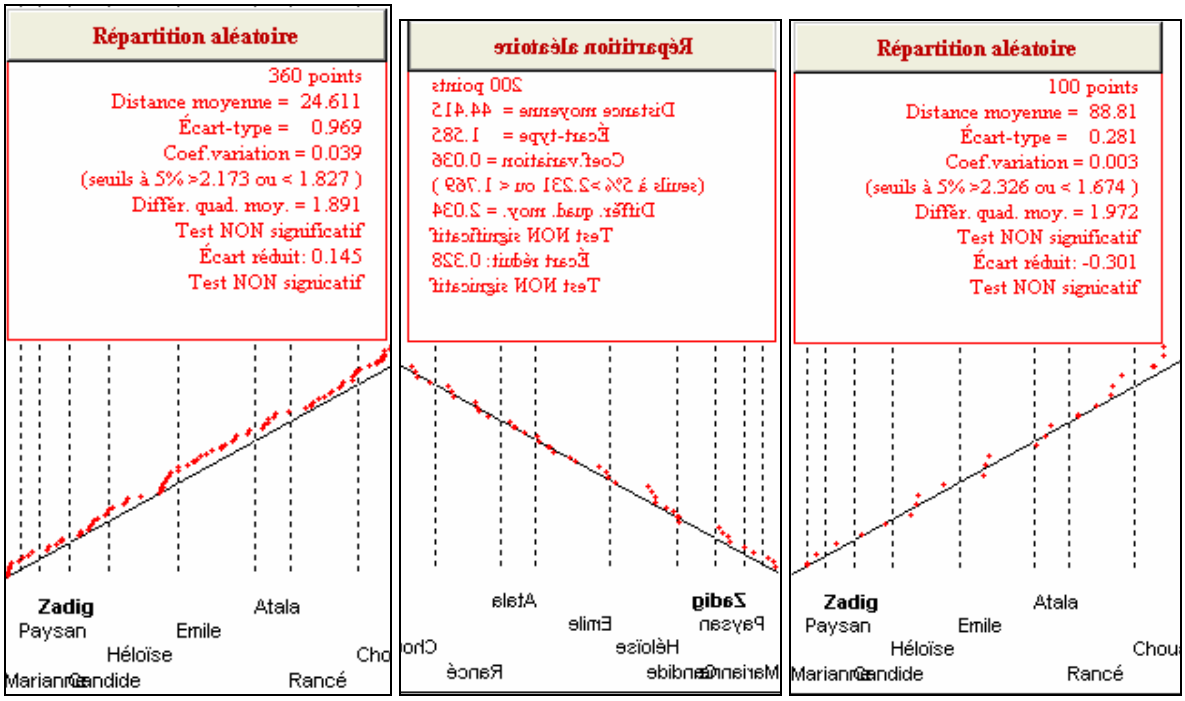


Figure 4. Test des formules avec des données aléatoires

Un générateur fournit des nombres au hasard, qu'on va considérer comme les positions du mot dans le corpus. Le résultat montre alors les points s'aligner sur la diagonale et les tests s'accorder avec le hasard : l'écart réduit a tendance à se rapprocher de zéro et le calcul de *delta* se maintient généralement dans la fourchette où l'hypothèse nulle ne peut être écartée. La figure 4 en montre quelques exemples pour les fréquences 360, 200 et 100.

2.6. Liste des distributions irrégulières

Fort de cette expérience concluante, on peut soumettre l'ensemble des mots du corpus au test de Lafon, ce que propose une fonction du programme. Le calcul exclut toutefois les mots de trop basse fréquence pour lesquels il perd son sens et ceux de haute fréquence (f > 360) pour lesquels il perd sa légitimité pour les raisons contingentes qu'on a dites. Les résultats, enregistrés dans un fichier ASCII accessible à l'éditeur, apparaissent successivement en ordre

alphabétique, puis selon un tri hiérarchique, fondé sur la valeur décroissante de l'écart réduit. La liste ne retient que les mots distribués en rafales, où le seuil significatif est atteint. Les premières places de la liste hiérarchique sont évidemment occupées par les personnages principaux de chaque roman, le plus souvent par les noms propres et parfois même ceux qui apparaissent dans le titre, comme c'est le cas pour *Zadig*, *Raquin*, *Storitz*, *Émile*, *Bovary*, *Chouans*, *Indiana*, *Atala*. La vertu heuristique est plus intéressante dans la suite de la liste où l'on rencontre des thèmes supportés par des noms communs, comme le *projectile* dans le roman de Verne *De la Terre à la lune* ou le *train* de la *Bête humaine*. On ne peut s'étonner de l'absence presque systématique des mots-outils puisque le critère de la fréquence a éliminé la plupart. Mais la rareté des verbes, des adverbes et des adjectifs s'explique autrement. Ces catégories ont des privilèges ou des exclusives moins affirmés que les substantifs. Ils s'accommodent plus aisément de situations, de thèmes, ou de genres différents, au lieu que le substantif a tendance à s'attacher, voire à s'identifier au texte où il apparaît, en ignorant les autres.

Cependant un corpus trop hétérogène comme celui que nous avons choisi ne rend pas justice à la vertu discriminante d'un test dont la sensibilité est brutalisée par la thématique trop diverse de romans indépendants, appartenant à des époques et à des auteurs différents. Si le corpus est homogène – et cette qualité est généralement requise – la finesse du test échappe aux constatations triviales. C'est le cas du corpus de la dernière campagne américaine, dont rend compte le tableau 6.

LISTE HIÉRARCHIQUE RÉPARTITION IRRÉGULIÈREtest fréq mot35.91 37 nuclear_3 31.96 33 korea_2 30.40 27 message 29.09 106 tax_2 26.49 65 troop_2 24.35 104 iraq_2 23.48 36 threat_ 23.39 36 raise_1 22.53 66 health_ 19.59 64 weapon_ 18.06 26 medicar 17.63 24 educati 17.01 49 child_2 16.79 57 care_2 15.31 26 mass_3	15.10 48 saddam_2 14.03 53 vote_1 13.96 50 opponent_2 13.89 21 citizen_2 13.73 42 pay_1 13.54 44 hussein_2 12.95 33 north_2 12.72 106 war_2 12.40 21 iran_2 11.89 112 world_2 10.96 52 wrong_3 10.94 42 right_2 10.72 35 drug_2 10.68 36 free_3 10.36 33 border_2 10.18 51 decision_2 9.97 24 mistake_2 9.68 24 budget_2	9.58 34 cost_2 9.53 29 law_2 9.46 103 plan_2 9.39 240 our_5 8.49 30 problem 8.29 26 destruction_2 8.26 31 terror_2 8.03 244 will_1 7.42 140 would_1 7.39 39 continue_1 7.22 44 fight_1 6.75 25 respect_2 6.49 41 lead_1 6.40 54 today_2 6.37 64 help_1 6.12 51 strong_3 6.11 26 business_2
MOTS RÉGULIÈREMENT RÉPARTIS test fréq mot 0.08 24 actually_6 -0.54 25 administration 1.26 30 after_9 0.56 21 all_6 0.06 66 all_7 -0.39 41 also_6 -0.63 35 back_6 -0.13 29 before_ -0.70 37 being_1 0.40 66 come_1 0.14 24 congres 0.26 32 day_2	0.21 23 difference 0.20 35 down_6 -0.58 25 find_1 -0.31 41 first_3 0.36 60 give_1 -0.32 22 happen_1 -0.30 77 his_5 -0.24 35 last_3 -0.50 66 let_1 -0.18 39 like_9 -0.19 24 listen_1 -0.47 38 lot_2 -0.36 23 off_6 -0.14 27 only_6 0.40 28 pass_1 -0.18 29 policy_2	-1.08 25 reason_2 0.36 39 right_6 -0.18 38 same_3 -0.69 25 so_9 0.24 26 something_7 0.47 79 talk_1 0.12 21 then_6 -0.29 45 these_7 -0.36 29 together_6 0.48 53 try_1 -0.08 26 under_9 -0.02 87 us_5 -0.40 45 very_6 -1.21 50 well_6

Tableau 6. Le débat Bush-Kerry de l'élection présidentielle américaine

Rappelons que les deux candidats Bush et Kerry ont été opposés à trois reprises dans des face-à-face télévisés, dont le premier était consacré à la politique étrangère, le second à l'économie, et le troisième aux questions sociales. Le test de Lafon souligne en effet les mots qui ont trait aux opérations militaires et aux questions économiques ou sociales, mais aussi ceux qui marquent l'opposition de deux hommes et de deux partis. Ainsi la guerre en Irak et la paix mondiale sont évidemment réservées au premier débat (*Iraq, troop, weapon, Iran, terror, Korea*), mais Bush préfère parler de Saddam Hussein et Kerry de la guerre et de son prétexte fallacieux (war, threat, nuclear). En soulignant les mots distribués irrégulièrement, même sans en préciser la cause (que le test des spécificités peut établir ultérieurement), le test de Lafon fixe la ligne de front, les points chauds où le débat est engagé. Ce sont le plus souvent des substantifs, comme le montre le code 2 attaché à cette catégorie et omniprésent dans la liste irrégulière. Les codes des verbes, des adjectifs, des pronoms, des adverbes ou des mots grammaticaux (respectivement 1, 3, 5, 6 et 7à 9) sont plus fréquents dans la liste régulière.

3.

3.1. Calcul hypergéométrique des cooccurrences

Il est facile de superposer deux distributions et de les distinguer par la couleur. Quant à les comparer plus précisément, la chose est délicate. Certes, s'il s'agit d'histogrammes, fondés sur la division du corpus en textes séparés, comme le nombre d'éléments est identique dans les deux distributions, la comparaison trouve un appui commode, soit qu'on se serve du coefficient de corrélation pour établir l'accord ou le désaccord des deux séries, soit qu'on établisse tout simplement le quotient des deux effectifs ou fréquences relevés pour chaque élément de la série. Mais, à la réflexion, que deux mots se trouvent avec des fréquences comparables dans le même texte ne prouve pas qu'ils ont des liens étroits. Il est possible qu'on les rencontre à des endroits différents du texte et jamais ensemble. La relation établie par le coefficient est au mieux un lien entre les textes, mais non entre les mots. Ce dernier lien ne peut être affirmé que si les deux mots apparaissent non pas dans les mêmes textes, mais dans les mêmes passages. Et par passage, il faut entendre une unité courte, qui peut être la phrase, le paragraphe, au maximum la page. C'est le paragraphe que nous avons choisi pour la fonction « thématique » de notre logiciel[8], c'est la même unité que nous considérons ici[9]. Seront retenus les paragraphes qui contiennent à la fois les deux mots considérés. Mais comme cette contrainte peut être estimée trop forte, la cooccurrence est aussi évaluée au niveau plus large de la page. Ces deux méthodes sont proposées simultanément, et aboutissent à un écart réduit. Les passages où sont observées les cooccurrences étroites, dans le même paragraphe, peuvent en outre être montrés en clair à l'écran.

Le calcul de la cooccurrence théorique s'appuie sur le produit de deux probabilités : celle qui est attachée au premier mot et celle qui est propre au second. Chacune de ces probabilités relève du calcul hypergéométrique, les paramètres étant fixés comme suit :

T = Nombre total de mots dans le corpus ;

[8] Cette fonction recherche tous les contextes d'un mot choisi pour pôle. Mis bout à bout les extraits constituent une sorte de texte composite, que l'on compare au corpus entier selon les voies habituelles (loi normale ou calcul hypergéométrique). Il s'ensuit une liste de spécificités, retenant les corrélats qui s'associent le plus souvent au mot-pôle.

[9] C'est aussi la voie que recommande Pierre Lafon, après avoir exploré d'autres chemins, cf « Dépouillements… » : 161.

t = nombre moyen de mots dans une page (ou dans un paragraphe) ;

f = fréquence du mot considéré ;

k = 0 (absence de ce mot dans la page ou dans le paragraphe).

Comme on ne s'intéresse qu'à une valeur unique pour k, on se trouve dispensé des itérations habituelles, nécessaires lorsqu'on doit cumuler les probabilités individuelles pour aller de 0 à k ou de k à f. Il suffit d'appliquer la formule brute, ainsi développée :

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

Ainsi disposé, le calcul évalue une absence. Le complément à l'unité de cette probabilité sert à mesurer une présence, c'est-à-dire les chances de rencontrer le mot dans l'espace considéré. L'opération est conduite pour chacun des deux mots, et le produit des deux résultats indépendants mesure les chances de rencontrer la cooccurrence des deux mots à la fois dans la même page (ou le même paragraphe). En multipliant cette probabilité élémentaire par le nombre de pages (ou de paragraphes), on obtient l'effectif théorique des cooccurrences. Reste à comparer l'effectif réel à l'effectif théorique, ce dont rend compte le calcul classique de l'écart réduit.

L'exemple de la figure 7 – où nous reprenons notre premier corpus - illustre le couple inséparable *vices/vertus* et la liaison forte qui assez souvent rapproche bon gré mal gré les antonymes. Pour que la dispute s'instaure entre eux, il faut bien qu'ils soient en présence l'un de l'autre. Visuellement le parallélisme des deux courbes saute aux yeux. Impression confirmée par le calcul : dans un océan de deux millions de mots, les 46 occurrences de *vices* et les 101 occurrences de *vertus* ne devraient se rencontrer ni dans le même paragraphe, ni dans le même page (l'effectif attendu est proche de zéro). Or la cooccurrence est observée respectivement 11 et 16 fois. Le même tropisme précipite l'un contre l'autre les deux partenaires au singulier : dans la bouche des moralistes le *vice* aime se frotter à la *vertu* (respectivement 11 et 15 fois). L'effet de style, même facile, est ainsi assuré. Encore faut-il que l'affrontement soit équilibré et que le singulier ne soit pas opposé au pluriel, ce qui rendrait l'antithèse boiteuse : le *vice* évite les *vertus* (une cooccurrence seulement) et la *vertu* les *vices* (un seul exemple également). Inversement les membres de la même famille n'ont rien qui les rapproche : *vice* et *vices* ne se rencontrent qu'une fois dans le même paragraphe, et de même la *vertu* ignore les *vertus*.

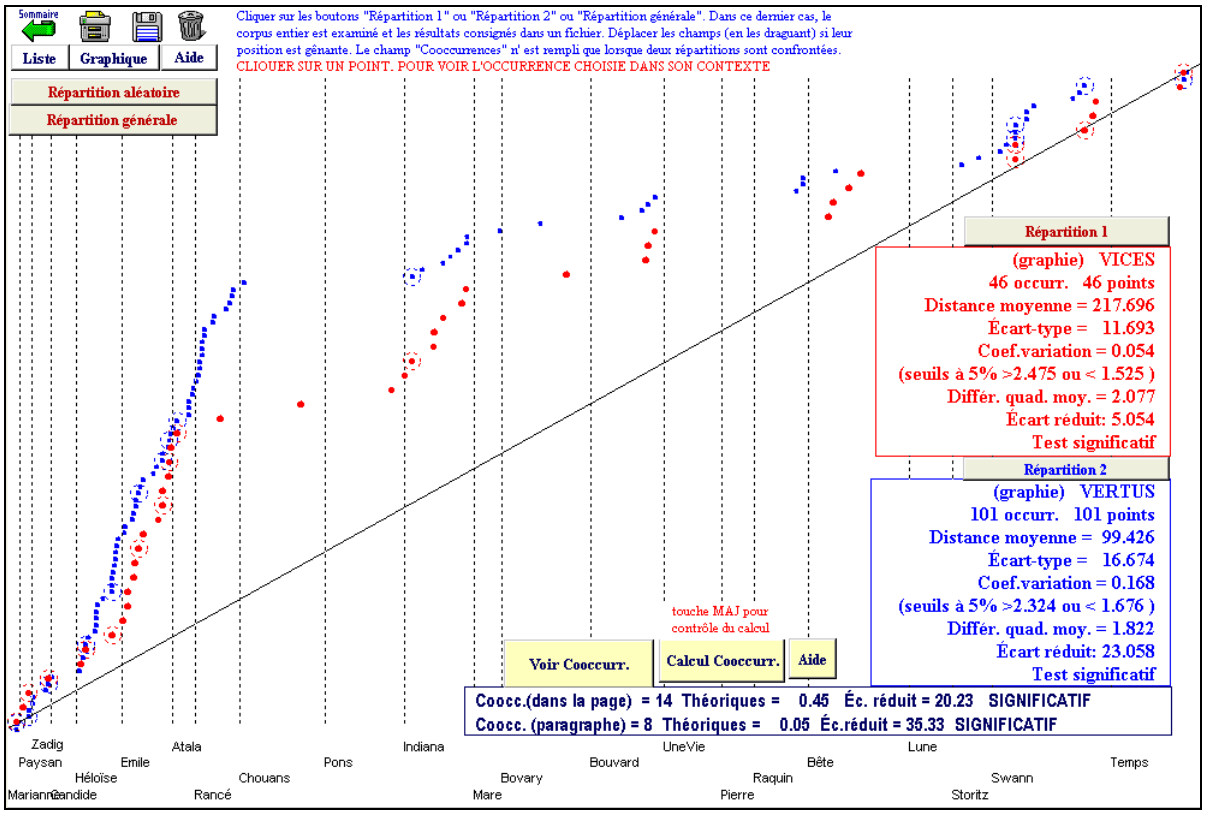


Figure 7. Comparaison de deux distributions. Vices *et* vertus

3.2. Comparaison avec d'autres modèles

Deux modèles classiques peuvent être comparés au calcul hypergéométrique - auquel on a donné la préférence, parce qu'en proposant un effectif théorique, il autorise le calcul d'un écart et d'une probabilité. Les autres indices n'ont pas cet avantage, même si leurs résultats s'accordent parfaitement avec notre modèle.

Le premier indice, ou *Rapport de Vraisemblance (RV)* a été proposé par Dunning en 1993. Il s'appuie sur quatre paramètres :

- *a* : nombre de cooccurrences des deux mots dans le champ exploré ;
- *b* : nombre d'occurrences du premier mot en l'absence du second ;
- *c* : nombre d'occurrences du second mot en l'absence du premier ;
- *d* : nombre d'occurrences des autres mots.

$$RV = -2 \log L = 2 (s1-s2)$$

pour $s1 = a \log a + b \log b + c \log c + d \log d + (a+b+c+d)\log(a+b+c+d)$

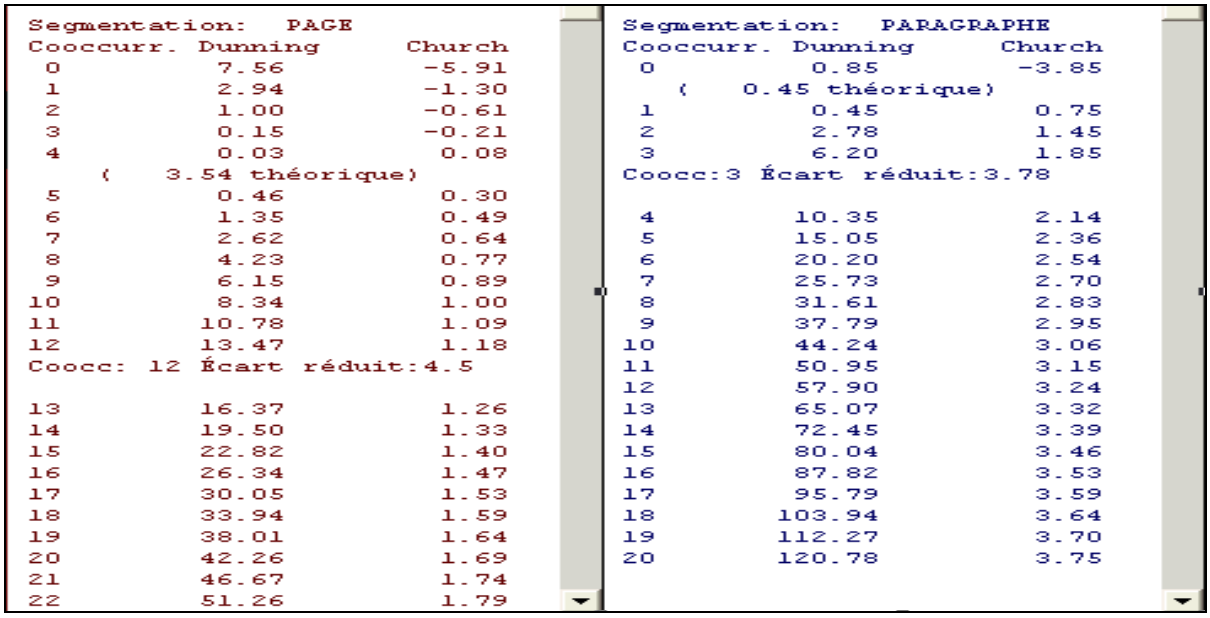
$s2 = (a+c)\log(a+c) + (b+d)\log(b+d) + (a+b)\log(a+b) + (c+d)\log(c+d)$

Le second indice est connu sous le nom *d'Information Mutuelle* (Church et Hovy, 1992). Là encore on a quatre données :

- *n* : nombre total de mots dans le champ exploré ;
- *n12* : nombre de cooccurrences, d'où $p12 = n12 / n$;
- *n1*: fréquence du premier mot, d'où $p1 = n1 / n$;
- *n2* : fréquence du second mot, *d'où* $p2 = n2 / n$;

La formule est très simple : $IM = \log (p12 / (p1*p2))$.

Notre programme met en regard les trois indices et l'on vérifiera, dans le tableau 8, que l'effectif théorique établi par la loi hypergéométrique correspond toujours à la valeur minimale des deux autres indices.



Segmentation: PAGE		
Cooccurr.	Dunning	Church
0	7.56	-5.91
1	2.94	-1.30
2	1.00	-0.61
3	0.15	-0.21
4	0.03	0.08
(3.54 théorique)		
5	0.46	0.30
6	1.35	0.49
7	2.62	0.64
8	4.23	0.77
9	6.15	0.89
10	8.34	1.00
11	10.78	1.09
12	13.47	1.18
Coocc: 12 Écart réduit:4.5		
13	16.37	1.26
14	19.50	1.33
15	22.82	1.40
16	26.34	1.47
17	30.05	1.53
18	33.94	1.59
19	38.01	1.64
20	42.26	1.69
21	46.67	1.74
22	51.26	1.79

Segmentation: PARAGRAPHE		
Cooccurr.	Dunning	Church
0	0.85	-3.85
(0.45 théorique)		
1	0.45	0.75
2	2.78	1.45
3	6.20	1.85
Coocc:3 Écart réduit:3.78		
4	10.35	2.14
5	15.05	2.36
6	20.20	2.54
7	25.73	2.70
8	31.61	2.83
9	37.79	2.95
10	44.24	3.06
11	50.95	3.15
12	57.90	3.24
13	65.07	3.32
14	72.45	3.39
15	80.04	3.46
16	87.82	3.53
17	95.79	3.59
18	103.94	3.64
19	112.27	3.70
20	120.78	3.75

Tableau 8. Accord des indices de Dunning et Church avec le calcul hypergéométrique

3.3. Liste triée des cooccurrences

On a pourvu à un traitement répétitif qui explore tous les couples, au moins ceux qui mettent en cause des substantifs. Comme il s'agit seulement de classer les cooccurrences les plus étroites, qui échappent sans conteste au hasard, le *Rapport de Vraisemblance* de Dunning offre un raccourci acceptable. L'œuvre intégrale de Gracq a fourni la matière à cette exploration, dont rend compte le tableau 9.

La cooccurrence est ici envisagée au niveau du paragraphe sans tenir compte de la place et de l'ordre des substantifs en présence. La syntaxe joue de ce fait un rôle atténué et la phraséologie des expressions figées et récurrentes reste plus discrète que dans le programme des segments répétés d'André Salem. Certains syntagmes quasi lexicalisés apparaissent toutefois en tête de liste, comme l'inévitable *chemin* (de) *fer* (ou *chef* (d')*œuvre, (à) perte (de) vue, (au) ras (du) sol, champ (de) bataille, coup (d')œil, point (de) vue, œuvre (d')art, (en) fin (de) compte).* Mais en dehors de ces scories qui viennent d'une lemmatisation incomplète, les accouplements se font par affinités sémantiques, avec des sèmes communs que les synonymes se partagent (*odeur parfum, timbre voix, cité ville*) et que les antonymes se disputent (*droite gauche, bas haut, auteur lecteur, lumière ombre, nord sud, bruit silence*). Parfois symétrique, la relation est plus souvent hiérarchique ou métonymique et manifeste la liaison de la partie au tout (*rue ville, maison village, personnage roman, rayon soleil, livre page, terrasse café, arbre branche, pan mur*), de la cause à l'effet (*bruit vague, feuille vent, bras geste, cri oiseau*) ou plus mollement un lien de proximité spatiale ou temporelle (*ciel nuage, ciel horizon, lune nuit, haie route, dîner soir, route voiture*).

LISTE HIÉRARCHIQUE		
test mot1 mot2	87.09 chemin route	62.20 brume soleil
	86.34 balzac stendhal	62.12 herbe prairie
284.66 perte vue	85.12 balzac roman	61.49 gare train
199.77 noir rouge	84.47 bruit oreille	60.84 lecteur roman
177.25 droite gauche	81.72 midi soleil	60.52 long rue
176.14 chef oeuvre	81.60 écrivain public	59.42 sable vent
168.34 ras sol	78.57 geste main	58.94 ton voix
156.33 coup oeil	78.29 air courant	58.64 bord lac
153.88 mer vent	78.15 mer plage	58.53 dîner soir
146.29 bois sous	75.90 berge eau	58.35 maison village
143.30 chemin fer	75.79 lecteur lecture	57.88 main visage
139.72 rayon soleil	75.06 mot sens	57.82 bord route
137.79 long route	72.67 cri oiseau	57.79 bord eau
134.45 heure matin	72.03 mot passe	56.86 nord sud
128.91 point vue	71.11 mur pan	55.89 mère père
125.83 bataille champ	70.78 compte fin	55.88 odeur parfum
122.08 bas haut	67.82 café terrasse	55.56 hauteur mi
112.53 horizon mer	67.57 livre page	55.11 bras geste
110.05 art oeuvre	67.23 lune nuit	54.56 fleur jardin
99.96 ciel nuage	66.51 ciel étoile	54.49 cité ville
99.34 feuille vent	66.38 mer oiseau	53.85 quartier rue
96.45 personnage roman	66.32 écrivain lecteur	52.49 bleu ciel
95.37 rue ville	65.98 bruit vague	52.44 arbre branche
92.07 ciel horizon	65.81 horizon ligne	52.15 eau feuille
91.14 auteur lecteur	64.49 esprit image	51.47 bruit silence
89.86 lumière ombre	64.49 horizon ras	51.29 fête rose
89.22 timbre voix	64.12 plage sable	50.97 haie route
	63.32 coup vent	
	62.38 route voiture	

Tableau 9. Liste triée des cooccurrences les plus étroites dans l'œuvre de Gracq

Ces faits sont prévisibles : les mots d'un texte sont aimantés par une attraction universelle, et la force qui s'exerce entre deux éléments est inversement proportionnelle à leur distance[10].

Resterait à représenter dans l'espace l'ensemble de ces relations et à s'aventurer dans la proxémie et la théorie des graphes. Notre exposé s'arrêtera provisoirement sur ce seuil, que d'autres ont franchi avec succès, à l'image de ce que Bruno Gaume a entrepris pour les verbes du *Grand Robert*[11], même si les cooccurrences dans le texte intégral ne bénéficient pas de la circularité qu'on observe dans les définitions du dictionnaire, ce qui limite la combinatoire.

3.4. La proxémie lexicale traitée par Alceste

Mais sans attendre on peut recourir à Alceste pour l'analyse de ces relations. Là aussi pour éviter les contraintes de la syntaxe, bornons-nous à la même catégorie : celle des substantifs lemmatisés, qui sont relativement indépendants dans la chaîne du discours, surtout quand leurs cooccurrences sont observées au niveau plus large de la page. On fournit donc à Alceste les 5000 pages que compte l'œuvre de Gracq, chacune étant réduite à une suite de substantifs. Les jalons des textes, marqués comme « variables étoilées », ne sont pas pris en compte dans

[10] Ajoutons qu'il ne s'agit pas seulement des couples lexicaux. Le programme des séquences et des cooccurrences s'applique aux graphies et aux lemmes, mais aussi aux codes grammaticaux et aux structures syntaxiques. On peut par exemple mettre en lumière les contraintes de la syntaxe et de la concordance des temps, et montrer graphiquement que le conditionnel est plus proche de l'imparfait que du futur. Au niveau des textes les effectifs observés pour les temps et les modes confirment cette observation : le coefficient de corrélation est positif pour la première paire (+0,51) et négatif pour la seconde (-0,35).

[11] « Désambiguïsation par proximité structurelle », *Taln 2004*, Fès. On peut voir les graphes ainsi obtenus sur le site du CRISCO, à l'adresse http://www.crisco.unicaen.fr/

l'analyse, même s'il apparaissent en superposition quand les jeux sont faits, la décision étant prise sans eux. Le résultat des votes est proposé dans la figure 10. Huit classes ont été distinguées, auxquelles nous avons donné un nom qui les résume au mieux. Les thèmes de la nature (campagne, ville et mer) sont attendus chez un écrivain qui passe pour être le meilleur paysagiste de notre littérature, comme aussi les débats critiques (sur le roman, la poésie et le surréalisme) chez un critique averti qui du haut du Montglonne contemple depuis longtemps le panorama littéraire. Il n'y pas de surprise dans cette analyse, sinon dans le fait qu'elle soit issue d'un traitement tout automatique et d'un vote populaire où chaque nom disposait d'une voix dans l'urne du village. Quand interviennent les textes, en superposition, dans un vote à deux degrés, où les fréquences se substituent aux séquences, le suffrage universel est confirmé, comme si le Sénat des corps constitués devait refléter l'image de l'Assemblée. Quatre partis prennent place aux quatre coins du graphique, selon qu'il s'agit de théâtre, de roman, de critique ou de réflexions personnelles (que Gracq désigne lui-même comme des «fragments »).

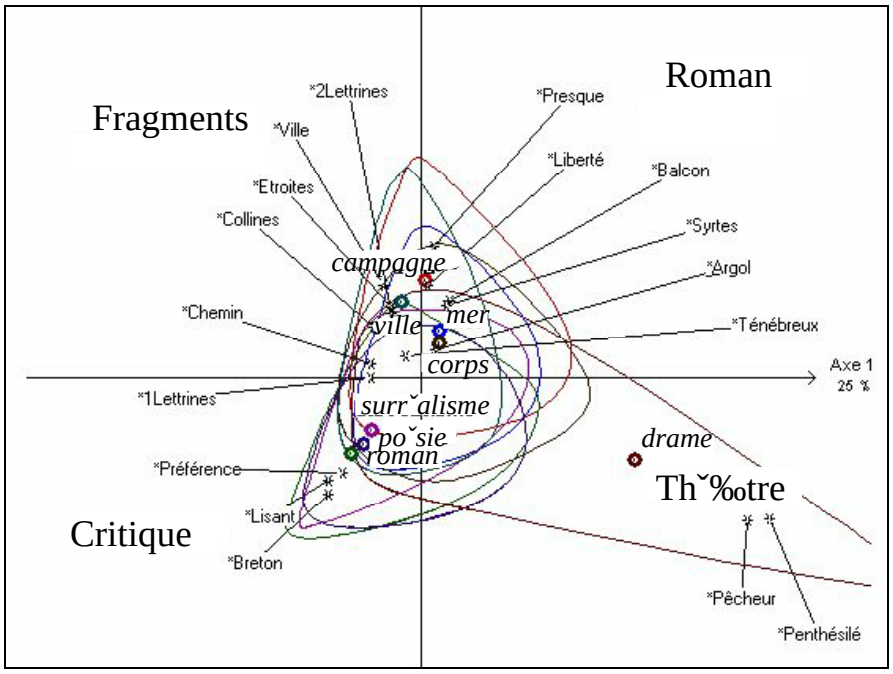


Figure 10. Analyse factorielle faite par Alceste sur les substantifs dans l'œuvre de Gracq

Il est heureux que les fréquences soient ainsi en accord avec les séquences, dont elles sont le reflet grossier et grossi. Trop souvent on s'est contenté de l'image abstraite renvoyée par les premières. En suivant les mots un à un et pas à pas dans leur environnement immédiat, l'étude des secondes favorise un juste retour au **texte,** je veux dire au microtexte, en écartant provisoirement les **textes**, je veux dire les divisions des textes et les barrières où ils encagent les mots.

Références

Church, K., and Hanks, P. (1990). Word association norms, mutual information and lexicography. In *Computational Linguistics*, 16(1) : 22-29.

Dunning T., Acurate Methods for the Statistics of Surprise and Coincidence. In Suzan Armstrong, *Using Large corpora*, The MIT Press,Cambridge, Massachussets : 61-74.

Kraif O. Extraction automatique de correspondances lexicales : évaluation d'indices et d'algorithmes. http://www.atala.org/doc/actes_taln/AC_0112.pdf

Gaume B. (2005). La proxémie : vers un modèle de sémantique lexicale pour un Traitement Automatique des Langues à Ergonomie Cognitive. http://www.limsi.fr/Individu/habert/04-05/index.html

Gaume B., Duvignau K. (2004). Pour une ergonomie cognitive des dictionnaires électroniques. In *Document numérique*, n° 3/2004, Vol.8 : 157-181.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion, Genève-Paris.