La répétition dans la phrase. Étude statistique

Etienne Brunet

La répétition dans la phrase. Etude statistique.

Étienne Brunet, *Bases, Corpus et Langage(CNRS)*, Nice

Le problème de la répétition paraît proche de la richesse lexicale, question que la lexicométrie a souvent abordée, jusqu'à satiété. Éviter la répétition, c'est gagner en variété et enrichir le texte. Ce qui diffère cependant c'est l'empan: la répétition a une portée courte et se limite généralement à la phrase. Certes en poésie ou dans la chanson, la répétition peut être recherchée à un niveau supérieur, de la strophe ou du poème. Et de même l'anaphore est prisée dans les couplets de l'art oratoire. Au delà, au niveau d'un texte entier, la répétition devient moins saisissable, même si certains écrivains comme Zola ont tenté d'adapter la technique du leitmotiv, même si l'on sent souvent dans l'œuvre des reprises de thèmes, des rappels de teinte, des isotopies sous-jacentes. Le point de vue diffère aussi en ce que le problème stylistique de la répétition n'échappe ni à la conscience, ni à la volonté de l'écrivain, au lieu que la richesse lexicale ne représente souvent qu'une caractéristique a posteriori, un rapport mathématique entre V (le nombre de vocables différents) et N (le nombre d'occurrences), dont l'écrivain ne se préoccupe guère en tant que tel. Alors que la musique est faite congénitalement de répétitions et variations, l'écrivain a horreur du déjà dit et le lecteur plus encore. Un vieux principe de rhétorique proscrit en effet la répétition dans la phrase, du moins lorsqu'il s'agit de mots-pleins et que la distance est trop courte d'une occurrence à l'autre. Du maître d'école qui pourchasse les répétitions, à l'écrivain qui multiplie les substituts et les périphrases, c'est la même tradition de variété, dont le principe ne doit rien à la logique et tout à l'esthétique. Quand ce principe est transgressé, ce que recommande Pascal dans certains cas[1], il s'agit d'un écart, doté d'un effet marqué ("C'en est la marque", dit Pascal).

Or Giraudoux - chez qui nous nous proposons de puiser nos premiers exemples - recherche assez souvent, parfois par plaisanterie, cet effet, qu'il renforce d'ailleurs en rapprochant jusqu'à la contiguïté les occurrences du terme répété dans une construction superlative inspirée de la Bible ou de l'Antiquité. Sur le modèle du *Cantique des Cantiques* ou du *Roi des Rois*, on trouve sous sa plume *l'ombre de l'ombre, le zénith du zénith, le silence des silences, l'heure des heures, l'oncle de l'oncle* ou avec un calembour *le signe du cygne*. Cet effet que Giraudoux appelle un écho[2] et que Valéry désigne en recourant à l'image visuelle du miroir[3] peut se reconnaître dans la variante du relatif "*sauver celle qui sauve la ville*"[4], ou de la conjonctive "*je crois que je crois aux dieux*"[5,] ou de la juxtaposition «*Selon que les nuages étaient blanc gris ou blanc blanc*[6] *»*. La répétition du nom là où un pronom personnel de rappel pourrait suffire marque souvent l'insistance, la force de la relation

---

[1] Pascal, Pensée 48, édition Brunschvicg.. « Quand dans un discours se trouvent des mots répétés, et qu'essayant de les corriger, on les trouve si propres qu'on gâterait le discours, il les faut laisser, c'en est la marque; et c'est là la part de l'envie, qui est aveugle, et qui ne sait pas que cette répétition n'est pas faute en cet endroit, car il n'y a point de règle générale.»

[2] Quand le chevalier se présente: "Hans Von Wittenstein zu Wittenstein", Ondine applaudit: "Que c'est joli l'écho dans un nom!".

[3] "Ô forme de ma forme", *La Jeune Parque*, Pléiade p. 109. "Je n'aime que le travail du travail". *Variétés*. Pléiade, p. 1500

[4] *Judith*, p. 250.

[5] *Électre*, I, scène 3, p. 21 .

[6] *Siegfried et le Limousin*. On peut citer encore la réplique célèbre de Jouvet dans *Drôle de Drame*: « Moi, j'ai dit bizarre… bizarre ? Comme c'est étrange…Pourquoi aurais-je dit bizarre… bizarre… », qui s'inspire probablement de *Juliette au pays des Hommes*, p. 27. "Eleveur d'animaux bizarres ? répéta-t-il. Expression bizarre, mademoiselle". En 1937, à la sortie du film de Carné, Jouvet et Giraudoux sont des intimes et l'on peut imaginer cette contribution de l'acteur aux dialogues de Prévert.

réflexive qui lie le sujet à lui même. Ainsi dans le roman *Églantine*, la suffisance épanouie de Moïse s'étale dans les épiphores de ce type "*La plupart des secrets de Moïse étaient des secrets même pour Moïse*"[7]. Dans tout le roman Moïse fait l'objet de 28 répétitions, contre 18 pour Eglantine et 17 pour Fontranges dont la personnalité s'affirme plus discrètement[8].

1 . La répétition dans un roman de Giraudoux -

En dehors de ces exemples manifestement voulus par l'auteur, la répétition a souvent le caractère de propension thématique ou de nécessité syntaxique. Le phénomène de spécialisation lexicale s'impose en effet avec force au niveau de la phrase. Dès que l'énoncé commence, l'éventail des possibilités de choix se referme, et plus l'idée se précise, plus la zone de disponibilité lexicale se rétrécit. Si la phrase s'ouvre avec le mot *cheval*, la probabilité d'emploi des mots *tiercé, course, écurie* se multiplie tandis que s'annulent les chances des éléments sémantiques qu'on associe difficilement au cheval. On peut même penser que si les contraintes thématiques jouent sans restriction, ce sont les mots déjà actualisés dans le début de l'énoncé dont les chances augmentent le plus dans la suite de la phrase: la logique du discours conduit à la répétition.

1 - Pour apprécier ce phénomène on peut imaginer un discours aléatoire indépendant de toute aimantation sémantique comme aussi de toute contrainte syntaxique, et calculer ce qu'y serait le nombre théorique des répétitions. Pour établir le modèle il faudrait connaître l'effectif des phrases, la longueur et le contenu de chacune. On doit être également en possession du tableau de distribution des mots du texte. Mais pour simplifier le raisonnement contentons-nous d'abord d'un seul mot, et cherchons à établir l'effectif théorique des emplois répétés de ce mot dans une même phrase. Il est évident que les chances de répétition varient avec la longueur de la phrase et la fréquence du mot. Si la phrase comporte 60 mots-occurrences (*n*) et si le mot considéré à la probabilité $p = 0{,}001$, on a l'espérance mathématique $np = 0{,}001 \times 60 = 0{,}06$ de voir apparaître une fois le mot dans la phrase. Pour espérer une seconde apparition du mot il reste 59 places possibles auxquelles la probabilité *p* reste attachée. L'espérance mathématique des répétitions du mot est donc égale au produit des probabilités composées soit: $np * (n-1)p = (n^2 - n) * p^2 = 0{,}0035$.
Mais il faut prendre garde que l'ordre d'apparition des deux éléments d'une répétition n'importe pas et qu'il convient d'envisager non le nombre d'arrangements n^2-n mais le nombre de combinaisons $(n^2 - n) /2$ et l'on aura donc $(n^2 - n) /2 * p^2 = 0{,}0018$ répétitions. Pour connaître le nombre de répétitions d'un mot dans un texte, il reste à cumuler les résultats obtenus pour chaque phrase particulière du texte, ce qu'on peut résumer dans la formule

$$\sum_2^n C_n^2 mp^2 \quad \text{ou} \quad \sum_2^n \frac{n^2 - n}{2} mp^2$$

n désignant le nombre de mots dans la phrase, et *m* l'effectif des phrases ayant *n* mots. On trouvera dans le tableau ci-dessous, le détail de ces opérations appliquées à la préposition "*sur*" dans *Églantine,* où la probabilité *p* attachée à ce mot est de 346/52397 = 0,006603 (d'où $p^2 = 0{,}006603^2 = 0{,}0000436$). Le calcul peut être simplifié si l'on met p^2 en facteur; il suffira dès lors de calculer le nombre total d'issues possibles soit :

$$\sum_2^n \frac{n^2 - n}{2} m = G = 991403$$

[7] *Églantine*, p. 58.
[8] Le nom de Moïse se prête d'autant mieux à la répétition qu'il intervient dans des phrases généralement plus longues, plus ornées, plus orientales, à l'image de Moïse lui-même. La moyenne des phrases est en effet de 26,57 mots, 26,08 et 25,51 dans les chapitres 2 et 3 et 6 consacrés à Moïse tandis qu'elle tombe à 22,99, 24,82 et 24,98 dans les chapitres 1, 4 et 7 où évolue Fontranges.

et de le pondérer par la probabilité p^2 des issues favorables pour un mot donné ayant la probabilité p, soit

$Gp^2 = 991\ 403 * 0{,}0000436 = 43{,}23$ répétitions

Tableau 1. Calcul de l'effectif théorique des répétitions de la préposition "sur" dans le roman Églantine

n	m	n	m	n	m	n	m	n	m	n	m
2	29	21	52	40	20	59	5	78	1	102	1
3	50	22	49	41	21	60	2	79	1	103	1
4	55	23	44	42	23	61	6	80	1	106	1
5	44	24	46	43	18	62	5	81	2	108	1
6	51	25	41	44	17	63	3	82	3	109	1
7	48	26	36	45	14	64	6	83	2	11O	1
8	57	27	36	46	23	65	6	84	1	123	1
9	68	28	33	47	23	66	7	85	3	136	1
10	59	29	35	48	13	67	3	86	2	155	1
II	74	30	39	49	13	68	7	87	2	168	1
12	66	31	32	50	12	69	4	88	1	180	1
13	43	32	36	51	11	70	2	89	3		
14	63	33	30	52	15	71	6	90	1		
15	60	34	30	53	13	72	2	91	1		
16	70	35	24	54	8	73	3	92	1		
17	39	36	21	55	3	74	3	93	1		
18	48	37	28	56	8	75	4	94	1		
19	51	38	30	57	9	76	1	98	2		
20	47	39	18	58	8	77	2	99	1		

p = probabilité du mot "sur"
= f / N = 346 / 52397 = 0,006603
$p^2 = 0{,}0000436$
n = nombre de mots dans la phrase
m = effectif des phrases ayant n mots

Tableau 2. Répétitions observées et théoriques dans certains mots du roman Eglantine

mot	fréquence	répétitions observées	répétitions théoriques
plus	362	63	47,32
même	206	27	15,32
comme	209	18	15,77
tous	123	28	5,46
tout	163	7	9,59
toute	57	6	1,17
toutes	61	5	1,34
peu	88	19	2,80
si	121	11	5,28
jamais	87	10	2,73
rien	31	2	0,35
très	11	2	0,04
chaque	59	7	1,26
déjà	37	1	0,49
loin	13	1	0,06
moins	47	3	0,79
toujours	49	2	0,87
trop	29	4	0,30
temps	45	5	0,73
dos	17	5	0,10
bois	10	2	0,04
bras	23	2	0,19
corps	41	3	0,61
mois	17	1	0,10
premier	20	4	0,14
pensée	20	1	0,14
regard	17	3	0,10
soir	42	5	0,64
signe	7	3	0,02
vérité	11	3	0,04
vraie	11	1	0,04
petits	13	1	0,06
suis	5	3	0,01
voyait	44	2	0,70
savait	39	1	0,55

2 - Or dans le roman on a compté 78 répétitions de la préposition "sur" à l'intérieur d'une même phrase. La différence 34,77 ne peut être aléatoire et correspond à un X2 de 27,97 (soit une probabilité très inférieure à 0,001). Or cet écart entre le discours et le hasard se retrouve dans la quasi totalité des mots, qu'il s'agisse de mots-pleins ou de mots-outils. La liste reproduite ci-dessus (tableau 2) est significative à cet égard : une fois seulement l'effectif réel reste en deçà de la valeur théorique ("*tout*" : 7 répétitions observées contre 9,59 attendues). Voir tableau 2.

3 – À vrai dire, ce sondage peut laisser quelque doute si l'on songe que la grande majorité des mots ne font l'objet d'aucune répétition et que la somme des valeurs qui représentent leurs chances de répétition peut constituer une masse appréciable qui compenserait les excédents constatés dans la liste ci-dessus. Il n'en est rien pourtant car ces mots sont peu fréquents et leur chance de répétition dans une même phrase est négligeable (elle est naturellement nulle pour les mots qui n'apparaissent qu'une fois dans le texte). Ainsi en étendant l'enquête à tous les mots du dictionnaire d'*Eglantine* à partir de la lettre p et jusqu'à la fin, on observe que la classe des mots de fréquence 2 - dont l'effectif est de 440 - ne

devrait compter aucune répétition. Malgré 991 403 issues favorables sollicitées 440 fois, le résultat est proche de zéro (exactement 0,636) car la probabilité p^2 est extrêmement faible :

$$p^2=(2/523297)^2 =1,46*10^{-11}.$$

De même la classe de fréquence 3 ne devrait fournir que 0,676 répétition et l'ensemble des 50 premières fréquences totalise un effectif théorique de 22,53 répétitions. On en relève en fait 161. La disproportion est flagrante également lorsqu'on tient compte des fréquences plus élevées (386 répétitions observées contre 146 attendues). Qu'il s'agisse de contrainte syntaxique ou d'influence thématique, la spécialisation lexicale se manifeste au niveau des phrases par un net excédent de répétitions.

Ainsi plutôt que de richesse lexicale, on devrait parler de pauvreté lexicale; quand l'homme puise dans la langue pour actualiser son discours, les mots qui lui viennent à l'esprit sont ceux du voisinage qu'un lien unit à la situation, au sujet, au genre, au registre. Certes un scrupule de sa conscience stylistique l'invite à s'abstenir de coucher dans la même phrase un mot-plein qui s'y trouverait déjà, comme s'il s'agissait d'un inceste. Mais cette réticence se tait souvent et cède devant l'urgence, quand le substitut manque ou que la phrase trop engagée ne permet plus d'éviter l'interdit. Si le hasard présidait seul à la fourniture du matériel lexical, la variété serait plus riche et les redites plus rares. C'est d'ailleurs ce qui rend malaisée l'application au langage des lois classiques de probabilité, le hasard des tirages aléatoires pouvant difficilement servir de modèle à une opération humaine qui manifeste des choix, des refus et des préférences et glisse la liberté dans les mailles de la contrainte. Parmi les facteurs qui structurent le langage (époque, auteur, genre, registre, sujet), il est souvent difficile de démêler les fils, parce qu'ils sont entremêlés et qu'ils influent les uns sur les autres. Le sujet, le registre et le genre ne sont pas indépendants et doivent être compatibles. De même le genre et l'auteur ont partie liée ; si l'écrivain choisit librement un genre, il n'est pas libre d'en détruire les lois. Enfin la dérive du temps s'impose plus ou moins à toutes les autres influences. Ainsi au lieu de s'opposer, les facteurs coulent souvent dans le même sens. Ce qui fait l'intérêt de la présente recherche, c'est qu'on assiste à une lutte frontale entre deux forces, l'une, logique, prégnante et centripète, attachée à la dynamique du discours, l'autre, consciente, réticente et centrifuge, liée à un scrupule esthétique.

2. Extension de l'enquête à l'œuvre de Giraudoux

1 - Cependant notre enquête fondée sur un seul roman d'un seul auteur risque de conduire à des résultats fragiles. Ajoutons quelques romans et tout le théâtre de Giraudoux, sans craindre la longueur des calculs. Il nous faut constituer la table de distribution des segments[9] et relever les effectifs des classes de longueur 2, 3, n mots (tableau 3). Chaque binôme de ce tableau entre alors dans une sommation qui établit le nombre G de combinaisons où un mot sort deux fois d'une urne variable qui contient n mots, n variant de 2 à la longueur maximale. À chaque pas de la progression, le résultat est multiplié par l'effectif de la classe considérée. On aboutit à un total de 8 728 043 combinaisons, offertes à chaque mot selon sa fréquence.

[9] Nous désignons par segment une unité linguistique qui recouvre généralement le paragraphe. Mais les nécessités du traitement informatique et l'obligation de restituer des passages ni trop courts, ni trop longs ont déplacé parfois les lignes de démarcation. Au reste ce formatage plus régulier est un avantage dans la présente recherche : le souci d'éviter les répétitions déborde le cadre d'une simple phrase mais il se dilue et se perd quand le paragraphe s'étend trop largement. La reprise d'un terme n'est sentie comme une répétition que si les deux occurrences sont visibles dans une fenêtre glissante assez étroite, de trois ou quatre lignes au maximum.

Figure 3. Tableau de distribution de la longueur des segments

Longueur	Effectif						
2	6363	34	209	66	26	98	3
3	4308	35	191	67	31	99	10
4	3122	36	225	68	35	100	8
5	3458	37	163	69	28	101	9
6	3011	38	159	70	35	102	8
7	2848	39	146	71	27	103	4
8	2510	40	147	72	27	104	4
9	2292	41	119	73	27	105	4
10	1991	42	114	74	19	106	8
11	1609	43	103	75	22	107	5
12	1493	44	117	76	14	108	2
13	1334	45	101	77	17	109	2
14	1166	46	98	78	10	110	6
15	1033	47	81	79	18	111	5
16	953	48	76	80	26	112	4
17	849	49	83	81	18	113	4
18	785	50	78	82	16	114	3
19	643	51	84	83	13	115	4
20	656	52	73	84	12	116	7
21	627	53	83	85	12	117	9
22	546	54	63	86	11	118	4
23	519	55	65	87	12	119	6
24	427	56	47	88	7	120	7
25	428	57	46	89	5	121	4
26	393	58	57	90	13	122	1
27	376	59	39	91	14	123	2
28	324	60	43	92	9	124	6
29	342	61	38	93	7	125	5
30	285	62	39	94	7	126	6
31	241	63	45	95	11		
32	246	64	37	96	7		
33	257	65	29	97	5		

2 - Reste à proposer une liste limitative de mots, afin de voir s'ils se prêtent ou non à la répétition. Sachant que les basses fréquences – et les hapax par définition – ont peu de chances vu leur rareté de figurer deux fois dans le même passage étroit, on peut les éliminer. À l'opposé les très hautes fréquences, qui concernent les mots grammaticaux, peuvent être provisoirement écartées parce que la répétition des mots-outils est inévitable et n'est pas sentie comme une faute. Restent donc les mots-pleins et parmi eux la catégorie la plus sensible à la répétition : celle des substantifs les plus fréquents. Il y en a 257 dans notre enquête sur Giraudoux. Pour chacun le calcul devient simple : pour être répété chacun doit présenter sa probabilité *p* extraite de sa fréquence *(p = f/N)* et tenter sa chance deux fois (p^2).

$$theo = G\, p^2$$

Ainsi pour le mot *homme* (fréquence 1102) qui vient en tête, on obtient :

$$8734960 * (1102/638452)^2 = 26{,}00$$ répétitions .

À cet effectif théorique, on oppose les répétitions réellement observées : 52, soit le double de ce qui était attendu.

Le tableau 4 montre que l'*homme* n'est pas seul dans ce cas. La *femme* l'accompagne avec un surplus de 13 répétitions, et toute la série suit dans le même sens, le *jour* comme la *nuit*, la *mort* comme la *vie*[10]. Sur la quarantaine de noms représentés dans l'extrait, un seul (*monsieur*) est négatif. La même proportion se maintient dans la suite, où l'effectif théorique devient négligeable et s'abaisse en dessous de 1. Au total les 257 mots de la liste, qui totalisent 43587 occurrences (sur les 638452 du corpus), ne devraient produire que 279 répétitions alors qu'on en relève quatre fois plus, soit 1319.

[10] Précisons qu'il s'agit de lemmes, les variations de nombre ou de genre grammatical étant sans influence sur la perception d'une répétition.

Tableau 4. Répétitions relevées et analysées dans le corpus Giraudoux

Fréq	*réel*	*théo*	*écart*		*fréqu.*	*réel*	*théo*	*écart*	
1102	52	26,00	25,05	homme	320	6	2,19	3,81	soleil
1043	37	23,29	13,71	femme	316	4	2,14	1,86	minute
977	46	20,44	25,56	jour	312	4	2,08	1,92	mari
598	66	7,66	58,34	mot	297	11	1,89	9,11	voix
665	24	9,47	14,53	vie	298	6	1,90	4,10	pas
575	29	7,08	21,92	mort	276	16	1,63	14,37	oiseau
652	41	9,10	31,90	main	288	8	1,78	6,22	temps
540	5	6,24	-1,24	monsieur	279	10	1,67	8,33	visage
576	27	7,10	19,90	heure	273	6	1,60	4,40	président
607	21	7,89	13,11	fois	254	11	1,38	9,62	roi
587	18	7,38	10,62	oeil	264	4	1,49	2,51	matin
542	35	6,29	28,71	nom	245	4	1,29	2,71	pied
510	6	5,57	0,43	monde	242	5	1,25	3,75	terre
493	13	5,20	7,80	nuit	233	9	1,16	7,84	île
488	9	5,10	3,90	fille	234	8	1,17	6,83	mère
400	17	3,43	13,57	guerre	222	14	1,06	12,94	ombre
387	9	3,21	5,79	soir	232	4	1,15	2,85	être
350	16	2,62	13,38	tête	223	8	1,06	6,94	lieu
345	8	2,55	5,45	père	227	3	1,10	1,90	moment
341	5	2,49	2,51	ville	220	4	1,04	2,96	madame

La répétition du même au même est un cas particulier de la **cooccurrence**. La lexicométrie s'intéresse de plus en plus à ce domaine qui met les mots en relation sémantique et, dans chaque phrase d'un corpus donné, relève les rencontres des substantifs les plus fréquents. Or dans le traitement des matrices carrées qui croisent la coprésence de ces substantifs, la diagonale, réservée à la cooccurrence réflexive d'un mot avec lui-même, est généralement évacuée ou arbitrairement et uniformément pourvue d'une valeur neutre, comme la moyenne ou le minimum de la série. Il y a de bonnes raisons pour compléter le tableau et remplir correctement les cases de la diagonale avec les chiffres réels. Bien loin d'y trouver le minimum, on observe généralement à cet endroit la valeur maximale. Pour s'en tenir aux trois substantifs les plus fréquents du corpus Giraudoux, la cooccurrence réflexive *jour-jour* devance toutes celles qui sont relatives au même mot (46 occurr.), et le deuxième rang est observé pour *homme-homme* (52 occurr.) et *femme-femme* (37 occurr.) derrière le couple *homme-femme* (119 occurr.). Cette aimantation qui accouple le même au même est encore plus visible si l'on prend appui non sur les valeurs absolues mais sur les probabilités. Le tableau 5 fondé sur l'écart réduit place beaucoup de répétitions dans le lot de tête et la répétition *mot-mot* y devance même le couple *homme-femme.*

Tableau 5. Mesure statistique de la cooccurrence (écart réduit) chez Giraudoux
(en gras les répétitions, en maigre les simples cooccurrences)

22.49 ciel étoile	**14.38 tendresse tendresse**	12.06 air eau	10.71 animaux arbre
21.91 mot mot	14.29 matin midi	**12.06 amitié amitié**	10.66 amitié amour
20.06 femme homme	14.29 frère soeur	**11.95 nom nom**	10.65 mois semaine
19.51 époux époux	14.24 passé souvenir	11.91 femme mari	**10.59 poète poète**
19.01 grand père	13.99 chemin route	11.89 guerre paix	10.53 fenêtre jardin
18.66 bruit bruit	13.86 lune soleil	11.64 mort vie	10.46 lit mariage
18.33 rêve sommeil	**13.77 eau eau**	**11.64 oiseau oiseau**	10.46 feu lumière
18.05 âme corps	13.71 raison tort	11.57 fils père	10.40 sourire visage
18.00 paix paix	13.70 jour naissance	**11.52 fleur fleur**	10.39 âme forme
17.25 île oiseau	13.64 aventure goût	**11.37 peau peau**	10.38 fils mère
16.95 chien mer	**13.58 tour tour**	**11.36 arbre arbre**	**10.36 pensée pensée**
16.29 mensonge vérité	13.54 bras jambe	**11.30 face face**	10.30 bouche oeil
16.20 conseil salle	13.53 dent or	11.25 année mois	10.30 doigt main
16.18 arbre oiseau	**13.33 silence silence**	11.23 mère père	10.29 mer vent
16.08 reine roi	**13.28 sommeil sommeil**	11.12 animaux chien	10.27 année semaine
16.06 travail travail	**13.00 ombre ombre**	**11.10 main main**	10.23 heure matin
15.34 cri cri	12.96 coup feu	11.08 étoile lune	**10.18 nez nez**
15.26 oeil regard	12.79 lune nuit	**11.02 vertu vertu**	10.13 couleur vêtement
15.25 fin monde	**12.66 travers travers**	**10.94 corps corps**	**10.08 cause cause**
15.01 oncle père	12.48 oeil visage	**10.81 cheval cheval**	10.06 pied tête
14.66 conseil président	**12.33 gens gens**	10.78 fleur parfum	10.04 mariage souvenir
14.50 monsieur président	**12.17 sourire sourire**	10.77 cheveu lèvre	10.00 nature poète

Avant de conclure à quelque négligence surprenante d'un écrivain qui passe pour un modèle d'élégance, il convient de vérifier si d'autres plumes ont la même faiblesse.

3 – Extension à d'autres écrivains

Mis à part Pascal, les écrivains sont peu explicites sur le chapitre de la répétition. Ils déclarent volontiers la guerre à l'adjectif ou à l'abstraction et s'en font gloire (c'est le cas de Giraudoux). Mais ils se taisent sur la répétition, estimant sans doute qu'il s'agit là d'une tâche ancillaire de nettoyage. En l'absence de déclaration explicite, on part de l'hypothèse que tous les écrivains n'ont pas une sensibilité égale à l'endroit de la répétition. Certains peuvent être indifférents à une contrainte scolaire dépassée, quand d'autres s'appliquent au respect d'un principe tenu pour la marque du style. Partant de Giraudoux et du résultat global observé dans son oeuvre, on se propose de mettre en œuvre les mêmes programmes pour l'ensemble des écrivains français.

Tableau 6. Les répétitions chez Gracq (œuvre complète)

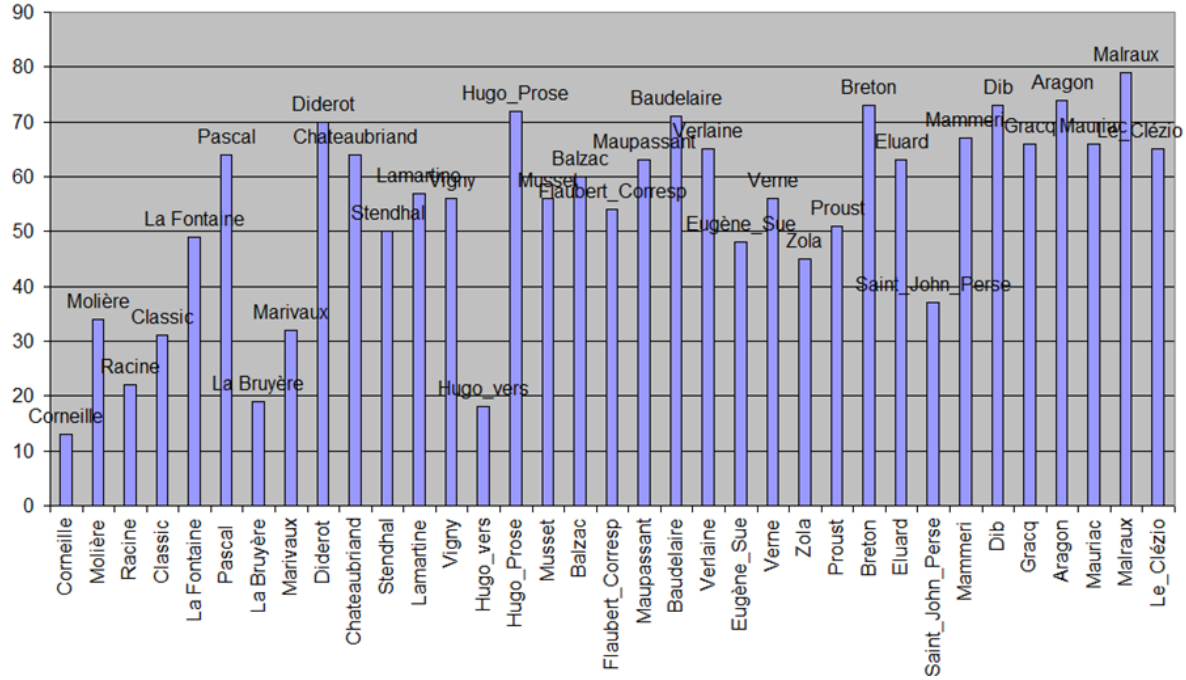
Fréqu.	Répét.	Théoriq.	Ecart	Réduit	LEMME
Total	2925	978.32	1946.68	excédents=	66.55%
1199	41	45.54	-4.54	-0.67	oeil
1105	123	38.67	84.33	13.56	temps
973	62	29.98	32.02	5.85	jour
969	30	29.74	0.26	0.05	fois
946	57	28.34	28.66	5.38	vie
914	54	26.45	27.55	5.36	nuit
856	31	23.20	7.80	1.62	coup
803	29	20.42	8.58	1.90	chose
786	35	19.56	15.44	3.49	ville
749	48	17.76	30.24	7.18	monde
742	27	17.43	9.57	2.29	main
726	45	16.69	28.31	6.93	mer
714	33	16.14	16.86	4.20	homme
702	30	15.60	14.40	3.65	route
682	37	14.72	22.28	5.81	heure
654	43	13.54	29.46	8.01	eau
642	16	13.05	2.95	0.82	moment
608	56	11.70	44.30	12.95	rue
599	20	11.36	8.64	2.57	lumière
594	22	11.17	10.83	3.24	terre
584	27	10.79	16.21	4.93	voix
565	20	10.10	9.90	3.11	air
554	16	9.71	6.29	2.02	soleil
553	25	9.68	15.32	4.93	maison
540	16	9.23	6.77	2.23	travers
535	16	9.06	6.94	2.31	instant
527	12	8.79	3.21	1.08	coeur
507	27	8.13	18.87	6.62	tête
499	20	7.88	12.12	4.32	silence
483	18	7.38	10.62	3.91	fond
479	10	7.26	2.74	1.02	long
474	31	7.11	23.89	8.96	mot
471	27	7.02	19.98	7.54	image
468	23	6.93	16.07	6.11	esprit
461	18	6.72	11.28	4.35	livre
452	16	6.46	9.54	3.75	forêt
449	28	6.38	21.62	8.56	siècle
449	15	6.38	8.62	3.41	côté

Le tableau 6, consacré aux substantifs les plus fréquents dans l'œuvre de Gracq, montre la même propension aux répétitions. Mis à part le premier (*œil*), tous les éléments sont en excédent, avec un écart le plus souvent significatif. Au total on rencontre environ 3000 répétitions quand le hasard n'en produirait que 1000, soit un excédent de 2000 (ou 66%). Ce taux de répétition n'est pas très différent de celui de Giraudoux (77%). Appliqué à d'autres écrivains le même calcul produit les mêmes observations qui confirment une loi générale. Reste à saisir et à expliquer les variations qu'on observe d'un auteur à l'autre. La figure 7 passe en revue 35 représentants majeurs de la littérature française, en projetant sur un histogramme le taux de répétition mesuré dans leur œuvre.

Il serait étonnant que la chronologie soit absente du débat. Car le refus de la répétition n'a rien d'une règle absolue. C'est un simple conseil, facile à contester ou à ignorer et apte à évoluer. Déjà Pascal limite la portée de cette prescription et la soumet à la nécessité d'être clair. Les auteurs du siècle classique respectent plus volontiers l'abstinence et Pascal est le seul à dépasser 50%. À partir de Diderot le principe s'assouplit et au XXe siècle le taux tourne autour de 70%.

On soupçonne aussi l'influence du **genre**. Car le taux de répétition pour des poètes (Corneille, Racine, Saint-John Perse) est inférieur à celui de prosateurs. Malheureusement le genre est souvent masqué par le mélange de la prose et des vers dans beaucoup de monographies. La distinction est faite dans le cas de Hugo et le corpus en vers s'y différencie très nettement du corpus en prose.

Figure 7. Le taux de répétition chez les écrivains



4 – Réflexions sur la typologie des répétitions et leur traitement

1 – Les matériaux accumulés montrent une différence entre les mots pleins et les **mots grammaticaux**. Ces derniers ont fait l'objet d'une exploration analogue, en ajustant le paramètre de la longueur. Car ces mots-outils ont une telle fréquence que leur répétition est inévitable. On a donc rétréci le champ à la limite du syntagme, ou plus précisément de l'espace compris entre deux ponctuations. Dans cette fenêtre étroite, il reste cependant des restes de scrupule, des réticences à employer le même mot deux fois de suite *(la France **de De** Gaulle*), même si trois constructions syntaxiques sont parfaitement acceptées: ***vous vous***, ***nous nous*** et ***en en*** (« *approcher un problème **en en** faisant le tour* »). Avec un ou deux mots intercalés, la répétition des mots-outils passe inaperçue s'il s'agit de coordination ou de juxtaposition (*les droits et les devoirs*, à côté de *les droits et devoirs)*. Mais on évite la répétition hiérarchique, la construction en cascade qui distribue à plusieurs niveaux la même préposition *à, de, en, par, sur, pour* (exemple *les maisons en bois en progrès en France)*[11]. Malheureusement ces chevauchements étagés sont difficiles à distinguer des constructions plates, juxtaposées ou coordonnées, du type *en France, en Belgique ou en Hollande*, qu'aucun interdit ne frappe.

On a tenté cependant les calculs, sans obtenir des résultats clairs[12] : l'observation s'écarte peu du modèle. À peine observe-t-on certains évitements logiques (les subordonnants, les relatifs et même les coordinations) ou certains excédents assez explicables, en particulier ceux des pronoms personnels et spécialement des **pronoms de rappel**. La statistique montre que leur répétition est multipliée, ce qui se comprend quand on songe au statut de ces mots, voués au remplacement et donc indirectement à la répétition.

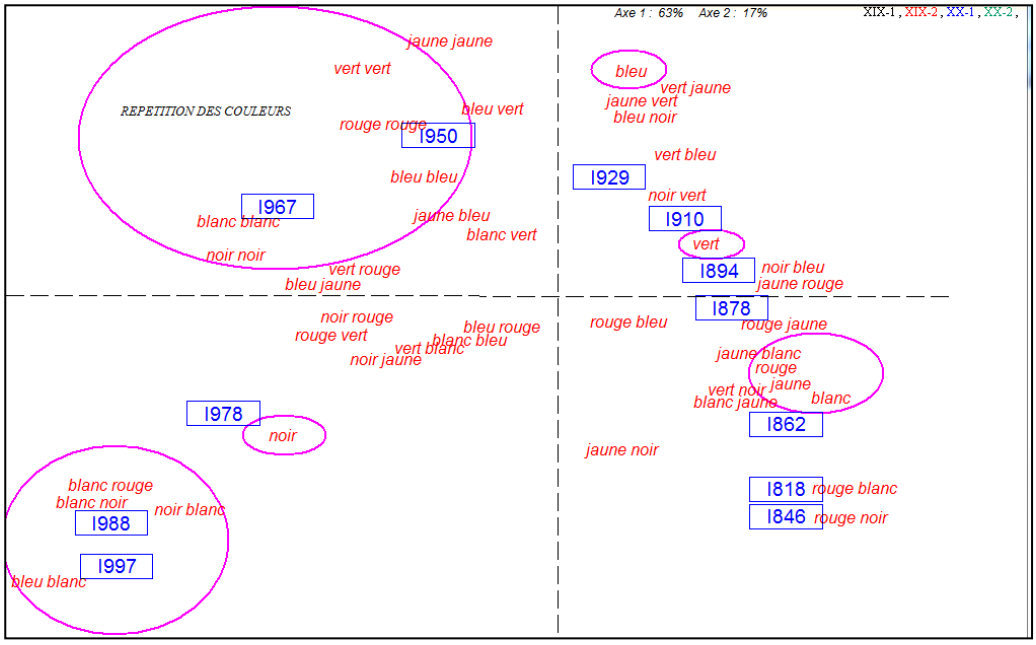
2 – Les **répétitions contiguës** où le même mot est repris immédiatement ne sont guère admises par la syntaxe (en dehors des constructions réflexives *nous nous* et *vous vous*).

[11] D'où des effets comiques obtenus par accumulation ou coordination de niveaux différents : *De Lattre de Tassigny de corvée de chiottes demain. Je vais à Paris et à vélo. Mon vélo est en panne et en aluminium*.

[12] La définition de la répétition est ambiguë lorsque le mot est réitéré plusieurs fois, par exemple 4 fois: ou bien on compte 1 (présence/absence) ; ou bien on détaille les « arrangements » (1 et 2, 1 et 3, 1 et 4 , 2 et 3, 2 et 4, 3 et 4) , soit 6 au total ; ou bien on compte les apparitions d'un mot reprenant la première occurrence, soit 3.

Quand on les rencontre dans un texte, il y a nécessairement un effet de sens voulu comme tel. L'effet le plus commun est celui de l'insistance (c'est *très très* beau), qui peut aller jusqu'à la surenchère, le second terme étant comme un exposant du premier : *Tout ce qui doit être blanc est blanc pur, ce qui doit être rose est rose rose*[13], ou jusqu'à la suspension hébétée ou contemplative, analogue à l'arrêt sur image au cinéma. On trouve parfois dans le nouveau roman cet arrêt sur le mot, qui est aussi un arrêt sur la chose, comme en témoigne ce relevé exhaustif perpétré dans le *Déluge* de Le Clézio : *gravier gravier, eau eau, Simon Simon Simon, chanter chanter, faces faces, très très, blanc blanc, orange orange, vert vert vert vert vert, émeraude émeraude, noir noir noir noir noir noir noir noir*[14]. Intrigué par cette propension à épaissir la couleur en apposant couche sur couche, on a cherché dans les milliards de mots engrangés par *Google Books* d'autres exemples de peinture au couteau. Il s'en est trouvé environ un millier pour chacune des couleurs : *blanc blanc* 1362, *bleu bleu* 974, *rouge rouge* 1307, *noir noir* 909, *jaune jaune* 1440, *vert vert* 882. L'équilibre des couleurs est respecté, mais la répartition chronologique de tels redoublements ne l'est pas, étant concentrée dans la seconde moitié du XXe siècle. L'analyse factorielle de la figure 8 rend compte des couleurs pures, mélangées ou redoublées qui ont la faveur de 1800 à 2000 et qui rassemblent 50 millions d'occurrences. Sans surprise on constate que le 19e siècle est plus coloré que le 20e, à l'exception des reprises ton sur ton qui redoublent dans le quadrant supérieur gauche, à proximité des tranches 1950 et 1967.

Figure 8. Analyse factorielle des couleurs dans Google Books (domaine français)



3 - Proche des répétitions immédiates est la construction où le même nom encadre une préposition sur le **modèle *corps à corps***. Ce sont généralement des expressions à demi figées où les composants disposent d'une semi-liberté. La préposition centrale est le plus souvent *à* (*bouche à bouche*), ou *en* (*de temps en temps*). Mais on peut trouver *pour* (*œil pour œil*), *sur*

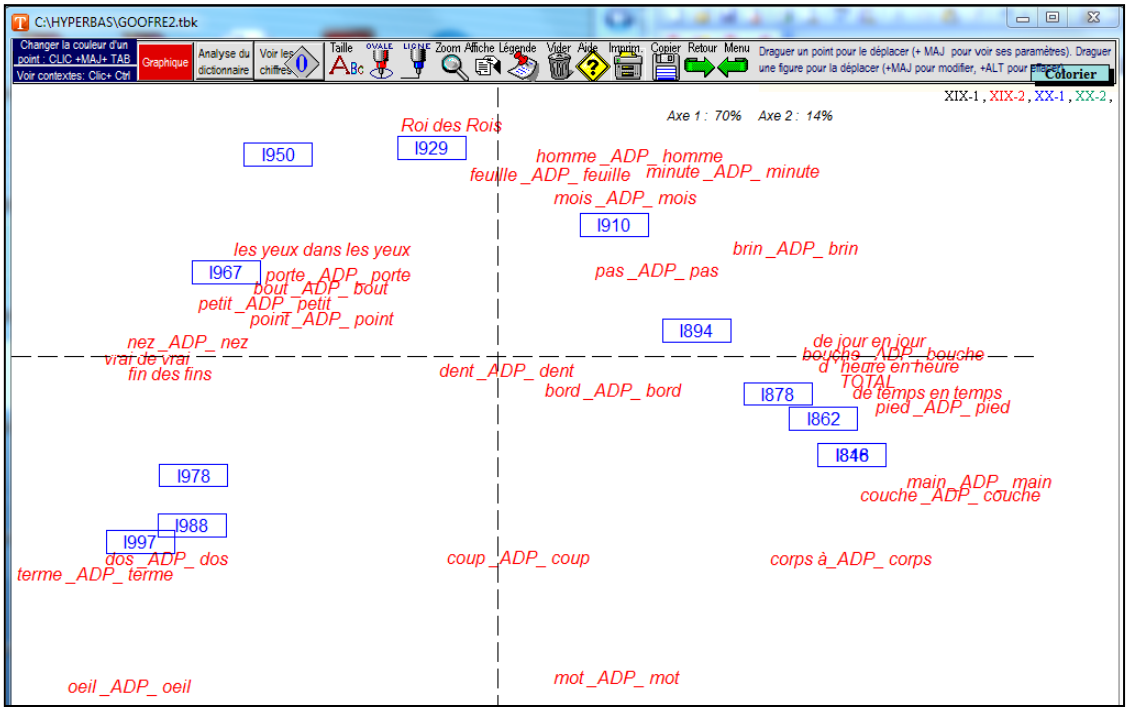
[13] Giraudoux, *Portugal*, p.21.

[14] Le Clézio abandonne assez vite ce procédé, les reprises de terme devenant de plus en plus rares dans la suite de ses oeuvres: *9 Procès, 30 Fièvre, 37 Déluge, 22 Extase, 30 Fuite, 26 Guerre, 16 Mydriase, 10 Voyage, 1 Prophéties, 7 Mondo, 4 Inconnu, 1 Icebergs, 0 Arbres, 3 Désert, 0 Villes, 0 Ronde, 6 Chercheur, 0 Angoli, 15 Rodrigues, 1 Rêve, 6 Printemps, 0 Sirandanes, 2 Onitsha, 6 Étoile, 1 Pawana, 7 Diego, 2 Quarantain, 5 Poisson, 2 Fête, 0 Nuages, 0 Hasard.*

(*coup sur coup*), *par (deux par deux), après* (*minute après minute*), *dans (main dans la main),* ou *de* (*vrai de vrai*). Parfois la locution fait intervenir une préposition supplémentaire (*de...en, de…à*). La valeur est souvent adverbiale (*pas à pas*), parfois nominale (faire du *porte à porte*). Quant au mot répété, ce peut être un adverbe (*peu à peu*), un adjectif (*petit à petit,* mais généralement c'est un substantif concret emprunté à l'environnement spatial ou temporel, et tout particulièrement au corps humain, lequel fournit une série très riche où s'exprime la confrontation ou le contact (*tête à tête, œil pour œil, dent pour dent, nez à nez, main dans la main, pied à pied, dos à dos, bouche à bouche, corps à corps*). La création de telles locutions repose sur un effet d'inattendu, volontairement recherché. La répétition d'un mot à si courte distance étant statistiquement improbable, on spécule sur la surprise et l'expressivité. Puis, au fil du temps, l'expression tend à se figer et, perdant ses déterminants, sa forme se fait lapidaire, comme il arrive aux dictons ou maximes populaires.

Là aussi le corpus de *Google Books* peut aider à mesurer dans le temps la fortune de cette forme de répétition. On aurait souhaité proposer à l'automate une structure à trois variables, de la forme *substantif + préposition + substantif* en précisant seulement que les deux substantifs devaient être identiques. Malheureusement le site *Culturomics*[15] qui distribue les données statistiques de *Google Books* accepte les variables aussi bien que les constantes, mais ne permet pas d'imposer le filtre qui respecterait la contrainte requise. Seule la préposition au centre du triplet a été exprimée sous la forme générale et codée *_ADP_*. Mais il a fallu désigner nommément les mots qui l'encadrent. C'est donc un échantillon d'une trentaine de locutions qu'on trouvera dans l'analyse factorielle de la figure 9.

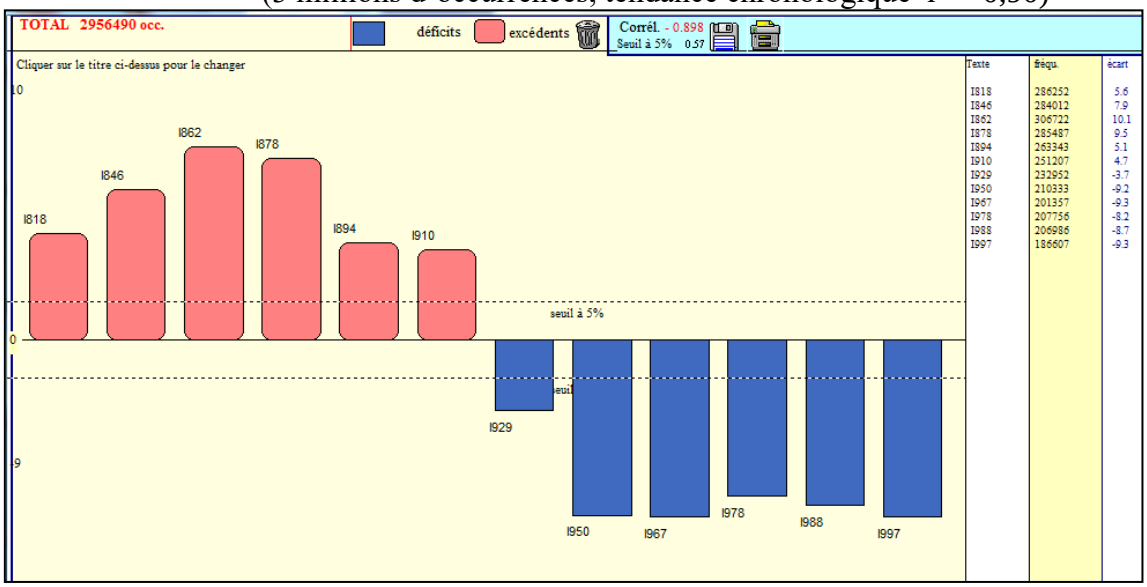
Figure 9. Analyse factorielle des locutions à redoublement du type *corps à corps*



[15] Le site Cuturomics (https://books.google.com/ngrams) dessine la courbe d'emploi de tout mot ou locution, pour un corpus donné et entre deux dates choisies (par défaut de 1800 à 2000). Le corpus français (soit 89 milliards de mots) correspond aux données de *Google Books*, dans l'état où elles se trouvaient en 2012. Pour un usage plus élaboré de telles données, on les a transférées dans une base autonome, GOOFRE.tbk, téléchargeable à l'adresse http://logometrie.unice.fr .

Quoique quasiment toutes les locutions soient présentes dès 1800, le moule est encore disponible pour de nouvelles créations (*vrai de vrai* ne se rencontre guère avant 1850) et les structures *de A à A* ou *de B en B* peuvent s'appliquer à n'importe quel élément spatial pour exprimer un processus progressif ou distributif (*de bloc à bloc, de proche en proche*). Le déclin est pourtant visible : non seulement les créations nouvelles sont rares, mais les locutions les plus fermement établies sont en perte de vitesse. Le quadrant inférieur gauche réservé aux tranches les plus récentes est presque vide et le centre de gravité de l'ensemble (correspondant à la position du point TOTAL) est déporté nettement à droite, là où campe le XIXe siècle. On peut d'ailleurs pour plus de clarté regrouper en un seul histogramme toutes les variétés rencontrées : la courbe de la figure 10 est régulièrement descendante et ne permet pas le doute.

Figure 10. Déclin des expressions à redoublement de 1800 à 2000
(3 millions d'occurrences, tendance chronologique r = -0,90)



À quoi peut-on rattacher cette sorte de glaciation qui se contente d'une expression plus neutre, plus froide, plus tournée vers le constat que vers le style ? Sans doute moins à l'évolution de la langue qu'à un changement dans la composition du corpus. Les publications les plus récentes, qui sont aussi les plus nombreuses, n'ont pas été soumises au tri de l'histoire : c'est le tout-venant de l'édition, où pullulent les ouvrages d'information, les traités techniques et les sujets les plus divers. Les livres plus anciens ont survécu à l'oubli et à la perdition parce que, leur intérêt se maintenant, des rééditions ont eu lieu qui ont augmenté leur chance de survie. C'est là le privilège des œuvres littéraires, rarement le cas des publications techniques, que le progrès condamne très vite. Or un effet de sens, un trait de style, étant attaché à cette sorte de répétition, au moins lors de sa création, sa raréfaction ne ferait que suivre la sous-représentation de la littérature dans les publications actuelles. Il est probable que les mêmes raisons expliqueraient le déclin des dictons et des clichés.

Ces cas extrêmes où la coocurrence du même au même est observée à très courte distance (dans la contiguïté ou avec un seul mot intercalaire), le hasard ne peut être invoqué et la coïncidence ne peut être qu'intentionnelle et résulter d'un effet de sens, même s'il faut remonter à l'inventeur d'une expression figée. Dans le schéma d'urne, l'obtention d'un même mot deux fois de suite (ou deux fois sur trois si l'on accepte un mot intercalaire) est beaucoup plus improbable qu'une répétition observée dans un cadre plus large où l'on dispose d'un certain nombre de tirages, qui multiplient les chances (autant de coups qu'il y a de mots dans la phrase). Les autres cas de répétition n'offrent guère de prise au traitement automatique et ne permettent guère de séparer la répétition volontaire de celle qui ne l'est pas. On bute

nécessairement sur les multiples variétés et valeurs des répétitions. Ainsi les répétitions évitées ou biffées échappent à l'emprise de l'automate, n'étant visibles que dans le manuscrit., (à moins de les saisir une à une les variantes et d'en faire la synthèse). Aucun moyen direct d'épingler les répétitions disgracieuses qui échapperont toujours à son jugement, comme celles qui sont voulues et portent un effet de sens. Difficile aussi d'ouvrir l'empan et d'élargir le cadre d'étude au delà de la phrase. Les retours ou résurgences de thèmes ou de termes ont souvent besoin d'un espace plus large et plus flou pour être repérés. Force est de se concentrer sur une définition étroite de l'objet d'étude (le lemme répété) dans un cadre délimité (la phrase). Le rôle de l'automate se borne alors à renifler le texte et à signaler les écarts, les anomalies, les concentrations ou les manques, comme ferait un chien d'arrêt ou un détecteur de métaux.

La répétition d'un mot est un cas particulier d'une itération qui peut s'appliquer à bien d'autres objets du langage : une sonorité, une rime, une construction syntaxique, un fait de rythme, une figure stylistique, un tic d'écriture. Tout l'univers lexicométrique est fondé sur la répétition. La fréquence d'un mot dans un texte, c'est le nombre qui mesure sa répétition. Or la fréquence des mots et des segments a été considérée généralement à l'échelle large du texte et du corpus, à un niveau où la répétition n'est sensible ni à l'oreille ni à la mémoire. Il convient de se situer aussi à l'échelle du micro texte et de la séquence courte qui s'arrête à la fin de la phrase ou du paragraphe et où se limite la portée de la rémanence mémorielle. Dans cette enceinte étroite et fermée les mots se voient, s'écoutent, se parlent. C'est là, dans ces rencontres furtives ou insistantes, que la lexicométrie a tendance à s'engager présentement. Mais dans ce cadre la répétition n'est qu' un cas particulier et secondaire d'un observable beaucoup plus riche : la cooccurrence[16]. Le sens et le style ne dépendent guère de la relation trouble, et jugée parfois haïssable, qu'un mot peut avoir avec lui-même mais des rapports et connivences multiples qui lient les mots entre eux et en font une chaîne.

Summary:
The entire lexicometric universe is based on repetition. Indeed, the frequency of a word in a text is the measure of its repetition. However, the frequency of words and segments has been generally considered at the broad scale of the text and the corpus, at a level where repetition is neither detectable to the ear nor to memory. It is therefore useful to also examine frequency at the level of the microtext, of the short sequences that terminate at the end of sentences and paragraphs and that fall within the range of the perdurance of memory. We propose performing a statistical investigation at the microtext level, first of a novel by Giraudoux and then of his entire body of work, before undertaking an analysis of all French writers and ultimately the entire French corpus within Google Books.

[16] Sur cette question de la cooccurrence, voir le numéro 11 de la revue *Corpus, La cooccurrence. Du fait statistique au fait textuel*, Université de Nice, 2012, 252p.