J.S. LIENARD

LE DICTIONNAIRE DES ELEMENTS PHONETIQUES ET SES APPLICATIONS A LA LINGUISTIQUE

SEPTEMBRE 1966 N° 22bis

G A M

BULLETIN DU GROUPE D'ACOUSTIQUE MUSICALE

Faculté des Sciences - 8 rue Cuvier - Paris 5°

LE DICTIONNAIRE DES ELEMENTS PHONETIQUES ET SES APPLICATIONS A LA LINGUISTIQUE

par J.S. LIENARD

Le discours étant formé, du point de vue acoustique, par une succession de ces " dominos " que sont les unités phonétiques, l'analyse et la synthèse de la parole rendent nécessaire l'établissement d'un dictionnaire dans lequel on puisse trouver la forme sonagraphique de chaque unité. Le dictionnaire se présente sous la forme d'un tableau carré de 28 lignes et 28 colonnes, si l'on prend comme base de l'étude les 28 phonèmes de la sténographie Duployé. Ainsi la forme RE se trouvera à l'intersection de la ligne R et de la colonne E et chacune des 784 possibilités se trouvera à une place logique.

I - LE REPERTOIRE

Etablir un dictionnaire est une tâche de longue haleine, qui commence par un recensement des matériaux à notre disposition : peut-être certaines unités n'ont-elles pas d'existence réelle; peut-être certaines sont-elles plus importantes ou plus fréquentes que les autres; peut-être la répartition change-t-elle complètement d'un texte à l'autre.... Nous avons donc cherché la fréquence d'occurence de chaque unité dans des textes variés, supposés lus à voix haute avec la diction du langage parlé usuel.

Une méthode artisanale consiste à tracer un bâton dans la case réservée à chaque unité, toutes les fois que l'on rencontre celle-ci; au bout de quelques pages de texte (une page imprimée contient de 500 à 1000 unités) le décompte des bâtons permet de connaître la répartition des unités dans le texte considéré. Comme le discours est généralement continu on n'observe pas de séparation entre les mots d'une même phrase, non plus que d'une phrase à l'autre, et une expression comme : " l'icophone est au point " se traduit par la suite : LI IK KO OF FO ON Nè èT TO OP PO OIN. D'autre part on a cherché à grouper dans le coin supérieur gauche les chiffres les plus élevés et c'est pourquoi on a adopté, au lieu de l'alphabet, l'ordre suivant, dans lequel les phonèmes apparaissent par fréquence décroissante : R E L A S T - I é D O K etc...

Ce travail de tri se révèle rapidement fastidieux heureusement il constitue un bon terrain d'application pour le calcul automatique. La CAB 500 du Conservatoire des Arts et Métiers nous a permis, grâce au programme rédigé par Mr TEIL, d'obtenir sans fatigue les tableaux relatifs à une quinzaine de textes différents.

...../

On trouvera ci-contre un texte de Stendhal frappé sur CAB 500 en symboles phonétiques, représentant les données fournies à la machine. Après un temps de calcul variable suivant la longueur du texte (le calcul se fait à mesure de l'introduction des données), la machine frappe elle-même le nombre total d'unités phonétiques rencontrées, le tableau des fréquences d'occurence en millièmes et le classement des unités par fréquence décroissante : le premier chiffre est la fréquence absolue, le second est la fréquence relative en millièmes, c'est-à-dire le rapport du 1er chiffre eu nombre total d'unités.

De toutes ces études il ressort que les principales unités phonétiques sont à peu près toujours les mêmes, et cela justifie l'établissement d'une liste-type à partir de l'ensemble des textes étudiés. Voici le début de cette liste :

rang	unité phonétique	fréquence 900
1	DE	17
2	LA	15
3	AR	14
4	IL	11
5	LE	10
6	Lè	10
7	èL	9
8	PA	9
9	EL	9
...	...	... etc..

La liste ne comprend pas toutes les unités attendues : sur les 784 possibilités, près de 200 sont restées inutilisées, et plus de 300 se rencontrent moins d'une fois pour mille. Dans un texte moyen de quelques pages, possédant une certaine unité de vocabulaire, on dénombre rarement plus de 400 unités phonétiques différentes, c'est-à-dire environ la moitié du répertoire.

Il semble donc que nous ayons là les éléments nécessaires pour construire rapidement le dictionnaire, en commençant par les formes les plus courantes. Il n'en est rien, car nous nous heurtons à une propriété de la théorie de l'information, selon laquelle les éléments les plus rares sont justement ceux qui transportent le maximum d'information. Pour s'en convaincre il suffit par exemple de chercher deux mots : l'un comprenant l'unité AR (forme courante) et l'autre comprenant l'unité SB (forme rare). Pour le premier on a l'embarras du choix; pour le second on ne trouve guère que "SBire", ou "iSBa". Autrement dit la simple forme SB équivaut à elle seule aux mots sbire ou isba, alors que l'on ne peut en dire autant de la forme AR. Finalement le dictionnaire devra être complet pour être vraiment utilisable.

..../

	R	E	L	A	S	T	=	I	,	D	O	K	N	P	M	Δ	/	U	V	Y	B	J	Z	W	1	G	X	F
R	0	1	4	8	6	3	10	8	4	8	2	2	1	1	1	7		2	0	2	0	1	1	0	0	0	0	0
E	5		9		12	3				7		2	3	6	8	0		0	3		0	0	0	1		1		1
L	1	8		12	0	0	8	5	1	1	2	1		1	1	2	0	6		0	0	0	0	1	1	0	0	2
A	12		6	0	9	6	0	0	0	3	0	4	3	7	4	0		0	4	2	2	3	2					0
S		6	0	3	1	2	7	4	2	2	3	2		2	0	3	2	2		8				1	0			
T	6	5		9	0		5	6	6		3			0	0	2	0	1	0	0				2	1			
=	18		7		5	12		1		3		1	3	2	1				2	0	0	2	4			1		
I	3		12	2	7	3	0		0	6	0	5	3	0	2	0		0	1	4	1	0	2			0	0	1
,	3		2	1	0	4		0		3	0	3	0	3	1	1	0				0	0		0	0	0	0	1
D	1	22		1			4	3	2		2				6	3	1	4						1	2			0
O	9		1	6	2	0				2			1	4	4				1			2	2		0			
K	1	4	0	2	0	0	0	7	1	0	4	0				0	4	1		0		0		0	0			0
N		2	0	4		0	0	2	2	0	0			0	0	1	0	2		3			0			0		0
P	5	1	5	7	0	0	4	1	0		3			0		1		3		0				2	0			
M		4		4		0	6	2	0	1	3					2	0	2										
Δ			3	0	6	5		0	0	3		2	1	3					0		1	1						2
/			1	0	1	3	0		0	2		1	1	1		0			0		0							1
U	5		1		1	0		5		2	1	0	1	1	0				0		2	0	2					1
V	0	0		0			4	2	1	0	2					1		0		0				1	0			
Y		6	0	1			6	0	2	0	0	0				0	4								1			
B	0		2	1		1	1	0	0			2						0		0				0				
J		2		2				1	0	0						1		0					0	1				
Z		0	0	2		0	0	2	1	0	0	1		0		2	2	1		0					1			
W	6				0	0	0	1	0	0	0		0	0					1		0	0	1					
1			1	0	0	2		0		2		0		2	0			0			0	0	0			0		
G	2		0	0												0		0					0					
X		0		0			0	0	0	0	0	1				0												
F	2	0		1			3	1			1							0		0								

DE	57	22
=R	46	18
LA	32	12
=T	32	12
ES	31	12
IL	31	12
AR	30	12
R=	26	10
AS	24	9
TA	24	9
EL	23	9
OR	23	9
RA	22	8
RI	22	8
LE	22	8
EM	21	8
L=	21	8
RD	20	8
SY	20	8
RΔ	19	7
AF	19	7
PA	19	7
ED	18	7

=	...è
,	....é
Δ	...an
/	...on
Y	...i mouille
W	...ou
1	...in
G	...g dur
X	...ch

STENDHAL LE ROUGE ET LE NOIR CHAPITRE 1 DEUXIEME PARTIE

T=L = LE M=R DE V=RY=R MESYE DE R=NAL
APR=Z AVOAR TRAV=RS, LA RU D 1 PA GRAV IL ΔTR A LA M=RI , DISPAR=T O Z YE DU VOAYAJER M= SΔ PA PLU O SI SELUI SI K/TINU SA PROMENAD IL AP=RSOAT UN M=Z/ D AS, B=L APARΔS , A TRAV=R Z UN GRIY DE F=R ATENΔT A LA M=Z/ D= JARD1 MANYIFIK
O DELA S =T UNE LINYE D ORIZ/ FORM, PAR L= KOLIN DE LA BWRGONY , KI SΔBLE F=T A SW= PWR L PL=ZIR D=Z YE
S=IE VU F=T WBLIY, O VOAYAJER L ATMOSF=R ΔP=ST, D= PETI Z 1T,R= D ARJΔ D/ T IL KOMΔS A =TR ΔP=ST,

/ LUI APRΔ KE S=T M=Z/ APARTY1T A MESYE DE R=NAL
S = T O B,N,FIS K IL A F= SUR SA GRΔD FABRIK DE KLW KE LE M=R DE V=RY=R DOA S=IE B=L ABITASY/ Δ FY=R DE TAY K IL AX=V Δ SE MOMΔ
SA FAMIY DIT / = T =SPANYOL ΔTIK , A SE K / PR,TΔ ,TABLI DΔ LE P=I BY1 AVΔ LA K/K=IE DE LWI KATORZE

DEPUI MIL UI SΔ K1Z IL RWJI D =IR 1DUSTRIY=L MIL UI SΔ K1Z L A F= M=R DE V=RY=R L= MUR Z ΔT=RAS KI SWTY=N L= DIV=RSE PARTI DE SE MANYIFIK JARD1 KI D ,TAJ Δ N ,TAJ D=SΔ JUSK O DW S/T OSI LA R,K/PΔS DE LA SYΔS DE MESYE DE R=NAL DΔ LE KOM=RSE DU F=R

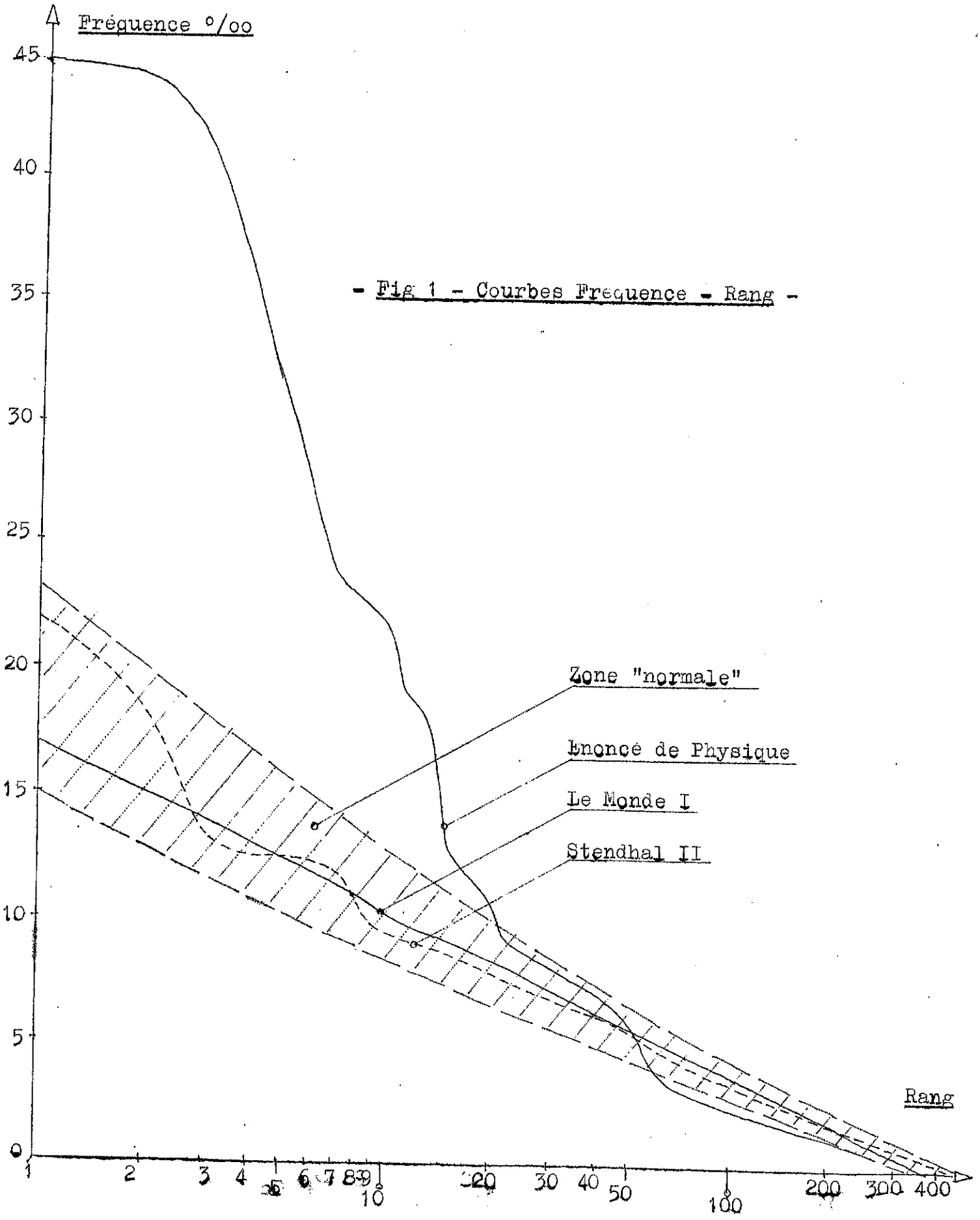
NE VW Z ATΔD, POI T A TRWV, Δ FRΔS S= JARD1 PITOR=SKE KI ΔIWR L= VIL MANUFAKTURY=R DE ALEMANY LAIPSIK FRΔKFOR NUR1B=R =IS=T,RA
Δ FRΔX K/T, PLUZ / BATI DE MUR PLUZ / ,RIS SA PROPRIY,T, DE PY=R RΔJ, L=Z UNZ O DESU D=Z OIREPLUZ / AKY=R DE DROAZ O R=SP= DE S= VOAZ1
L= JARD1 DE MESYE DE R=NAL RΔPLI DE MUR S/T ΔKOR ADMIR, PARSE K IL A AXET, O POA DE L OR S=RTL PETI MORSO DE T=R1 K IlZ OKUP
PAR =GZΔPLE S=TE SI A BOA D/ LA POZISY/ S1 GULY=R SUR LA RIV DU DW VWZ A FRAP, Δ N ΔIRΔT A V=RY=R , W VWZ AV, REMARK, LE N/ DE SOR=L ,KRI T Δ KARAKT=R JIGΔT=SKE SUR UN PLΔX KI DOMIN LE TOA =L OKUP= IL I Y A SIZ Δ L =SPAS SUR LEK=L / N ,L=V ΔSE MOMΔ LE MUR DE LA KATRIY=M T=RAS D= JARD1 DE MESYE DE R=NAL

MALGR, SA FY=RT, MESYE LE M=R A DU F=R BY1 D= D,MARX OPR= DU VYE SOR=L P=IZΔ DUR , ΔT=T, IL A DU LUI K/T, DE BO LWI D OR PWR OBTENIR K IL TRΔSPORTA S/ N UZIN AYER KΔT O RUISO PUBLIK KI FEZ=T AL, LA SI MESYE DE R=NAL O MOAY1 DU KR,DI D/T IL JWIT A PARI AOBTENU K IL FU D,TWRN,
S=TE GRAS LUI V1T APR= L=Z ,L=KSY/ DE MIL UI SΔ V1

IL A DON, A SOR=L KATR ARPΔ PWR 1 A S1 SΔ PA PLU BA SUR L= BOR DU DW ,KOAKE S=T POZISY/ FU BOKW PLUZ AVΔTAJEZ PWR S/ KOM=RS DE PLΔX DE SAP1 LE P=R SOR=L KOM K IL = RIX A U LE SEKR= D OBTENIR DE L 1PASYΔS , DE LA MANI DE PROPRIY,T=R KI ANIM= S/ VOAZ1 UNE SOM DE UI MIL FRΔ

IL = VR= KE S=T ARΔJEMΔ A ,T, KRITIK, PAR L= BON T=IE DE L ΔDROA
UN FOA S,T=T 1 JWR DE DIMΔX ILI Y A KATR Δ DE SELA MESYE DE R=NA L REVENΔ DE L ,ELIZE Δ DE TUM DE M=R VI DE LO1 LE VYE SOR=L ΔIWR, DE S= TROA FIS SWRIR Δ LE REGARDΔ
SE SWRIR A PORT, 1 JWR FATAL DΔ L AM DE MESYE LE M=R IL PΔS DEPUI LOR K IL U PU OBTENIR L ,XΔJ A M=YER MARX,

PWR ARIV,R A LA K/SID,RASY/ PUBLIK A V=RY=R L =SΔSY=L = DE NE PAZ ADOPT, TWT Δ BATISΔ BOKW DE MUR KELKE PIΔ APORT, D ITALI PAR S= MAS/ KI O PR1 TΔ TRAV=RSE L= GORJE DU JURA PWR GANY, PARI
UNE T=L INOVASY/ VODR=T A L 1PRUDΔ BATISER UN ,T=RN=L R,PUTASY/ DE MOV=Z T=T , IL SER=T A JAM= P=RDU OPR= D= JΔ SAJZ , MOD,R, KI DISTRIBU LA K/SID,RASY/ Δ FRΔX K/T,



- Fig 1 - Courbes Fréquence - Rang -

Mais l'étude des tableaux de répartition nous a montré une curieuse homogénéité dans les résultats relatifs à des textes aussi divers que "le Rouge et le Noir ", " Zazie dans le métro ", " Phèdre " etc.... Nous avons donc essayé de comprendre à quoi correspond la structure du répertoire, au moyen de deux outils : les courbes fréquence d'occurence en fonction du rang, et le coefficient de corrélation entre les tableaux de répartition.

II - LES COURBES FREQUENCE-RANG

Si, dans une des classifications précédentes, nous ne retenons que l'énoncé suivant : l'unité numéro 1 a pour fréquence relative 17 pour mille, l'unité n° 2 à 15 °/oo, l'unité n° 3 a 14 °/oo , ... l'unité de rang R a pour fréquence F °/oo etc.... nous pouvons porter sur un graphique le rang en abscisse, la fréquence en ordonnée, et construire une courbe en reliant tous les points obtenus. Cette courbe est décroissante par construction, puisque les unités phonétiques sont classées par fréquence décroissante. De plus l'utilisation d'une abscisse logarithmique, dans laquelle la distance entre 1 et 10 est la même qu'entre 10 et 100, montre que pour la plupart des textes cette courbe devient une droite à partir du rang 50 environ (fig.1). Pour certains textes, en particulier pour un éditorial du journal " Le Monde ", la répartition est parfaitement régulière du début jusqu'à la fin, et obéit à la relation $F = A - B \log R$, où A et B sont des coefficients constants. Les autres textes s'écartent légèrement de cette droite, mais jamais dans des proportions importantes : ils restent tous dans une zone relativement restreinte, à condition d'avoir une longueur d'au moins 1000 unités phonétiques. Hors de cette zone nous avons un " langage monstre ", absolument déséquilibré, c'est le cas d'un énoncé de physique comprenant un grand nombre d'équations différentielles composées de termes en $\frac{dx}{dy}$ (phonétiquement DéIKSUR DéIGRèK). La fréquence des premiers rangs est donc très élevée (40 °/oo), mais elle tombe au-dessous de la moyenne au delà du rang 50.

Nous avons donc un moyen, avec ce graphique, de savoir si un texte est normal quant à la répartition de ses unités phonétiques. Mais on peut se demander si les différences entre les diverses courbes, à l'intérieur de la zone " normale ", sont imputables aux différences de longueur, de genre ou de style entre les textes étudiés, ou à d'autres causes indéterminées. C'est pourquoi nous avons fait les essais suivants de caractérisation des textes.

Les deux courbes de la fig.2 sont relatives à deux parties du 1er chapitre du " Rouge et le Noir ". De longueurs différentes (1700 et 2500 unités), ces extraits présentent une quasi-identité de style et de sujet. Les courbes sont d'allure identique : la pente à l'origine, jusqu'au rang 10, est environ deux fois plus grande que par la suite; l'écart moyen est de l'ordre de 0,5 °/oo, n'atteignant 2 °/oo que dans les premiers rangs; en-

...../

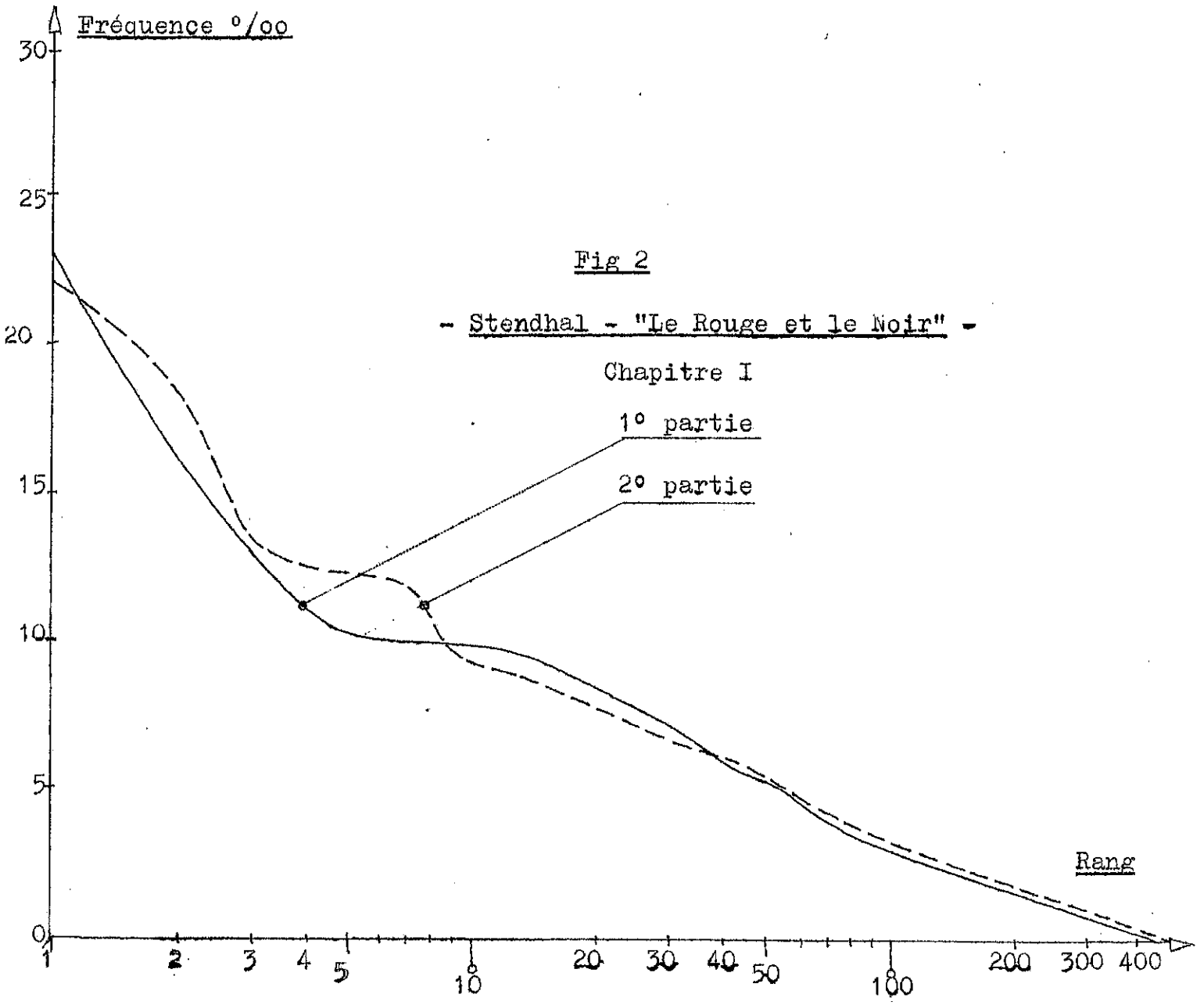


Fig 2

- Stendhal - "Le Rouge et le Noir" -

Chapitre I

Fig 3

- Maupassant - "Boule de Suif" -

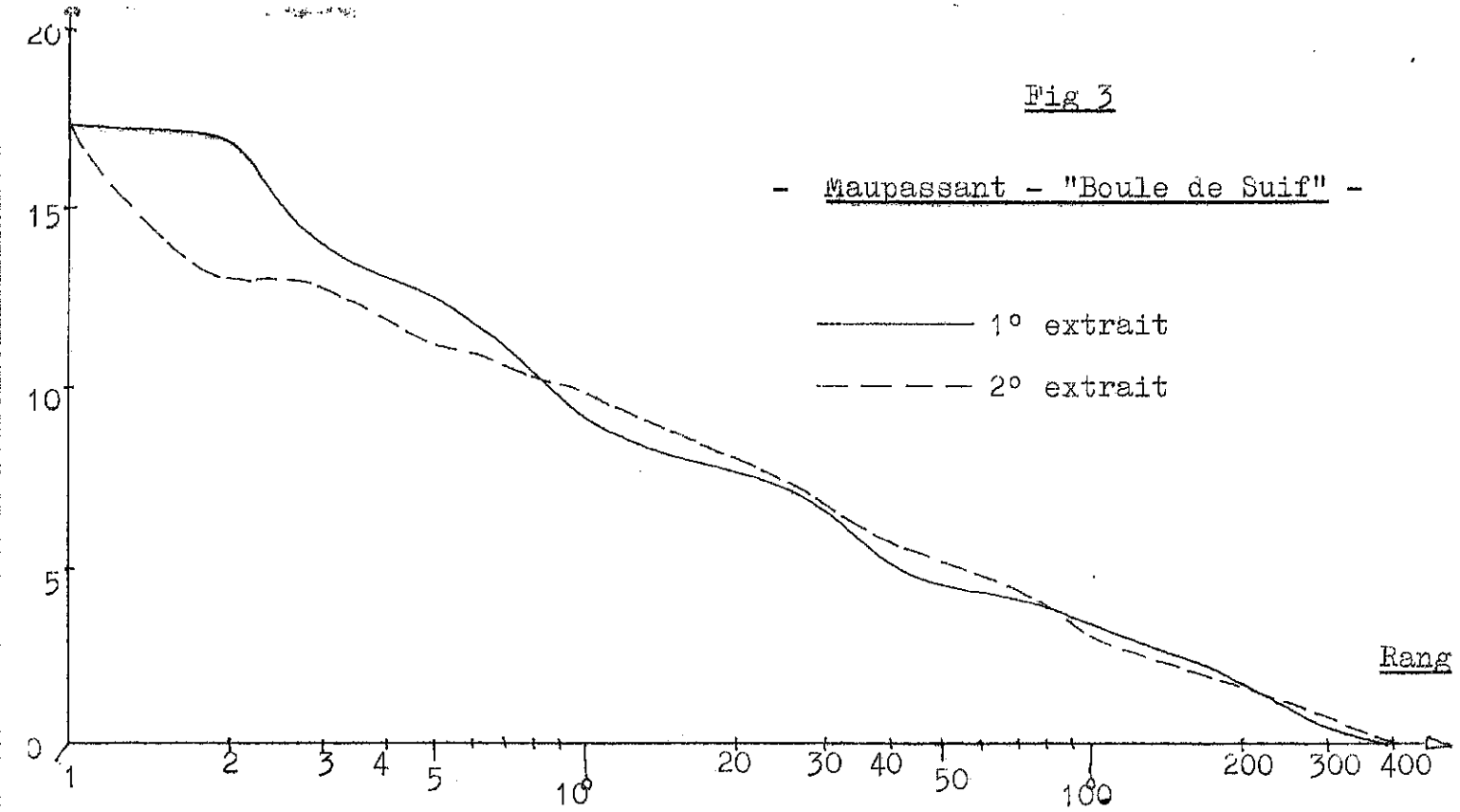
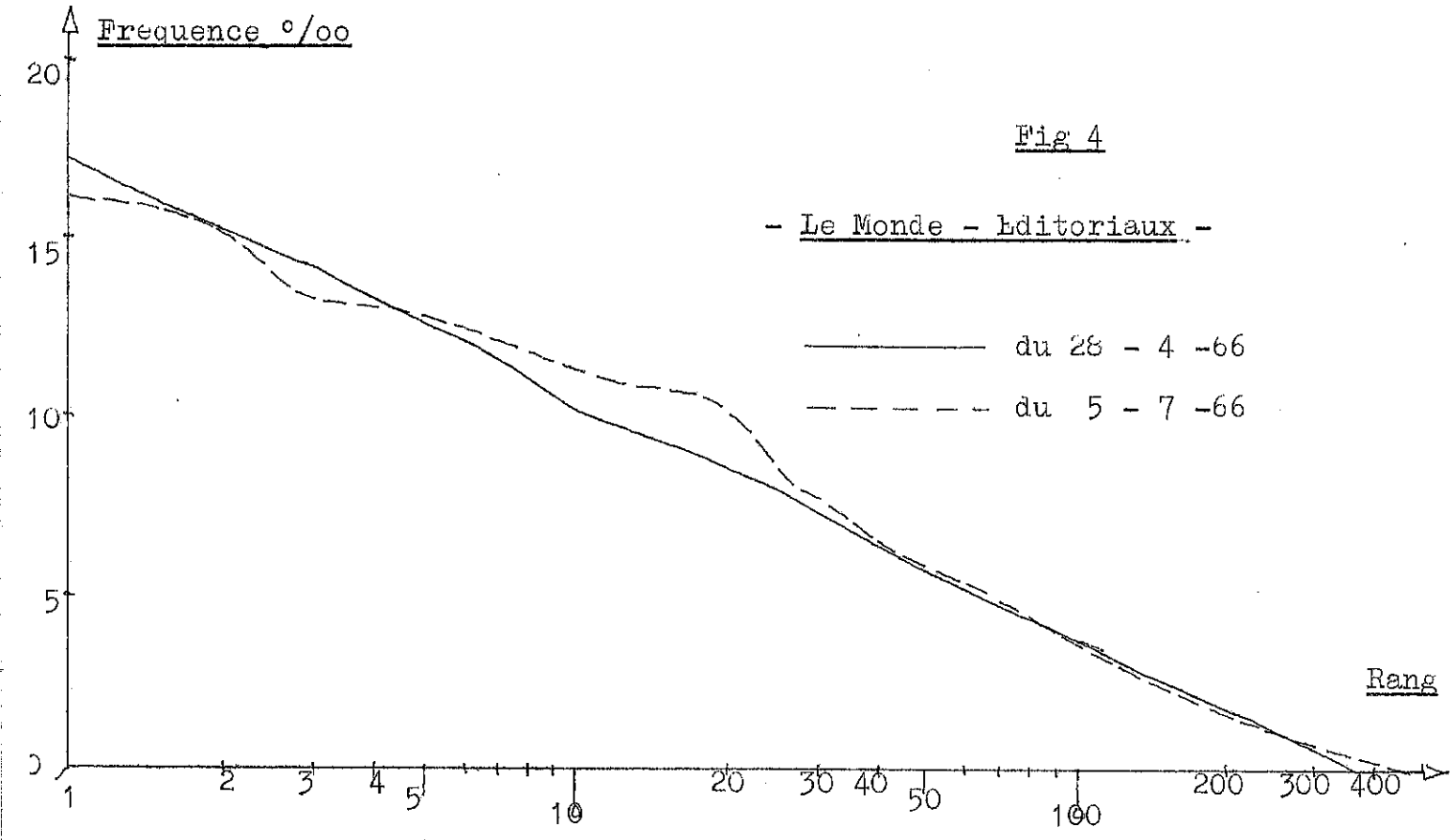


Fig 4

- Le Monde - Editoriaux -



fin les points de départ (22 et 23 °/oo) sont très proches, de même que les points d'arrivée (rangs 418 et 434).

La comparaison porte ensuite (fig.3) sur deux extraits de " Boule de Suif " de Maupassant, de longueurs comparables (1500 et 1700 unités), prélevés dans des chapitres différents de l'ouvrage. Les courbes sont très semblables dans leur allure générale (départ 17 °/oo, arrivée aux rangs 378 et 392) et plus linéaires que les précédentes, malgré un écart important au rang 2.

Les deux derniers textes (fig.4) ne présentent qu'une parenté de genre ; il s'agit de deux éditoriaux du journal " Le Monde " de longueurs 2000 et 2400 unités, portant sur des sujets différents. La première courbe est parfaitement linéaire; la seconde est beaucoup moins régulière, en particulier du rang 10 au rang 25, et les rangs d'arrivée sont différents (364 et 440). Pourtant du rang 30 au rang 300 ces courbes sont pratiquement confondues, et dans l'ensemble elles ont, entre elles, une parenté plus grande qu'avec une quelconque des courbes précédentes.

Ces comparaisons sont par trop sommaires pour permettre d'affirmer que la caractérisation d'un texte est possible à partir de la distribution de ses unités phonétiques; en tous cas, l'hypothèse est certes pas à rejeter. Une étude en cours nous permettra de préciser ses limites.

Il n'en reste pas moins que toutes les courbes présentent globalement une grande similitude, dont l'origine est à chercher vraisemblablement dans l'économie d'effort réalisée par le système de phonation dans l'exercice d'une langue donnée. Nous allons retrouver dans le paragraphe suivant, en utilisant un autre moyen d'approche, cette même idée d'une norme générale dont s'écarte chaque texte d'une manière caractéristique.

III - LE COEFFICIENT DE CORRELATION

La classification fréquence - rang ne tient compte que de la répartition numérique des unités phonétiques, quelles sue soient celles-ci; c'est-à-dire que dans cette classification les deux listes suivantes seront identiques :

rang.....		1	2	3	4	5	6
liste I	unité	DE	LA	AR	èR	IL	LE
	fréquence..	18	16	14	13	12	11
liste II	unité	LA	èR	DE	LE	AR	IL
	fréquence..	18	16	14	13	12	11

et pourtant elles sont différentes, puisque chaque unité phonétique se trouve à un rang différent dans l'une et dans l'autre.

..../

TABLEAU I

CORRELATION DE DIVERS TEXTES AVEC LA NORME

TEXTES	Coefficient	Longueur
Stendhal I		
- Le Rouge et le Noir, chap.I, 1° partie...	0,93	1700
Le Monde I		
- Editorial du 28/4/66.....	0,93	2000
Stendhal II		
- Le Rouge et le Noir, chap. I, 2° partie..	0,92	2500
Maupassant I		
- Boule de Suif, p.5 et 6	0,91	1500
Maupassant II		
- Boule de Suif, p. 22 et 23	0,91	1700
Eluard I		
- Et notre mouvement.....	0,88	600
Le Monde II		
- Editorial du 5/7/66	0,87	2400
Eluard II		
- Grandeur d'hier et d'aujourd'hui	0,85	1000
Racine		
- " Phèdre ", acte I, scène I	0,84	1500
Vian		
- L'écume des Jours, début chap. I	0,80	1000
Queneau		
-Zazio dans le métro, début chap. I	0,79	2400
Rimbaud		
- Les chercheuses de Poux.....	0,70	600
Hugo		
- Souvenir de la nuit du 4 Août.....	0,66	1000
Enoncé de Physique	0,41	800

Pour tenir compte du caractère individuel de chaque distribution nous pouvons calculer un coefficient de corrélation, ou de ressemblance entre deux tableaux. Ce coefficient aura la valeur 1 si les deux tableaux sont absolument identiques, et la valeur 0 s'ils n'ont entre eux aucune ressemblance. Ainsi, à titre d'exemple, les deux listes précédentes ont un coefficient de corrélation de 0,44.

Avant de mettre en machine le programme de calcul relatif à l'ensemble des textes, nous avons cherché à obtenir manuellement des ordres de grandeur et les résultats qui suivent appellent certaines réserves du point de vue mathématique : les distributions ne sont pas gaussiennes, le nombre d'échantillons considérés dans chaque tableau (25) est faible, les textes sont trop courts etc... Mais leur concordance avec les résultats précédents nous a incités à les présenter ici en attendant les conclusions de l'étude approfondie évoquée plus haut.

Nous avons d'abord cherché à tester la valeur de la méthode en l'appliquant à des textes comparables deux à deux : entre les deux tableaux relatifs au 1er chapitre du " Rouge et le Noir " le coefficient de corrélation est de 0,91, donc très élevé, comme prévu. Entre les deux extraits de " Boule de Suif " il est de 0,85 et de 0,80 entre les deux éditoriaux du " Monde ", ce qui recoupe parfaitement les conclusions du paragraphe II.

Il était tentant de classer nos textes suivant la plus ou moins grande ressemblance de leur distribution phonétique avec celle d'une moyenne ou norme, ou liste-type établie sur un ensemble de textes. Le tableau donne un tel classement par rapport à la liste type dont le début figure au paragraphe I; la longueur approximative des textes est indiquée en unités phonétiques, et ceux ayant servi à établir la liste type sont soulignés.

Comme on pouvait s'y attendre, les deux extraits du " Rouge et le Noir " ont des coefficients semblables, de même que les extraits de " Boule de Suif "; les éditoriaux du " Monde " sont comparables.

Un énoncé de physique, lu à vois haute, présente une distribution phonétique tout à fait anormale; son faible coefficient de corrélation avec la norme justifie la désignation de " langage-monstre " et sert de point de comparaison pour les autres textes.

Le fait que les 6 textes utilisés pour établir la norme s'échelonnent régulièrement entre 0,66 et 0,93 malgré leurs différences de longueur, et que les autres textes s'insèrent dans ce classement montre que tous les textes ont une sorte de fonds commun et justifie l'établissement d'une norme : chaque texte s'en écarte plus ou moins, mais dans une direction qui lui est propre; il est toujours plus proche de la norme que d'aucun autre texte. Qu'on en juge par le tableau II donnant le coefficient de corrélation de divers textes pris deux à deux.

..../

TABLEAU II

CORRELATION DE DIVERS TEXTES ENTRE EUX

	Coefficient
Stendhal I et II	0,91
Maupassant I et II	0,85
Le Monde I et II	0,80
Eluard I et II	0,75
Queneau et Vian	0,67
Queneau et Stendhal I	0,62
Queneau et Rimbaud	0,59
Rimbaud et Eluard II	0,47