MATHÉMATIQUES ET SCIENCES HUMAINES

M. PETRUSZEWYCZ

Quelques exercices d'utilisation des critères de A. A. Markov

Mathématiques et sciences humaines, tome 66 (1979), p. 51-90

<http://www.numdam.org/item?id=MSH_1979__66__51_0>

Math. Sci. hum. (17^{e} année, n°66, *1979*, p.51-90)

QUELQUES EXERCICES D'UTILISATION DES CRITERES DE A.A. MARKOV

M. PETRUSZEWYCZ

La lecture de l'article [1][(1)] de A.A. Markov et la reconstitution pas à pas de sa démarche, nous ont suggéré d'utiliser les paramètres du modèle markovien comme instruments de recherche linguistique. Dans un premier temps au moins, nous avons décidé de poursuivre nos recherches sur les textes mêmes en langue russe qu'avait utilisé Markov. Ce travail eût probablement été différent, fait par quelqu'un pour qui la langue russe aurait été à la fois langue maternelle et pratique quotidienne et en même temps le sujet d'étude, ce n'est pas notre cas.

Mais avant d'effectuer ces exercices, il convient de dire quelques mots de la distinction graphèmes/phonèmes. En effet, Chomsky et Miller ont écrit [27] qu'il était "peu vraisemblable que le résultat [l'interprétation de la différence entre p_1, probabilité d'avoir une voyelle après une voyelle, et p_0, probabilité d'avoir une voyelle après une consonne] eût été sérieusement modifié si l'analyse avait porté sur les phonèmes plutôt que sur les graphèmes". Il semble que cette opinion ne soit pas l'objet d'un consensus et on a donc envie de voir ce qui en est réellement, au moins pour des textes écrits en russe. Il est certain que le statisticien qui va, comme Markov, constituer les carrés élémentaires, part toujours d'un texte matérialisé dans son écriture, que celle-ci utilise l'alphabet de la langue ou un alphabet phonologique : c'est le même texte. Il ne s'agit absolument pas d'aborder l'étude de la langue parlée.

---

(1) Les références bibliographiques se trouvent sous le titre BIBLIOGRAPHIE p.91-97 dans ce numéro.

Il se pose par ailleurs un autre problème : il n'y a pas de transcription phonologique universelle. Nous avons utilisé pour le russe deux systèmes de transcription[1]. La seule difficulté, de notre point de vue, était l'affectation du yod : en transcription phonologique nous l'avons dans les deux cas compté comme consonne. Voici les résultats mais auparavant, rappelons le procédé de dénombrement et de calcul.

T étant la longueur du texte en graphèmes, il y a (T-1) couples, (T-2) triplets, etc. V étant le nombre de graphèmes vocaliques dénombrés, $v*$ désigne les doublets commençant par une voyelle et inversement $*v$ désigne les doublets finissant par un graphème vocalique.

$p_1 = \frac{vv}{v*}$ ou $\frac{vv}{(v*)-1}$ si le texte finit par un graphème vocalique.

$p_o = \frac{cv}{c*}$ ou $\frac{cv}{(c*)-1}$ si le texte finit par un graphème consonantique.

Il est facile de généraliser aux n-uplets, n étant la longueur de la séquence considérée.

Discours poétique.

a) graphèmes orthographiques :

T = 1715 signes ; nombre de graphèmes vocaliques : 740
nombre de successions de graphèmes vocaliques : 71

$$\delta = \frac{\text{nombre de VV}}{\text{nombre de V*}} - \frac{\text{nombre de CV}}{\text{nombre de C*}} = \frac{71}{740} - \frac{669}{975-1} \simeq -0{,}591$$

$$M = \frac{1+\delta}{1-\delta} \simeq 0{,}257$$

b) transcription phonologique comptant le yod comme consonne :

T = 1707 signes ; nombre de graphèmes vocaliques : 698
nombre de successions de graphèmes vocaliques : 19

$$\delta = \frac{19}{698} - \frac{678}{1009-1} = -0{,}646$$

$$M' = \frac{1+\delta}{1-\delta} = 0{,}215$$

---

(1) Nous remercions vivement M. J.P. Sémon qui a obligeamment transcrit selon son système [34] les six premières strophes d'*Evgenii Onegin* ; ainsi que Melle C.Prokhoroff qui a transcrit, selon le système enseigné par Mme Fougeron dans son cours de Phonétique russe pour la licence à Paris IV- Sorbonne, des fragments de *Kapitanskaja Dočka*.

Discours en prose. Trois fragments de 600 signes regroupés pour constituer un fragment de 1800 signes.

a) graphèmes orthographiques :

T = 1800 signes ; nombre de graphèmes vocaliques : 791

nombre de successions de graphèmes vocaliques : 103

$$\delta = \frac{103}{791} - \frac{687}{1009-1} \simeq -0,551$$

$$M = \frac{1+\delta}{1-\delta} \simeq 0,289$$

b) transcription phonologique comptant le yod comme consonne :

T = 1800 signes ; nombre de graphèmes vocaliques : 705

nombre de successions de graphèmes vocaliques : 36

$$\delta = \frac{36}{705-1} - \frac{668}{1094} = -0,559$$

$$M' = \frac{1+\delta}{1-\delta} = 0,282$$

Il nous paraît que ces résultats nous autorisent à travailler, ainsi que le fit Markov, sur les graphèmes. Bien entendu cette étude n'est valable réellement que pour le russe et par ailleurs ce n'est qu'un coup de sonde et non une étude statistique ; d'autres échantillons pourraient donner des valeurs plus différentes mais nous pensons que dans le cadre d'une recherche de type markovien on peut, sans trop de risque, faire l'économie de l'étape de la transcription phonologique.

Nous pouvons maintenant formuler nos hypothèses et présenter au lecteur quelques exercices d'utilisation des critères de Markov : soit $\frac{1+\delta}{1-\delta}$ que nous appelons M en son honneur précisément, soit C_m le critère de la chaîne d'ordre 2.

I - DEUX EXEMPLES D'UTILISATION DES PARAMETRES MARKOVIENS.

Nous avons fait l'hypothèse que les fréquences des doublets ou triplets de voyelles pouvaient caractériser un texte. Des considérations de deux ordres nous y ont poussé. Les premières sont d'ordre historique. Préparant la 4ème édition de son Cours de Calcul des Probabilités, Markov reproduit intégralement [1]. Il y ajoute un deuxième exemple d'application de sa théorie des chaînes au même domaine d'application : la succession des "voyelles" et des "consonnes" dans un texte littéraire russe. L'ampleur de cette deuxième application, 100000 lettres, a sans doute son origine dans les critiques qu'il a faites à Morozov et sous le coup desquelles il ne veut pas tomber lui-même. Mais nous renvoyons l'examen de ce problème : longueur du corpus - validité des résultats à plus tard. Validité peut s'entendre de deux façons : soit il s'agit de la "robustesse" du modèle de chaîne. Soit il s'agit de la stabilité des paramètres. Si Markov était très satisfait du résultat obtenu dans l'étude des deux corpus Aksakov, nous le sommes beaucoup moins de ceux que nous présentons plus loin. Ces précisions données, les considérations ci-dessus nous induisent à penser que Markov ne tenait pas pour seulement fortuit ou simplement commode son domaine d'application et nous allons sur deux exemples essayer de montrer que la chaine markovienne peut être un instrument de recherche.

I.1. Opposition poésie/prose chez Puškin.

Notre argumentation est d'ordre prosodique au sens traditionnel du terme. Nous n'essaierons pas de donner au lecteur une idée des rapports complexes qu'entretiennent la sémantique, le rythme, la scansion et la rime dans la poésie russe : le lecteur intéressé pourra se reporter aux ouvrages cités en bibliographie [34] et [38], seul ouvrage didactique disponible en français actuellement. Nous simplifions à l'extrême en disant que c'est une versification syllabo-tonique le plus souvent rimée, "fondée sur l'alternance régulière des syllabes accentuées (fortes- et des syllabes atones (faibles)... Dans un vers syllabo-tonique, le nombre des voyelles est fixe par suite du nombre fixe des syllabes - le nombre des consonnes, par contre, n'est limité par aucun dessin métrique", ni par d'autres considérations, au moins à l'époque de Puškin.

Nous avons en conséquence fait l'hypothèse que nous trouverions en moyenne, plus de voyelles dans un texte en prose que dans un texte en vers[1] puisque le prosateur ne compte pas les syllabes et donc aussi plus de doublets de voyelles. Nous avons aussi fait l'hypothèse que, l'assourdissement par les consonnes étant un trait prosodique russe, nous trouverions plus de triplets de consonnes en poésie qu'en prose.

Nous avons choisi de comparer *Evgenii Onegin* [32] au roman en prose du même auteur : *Kapitanskaja dočka* (La Fille du Capitaine) [33]. Nous avons procédé, ainsi que l'avait fait Markov, avec les mêmes conventions par relevés de dix lignes sur dix colonnes formant les tableaux élémentaires de cent lettres. Forts du résultat obtenu par Markov sur les textes d'Aksakov prouvant la stabilité de ses caractéristiques et d'autre part ayant des raisons de mettre en doute l'opinion trop répandue, à savoir que la statistique n'aboutit à des résultats valides qu'en s'appuyant sur un grand nombre de données, nous avons, dans un premier temps fait quelques essais sur des corpus courts. Nous avons choisi deux fragments d'*Evgenii Onegin*, tous deux d'une longueur de 5000 lettres, plus précisément nous fournissant 50 tableaux élémentaires. Le premier est le chapitre I du poème, de la strophe I à la strophe XX 5e vers ; le deuxième, choisi au hasard, comprend les strophes XX à XXXVIII du chapitre VI. Pour la prose nous avons choisi au hasard le chapitre VII (jusqu'à la page 190 de l'édition de référence). Nous avons ainsi pour la prose dépouillé 20000 lettres ou précisément constitué 200 tableaux élémentaires que nous avons divisé en quatre tranches égales de 50 tableaux. Puis nous avons effectué pour toutes les paires de comparaisons possibles : 50 tableaux de prose / 50 tableaux de

---

[1] M. Fougeron nous ayant entendu émettre cette hypothèse nous a fait remarquer que, si elle était vraie, alors on devrait trouver des termes synonymes comme град/город (grad/gorod) = ville ou младой/молодой (mladoj/molodoj) = jeune, plus fréquents les premiers en poésie qu'en prose. Mais ainsi exprimée cette hypothèse est à la fois plus forte et infiniment plus subtile que celle que nous avons faite. Nous avons cependant effectué quelques décomptes - à partir de [9]. Saluons au passage l'extraordinaire instrument de travail qu'est une concordance. Celle-ci est complète aux deux sens du terme : elle analyse l'intégralité de l'oeuvre de Puškin ; les entrées sont classiques mais elle cite, pour chaque entrée, chaque forme fléchie au moins une fois en contexte, les autres occurrences étant données sous forme d'index. Si град et младой apparaissent effectivement rarement en prose (5 sur 37 - toutes formes fléchies confondues et 3 sur 179) ce qui va dans notre sens il n'en est pas de même pour les formes longues : pour молодой en particulier on a un quasi équilibre 311/347. Une étude systématique pourrait être intéressante mais ne nous appartient pas.

poésie, les calculs markoviens. Nous avons alors constaté que si les chiffres absolus confirmaient notre hypothèse par contre certaines caractéristiques markoviennes ne présentaient pas la stabilité nécessaire. Nous avons attribué cette instabilité aux fluctuations d'échantillonnage et nous avons décidé de travailler sur des corpus plus longs c'est-à-dire sur 100 tableaux élémentaires de poésie et 100 tableaux de prose. Pour cela nous avons abandonné le deuxième fragment d'*Evgenii Onegin* et poursuivi la mise en tableaux du premier fragment jusqu'à avoir 100 tableaux (cela correspond au 9ème vers de la strophe XXXVII du chapitre I). Et nous avons comparé ce texte que nous désignerons par E a+b aux trois textes que nous pouvions constituer pour la prose en désignant par K 1 + 2 les 100 premiers tableaux, K 2 + 3 les tableaux 51 à 150, K 3 + 4 les 100 derniers tableaux. On nous objectera peut-être que nous appelons texte au sens ordinaire du terme quelque chose qui commence en plein milieu d'un mot mais comme il s'agit de décomptes de lettres, et non de mots, nous ne pensons pas que ce "flottement" initial, et final d'ailleurs, qui n'est jamais que de quelques unités ait beaucoup d'importance.

Nous présenterons d'abord les tableaux permettant de comparer les distributions des voyelles, des doublets de voyelles et des triplets de consonnes.

TABLEAU I. Distributions comparées des centaines de lettres selon leur nombre de voyelles dans les quatre textes de Puškin.

	Nombre de voyelles	37	38	39	40	41	42	43	44	45	46	47	48	49	50	$\bar{x}$	σ^2
Effectifs des centaines	E a + b	1	3	6	8	7	17	14	19	9	10	4	2			42,93	5,825
	K 1 + 2			1	3	3	11	13	16	23	14	5	8	2	1	44,53	4,669
	K 2 + 3			1	2	2	10	16	22	22	15	5	3	2		44,34	3,444
	K 3 + 4		1	0	5	1	9	14	24	18	14	8	4	2		44,35	4,227

On voit très nettement que notre hypothèse est vérifiée, les valeurs moyennes du nombre de voyelles par centaines de lettres étant pour les trois textes en prose supérieures à la valeur moyenne (représentée classiquement par la notation $\bar{x}$) correspondant à la poésie. Les trois distributions relatives à la prose se décalent vers les grandes valeurs de la variable par rapport à la poésie.Les coefficients d'asymétrie sont dans l'ordre de haut en bas : - 0,192 ; - 0,008 ; - 0,061 et - 0,282. Les coefficients d' aplatissement sont de même : 2,545 ; 2,886 ; 3,349 et 3,363. Les lecteurs préférant les représentations graphiques se rapporteront aux Fig. 1,2 et 3.

TABLEAU II. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles dans les quatre textes de Puškin.

	Nombre de doublets	0	1	2	3	4	5	6	7	8	9	10	11	12	$\bar{x}$	σ^2
Effectifs des centaines	E a + b	2	7	10	9	18	11	15	9	7	7	3	1	1	5,06	6,916
	K 1 + 2		4	4	9	8	18	24	9	13	4	4	2	1	5,820	5,387
	K 2 + 3		3	4	8	10	20	26	7	13	5	3	1		5,70	4,390
	K 3 + 4		4	4	12	11	18	19	10	11	6	3	2		5,60	5,240

Les coefficients d'asymétrie sont dans l'ordre de haut en bas : 0,261 ; 0,123 ; 0,015 et 0,126. Les coefficients d'aplatissements sont de même : 2,503 ; 2,934 ; 2,863 et 2,592.

Ici aussi les nombres parlent directement, de même que les graphiques correspondants : Fig. 1,2 et 3.

Par contre l'examen des distributions comparées des centaines de lettres selon leur nombre de triplets de consonnes ne nous donne pas les résultats attendus ; cela ne signifie pas que l'assourdissement par les consonnes ne soit pas un procédé prosodique russe mais simplement que notre corpus ne nous a pas permis de vérifier cette hypothèse. De plus cet assourdissement par les consonnes n'était peut-être pas recherché par Puškin dont on caractérise la langue comme "fluide" ; il le sera beaucoup plus tard par les Futuristes qui considéraient les groupements de consonnes comme caractéristiques du génie de la langue russe.

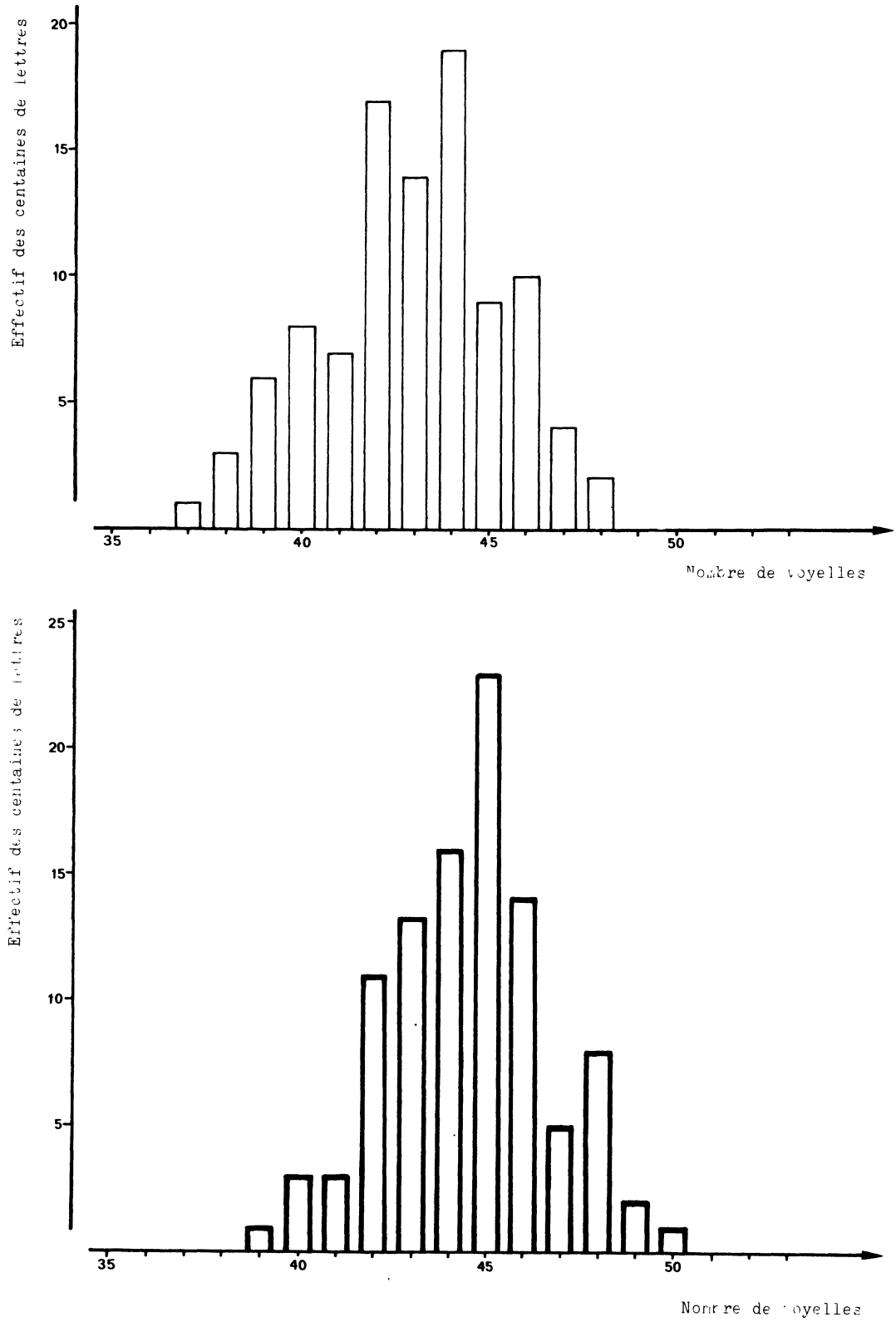


Figure 1. Distributions des centaines de lettres selon leur nombre de voyelles ;

en haut dans E a + b

en bas dans K 1 + 2.

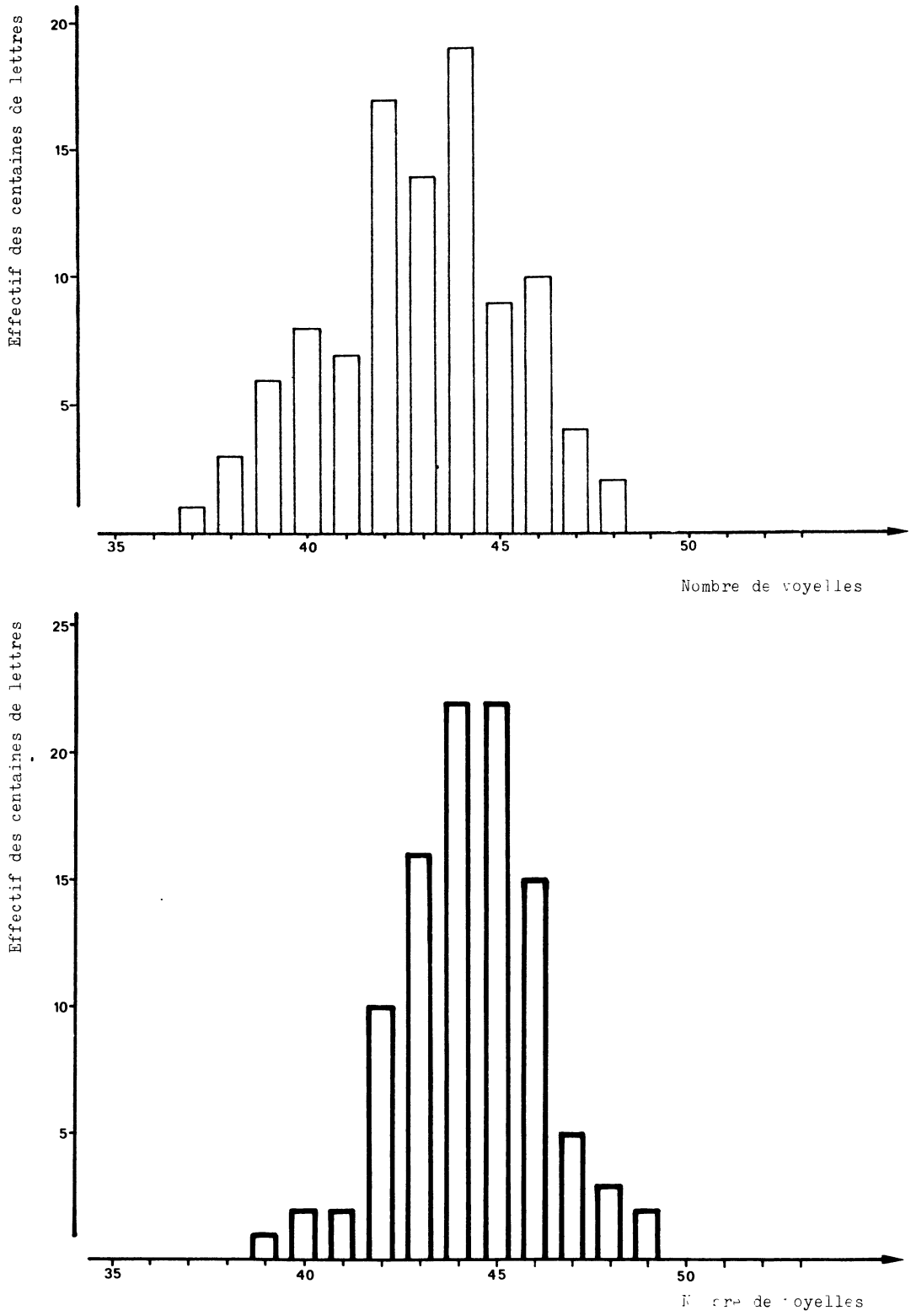


Figure 2. Distributions des centaines de lettres selon leur nombre de voyelles ;
en haut dans E a + b
en bas dans K 2 + 3 .

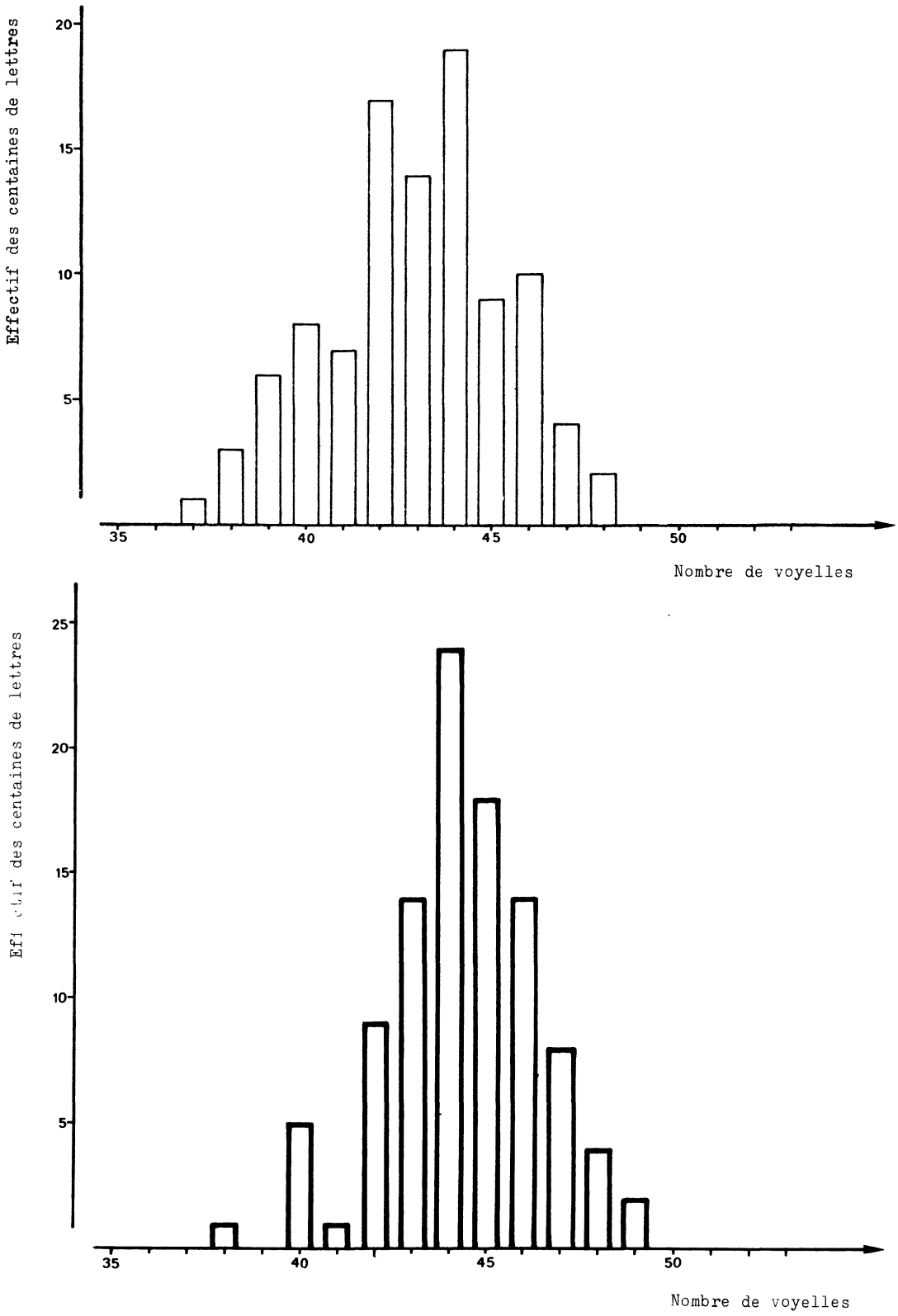


Figure 3. Distributions des centaines de lettres selon leur nombre de voyelles ;
en haut dans E a + b
en bas dans K 3 + 4 .

TABLEAU III. Distributions comparées des centaines de lettres selon leur nombre de triplets de consonnes sans les quatre textes de Puškin

	Nombre de triplets	0	1	2	3	4	5	6	7	8	$\bar{x}$	σ^2
Effectifs des centaines	E a + b	7	21	24	21	12	11	1	2	1	2,63	2,793
	K 1 + 2	8	19	27	12	21	8	3	1	1	2,66	2,844
	K 2 + 3	6	16	28	12	24	7	5	1	1	2,84	2,814
	K 3 + 4	6	20	25	18	14	9	6	2		2,750	2,887

Les coefficients d'asymétrie sont de haut en bas : 0,673 ; 0,529 ; 0,469 et 0,504 Les coefficients d'aplatissement sont de même : 3,290 ; 2,965 ; 2,892 et 2,549.

Nous avons alors calculé les caractéristiques markoviennes pour nos quatre textes. Nous allons détailler les calculs pour le texte en poésie E a + b, puis nous présenterons un tableau comparatif des résultats pour les quatre textes.

Longueur du texte en lettres relevées $L = 10000$

X = nombre de voyelles : 4299

Y = nombre de consonnes : 10000 - 4299 = 5701

x = nombre de doublets de voyelles : 506

$$p_1 = \frac{506}{4299} \simeq 0,11770 \qquad q_1 = 1 - p_1 \simeq 0,88228$$

Z = nombre de consonnes précédant une voyelle : $X - x = 4299 - 506 = 3793$

Q = nombre de doublets de consonnes : $Y - Z = 5701 - 3793 = 1908$

$$p_o = \frac{Z}{Y - 1} = \frac{3793}{5700} \simeq 0,66543 \qquad q_o = 1 - p_o \simeq 0,33456$$

$$\delta = p_1 - p_o \simeq - 0,54772$$

Coefficient de dispersion de la chaîne simple : $\frac{1 + \delta}{1 - \delta} \simeq 0,29221$; la valeur de ce coefficient que nous appelons M en l'honneur de Markov nous autorise à adopter le modèle de chaîne. Ici, comme Markov l'avait déjà montré cela n'apporte rien mais ce sera intéressant pour les autres textes.

Nous essayons alors le modèle de chaîne multiple.

x' = nombre de triplets de voyelles : 56

y' = nombre de triplets de consonnes : 265

$$p_{11} = \frac{x'}{x} = \frac{56}{506} \simeq 0,11067 \qquad q_{oo} = \frac{y'}{Q} = \frac{265}{1908} \simeq 0,13888$$

$$\varepsilon = \frac{p_{11} - p_1}{q_1} \simeq \frac{-\ 0,00705}{0,33468} \simeq -\ 0,00799$$

$$\eta = \frac{q_{oo} - q_o}{p_o} \simeq \frac{-\ 0,19580}{0,66532} \simeq -\ 0,29427$$

Coefficient de dispersion de la chaîne multiple :

$$C_m = \frac{1 + \delta}{1 - \delta} \left\{ \frac{1 + \varepsilon}{2(1 - \varepsilon)} + \frac{1 + \eta}{2(1 - \eta)} \right\} + \frac{(q - p)(\eta - \varepsilon)}{(1 - \varepsilon)(1 - \eta)} \simeq 0,19277 \quad \text{où } p \simeq 0,4299 \quad q \simeq 0,5701$$

Le tableau IV présente l'ensemble des résultats obtenus pour les quatre textes ; pour chaque caractéristique nous donnons la valeur décimale tronquée.

Nous voyons que les valeurs de p_1 (probabilité de voir apparaître une voyelle après une voyelle) pour les textes en prose sont toutes supérieures à la valeur de p_1 pour la poésie ; ceci confirme bien notre hypothèse. L'autre probabilité d'apparition de la voyelle, mais après une consonne, p_o, présente elle aussi des valeurs plus élevées dans les textes en prose que dans le roman en vers. Comme ces dernières sont supérieures aux premières cela entraîne le sens négatif des différences pour δ (delta grec) qui établit une relation entre les deux probabilités. Quant au coefficient de dispersion dont la valeur serait d'autant plus proche de l'unité que voyelles et consonnes seraient réparties au hasard, c'est-à-dire que leur succession n'obéirait pas à un modèle de chaîne, ses valeurs nous permettent d'affirmer que tous les textes obéissent au modèle.

Quant nous passons au modèle de chaîne multiple, nos hypothèses ne se vérifient pas de façon homogène. Pour p_{11}, probabilité de transition qui fait passer le système qui a présenté les états V et V aux moments $t - 2$ et $t - 1$ à l'état V à l'instant t, nous constatons que la valeur attachée au texte en prose K 3 + 4 est supérieure à celle de la poésie.

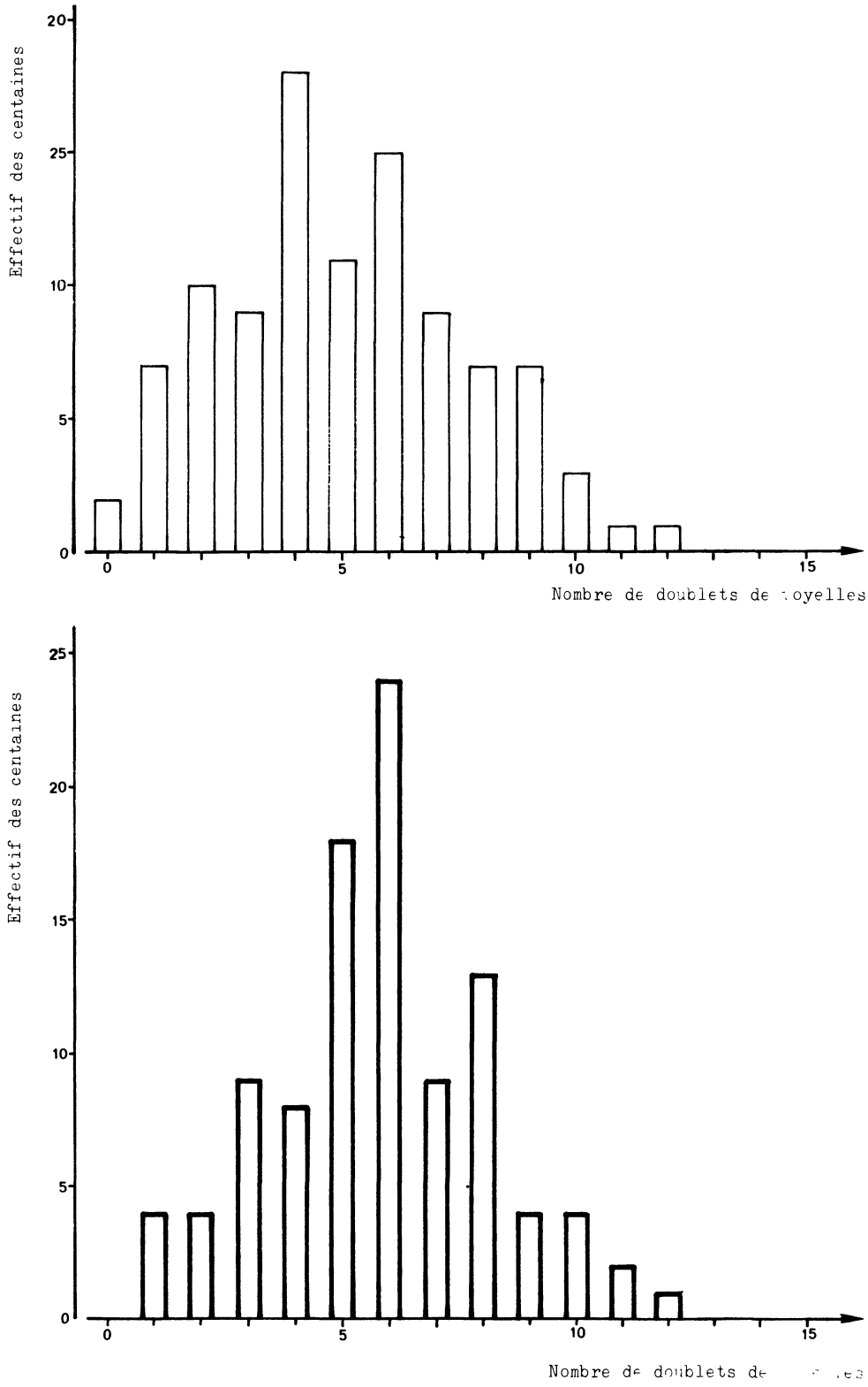


Figure 4. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles ;

en haut dans E a + b
en bas dans K 1 + 2

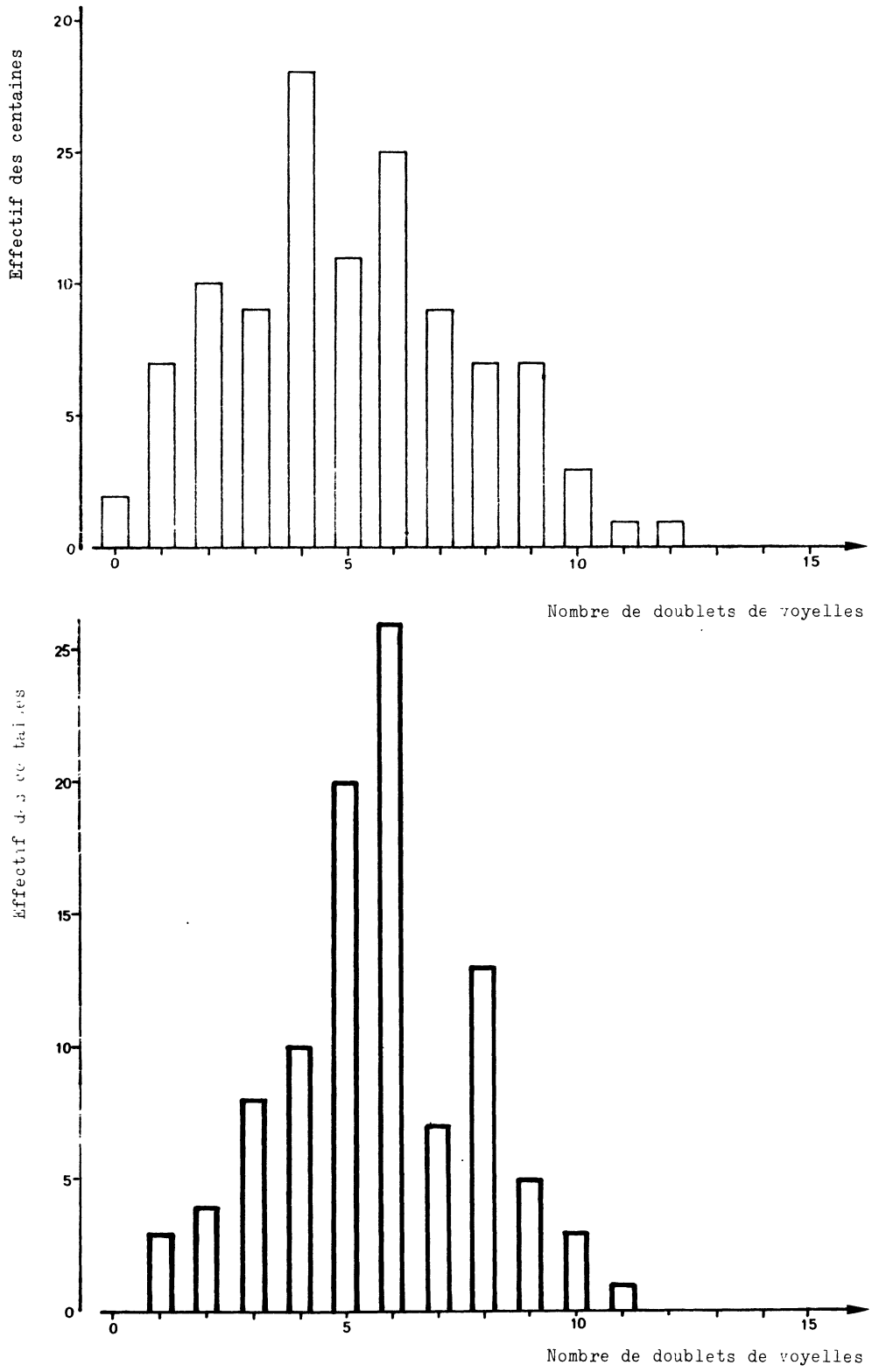


Figure 5. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles ;

en haut dans E a + b
en bas dans K 2 + 3 .

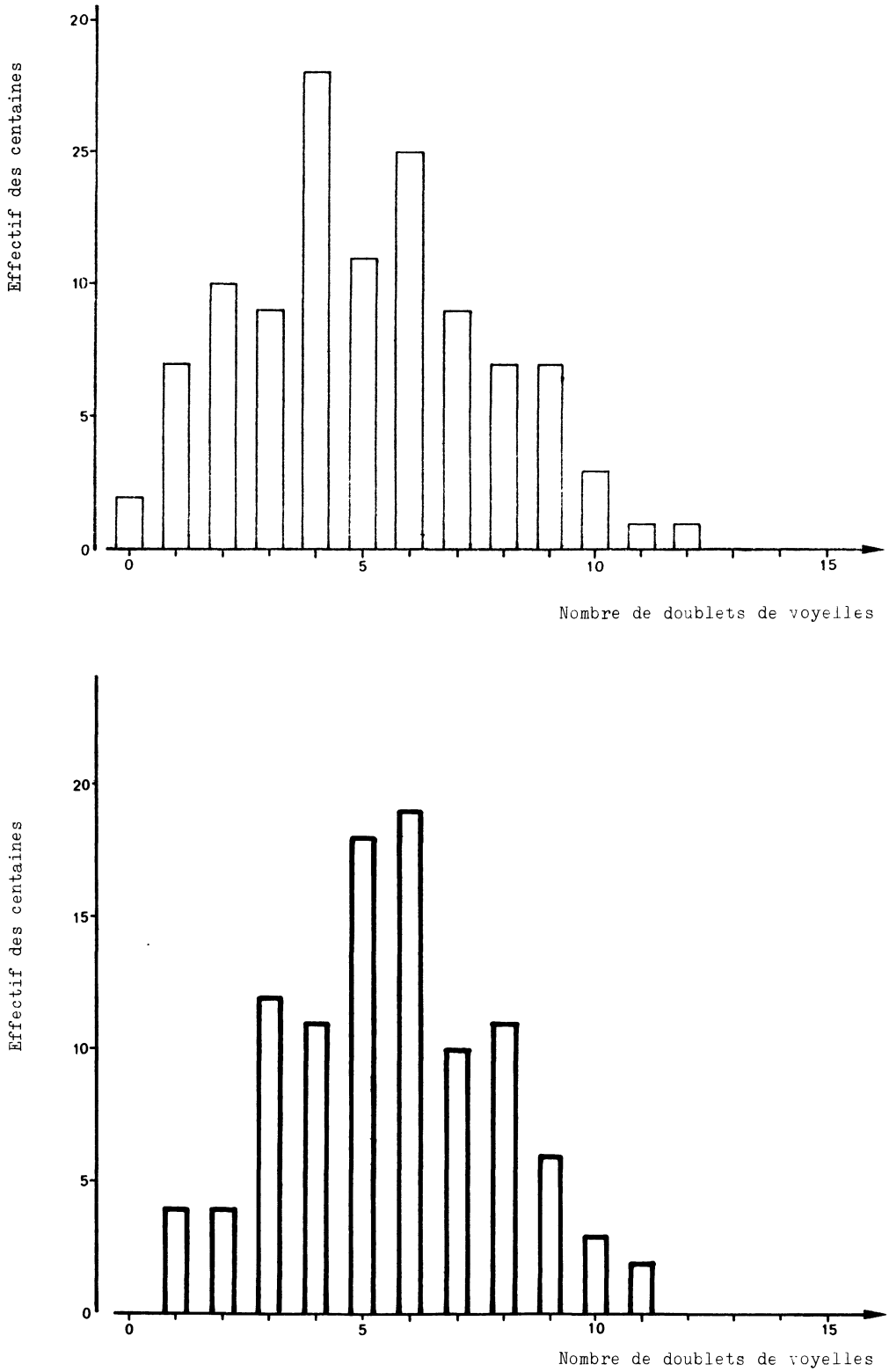


Figure 6. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles ;

en haut dans E a + b
en bas dans K 1 + 2 .

TABLEAU IV. Valeurs des caractéristiques markoviennes pour les quatre textes comparés de Puskin.

Caractéristiques markoviennes	Poésie E a + b	Prose K 1 + 2	K 2 + 3	K 3 + 4
$p_1 = \frac{\text{nombre vv}}{\text{nombre v}}$	0,117	0,130	0,128	0,126
$p_o = \frac{\text{nombre cv}}{\text{nombre c}}$	0,665	0,698	0,694	0,696
$\delta = p_1 - p_o$	0,547	0,567	0,565	0,569
$M = \frac{1 + \delta}{1 - \delta}$	0,292	0,276	0,277	0,274
$p_{11} = \frac{\text{nombre vvv}}{\text{nombre vv}}$	0,110	0,110	0,112	0,114
$q_{oo} = \frac{\text{nombre ccc}}{\text{nombre cc}}$	0,138	0,158	0,166	0,165
$\varepsilon = \frac{p_{11} - p_1}{q_1}$	- 0,007	- 0,024	- 0,018	- 0,014
$\eta = \frac{q_{oo} - q_o}{p_o}$	- 0,294	- 0,205	- 0,200	- 0,199
C_m	0,192	0,207	0,209	0,207

L'événement VVV étant relativement rare il se pourrait que cette valeur soit due à une fluctuation d'échantillonnage, que nos corpus n'aient pas été assez longs pour obtenir la stabilité. Aussi, bien qu'elle aille dans le sens de notre hypothèse hésitons-nous à lui accorder une grande valeur.

Par ailleurs l'indice de dispersion de la chaîne multiple C_m (dont nous avons redonné l'expression ci-dessus) bien qu'il constitue un meilleur ajustement que M, ce qui est logique et satisfaisant, nous semble difficile à interpréter car nous observons un renversement du sens des valeurs par rapport à celles de M, dans la comparaison prose/poésie.

I .2. OPPOSITION DISCOURS/REDACTION CHEZ LENINE.

Cependant encouragée par les résultats obtenus dans la comparaison des textes en prose et poétique de Puškin nous avons avancé une hypothèse plus intuitive que la précédente : les doublets et triplets de voyelles sont plus fréquents dans le langage oral que dans le langage écrit d'un même auteur russe. Cette hypothèse reposait au départ sur le raisonnement suivant : tout orateur, même très doué, lorsqu'il parle sans préparation a tendance à répéter la conjonction de coordination qui en russe est un graphème vocalique. Pour constituer le corpus il nous a paru intéressant de tenter l'expérience sur un homme réputé pour ses écrits théoriques mais en même temps grand tribun : Lénine. Il nous fut conseillé de retenir les années 1917-1918 où nous trouverions à comparer un important écrit théorique : *L'Etat et la Révolution* à l'un ou l'autre des nombreux discours à peu près contemporains de sa rédaction. Concrètement le choix s'est avéré assez difficile à faire : il n'est pas aisé de trouver un texte de 10000 lettres de Lénine qui ne soit coupé de citations longues de K. Marx. Finalement, ne pouvant comme Yule l'a montré [43] réaliser un véritable échantillon aléatoire, nous avons choisi un extrait du paragraphe trois : "Première phase de la société communiste" du chapitre cinq : "Bases économiques du dépérissement de l'Etat" [21]. Bien plus difficile encore a été le choix d'un discours susceptible de représenter le langage parlé de Lénine. Il nous a semblé que le texte intitulé *Réponses aux billets* qui suit le discours fait par Lénine au "Congrés extraordinaire des cheminots de toute le Russie" [22] et non le discours lui-même, pouvait être considéré comme ce qu'il y avait de plus proche du véritable discours oral. Un rapide décompte sur les quarante premières lignes (dans la même édition) des deux textes a confirmé notre hypothèse. Si nous appelons v' = "и" (cyrillique) conj. et v = tous les autres graphèmes vocaliques y compris "и" dans un mot on a

dans le texte écrit : $\underset{vv'}{5} + \underset{v'v}{1} = 6$

dans le discours : $\underset{vv'}{11} + \underset{v'v}{3} = 14.$

Reste le problème de la publication qui a pu nécessiter des "retouches" mais nous pensons qu'elles ont été légères si on en juge par le critère stylistique introduit par G.U. Yule : la longueur des phrases. Ce texte n'a permis de constituer qu'un corpus de 5000 lettres.

Yule [43], en présence de deux corpus qu'il savait d'auteurs différents montre que certaines caractéristiques de la longueur des phrases considérée comme variable statistique sont différentes. Bien entendu il prend en compte, dans la mesure du possible, une certaine homogénéité de sujet ; nous avons, comme lui "fait de notre mieux" sur ce point. Mais il ajoute à l'unité de sujet "des circonstances identiques". C'est sur cette condition qui, précisément, n'est pas réalisée dans notre cas que nous nous appuyons. Nous allons essayer de montrer que le même auteur, parlant approximativement des mêmes sujets, mais dans des circonstances très différentes : discours non préparé - rédaction d'un ouvrage théorique - tient deux discours présentant des caractéristiques permettant de les distinguer.

Nous présentons le Tableau V des distributions comparées de la longueur des phrases, mesurée par le nombre de mots dans les corpus considérés. Nous appelons L_{1a} les 5003 premières lettres consécutives du discours écrit et L_{1b} les 5121 suivantes (leur total couvre donc un peu plus que les 10000 lettres relevées pour l'étude markovienne puisqu'il est de 10124 lettres). L_o correspond aux 5019 premières lettres consécutives du discours parlé. Il n'est pas surprenant de constater que des corpus d'environ 5000 lettres se distinguent par les caractères suivants : L_o présente la phrase la plus courte et en même temps *le plus grand nombre de phrases* : donc une longueur moyenne de 20,16 mots ou en lettres 116,72. L_1 contient la phrase la plus longue et présente une longueur moyenne de 35,72 mots ou en lettres 227,40.

TABLEAU V . Caractéristiques de Yule des longueurs de phrase dans les trois textes étudiés.

	L_o	L_1	
		a	b
longueur moyenne	20,16	35,72	24,97
longueur de la phrase médiane (en mots)	17	26	23
longueur modale	63	98	63
Q_1 (longueur en mots)	10	23	14
Q_3 (longueur en mots)	28	46	38
$Q_3 - Q_1$ (longueur en mots)	18	23	24
D_3 (longueur en mots)	40	47	44

Ces résultats nous autorisent à considérer les deux discours comme "différents".

Nous avons alors appliqué la procédure markovienne aux trois corpus. Les tableaux élémentaires, 50 pour chaque corpus, nous ont permis d'établir les distributions comparées des centaines de lettres selon leur nombre de voyelles ou de doublets de voyelles.

TABLEAU VI . Distributions comparées des centaines de lettres selon leur nombre de voyelles dans les trois textes de Lénine.

	Nombre de voyelles	37	38	39	40	41	42	43	44	45	46	47	48	49	$\bar{x}$	σ^2
Effectif des centaines	L_o	1	0	0	0	2	2	6	11	13	5	5	4	1	44,72	4,466
	L_{1a}		2	0	1	5	7	9	9	6	5	4	2		43,44	5,190
	L_{1b}	1	1	1	4	4	5	9	7	6	5	6	0	1	43,48	6,609

Les coefficients d'asymétrie sont, de haut en bas, respectivement : - 0,722 ; - 0,221 ; - 0,275. Les coefficients d'aplatissement sont de même : 4,914 ; 2,825 ; 2,600.

On voit aisément que notre hypothèse se vérifie : la valeur moyenne du nombre de voyelles étant plus élevée dans le discours que dans les deux textes (ici rappelons que nous sommes revenus à des corpus égaux ou du moins nous ayant permis d'établir pour chacun 50 tableaux élémentaires). On peut se reporter aux Fig. 7 et 8 pour les graphiques correspondants.

TABLEAU VII . Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles dans les trois textes de Lénine.

	Nombre de doublets de voyelles	2	3	4	5	6	7	8	9	10	11	12	13	14	$\bar{x}$	σ^2
Effectif des centaines	L_o			10	5	7	7	6	8	4	3				6,98	4,779
	L_{1a}		6	4	10	9	7	5	4	2	2	0	0	1	6,38	5,635
	L_{1b}	5	3	6	6	5	4	6	9	1	4	1			6,52	7,689

Les coefficients d'asymétrie sont de haut en bas, respectivement : 0,125 ; 0,749 ; 0,011. Les coefficients d'aplatissement sont de même : 1,789 ; 3,506 ; 1,900.

TABLEAU VIII. Distributions comparées de la longueur des phrases mesurée en mots dans L_o, L_{1a} et L_{1b}

Longueurs ordonnées en nombre de mots			Effectifs			Effectifs cumulés		
L_o	L_{1a}	L_{1b}	L_o	L_{1a}	L_{1b}	L_o	L_{1a}	L_{1b}
3			1			1		
		4			1			1
5			3			4		
6		6	2		1	6		2
7			1			7		
8	8	8	3	2	1	10	2	3
10		10	1		2	11		5
11			3			14		
12		12	1		1	15		6
13	13	13	3	1	1	18	3	7
14		14	1		1	19		8
15		15	1		1	20		9
16	16	16	1	1	2	21	4	10
17			1			22		11
18		18	2		1	24		
19		19	1		1	25		12
	20	20		2	3		6	13
		21			1			14
22		22	2		1	27		15
23	23	23	2	1	1	29	7	16
	24	24		1	1		8	17
	25			1			9	18
26	26	26	1	1	2	30	10	19
27			1			31		20
28		28	2		1	33		
	29			1			11	22
30			1			34		
31			1			35		23
32		32	1		1	36		
34			1			37		
								24
38		38	1		1	38		25
	39	39		1	1		12	26
40	40		1	2		39	14	
41	41		2	1		41	15	
		43			1			27
		44			1			28
46	46			1			16	
47	47	47	1	1	1	43	17	29
	48			1			18	
		50			2			31
	53			1			19	
	58			1			20	
62			1			44		
	65	65		1	1		21	32
	97			1			22	

Nous obtenons ici aussi des résultats qui vont dans le sens de notre hypothèse bien que les écarts observés entre les valeurs moyennes du nombre des doublets soient moins nets que ceux observés dans les distributions des voyelles ; ceci est visible aussi sur les Fig. 9 et 10 correspondant à ces distributions. Mais ici aussi il nous faut nous montrer prudents car nous n'avons réalisé que deux comparaisons et sur des corpus relativement courts.

Après cela nous avons effectué pour chacun des corpus les décomptes directs permettant de calculer les caractéristiques markoviennes et en particulier de connaître les valeurs du coefficient de dispersion pour nous assurer que nous n'étions pas dans un cas d'approximation par la loi normale.

Pour le corpus L_o : longueur du texte en lettres relevées : 5000

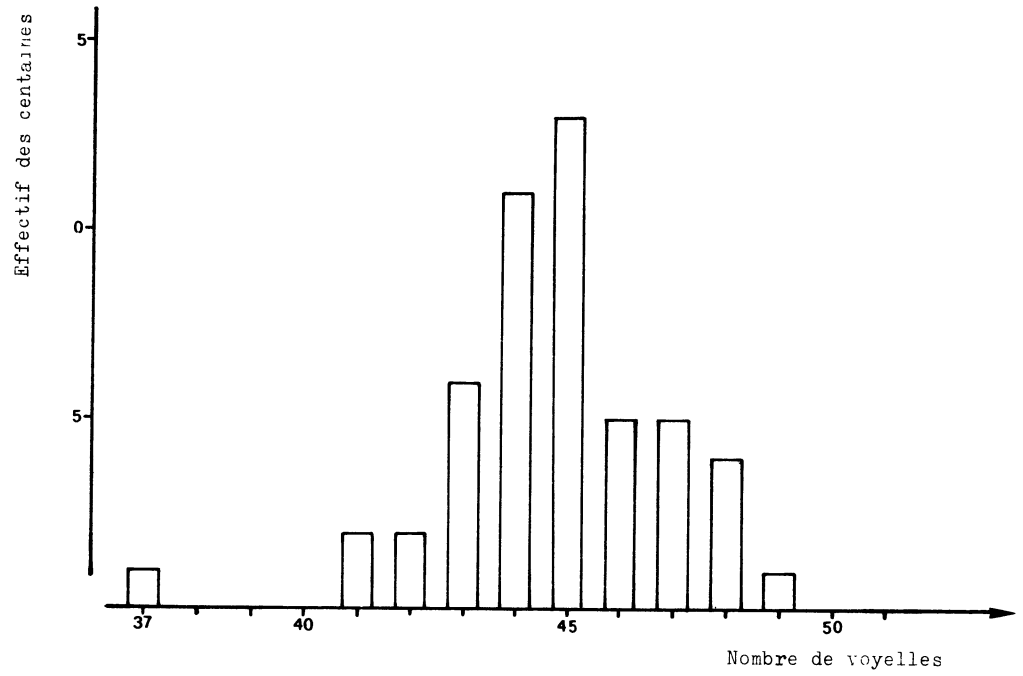
X = nombre de voyelles : 2236

Y = nombre de consonnes : 5000 - 2236 = 2764

x = nombre de doublets de voyelles : 349

TABLEAU IX. Les caractéristiques markoviennes des trois textes de Lénine.

Caractéristiques markoviennes	L_o	L_{1a}	L_{1b}
$p_1 = \frac{\text{nombre vv}}{\text{nombre v}}$	0,156	0,146	0,151
$p_o = \frac{\text{nombre cv}}{\text{nombre c}}$	0,682	0,660	0,652
$\delta = p_1 - p_o$	- 0,526	- 0,514	- 0,504
$M = \frac{1 + \delta}{1 - \delta}$	0,310	0,321	0,332
$p_{11} = \frac{\text{nombre vvv}}{\text{nombre vv}}$	0,111	0,109	0,097
$q_{oo} = \frac{\text{nombre ccc}}{\text{nombre cc}}$	0,173	0,187	0,206
$\varepsilon = \frac{p_{11} - p_1}{q_1}$	- 0,052	- 0,042	- 0,063
$\eta = \frac{q_{oo} - p_o}{p_o}$	- 0,210	- 0,229	- 0,217
σ_m	0,227	0,229	0,236



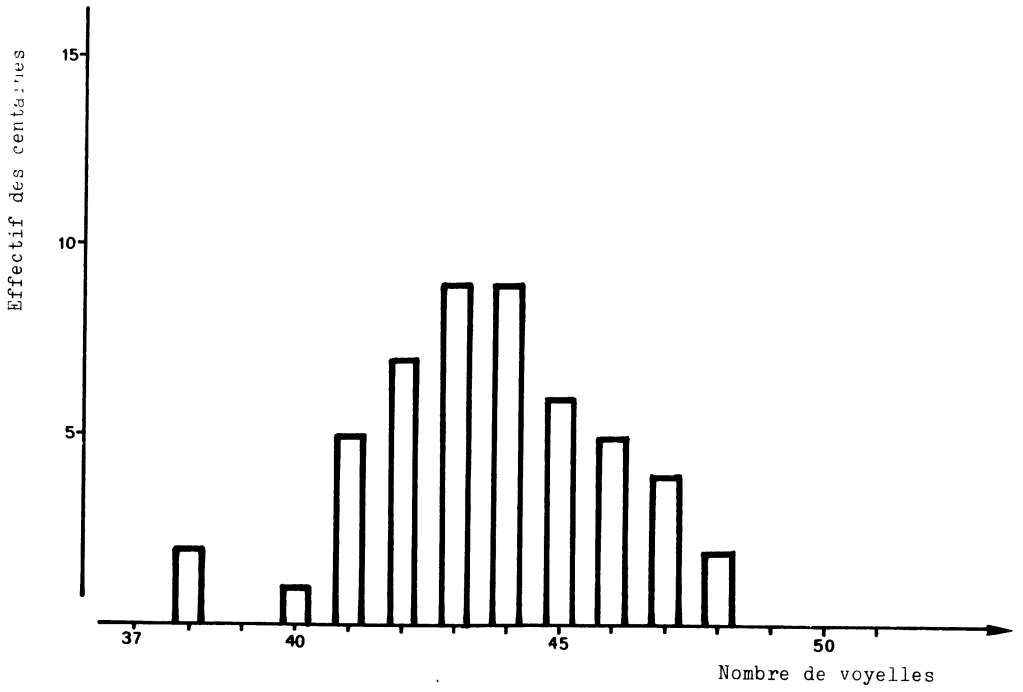
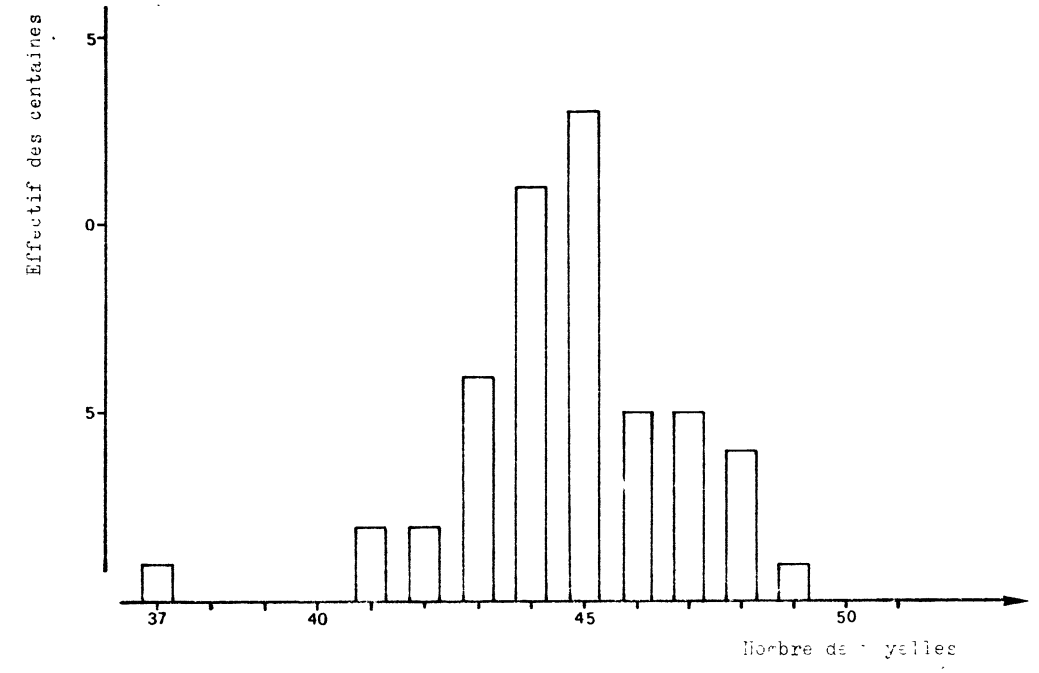


Figure 7. Distributions comparées des centaines de lettres selon leur nombre de voyelles ;
en haut dans L_o = discours
en bas dans L_{la} = rédaction .



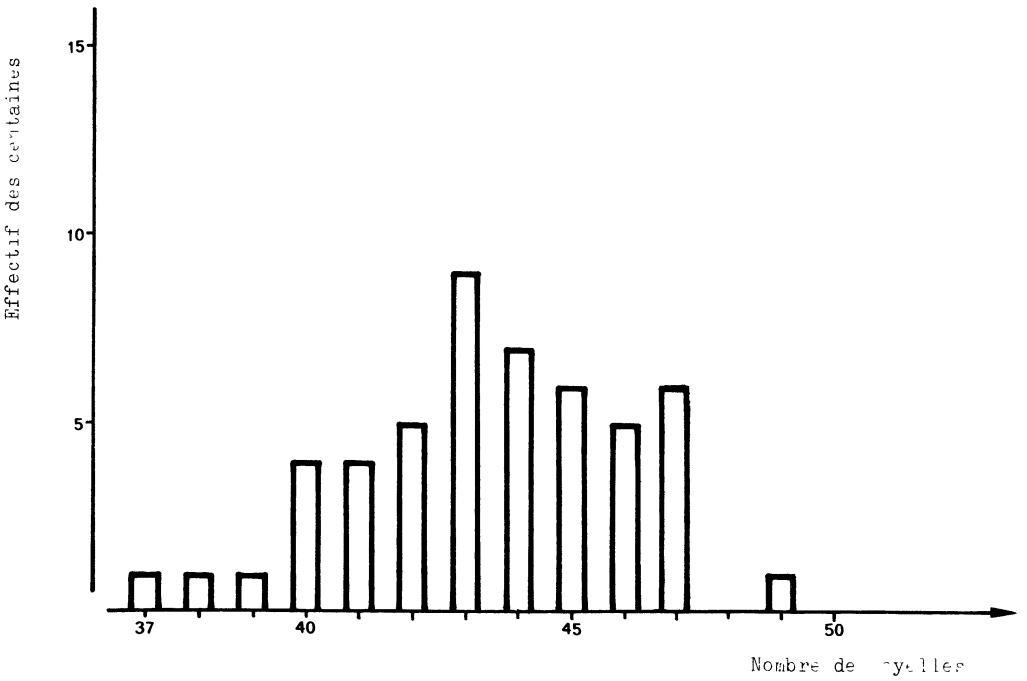


Figure 8. Distributions comparées des centaines de lettres selon leur nombre de voyelles ;
en haut dans L_o = discours
en bas dans L_{lb} = rédaction .

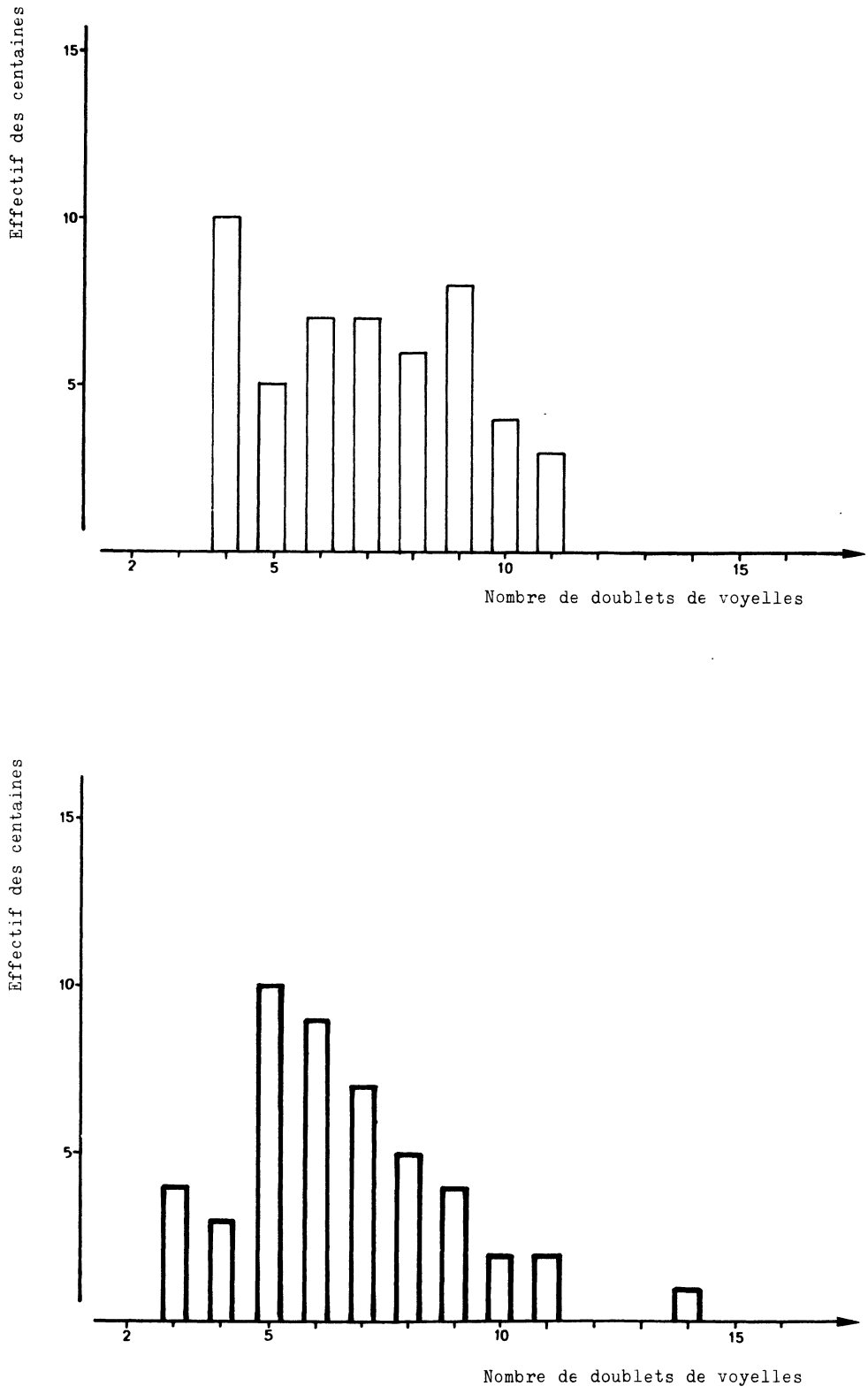
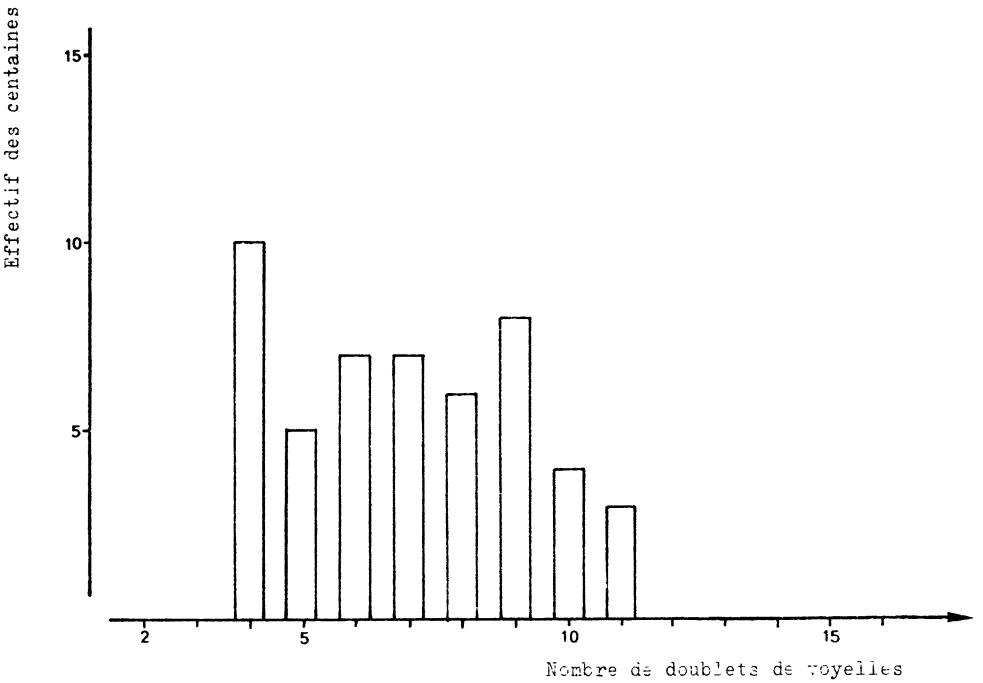


Figure 9. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles ;
en haut dans L_o = discours
en bas dans L_{la} = rédaction .



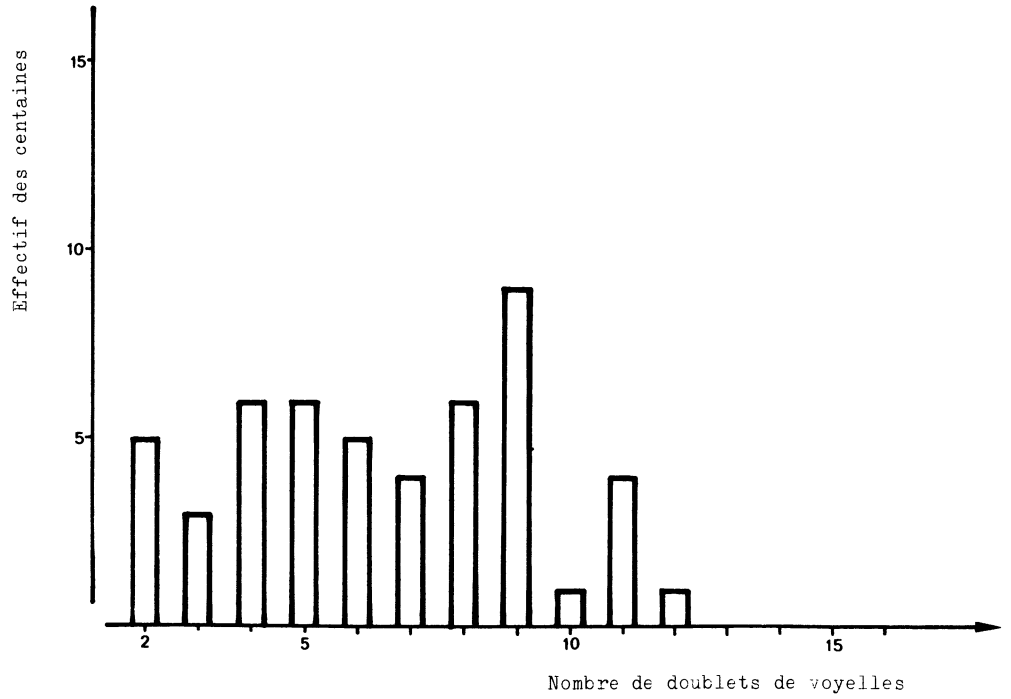


Figure 10. Distributions comparées des centaines de lettres selon leur nombre de doublets de voyelles ;
en haut dans L_o = discours
en bas dans L_{1b} = rédaction .

$p_1 = \frac{349}{2235} \simeq 0,15615$

$Z = X - x$ = nombre de cv : 2236 - 349 = 1887

Q = nombre de cc : 2764 - 1887 = 877

$p_o = \frac{1887}{2764} \simeq 0,68270$

$\delta = p_1 - p_o \simeq 0,15615 - 0,68270 \simeq - 0,52655$

$\frac{1 + \delta}{1 - \delta} \simeq \frac{0,47345}{1,52655} \simeq 0,31014$

nombre de triplets de voyelles : $x' = 39$

nombre de triplets de consonnes : $y' = 152$

$p_{11} = \frac{x'}{x} = \frac{39}{349} \simeq 0,11174$ $\qquad$ $q_{oo} = \frac{y'}{Q} = \frac{152}{877} \simeq 0,17331$

$p = \frac{2236}{5000} \simeq 0,447$ $\qquad$ $q = 1 - p \simeq 0,553$

On verra les valeurs des caractéristiques des autres corpus dans le Tableau IX.

Dans le cas de la comparaison tentée le modèle de chaîne simple nous donne des résultats satisfaisants et cette fois-ci p_{11} confirme aussi notre hypothèse mais par contre ε et C_m ne sont pas stabilisés.

II.1. UN AUTRE EXERCICE : COMPARAISON POESIE/PROSE ENTRE UN AUTEUR CLASSIQUE A.S. PUŠKIN ET UN AUTEUR FUTURISTE V.V. HLEBNIKOV.

Nous ne pouvons donner ici - ce serait à la fois hors de notre compétence et de notre sujet - un aperçu de la poétique des Futuristes. Nous renvoyons le lecteur aux ouvrages qui en traitent : [19], [20], [26], [35] par exemple. Ce qui nous intéresse c'est le fait qu'il s'en déduit que les Futuristes ont consciemment tenté d'altérer les fréquences des graphèmes vocaliques et consonantiques dans la langue russe. Nous allons essayer de voir, par l'utilisation des critères markoviens, dans quelle mesure ils y ont réussi et se distinguent ainsi des auteurs classiques. Et nous avons choisi de comparer Puškin et Hlebnikov.

Nous avons constitué pour ce dernier, deux corpus à partir des textes présentés par [20] mais en les complétant selon l'édition des Oeuvres complètes [47] quand il y a eu coupure : l'un comprend des textes en prose, l'autre des textes en vers. Les deux corpus sont de 15000 graphèmes car nous allons les comparer à des données de même longueur pour Puškin, données qui comprennent les 10000 graphèmes étudiés en I.1. Il s'agit donc pour le texte en prose de *La Fille du Capitaine* et pour la poésie d'*Evgenii Onegin* : les 10000 premiers graphèmes consécutifs de I.1. auxquels on a ajouté 5000 autres graphèmes consécutifs pris au hasard dans la suite de l'oeuvre, l'ensemble analysé comme formant un tout. Pour Hlebnikov voici comment sont constitués les deux corpus :

Prose : 1) *Nous et les maisons* (texte de théorie urbanistique daté de 1914/15).

2) *Les enfants de la loutre* (ce poème épique comprend des parties en prose et des parties versifiées ; ici nous prenons ce que l'auteur a appelé la "Première voile" ; daté de 1910/12).

3) *Ka* (ce roman a été choisi pour permettre une comparaison interne entre prose 'normale' et prose 'zaoum'. Ici nous prenons le texte normal ; daté de 1916).

4) *Notre base* (texte de théorie littéraire très important puisqu'il énumère les 'langages' ou 'procédés' qui constituent l'essentiel de la poétique futuriste aux yeux de l'auteur ; daté de 1919/20).

Ces quatre extraits sont de longueur sensiblement égale et sont analysés comme formant un tout.

Poésie : 1) *Trois poèmes* (pièces relativement courtes, caractéristiques des théories de l'auteur sur les 'racines' ; datées de 1913).

2) *La guerre dans une souricière* (poème politique où l'auteur exprime vigoureusement ses positions pacifistes ; daté de 1915/17).

3) *Les enfants de la loutre* (partie versifiée : la "troisième voile" ; daté de 1910/12).

4) *Ladomir* (long poème philosophique ; daté de 1922).

Ces quatre parties sont un peu moins bien équilibrées que celles du discours en prose, l'extrait de Ladomir étant un peu plus long que chacun des trois autres ; elles constituent un corpus analysé comme un tout.

Les théories de l'auteur nous ont confronté à deux types de difficultés : l'utilisation du 'langage zaoum' qui concrètement se ramène à la présence d'onomatopées et l'utilisation de formules algébriques dans le texte. Des deux procédés nous avons des exemples courts dans nos corpus. Nous avons accepté des onomatopées courtes (séquence de cinq voyelles dans le choix des poèmes du fragment 1. Exemple :

Потужить за лесом совкой
Ай! Ах, на ...

Potužit' za lesom sovkoj
Aj! Ah, na ...

S'attrister en chouette au bois (1)
Las, hélas! sur ...

car dans le fragment 2 nous observons :

...где волк воскликнул кровью:
"ей ! я юноши ...

gde volk voskliknul krovjo :
"ej ! ja jonoši ...

où le loup a glapi dans le sang :
"Hé! d'un adolescent je ...

Il en est de même pour les graphèmes consonantiques ; ceci n'est qu'un exemple. Il faut préciser cependant qu'aucune de ces séquences remarquables ne naît de la mise bout à bout de nos fragments, elles sont toutes intérieures à ceux-ci.

Le problème des onomatopées a paru assez important pour que lui soit consacré une étude interne que l'on trouvera ci-dessous en III.2. Comme L. Schnitzer, mais pour d'autres raisons, nous avons écarté les textes comportant des expressions et même de véritables développements algébriques et nous n'avons qu'un exemple d'emploi du signe = que nous avons transcrit en toutes lettres (равно : ravno : égale) dans *Ka*. Nous avons aussi écrit en toutes lettres les nombres rencontrés comme nous l'avions fait pour les dates dans la prose de Puškin. Nous n'allons pas ici détailler les calculs markoviens et donnons les résultats sous forme de tableaux.

(1) Trad. L. Schnitzer.

Tableau X. Les caractéristiques markoviennes pour les corpus comparés de Puškin et Hlebnikov (15000 graphèmes de prose).

Caractéristiques markoviennes	Puškin	Hlebnikov
p	0,677	0,668
$p_1 = \frac{VV}{V*}$	0,132	0,146
$p_o = \frac{CV}{C*}$	0,714	0,686
$\delta = p_1 - p_o$	-0,582	-0,540
$M = \frac{1 + \delta}{1 - \delta}$	0,264	0,309
$p_{11} = \frac{VVV}{VV*}$	0,119	0,101
$q_{oo} = \frac{CCC}{CC*}$	0,171	0,164
$\varepsilon = \frac{p_{11} - p_1}{q_1}$	-0,015	-0,052
$\eta = \frac{q_{oo} - q_o}{p_o}$	-0,160	-0,217
C_m	0,169	0,198

En prose on constate la présence d'un plus grand nombre de consonnes (8317) chez un futuriste que chez Puškin (8230). Les valeurs de M nous permettent dans les deux cas d'adopter l'hypothèse d'une chaîne sous-jacente d'ordre 1. Elles sont très différentes mais toutes lés deux comprises dans l'intervalle de probabilité que nous avons pu déterminer pour la prose russe [45] dont nous rappelons les bornes : $0,089 < M < 0,489$. L'indicateur de chaîne d'ordre 2 est dans les deux cas plus satisfaisant que M, comme dans l'exemple traité par Markov lui-même (cf. dans ce numéro p.27).

Ces résultats tiennent peut-être à la constitution du corpus futuriste mais nous devons constater que si M et C_m discriminent effectivement les deux auteurs, les disparités que notent les autres caractéristiques markoviennes ne sont pas aussi marquées qu'on aurait pu s'y attendre.

Considérons maintenant le corpus de poésie.

Tableau XI. Les valeurs des caractéristiques markoviennes pour les corpus comparés de Puškin et de Hlebnikov (15000 graphèmes de poésie).

Caractéristiques markoviennes	Puskin	Hlebnikov
p	0,643	0,651
$p_1 = \frac{VV}{V*}$	0,122	0,135
$p_o = \frac{CV}{C*}$	0,659	0,663
$\delta = p_1 - p_o$	-0,538	-0,527
M	0,300	0,309
$p_{11} = \frac{VVV}{VV*}$	0,093	0,132
$q_{oo} = \frac{CCC}{CC*}$	0,152	0,185
$\varepsilon = \frac{p_{11} - p_1}{q_1}$	-0,032	-0,004
$\eta = \frac{q_{oo} - q_o}{p_o}$	-0,286	-0,229
C_m	0,279	0,305

Nous ne pouvons donc pas conclure ici que M soit discriminant ; comme il s'agit de corpus poétique, nous pouvons admettre que M soit insuffisant si C_m répond à notre attente ce qui est le cas. Par ailleurs on peut noter l'importance de la valeur de p_{11} c'est-à-dire en définitive du rôle des

triplets de voyelles dans l'oeuvre poétique du futuriste ainsi que celui moins marqué cependant des triplets de consonnes. Mais par contre nous ne savons pas interpréter la quasi égalité des valeurs de M et de C_m pour Hlebnikov.

S'intéresser à C_m c'est considérer les séquences de graphèmes de longueur trois ou plus ; c'est ce que nous allons faire mais auparavant nous allons exposer les résultats de la comparaison chez Hlebnikov entre langage naturel et langage 'zaoum'.

II.2. LANGAGE 'NORMAL' ET LANGAGE 'ZAOUM' EN PROSE ET EN POESIE.

Pour cette étude nous avons constitué deux corpus, l'un de prose l'autre de poésie ; ces corpus sont courts (600 graphèmes toujours relevés selon les conventions de Markov) car nous avons pris le plus long fragment de prose 'zaoum' dont nous disposions (*Ka* : p.144 de [20]) encadré de longueurs convenables de texte en langue naturelle car le procédé ne peut être tiré de son contexte sinon on altère l'oeuvre. Les autres fragments ont donc été choisis conventionnellement de cette longueur. Pour la prose il s'agit du début même de l'oeuvre (*Ka*).

Pour la poésie on a comparé un fragment de poésie zaoum : *Nuits de Galicie*[1] à deux fragments de poésie normale : les 600 premiers graphèmes des *Enfants de la loutre* et 600 graphèmes pris au hasard dans *Ladomir*. Les résultats sont présentés au tableau XII.

---

(1) Pour les onomatopées utilisées dans ce poème l'auteur s'est explicitement référé aux relevés d'incantations magiques contenues dans l'oeuvre de Caharov [48]. Il s'en est sans aucun doute inspiré mais en dehors des interjections les plus courantes il ne semble pas qu'il les ait reproduites in extenso. Pour la prose nous n'avons pas d'autres indications que celles de l'auteur sur le fait que certaines séquences "mantch" ce sont imposées spontanément à lui. Le reste est composition.

Tableau XII. Valeurs des caractéristiques markoviennes pour le langage normal et le langage zaoum en poésie et en prose.

Caractéristiques markoviennes	Prose		Poésie		
	zaoum	normal	zaoum	normal Enf.Loutre	normal Ladomir
p	0,270	0,265	0,276	0,254	0,252
$p_1 = \frac{VV}{V*}$	0,222	0,147	0,170	0,119	0,119
$p_o = \frac{CC}{C*}$	0,362	0,326	0,709	0,644	0,640
$\delta = p_1 - p_o$	-0,140	-0,178	-0,539	-0,526	-0,521
$M = \frac{1 + \delta}{1 - \delta}$	0,754	0,697	0,300	0,311	0,315
$p_{11} = \frac{VVV}{VV*}$	0,200	0,102	0,277	0,013	0,067
$q_{oo} = \frac{CCC}{CC*}$	0,210	0,193	0,115	0,190	0,208
ε	-0,563	-0,052	0,130	-0,120	-0,060
η	-0,237	-0,200	-0,260	-0,261	-0,240
C_m	0,264	0,200	0,123	0,164	0,170

S'il n'y a rien de très notable pour la poésie où les critères sont discriminants et 'normaux', la prose fait problème car M sort de l'intervalle de probabilité de la prose russe déterminé en [45]. Nous aurions pu être tentée d'interprêter par le caractère 'spécial' de l'échantillon mais nous allons voir en III.2. que l'étude de l'ordre selon les effectifs des n-uplets de graphèmes nous conduit a admettre que ce texte court d'écriture normale ne peut être considéré comme un échantillon 'spécial'.

III. ETUDE DE L'ORDRE DES *n*-UPLETS DE GRAPHEMES SELON LEURS EFFECTIFS.

Il est aisé de comprendre que le dénombrement[1] des doublets, des triplets, ..., des *n*-uplets de graphèmes est affecté par la présence d'une seule séquence exceptionnelle. On trouve par exemple dans *Ka* la huitième proposition d'Aménophis :

"8) р р р р а га-га.Га ! грав! Эньма мээиу-уиай! Аменофис .."

"8) r r r r a ga-ga. Ga! grav ! Enma mééou-ouiaiiii! Aménophis ..."

On observe une séquence de neuf voyelles ... CVVVVVVVVC ... . On symbolisera V^i une séquence de i voyelles. On observera donc nécessairement deux fois V^8 mais aussi trois fois V^7, etc... Le dénombrement des séquences V^i se partage entre les effectifs des séquences issues de V^{i+1} et des séquences V^i *encadrées* c'est-à-dire où apparaît à droite et à gauche un graphème de l'autre catégorie. On voit donc que les effectifs de ces dénombrements ne sont pas indépendants. Nous pensons que Markov avait sûrement vu le problème au niveau des doublets mais qu'il ne s'y est pas intéressé. Pour élaborer ses paramètres il a tenu un discours probabiliste. Nous allons essayer de montrer qu'il est possible parallèlement de rester à un niveau purement descriptif mais que cette façon d'appréhender un texte présente, nous semble-t-il aussi son intérêt.

III.1. COMPARAISON PUŠKIN-HLEBNIKOV (15000 graphèmes de prose et 15000 GRAPHEMES EN VERS.

Les doublets en prose

Hlebnikov		Puškin
5708	CV	5877
5707	VC	5876
2609	CC	2353
975	VV	893
14999		14999

L'ordre des séquences est le même pour les deux auteurs. Pour Hlebnikov l'alternance (CV ou VC) est un peu moins fréquente que chez Puškin et la compensation nécessaire est dûe plutôt aux doublets de consonnes ce qui est cohérent avec les théories de l'auteur. Les liaisons nécessaires entre les effectifs apparaissent dès que l'on calcule les différences entre termes successifs : entre les effectifs de CV et de VC la différence est au plus d'une unité.

(1) La plupart des décomptes ont été effectués par un programme qu'a rédigé H. Labesse (Département de Mathématique et Informatique appliquée aux Sciences Humaines de l'Université de Paris-Sorbonne (Paris IV)).

Les doublets en vers

Hlebnikov		Puškin
5627	CV	5650
5627	VC	5649
2864	CC	2918
881	VV	782
14999		14999

La comparaison se fait comme ci-dessus pour la prose. L'ordre des séquences est le même mais ici, contrairement à toute attente, le jeu des doublets de consonnes est beaucoup moins marqué chez Hlebnikov.

Les triplets

en prose			*en vers*	
Hlebnikov	Puškin		Hlebnikov	Puškin
4831	5089	CVC	4861	4940
3527	3927	VCV	3294	3175
2180	1949	VCC	2333	2474
2180	1950	CCV	2333	2474
876	787	VVC	765	709
876	787	CVV	765	709
429	403	CCC	531	444
99	106	VVV	116	73
14998	14998		14998	14998

Deux remarques s'imposent :

. la ressemblance est très forte entre les deux distributions prises deux à deux, l'ordre des séquences est le même ;

. la liaison se voit aisément : les séquences VCC et CCV sont également fréquentes à une unité près. Les déficits de Hlebnikov par rapport à Puškin quant aux séquences hétérogènes sont compensées, en prose, par des excédents des séquences soit parfaitement homogènes pour les consonnes, soit partiellement homogènes et symétriques (VVC et CVV). Il faut remarquer aussi un petit excédent des triplets de voyelles chez Puškin qui est une façon de voir la fluidité de sa prose.

Pour les vers on constate un léger contraste que traduit l'excédent des triplets de consonnes chez Hlebnikov et, ce qui étonnera davantage, un contraste mais plus léger des triplets de voyelles par rapport à Puškin.

Les quadruplets. Nous présentons les deux types de discours et les deux auteurs sous forme de tableau.

Tableau XIII. Les effectifs de séquences de graphèmes et leurs différences.

	Puškin			Hlebnikov	
	Vers	Prose		Prose	Vers
VCVC	2817	3433	CVCV	3007	2856
		(52)		(12)	
CVCV	2765	3408	VCVC	2995	2827
		(25)			(29)
CVCC	2175	1681	CCVC	1835	2034
		(52)		(11)	
CCVC	2123	1656	CVCC	1824	2005
	(53)	(57)		(38)	(136)
VCCV	2070	1599	VCCV	1786	1869
CVVC	643	686	CVVC	787	660
	(233)	(167)		(256)	(193)
VVCV	410	519	VCVV	531	467
		(52)		(11)	
VCVV	358	494	VVCV	520	437
		(25)			(30)
VCCC					
CCCV	404	350	VCCC=CCCV	394	464
	(54)	(57)		(38)	(136)
CCVV	350	293	VVCC	356	328
		(51)		(25)	(11)	(30)
VVCC	299	268	CCVV	345	298
	(233)	(167)		(256)	(193)
VVVC					
CVVV	66	101	VVVC=CVVV	89	105
CCCC	40	53	CCCC	35	67
VVVV	7	5	VVVV	10	11

Les deux distributions se ressemblent beaucoup. Il y a cependant d'assez petites différences. Pour les interpreter il faudrait tenir compte des liaisons nécessaires dont nous avons déjà parlé or, pour les quadruplets c'est un peu moins simple car il ne suffit pas de considérer les effectifs des séquences, simplement comme tout à l'heure, mais il faut s'intéresser à la différence des effectifs par groupement de séquences comme le montre la théorie. On lira dans le même numéro l'article de G.Th. Guilbaud p.99.
On a présenté dans le tableau ci-dessus les groupements et les différences. En prose on voit qu'il n'y a finalement qu'un seul renversement d'ailleurs numériquement très faible (+25, -12).

Nous donnons enfin au tableau XIV les effectifs,pour les deux auteurs et les deux types de discours, des *n*-uplets encadrés analysés dans les deux corpus.

Tableau XIV.Les séquences de n-uplets encadrées chez Puškin et Hlebnikov.

Prose.

Types de séquence	Puškin	Hlebnikov	Types de séquence	Puškin	Hlebnikov
C*VVVVV*C	0	2	V*CCCCC*V	0	1 (1)
C*VVVV*C	5	6	V*CCCC*V	53	33
C*VVV*C	96	81	V*CCC*V	297	360
C*VV*C	686	787	V*CC*V	1600	1786
C*V*C	5090	4832	V*C*V	3927	3528

Vers.

Types de séquence	Puškin	Hlebnikov	Types de séquence	Puškin	Hlebnikov
C*VVVVV*C	0	2	V*CCCCC*V	2	4
C*VVVV*C	7	7	V*CCCC*V	36	57
C*VVV*C	59	96	V*CCC*V	366	405
C*VV*C	643	660	V*CC*V	2070	1807
C*V*C	4863	4956	V*C*V	3104	3294

Ce tableau, que nous ne commenterons pas car les chiffres parlent d'eux-mêmes est très intéressant à considérer mais il ne faut pas perdre de vue qu'il contient moins d'information que les tableaux précédents.

(1) проницательным взглядом
pronicatel'nym vzgljadom
perçant regard

III.2. COMPARAISON CHEZ HLEBNIKOV ENTRE LANGAGE 'NORMAL' ET LANGAGE 'ZAOUM'.

Les doublets dans les deux types de langage.

Prose			Poésie	
Zaoum	Normal		Zaoum	Normal
210	226	VC	229	222
210	225	CV	229	222
119	109	CC	94	125
60	39	VV	47	30
599	599		599	599

L'ordre des séquences est le même pour les deux langages en prose l'alternance est un peu moins fréquente qu'en poésie et la compensation nécessaire est due plutôt aux doublets de consonnes quoi qu'il faille remarquer l'écart des doublets de voyelles en 'zaoum'.

Les triplets dans les deux types de langage.

Prose			Poésie	
Zaoum	Normal		Zaoum	Normal
162	190	CVC	195	194
115	137	VCV	145	122
94	88	VCC	84	99
94	88	CCV	83	99
48	35	VVC	34	28
48	35	CVV	34	28
25	21	CCC	10	26
12	4	VVV	13	2
598	598		598	598

Ici on doit noter la manifestation de la liaison (VCC et CCV ont mêmes effectifs à une unité près). La ressemblance entre les distributions prises deux à deux est frappante, l'écart notable porte surtout sur les triplets de voyelles, surtout en poésie.

Tableau XV . Les effectifs des séquences de graphèmes et leurs différences.

Vers			Prose	
Zaoum	Normal		Zaoum	Normal
121)1 120	110)6 104	VCVC CVCV	90)2 88	113)-4 117
74)0 74 (0) 74	90)7 83 (+9) 74	CVCC CCVC VCCV	74)3 71 (+4) 75	72)-5 77 (+7) 70
28 (3) 25)1 24	26 (8) 18)6 12	CVVC VVCV VCVV	42 (15) 27)2 25	31 (2) 19)-5 24
9 (-1) 10)1 9 (3) 6	25 (+9) 16)7 9 (7) 2	VCCC=CCCV CCVV VVCC CVVV=VVVC	19 (-4) 23)3 20 (14) 6	18 (+7) 11)-5 16 (12) 4
1	1	CCCC	6	3
7	0	VVVV	6	0

Les deux échantillons étant courts (600 graphèmes mais on sait que le modèle markovien est assez robuste pour être accepté sur des échantillons de 100 graphèmes) on observe ici une plus grande variabilité des différences entre effectifs ce qui est normal. La grande différence est locale, l'ordre restant stable et porte en prose sur les quadruplets de consonnes mais surtout dans les deux types de discours sur les séquences homogènes de voyelles.

En conclusion de ces deux études, il nous paraît important de retenir d'abord l'existence de contraintes de nature combinatoire qui font que les effectifs des différentes séquences ne sont pas indépendants. Autre fait important : les tableaux des effectifs se ressemblent. C'est tout particulièrement remarquable pour les vers et on peut se demander si la métrique en est responsable, les deux auteurs ayant des poétiques fort différentes. Enfin, quand il y a écart, comme on l'observe dans le discours en prose de Hlebnikov, cet écart est bien le résultat d'un effort conscient de l'auteur mais l'écart obtenu n'est qu'un écart local, presque ponctuel qui n'altère que peu les structures du russe écrit telles qu'elles sont appréhendées par l'étude des séquences de graphèmes du texte.

IV. EXTENSION AU FRANCAIS PAR LE MOYEN DE LA COMPARAISON DE L'ORDRE DES SEQUENCES DE GRAPHEMES.

A ce point de notre étude, il était tentant de voir ce qu'on obtenait en comparant des textes français classiques (deux fragments de 15000 graphèmes de Stendhal pris au hasard dans *Le Rouge et le Noir*, le début de *Le Côté de chez Swann* de Proust et toujours 15000 graphèmes consécutifs pris au hasard dans *La disparition* de Pérec. On sait que dans ce roman, l'auteur a volontairement évité tout emploi de la lettre é. Nous donnons les résultats au tableau XVI.

Tableau XVI. Les effectifs des séquences de graphèmes et leurs différences en français.

	Proust	Stendhal 1	Stendahl 2	Pérec
CVCV	2091	1994	2095	1509
	150	84	108	76
VCVC	1941	1910	1987	1433
VCCV	1969	2072	2047	2168
	135	195	233	148
CCVC	1834	1877	1814	2020
	150	84	108	76
CVCC	1684	1793	1706	1944
CVVC	1239	1218	1284	1324
446		499	474	671
	150	84	108	76
VCVV	793	719	810	653
VVCV	643	635	702	577
VVCC	775	750	744	879
	150	84	109	77
CCVV	625	666	635	802
446	135	498 195	473 232	671 147
CCCV=VCCC	490	471	403	655
VVVC=CVVV	179	168	162	131

Comme on pouvait s'y attendre l'ordre des effectifs est différent dans les deux langues, les alternances étant nettement plus fréquentes en russe qu'en français. Mais ce qui nous intéresse c'est la comparaison des auteurs français. Bien entendu jouent ici aussi les contraintes combinatoires : on en trouvera la théorie ci-dessous dans l'article de G.Th. Guilbaud. En français aussi on remarque la grande ressemblance des trois corpus classiques quant à l'ordre des séquences de graphèmes selon leurs

effectifs. Qu'elle soit si étroite dans le cas de Proust et Stendhal 2 est un effet d'échantillon sans doute. Par contre les écarts que fournit le texte de Pérec traduisent l'altération voulue et donc réussie par l'auteur. Malgré le procédé à la fois plus radical et surtout d'un effet plus constant au long du texte que les procédés futuristes russes, l'écart obtenu n'est pas ce à quoi nous nous attendions. En français, comme en russe, le "poids de la langue" est considérable et on voit quel genre d'effort un auteur peut faire lorsqu'il essaie d'en briser les mécanismes.

CONCLUSION.

Nous avons essayé de construire quelques exercices pour montrer que l'application directe, immédiate du modèle markovien à deux états, d'ordre 1 ou d'ordre 2 selon le type de discours semble promettre des résultats intéressants. Notre étude ne prétend pas montrer qu'un processus markovien est un "bon modèle" d'un langage naturel et ceci principalement parce que ce n'était sans doute pas l'intention de l'auteur. Mais il nous paraît que cette application, qui ne semblait qu'une commode et bonne "approximation d'un système ergodique" méritait qu'on s'y attache plus qu'on ne l'a fait jusqu'ici du point de vue linguistique. Nos résultats ne constituent pas une réponse à cette attente : ils sont en nombre beaucoup trop insuffisants d'une part et des hypothèses plus pertinentes devraient sans doute être formulées. Mais tels qu'ils sont ils nous conduisent à penser qu'il serait intéressant de réaliser soigneusement une série d'expériences en nombre suffisant pour pouvoir donner lieu à une étude statistique des résultats obtenus. Seuls les résultats d'une telle étude pourront nous permettre d'affirmer que les paramètres markoviens sont des instruments de recherche linguistique.