La théorie de l'information vingt ans après Guiraud

Étienne Brunet

La théorie de l'information vingt ans après Guiraud

Etienne Brunet

Tous les raffinements mathématiques dont on peut entourer le rapport *N / V* ne peuvent dissoudre l'ambiguïté initiale qui s'attache à la notion de richesse lexicale[1]. Il ne suffit pas de réduire l'influence de la longueur des textes comparés. Encore faudrait-il s'attarder à la qualité, concrète ou abstraite, des vocables rencontrés et ne pas confondre l'abondance des objets et l'abondance des mots. En outre, en s'en tenant à la seule structure lexicale, un même rapport *N / V* peut être obtenu de bien des façons différentes, suivant qu'un texte privilégie les fréquences moyennes ou extrêmes. Enfin ce rapport peut varier dans le déroulement d'un texte et ce mouvement peut être intéressant à observer.

C'est pourquoi on peut penser que pour rendre compte de la diversité lexicale d'un texte il convient non seulement d'explorer toutes les fréquences – ce qu'on fait partiellement dans la loi binomiale et dans l'indice de Yule-Herdan –, mais aussi de suivre la progression lexicale de ce texte à des intervalles réguliers. Une recherche dans cette direction a été menée par Étienne Évrard sur un corpus d'auteurs latins et exposée au 11e Colloque de l'ALLC, à Louvain. Évrard a songé à utiliser ici la

[1] Ce rapport n'est jamais brut car il serait trop dépendant de l'étendue du texte (c'est-à-dire du nombre d'occurrences N, qui croît plus vite que celui des vocables V). Diverses pondérations ont été apportées depuis que Guiraud a proposé la première :

$V = \sqrt{N}$ = constante 22

Voir P. Guiraud, *Problèmes et méthodes de la statistique linguistique*, p. 89. Parmi ces indices de richesse lexicale, on citera, parmi bien d'autres, l'indice Uber de D. Dugast, celui de Rubet, celui de Yule-Herdan, et le nôtre ou indice *w*. Toutes les formules proposées sont, à l'image de celle de Guiraud, des approximations expérimentales. Une seule, prônée par Ch. Muller, procède par raisonnement et prend appui sur une loi générale (formule binomiale).

notion d'entropie[2] qui évoluerait entre deux limites : la valeur log *N*, lorsque la diversité lexicale est à son maximum (lorsque toutes les occurrences appartiennent à des vocables différents) et la valeur zéro (lorsque le texte n'a qu'un seul vocable). De la forme initiale de l'entropie :

$$H = -\sum n_i \; p_i \; log(p_i)$$

(où *n* désigne l'effectif d'une classe de fréquence x_i, et p_i la probabilité attachée à un mot de fréquence x_i : $p_i = x_i / N$), Évrard aboutit à la formule :

$$H = \; log(N) - \; 1/N \sum n_i \; x_i \; log(x_i)$$

Prenons par exemple les 20 premières pages de la *Faute de l'abbé Mouret*. L'on y trouve 982 vocables de fréquence 1, 285 pour *f* = 2, 114 pour *f* = 3, 76 pour *f* = 4, etc. Le cumul de toutes les classes jusqu'à la plus élevée (*f* = 74) aboutit à un total de 4 309 occurrences dont le logarithme (décimal) est de 3,6344. Le terme $n_i \; x_i . \; \log(x_i)$ vaut : 2 789,22 et pondéré par *N* : 0,6473. L'entropie a donc pour valeur :

$$H = \; 3{,}6344 \; - 0{,}6473 = 2{,}9871$$

On a procédé à de tels calculs non seulement pour les 20 premières pages du texte en question, mais aussi pour le roman entier, saisi à intervalles réguliers, de 10 en 10 pages, et même pour tous les textes qui constituent le cycle des *Rougon-Macquart* (en y adjoignant deux romans antérieurs : *Thérèse Raquin* et *Madeleine Férat*). Comme on ne saurait restituer ici les 200 000 cases du tableau des effectifs, on s'est contenté de reproduire la progression de *N* dans le tableau 1 et celle de *V* dans le tableau 2. Avant d'aborder les résultats consignés dans le tableau 5, il convient d'avertir le lecteur que les calculs sont considérables puisqu'à chaque pas, pour chaque texte, il faut réordonner tout le tableau de distribution des fréquences. Imaginons le discours comme un sablier dont le contenu se répand. La présente expérimentation consiste à boucher le sas à intervalles réguliers et à examiner la position de chaque grain de sable dans la pyramide qui s'est constituée et dont la base représente la donnée *V*, le volume la donnée *N* et la hauteur la fréquence la plus élevée. Un même volume peut être obtenu avec une base plus ou moins large et

[2] Là encore Guiraud a été le premier linguiste en France à utiliser la notion d'entropie et à mettre en œuvre la théorie de l'information. Voir *Problèmes*... p. 79.

une hauteur plus ou moins grande. Chaque colonne de grains empilés représente une classe de fréquence, les hapax étant les grains libres qui sont répandus à la périphérie et qui n'en supportent (provisoirement) aucun autre. Or il existe bien des formes de pyramides, de la plus plate à la plus aiguë, et l'entropie mesure précisément la forme de cette pyramide et principalement le rapport entre la surface de base et la hauteur.

Mais le sable est mouvant et la pyramide changeante. Généralement aplatie au départ, elle prend ensuite une forme plus élancée et Étienne Évrard tente d'imposer une mesure à cette évolution. Avant de suivre ses pas, nous avons pris quelques précautions du côté des hautes fréquences dont l'influence est prépondérante sur l'entropie comme elle l'est sur l'indice de Yule-Herdan. Observons en effet ce qui se passe dans le cas du *Docteur Pascal* (tableau 3). La sommation $S = n_i\ x_i.\ \log(x_i)$ atteint un maximum de 105 002 quand le texte touche à sa fin, à la 310e page, et que le total des occurrences est de 65 620. Si l'on prête attention, on remarque qu'il manque à l'appel 54 391 occurrences, puisque l'étendue de ce texte est en réalité de 120 011. C'est que nous avons écarté volontairement une trentaine de mots très fréquents dont la masse eût troublé les mesures de l'entropie[3]. Prenons l'exemple de la préposition DE qui a 6 663 emplois dans le *Docteur Pascal*. La part de ce seul mot dans la sommation eût été écrasante soit :

$$6\,663\ \times \log 6\,663 = 25\,477$$

c'est-à-dire le quart du total obtenu. Même en écartant ces mots encombrants, le calcul reste sensible aux mots fréquents. Ceux qui ont plus de 100 occurrences (il y en a 111 dans le texte considéré, sans compter les 36 unités écartées) représentent la moitié du total : 53 086/105 002 = 0,51[4].

[3] Voici 1a liste des mots écartés : A, AVOIR, CE, DANS, DE, DES (et DES), DU (et DU), ELLE, EN, ET, ETRE, IL, ILS, JE, L', LA (et LA), LE, LES, LUI, NE (et NE), PAS, PLUS, POUR, QUE, QUI, SA, SE, SES, SON. UN, NE, VOUS. L'ensemble de ces 36 vocables représente 44% des occurrences du corpus.

[4] Bien entendu les fréquences supérieures à 100 entrent dans le calcul avec leurs valeurs et leurs effectifs propres. Leur influence sur le terme $n_i\ x_i.\ \log(x_i)$ croît avec l'étendue :

Pages	10	50	100	150	200	250	310
Influence	0%	2%	27%	33%	40%	42%	51%

Tableau 1 : Les occurrences de 10 en 10 pages

	RAQ	PER	POR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEB	PAS
10	1729	1633	2159	2251	2175	1976	2178	1978	1985	1984	2021	2078	2121	2076	2107	2115	2102	2258	2206	2076	2061	2141
20	3326	3389	4413	4517	4374	3997	4309	4059	4176	3943	4135	4088	4194	4227	4255	4182	4278	4561	4357	4233	4127	4322
30	4970	5366	6624	6555	6673	6121	6396	6052	6293	6072	6219	6157	6377	6302	6437	6391	6401	6771	6602	6350	6253	6529
40	6688	7121	8869	8746	8847	8190	8528	8062	8467	8005	8249	8167	8540	8483	8586	8658	8532	8979	8895	8546	8436	8737
50	8309	9165	11065	10964	11112	10276	10665	10154	10671	10118	10261	10178	10586	10659	10675	10848	10643	11196	11157	10781	10596	10811
60	9775	10692	13250	13159	13315	12430	12836	12212	12883	12005	12296	12192	12502	12769	12784	13017	12655	13314	13282	12971	12732	12911
70	11500	12714	15502	15313	15435	14496	14954	14185	15129	14056	14252	14330	14675	14902	14827	15109	14928	15520	15427	15171	14916	14962
80	12983	14401	17708	17503	17533	16632	17261	16311	17338	16051	16286	16341	16750	17085	16929	17204	17034	17578	17437	17194	17004	17206
90	14353	16502	19776	19749	19761	18759	19561	18466	19470	18084	18385	18312	18878	19129	19041	19416	19169	19637	19606	19402	19101	19375
100	15993	18641	21870	21975	21940	20903	21873	20540	21692	20210	20464	20373	20966	21276	21133	21614	21334	21756	21682	21515	21281	21588
110	17744	20260	23906	24166	24132	23000	23971	22682	23916	22373	22476	22422	22930	23353	23297	23796	23400	23892	23651	23642	23343	23799
120	19595	22356	26022	26359	26253	25088	26178	24739	26102	24264	24587	24426	25035	25447	25304	25962	25406	25944	25816	25747	25463	25975
130	21039	23823	28196	28495	28523	27146	28520	26726	28327	26335	26768	26436	27201	27515	27386	28121	27563	28065	27870	27884	27533	28007
140	22993	25995	30402	30652	30675	29254	30771	28796	30514	28380	28899	28478	29295	29650	29483	30266	29664	30162	30102	30033	29530	30012
150	24527	28181	32551	32842	32780	31420	32982	30837	32774	30461	31026	30455	31354	31671	31657	32490	31830	32200	32399	32095	31595	32330
160	26230	29987	34678	35054	34936	33375	35211	32873	34964	32517	33120	32519	33403	33717	33715	34648	33955	34335	34555	34247	33641	34586
170	28134	32018	36908	37379	37236	35324	37384	34901	37161	34586	35312	34697	35499	35798	35802	36774	36138	36422	36760	36335	35525	36776
180	29816	33829	39137	39636	39586	37441	39675	37063	39348	36485	37433	36763	37645	37930	37786	38971	38225	38554	38915	38486	37664	38924
190	31395	36056	41416	41747	41806	39536	41849	39126	41492	38469	39711	38756	39749	40140	39802	41093	40335	38554	41132	40609	39694	41105
200	33206	37979	43754	43748	44095	41659	44029	41190	43673	40588	41775	40704	41777	42263	41908	43321	42427	0	43310	42728	41782	43203
210	34799	39993	46011	45742	46241	43715	46015	43179	45890	42521	43973	42758	43921	44385	43995	46578	44566	0	46628	44792	43913	45377
220	36587	42100	48128	47856	48473	45689	48078	45106	47856	44545	46228	44801	45948	46488	46012	47731	46628	0	47793	46934	45979	47582
230	37866	43955	50263	49839	50656	47733	50259	47087	49983	46758	48269	46880	48087	48666	48047	50049	48686	0	49963	49106	48132	49763
240	0	45999	52397	51958	52766	49684	52384	49078	52127	48794	50370	49069	50215	50773	50199	52276	50799	0	52122	51252	50257	51890
250	0	47696	54527	54014	54913	51701	54619	51224	54298	50888	52438	51153	52249	52970	52309	54447	52762	0	54270	53424	52442	54040
260	0	49763	56609	56206	57031	53674	57006	53326	56480	52849	54599	53265	54317	55141	54274	56753	54874	0	56496	55529	54678	56203
270	0	51797	58787	58361	59197	55765	59259	55364	58603	54900	56739	55389	56351	57307	56416	58993	57006	0	58692	57615	56789	58297
280	0	53528	60920	60593	61227	57818	61507	57457	60852	56949	58829	57449	58459	59309	58577	61254	59295	0	60925	59743	58882	60400
290	0	55619	63170	60778	63348	59876	63751	59548	62947	59142	60891	59500	60466	61367	60704	63502	61515	0	63184	61808	61057	62553
300	0	0	65230	0	63921	62044	65970	61616	65087	59546	62955	61530	62435	63530	62806	65626	63723	0	65449	63929	63151	64827
310	0	0	67390	0	0	64114	68113	63641	67262	0	65018	63557	64441	65684	64905	67811	65872	0	67569	66101	65277	65620
320	0	0	67902	0	0	64869	68951	65825	69544	0	67158	65653	66489	67863	66935	69861	67923	0	69743	68187	67465	0
330	0	0	0	0	0	0	0	67879	71809	0	69265	67665	68674	68722	68909	72003	70015	0	71907	70308	69529	0
340	0	0	0	0	0	0	0	69945	74011	0	71444	69695	70824	0	71052	74346	72128	0	73002	72595	71772	0
350	0	0	0	0	0	0	0	72039	76187	0	73610	71785	72886	0	73165	76651	74082	0	0	74784	73981	0
360	0	0	0	0	0	0	0	73468	78417	0	75788	73842	75099	0	75341	77252	76233	0	0	76926	76196	0
370	0	0	0	0	0	0	0	0	80610	0	77975	75970	77192	0	77499	0	78464	0	0	79062	78385	0
380	0	0	0	0	0	0	0	0	82834	0	80084	78008	79374	0	79484	0	80605	0	0	81234	80516	0
390	0	0	0	0	0	0	0	0	85039	0	82141	78642	81554	0	81631	0	82719	0	0	83015	82647	0
400	0	0	0	0	0	0	0	0	87344	0	82310	0	83678	0	83742	0	84864	0	0	0	84810	0
410	0	0	0	0	0	0	0	0	89488	0	0	0	85790	0	85857	0	86951	0	0	0	86991	0
420	0	0	0	0	0	0	0	0	91706	0	0	0	86645	0	88075	0	88991	0	0	0	89033	0
430	0	0	0	0	0	0	0	0	92591	0	0	0	0	0	90157	0	91128	0	0	0	91093	0
440	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92292	0	93316	0	0	0	93249	0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	94355	0	94299	0	0	0	95370	0
460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96246	0	0	0	0	0	97444	0
470	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99657	0
480	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101807	0
490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103968	0
500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	106170	0
510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108365	0
520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	108657	0

Tableau 2 : Les vocables de 10 en 10 pages

	RAQ	PER	POR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEB	PAS
10	906	890	1073	1083	996	860	1041	969	994	903	988	975	1025	1044	982	1127	986	1141	1078	1073	1051	1061
20	1368	1331	1656	1730	1552	1349	1674	1559	1547	1413	1583	1540	1609	1633	1601	1701	1604	1923	1619	1728	1654	1605
30	1763	1833	2096	2229	2074	1731	2078	2045	1941	1832	2093	2016	2060	2060	2049	2333	2064	2457	2084	2275	2118	2179
40	2094	2161	2557	2657	2541	2060	2411	2421	2300	2161	2432	2377	2460	2485	2433	2775	2575	2931	2432	2738	2524	2573
50	2316	2503	2958	3117	2938	2269	2745	2725	2679	2522	2710	2701	2811	2848	2731	3161	2965	3275	2763	3181	2952	2924
60	2517	2705	3362	3471	3272	2583	3022	2983	2985	2829	2980	2951	3051	3134	3029	3529	3254	3559	3033	3544	3215	3238
70	2756	2878	3696	3781	3515	2778	3233	3259	3237	3097	3194	3184	3310	3454	3394	3803	3635	3818	3276	3865	3498	3492
80	2936	3037	4005	4087	3742	3038	3576	3583	3506	3264	3386	3372	3539	3740	3638	4055	3922	4011	3481	4079	3730	3700
90	3104	3286	4231	4360	3972	3257	3897	3839	3737	3443	3596	3609	3775	3939	3815	4257	4182	4176	3791	4327	3918	3907
100	3269	3498	4411	4645	4266	3457	4139	4025	3963	3699	3766	3793	3988	4142	3987	4483	4416	4329	3971	4513	4084	4263
110	3428	3605	4585	4864	4474	3600	4291	4238	4153	3935	3932	3952	4141	4310	4159	4666	4603	4486	4111	4708	4225	4449
120	3563	3858	4776	5035	4671	3727	4435	4463	4301	4091	4116	4095	4349	4478	4324	4886	4758	4600	4299	4851	4368	4612
130	3674	3935	4958	5173	4868	3849	4588	4643	4480	4224	4300	4227	4510	4578	4526	5057	4982	4734	4451	5024	4510	4718
140	3769	4059	5155	5333	5018	4001	4864	4789	4636	4333	4473	4368	4704	4721	4684	5211	5146	4828	4639	5229	4634	4871
150	3866	4173	5298	5465	5189	4143	5030	4922	4802	4448	4606	4523	4877	4841	4873	5368	5298	4930	4779	5380	4784	4976
160	3953	4283	5426	5593	5297	4276	5177	5102	4940	4540	4752	4638	5013	4942	4997	5518	5475	5018	4924	5537	4924	5120
170	4091	4386	5527	5737	5460	4400	5260	5236	5086	4657	4935	4759	5127	5047	5117	5669	5634	5162	5063	5687	5042	5238
180	4196	4479	5640	5880	5674	4539	5406	5372	5207	4769	5060	4893	5250	5157	5279	5813	5725	5289	5189	5787	5146	5368
190	4269	4619	5766	5967	5801	4659	5491	5485	5319	4877	5209	5039	5370	5275	5428	5943	5832	5289	5276	5920	5257	5477
200	4357	4687	5885	6084	5921	4782	5582	5592	5417	5002	5327	5130	5494	5411	5558	6107	5976	0	5377	6027	5356	5596
210	4434	4763	5987	6165	6057	4894	5657	5697	5518	5083	5441	5216	5607	5509	5661	6219	6081	0	5441	6151	5438	5696
220	4517	4840	6083	6228	6208	5015	5730	5779	5645	5169	5570	5339	5671	5602	5809	6343	6174	0	5549	6284	5573	5786
230	4571	4910	6194	6335	6382	5117	5831	5868	5728	5260	5679	5436	5781	5677	5902	6471	6298	0	5639	6395	5661	5847
240	0	4994	6312	6448	6474	5216	5923	5954	5837	5338	5781	5562	5924	5741	6052	6607	6403	0	5741	6493	5749	5923
250	0	5068	6398	6547	6612	5329	5999	6086	5920	5418	5861	5634	6020	5824	6200	6710	6490	0	5837	6593	5829	5993
260	0	5195	6515	6609	6682	5452	6075	6181	6012	5488	5980	5716	6099	5925	6301	6809	6586	0	5894	6666	5919	6078
270	0	5259	6604	6715	6802	5569	6186	6276	6068	5569	6068	5809	6178	6000	6377	6955	6681	0	5983	6740	5979	6162
280	0	5312	6660	6803	6858	5643	6290	6403	6180	5645	6146	5868	6297	6054	6451	7090	6799	0	6063	6823	6088	6229
290	0	5361	6744	6807	6918	5699	6348	6500	6254	5745	6254	5920	6347	6138	6534	7209	6891	0	6114	6914	6178	6321
300	0	0	6814	0	6989	5807	6415	6575	6330	5771	6331	5973	6408	6217	6633	7293	6990	0	6163	7011	6244	6404
310	0	0	6886	0	0	5904	6473	6653	6421	0	6422	6049	6463	6295	6726	7419	7093	0	6241	7093	6312	6433
320	0	0	6902	0	0	5916	6484	6780	6518	0	6499	6103	6552	6341	6799	7529	7167	0	6349	7168	6384	0
330	0	0	0	0	0	0	0	6869	6623	0	6590	6154	6639	6365	6892	7628	7223	0	6445	7234	6459	0
340	0	0	0	0	0	0	0	6972	6709	0	6680	6205	6721	0	6957	7690	7288	0	6482	7297	6527	0
350	0	0	0	0	0	0	0	7082	6780	0	6760	6278	6809	0	7028	7796	7363	0	0	7371	6616	0
360	0	0	0	0	0	0	0	7175	6870	0	6848	6337	6897	0	7098	7827	7417	0	0	7456	6689	0
370	0	0	0	0	0	0	0	0	6939	0	6949	6442	6955	0	7168	0	7490	0	0	7523	6742	0
380	0	0	0	0	0	0	0	0	7027	0	7030	6537	7059	0	7243	0	7544	0	0	7577	6814	0
390	0	0	0	0	0	0	0	0	7104	0	7081	6553	7124	0	7307	0	7604	0	0	7658	6869	0
400	0	0	0	0	0	0	0	0	7204	0	7098	0	7242	0	7391	0	7667	0	0	0	6932	0
410	0	0	0	0	0	0	0	0	7255	0	0	0	7314	0	7444	0	7733	0	0	0	7038	0
420	0	0	0	0	0	0	0	0	7324	0	0	0	7333	0	7538	0	7788	0	0	0	7110	0
430	0	0	0	0	0	0	0	0	7342	0	0	0	0	0	7602	0	7838	0	0	0	7162	0
440	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7649	0	7894	0	0	0	7243	0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7693	0	7918	0	0	0	7352	0
460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7737	0	0	0	0	0	7418	0
470	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7505	0
480	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7600	0
490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7655	0
500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7725	0
510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7773	0
520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7788	0

Tableau 3 : L'entropie. Détail du calcul pour le *Docteur Pascal*

	N	V	V/√N	H réel	H théo	log(N)	S	S/N
1	2141	1061	22.9301	2.8577	2.8607	3.3306	1012	0.4729
2	4322	1605	24.4136	2.9458	2.9807	3.6357	2981	0.6899
3	6529	2179	26.9671	3.0446	3.0401	3.8148	5028	0.7702
4	8737	2573	27.5270	3.0815	3.0770	3.9414	7512	0.8599
5	10811	2924	28.1219	3.1132	3.1015	4.0339	9953	0.9207
6	12911	3238	28.4968	3.1308	3.1202	4.1110	12654	0.9801
7	14962	3492	28.5482	3.1404	3.1346	4.1750	15479	1.0346
8	17206	3700	28.2073	3.1444	3.1472	4.2357	18777	1.0913
9	19375	3907	28.0687	3.1495	3.1573	4.2872	22043	1.1378
10	21588	4263	29.0141	3.1799	3.1658	4.3342	24919	1.1543
11	23799	4449	28.8392	3.1852	3.1730	4.3766	28354	1.1914
12	25975	4612	28.6162	3.1909	3.1790	4.4146	31784	1.2237
13	28007	4718	28.1919	3.1904	3.1840	4.4473	35201	1.2569
14	30212	4871	28.0239	3.1955	3.1887	4.4802	38812	1.2847
15	32330	4976	27.6743	3.1926	3.1926	4.5096	42577	1.3170
16	34586	5120	27.5309	3.1959	3.1963	4.5389	46449	1.3430
17	36776	5238	27.3139	3.1979	3.1995	4.5656	50298	1.3677
18	38924	5368	27.2084	3.2040	3.2023	4.5902	53957	1.3862
19	41105	5477	27.0144	3.2082	3.2048	4.6139	57780	1.4057
20	43203	5596	26.9228	3.2101	3.2070	4.6355	61581	1.4254
21	45377	5696	26.7394	3.2103	3.2091	4.6568	65641	1.4466
22	47582	5786	26.5251	3.2096	3.2109	4.6774	69843	1.4679
23	49763	5847	26.2108	3.2061	3.2126	4.6969	74186	1.4908
24	51890	5923	26.0016	3.2055	3.2140	4.7151	78334	1.5096
25	54040	5993	25.7802	3.2077	3.2154	4.7327	82411	1.5250
26	56203	6078	25.6378	3.2075	3.2166	4.7498	86677	1.5422
27	58297	6162	25.5210	3.2106	3.2176	4.7656	90651	1.5550
28	60400	6229	25.3454	3.2104	3.2186	4.7810	94866	1.5706
29	62553	6321	25.2733	3.2126	3.2195	4.7962	99059	1.5836
30	64827	6404	25.1520	3.2159	3.2204	4.8118	103452	1.5958
31	65620	6433	25.1128	3.2169	3.2206	4.8170	105002	1.6002

$S = \sum n_i x_i \log(x_i)$ r = 0.9935 pour a = -0.02415 et b = 0.55823
H max = $- b^2 / 4a$ = 3.2259
x max = -b/2a = 11.5576

TABLEAU DE DISTRIBUTION pour le texte entier
(pour f= 1 à 100)

2210	991	642	416	282	236	175	147	125	103	76	84	68	48	44	43	31	36	30	29
31	26	30	20	18	19	24	7	10	9	11	12	18	9	9	10	11	9	10	9
13	8	8	5	12	6	5	5	3	5	6	3	5	4	4	6	2	4	1	3
5	7	4	3	5	6	2	2	4	5	3	1	3	2	4	3	2	1	0	4
2	1	4	0	1	4	1	1	1	2	1	0	1	3	1	1	1	2	3	1 1 1

Quant à l'unité de mesure choisie – tranches de 10 pages – elle est arbitraire, comme l'eût été l'adoption des lignes, des mots ou des lettres. Mais comme il s'agit d'une édition homogène (celle de la Pléiade), la mesure est à peu près constante[5] et le choix imposé par des contraintes

[5] A deux réserves près : d'une part les deux premiers textes (*Thérèse Raquin* et *Madeleine Férat*) sont empruntés à une autre édition, et d'autre part là où le dialogue est

techniques[6] en vaut un autre. De toute façon le nombre d'occurrences a été relevé dans chaque tranche pour servir de pondération lorsqu'il en serait besoin.

On pourrait s'attarder sur le tableau de distribution à chaque palier de l'observation. Rarement pareille occasion a été donnée d'étudier dans le détail et au ralenti le mouvement des fréquences au fur et à mesure qu'un discours s'organise dans la durée. Mais l'information est si abondante qu'on ne saurait en rendre compte sans faire appel à des programmes sophistiqués de traitement graphique. Seule une image animée pourrait montrer en effet la forme changeante de la pyramide des mots et les modifications incessantes – amoncellements, excroissances, tassements, écroulements – que la consistance variable de la matière lexicale peut engendrer à tel ou tel point de la surface. Des ondes se créent, l'effectif d'une classe se gonflant quand celui de la classe inférieure s'épuise, en attendant qu'un afflux nouveau comble les vides. Cette mécanique ondulatoire explique les variations d'effectifs qui affectent chaque classe de fréquence et qui créent parfois des baisses provisoires dans un mouvement général ascendant. Voici à titre d'exemple l'évolution des effectifs de la classe 20 dans la *Faute de l'abbé Mouret* :

0,0,2,1,6,9,8,9,13,10,8,13,12,19,15,17,14,15,14,17,16,21,18,17,25,25,36,30,29,31,27,29,29[7].

La seule classe qui paraît échapper à la baisse est celle de la fréquence 1. Dans la *Faute de l'abbé Mouret*, l'effectif des hapax croît pratiquement sans cesse de 665 pour la première tranche à 2 284 pour l'ensemble du texte. Mais il y a des à-coups, des accélérations et des ralentissements de la progression de cette classe qui est très sensible au renouvellement lexical. Ainsi, dans le même roman, l'épisode du *Paradou*, si riche en couleurs, en odeurs et en mots, produit un accès de

---

abondant, les alinéas se multiplient – à chaque réplique – et la densité en mots de chaque page s'affaiblit. Voir dans le tableau 7 le relevé des occurrences et dans le tableau 8 le relevé des vocables.

[6] Le programme prend ses données dans l'index où seule subsiste l'indication de la page mais non le numéro d'ordre de chaque mot.

[7] Pour saisir en mouvement les phénomènes de transit qui se produisent dans les classes de fréquence, on s'est intéressé, à titre d'exemple, aux gains et aux pertes de la classe 10 au fur et à mesure que se développent les 22 textes du corpus. Il s'agit certes d'un mouvement ascendant, les effectifs se gonflant de la première à la dernière tranche. Mais c'est aussi un mouvement oscillatoire ou ondulatoire, fait de gains provisoires et de décrues passagères. La progression est hésitante et vibratile avec deux pas en avant pour un en arrière (593 écarts positifs pour 225 négatifs). Les classes de fréquence sont des lieux de passage, des vases communicants où se déploie la mécanique des fluides.

fièvre panique dans la classe des hapax. Cependant dans ce texte comme dans les autres, la fréquence 1 est sujette à fléchissement, lorsqu'on s'approche de la fin. On observe plusieurs exemples de cette décrue dans la *Fortune des Rougon* dont le plus net se situe dans les 40 dernières pages, le lot des hapax dans les 4 dernières tranches baissant de 2 376 à 2 370, puis à 2 367 avant de se fixer à 2 370. Et ce même essoufflement lexical apparaît dans la plupart des textes de notre corpus. Est-ce à dire que les possibilités lexicales de Zola s'épuisent chaque fois que l'auteur achève la rédaction d'un roman ? Il serait difficile de soutenir pareille thèse car au niveau de l'ensemble des *Rougon-Macquart*, on ne discerne pas chez Zola le signe de quelque défaillance de cet ordre. L'étude des hapax du corpus le montrera ultérieurement. Le tarissement de la fréquence 1 à la fin des textes est probablement lié à quelque loi de composition qui multiplie les reprises de thèmes et de termes dans les pages de conclusion.

Tableau 4 : La fréquence 1. Gains et pertes par tranches de 10 pages

	RAQ	PER	POR	CUR	VEN	CON	FAU	EXC	ASS	FAQ	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEP	PAS
10	584	565	699	702	624	510	665	622	648	535	634	610	661	689	608	757	606	735	691	724	703	703
20	215	160	281	292	261	223	317	291	261	237	302	282	315	266	284	228	324	406	255	329	283	259
30	166	236	192	228	248	168	153	248	172	194	263	248	183	206	209	341	228	239	171	270	211	286
40	106	108	235	169	240	144	107	170	133	143	136	154	167	149	152	191	274	207	133	194	162	130
50	73	118	137	230	161	30	144	105	148	138	90	126	146	134	112	195	159	98	103	186	177	128
60	65	57	166	131	135	142	98	49	102	127	96	93	57	135	111	136	121	66	103	138	55	125
70	68	17	116	102	45	56	55	90	100	98	65	27	77	65	163	75	137	83	83	108	99	85
80	39	34	108	135	72	114	109	177	102	41	32	43	69	111	86	107	105	47	57	60	64	45
90	55	64	46	85	65	80	109	55	74	47	53	109	43	53	29	12	115	25	118	84	58	35
100	49	82	25	86	95	51	81	51	84	89	44	44	77	46	23	43	91	43	48	38	23	161
110	38	21	24	70	61	17	6	87	44	62	28	61	39	20	47	49	42	-8	33	58	25	46
120	30	103	45	32	51	21	35	66	28	34	38	45	53	43	53	71	33	10	34	1	29	16
130	14	4	52	12	70	25	31	60	46	29	55	19	57	-2	70	37	77	30	25	28	16	7
140	-2	37	38	28	17	57	107	15	43	8	48	27	72	35	28	22	14	-17	43	59	31	30
150	1	31	23	2	45	64	51	17	29	23	33	42	52	21	58	35	44	-8	7	11	47	8
160	17	21	6	44	21	37	27	49	33	6	43	8	9	25	17	43	60	13	15	22	22	29
170	31	14	-9	5	22	23	8	8	49	2	59	17	24	22	10	45	40	27	4	42	36	0
180	40	11	14	22	42	45	37	30	22	22	17	34	4	29	26	25	3	38	43	2	1	3
190	10	48	23	10	9	36	29	12	11	31	48	52	30	12	61	18	13	0	-5	50	35	-4
200	20	-8	1	28	13	35	3	33	30	30	34	10	46	25	33	58	32	0	15	5	23	34
210	10	4	15	-14	4	19	4	-23	12	6	30	-3	21	14	13	-2	-2	0	-13	12	9	13
220	8	6	-13	-13	45	31	3	9	55	12	36	46	-1	12	36	33	6	0	30	9	36	17
230	2	9	26	21	81	18	18	39	17	7	19	15	13	6	25	35	38	0	15	12	6	-18
240	0	6	53	30	-8	30	19	9	10	18	21	23	43	-5	40	33	17	0	29	33	3	17
250	0	18	21	21	46	41	15	24	12	7	8	9	17	-1	25	23	12	0	29	-9	5	-16
260	0	39	31	-25	-1	43	7	35	18	12	23	2	-1	18	22	11	22	0	-14	8	31	19
270	0	-1	9	19	36	27	22	5	-13	25	24	9	15	13	3	34	25	0	3	26	-4	-6
280	0	6	-6	11	-14	-7	19	18	42	7	-7	1	25	-6	2	33	30	0	30	23	15	16
290	0	-2	18	-3	-6	-17	6	18	4	31	12	4	-20	6	18	30	10	0	-8	20	10	29
300	0	0	-6	0	6	35	5	9	22	3	16	-16	-1	10	22	25	-3	0	-1	2	-12	10
310	0	0	-3	0	0	18	-5	8	10	0	12	6	3	15	9	23	21	0	30	22	9	3
320	0	0	3	0	0	-5	-1	41	33	0	17	0	18	-10	6	22	10	0	22	-6	-12	0
330	0	0	0	0	0	0	0	7	29	0	25	3	5	1	6	18	1	0	19	-7	-2	0
340	0	0	0	0	0	0	0	28	20	0	12	0	12	0	-12	0	-1	0	-6	-4	-8	0
350	0	0	0	0	0	0	0	25	14	0	23	6	29	0	1	12	1	0	0	-8	30	0
360	0	0	0	0	0	0	0	17	39	0	23	-5	11	0	18	2	-3	0	0	21	14	0
370	0	0	0	0	0	0	0	0	-6	0	17	25	0	0	22	0	12	0	0	6	-2	0
380	0	0	0	0	0	0	0	0	25	0	11	34	34	0	5	0	6	0	0	-13	-4	0
390	0	0	0	0	0	0	0	0	22	0	4	-3	0	0	10	0	11	0	0	16	14	0
400	0	0	0	0	0	0	0	0	15	0	7	0	45	0	19	0	16	0	0	0	3	0
410	0	0	0	0	0	0	0	0	4	0	0	0	1	0	9	0	7	0	0	0	28	0
420	0	0	0	0	0	0	0	0	7	0	0	0	2	0	15	0	-1	0	0	0	19	0
430	0	0	0	0	0	0	0	0	-12	0	0	0	0	0	-3	0	-3	0	0	0	8	0
440	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-9	0	-7	0	0	0	22	0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-14	0	13	0	0	0	38	0
460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	4	0
470	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0
480	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0
490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4	0
500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0
510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-4	0
520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0

Tableau 5 : Valeur de l'entropie de 10 en 10 pages (multipliée par 1 000)

	RAG	FEN	FOR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GEP	OEU	TER	REV	BET	ARG	DEP	PAS
10	2809	2825	2862	2850	2814	2751	2842	2823	2828	2786	2830	2809	2841	2858	2820	2906	2822	2888	2859	2862	2853	2856
20	2923	2920	2975	2999	2926	2869	2987	2955	2932	2907	2952	2937	2955	2972	2974	3005	2957	3054	2949	2995	2985	2946
30	2994	3013	3028	3072	3000	2920	3027	3025	2978	2962	3026	3004	3012	3016	3029	3098	3011	3114	3019	3072	3040	3045
40	3032	3053	3077	3115	3059	2954	3054	3060	3020	2993	3051	3040	3055	3067	3066	3127	3074	3159	3039	3123	3078	3081
50	3049	3087	3118	3155	3098	2963	3078	3079	3057	3039	3068	3067	3087	3099	3082	3150	3110	3182	3071	3161	3118	3113
60	3065	3101	3152	3179	3124	2995	3093	3097	3081	3070	3083	3079	3101	3129	3106	3175	3125	3197	3090	3182	3131	3131
70	3088	3105	3175	3197	3138	3004	3095	3117	3094	3088	3090	3091	3116	3144	3136	3183	3156	3206	3102	3201	3146	3140
80	3102	3113	3192	3210	3147	3021	3132	3138	3109	3090	3095	3098	3128	3160	3148	3190	3175	3208	3109	3205	3157	3144
90	3116	3134	3203	3221	3158	3034	3160	3155	3120	3094	3103	3114	3145	3167	3149	3196	3184	3210	3135	3211	3158	3149
100	3124	3143	3211	3235	3181	3045	3174	3158	3126	3110	3107	3121	3154	3175	3152	3207	3191	3208	3142	3215	3161	3180
110	3134	3144	3216	3242	3190	3050	3176	3164	3134	3129	3116	3124	3158	3181	3154	3211	3194	3213	3143	3219	3159	3185
120	3138	3161	3221	3245	3198	3053	3176	3175	3137	3135	3128	3126	3169	3185	3160	3217	3195	3212	3152	3219	3158	3191
130	3144	3163	3226	3244	3206	3057	3177	3182	3144	3138	3138	3127	3172	3180	3170	3222	3205	3212	3157	3226	3163	3190
140	3145	3161	3234	3246	3208	3064	3195	3186	3147	3138	3142	3132	3179	3182	3175	3224	3209	3211	3168	3232	3166	3196
150	3148	3158	3235	3249	3211	3066	3198	3189	3152	3138	3146	3140	3185	3182	3182	3226	3210	3213	3173	3235	3171	3193
160	3150	3163	3239	3248	3208	3072	3201	3196	3156	3138	3150	3142	3190	3181	3185	3228	3214	3210	3180	3240	3176	3196
170	3159	3166	3238	3254	3210	3080	3196	3201	3160	3142	3160	3144	3191	3181	3187	3229	3219	3215	3186	3245	3180	3198
180	3161	3170	3239	3257	3221	3085	3199	3205	3160	3147	3165	3147	3193	3184	3195	3231	3214	3221	3191	3241	3183	3204
190	3162	3171	3241	3254	3224	3090	3195	3206	3164	3148	3170	3152	3195	3188	3198	3234	3214	3221	3190	3242	3186	3208
200	3164	3170	3243	3256	3226	3094	3195	3206	3163	3152	3171	3152	3197	3193	3202	3240	3218	0	3191	3243	3186	3210
210	3165	3167	3244	3254	3232	3100	3195	3210	3164	3151	3174	3153	3199	3194	3204	3243	3220	0	3187	3246	3186	3210
220	3166	3165	3246	3250	3235	3105	3194	3210	3167	3152	3176	3156	3198	3195	3211	3245	3221	0	3187	3252	3191	3210
230	3168	3166	3249	3255	3239	3110	3195	3209	3167	3155	3178	3158	3200	3194	3213	3248	3221	0	3190	3253	3192	3206
240	0	3168	3251	3257	3240	3112	3195	3211	3169	3156	3179	3163	3206	3193	3219	3252	3223	0	3193	3252	3194	3205
250	0	3172	3252	3257	3243	3118	3195	3217	3170	3156	3177	3164	3208	3195	3227	3252	3223	0	3195	3254	3196	3208
260	0	3177	3255	3257	3242	3122	3192	3218	3171	3154	3182	3165	3210	3199	3229	3252	3223	0	3196	3252	3197	3208
270	0	3179	3255	3259	3246	3127	3197	3222	3170	3155	3183	3166	3210	3200	3229	3257	3223	0	3200	3250	3198	3211
280	0	3180	3254	3260	3244	3128	3201	3227	3172	3156	3184	3166	3214	3199	3230	3261	3225	0	3202	3250	3204	3210
290	0	3177	3255	3260	3242	3129	3199	3231	3173	3160	3188	3165	3214	3200	3230	3264	3225	0	3200	3251	3206	3213
300	0	0	3256	0	3243	3133	3200	3231	3173	3162	3190	3165	3214	3201	3234	3265	3226	0	3195	3254	3207	3216
310	0	0	3257	0	0	3138	3200	3231	3175	0	3194	3166	3214	3200	3237	3268	3229	0	3195	3255	3208	3217
320	0	0	3258	0	0	3138	3200	3236	3177	0	3195	3165	3216	3199	3238	3270	3229	0	3200	3258	3211	0
330	0	0	0	0	0	0	0	3237	3180	0	3196	3164	3219	3198	3243	3272	3228	0	3204	3259	3212	0
340	0	0	0	0	0	0	0	3241	3183	0	3200	3164	3221	0	3243	3270	3228	0	3205	3258	3214	0
350	0	0	0	0	0	0	0	3245	3183	0	3200	3166	3222	0	3243	3274	3230	0	0	3259	3216	0
360	0	0	0	0	0	0	0	3249	3185	0	3202	3167	3225	0	3243	3275	3229	0	0	3262	3218	0
370	0	0	0	0	0	0	0	0	3186	0	3205	3172	3226	0	3244	0	3230	0	0	3263	3218	0
380	0	0	0	0	0	0	0	0	3188	0	3207	3174	3230	0	3245	0	3229	0	0	3262	3220	0
390	0	0	0	0	0	0	0	0	3189	0	3207	3175	3231	0	3245	0	3229	0	0	3265	3219	0
400	0	0	0	0	0	0	0	0	3193	0	3207	0	3235	0	3248	0	3229	0	0	0	3220	0
410	0	0	0	0	0	0	0	0	3193	0	0	0	3238	0	3248	0	3229	0	0	0	3224	0
420	0	0	0	0	0	0	0	0	3194	0	0	0	3238	0	3251	0	3230	0	0	0	3225	0
430	0	0	0	0	0	0	0	0	3194	0	0	0	0	0	3252	0	3229	0	0	0	3225	0
440	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3253	0	3230	0	0	0	3226	0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3254	0	3230	0	0	0	3230	0
460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3255	0	0	0	0	0	3231	0
470	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3234	0
480	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3238	0
490	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3239	0
500	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3241	0
510	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3242	0
520	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3243	0

Nulle part autant que dans l'orchestration du final, Zola ne donne si libre cours à son goût du leitmotiv. Peut-être s'agit-il d'un phénomène général qu'il serait intéressant d'observer chez d'autres romanciers, Balzac par exemple. En tous cas, chez Zola la tendance est régulière et tous ses romans laissent apparaître la stagnation des hapax dans les dernières pages. Une claire illustration en est donnée par le tableau 4 qui pour chaque texte gère de page en page le stock des hapax, en précisant à chaque étape les gains et les pertes. Au bas de chaque colonne – à la fin de chaque texte – les signes négatifs, c'est-à-dire les déficits, se multiplient.

Mais revenons à l'entropie dont l'avantage est de résumer d'un mot – d'un chiffre – les mille leçons du tableau des fréquences. Le résultat pour la *Faute de l'abbé Mouret* est de 3,200. Pour donner un sens à ce nombre il faut le comparer à ceux qu'on obtient dans les autres textes du corpus. Le tableau 5 consacre à chacun d'eux une colonne de résultats dont le dernier fixe l'entropie du texte à son point final, ce que nous résumons dans le tableau ci-après :

	H	rang 1	rang 2		H	rang 1	rang 2
RAQ	3,168	20	19	POT	3,175	19	20
FER	3,177	18	22	BON	3,238	9	10
FOR	3,258	4	6	JOI	3,198	16	15
CUR	3,260	3	2	GER	3,255	5	9
VEN	3,243	7	3	OEU	3,275	1	1
CON	3,138	22	21	TER	3,230	10	8
FAU	3,200	15	12	REV	3,221	11	7
EXC	3,249	6	5	BET	3,205	14	17
ASS	3,194	17	14	ARG	3,265	2	4
PAG	3,162	21	18	DEB	3,243	8	16
NAN	3,207	13	11	PAS	3,217	12	13

On a rapproché le classement obtenu selon l'entropie (rang 1) de celui que la loi binomiale nous avait proposé pour la richesse lexicale. La similitude est frappante et l'entropie semble donc bien mesurer, comme le pense Évrard, la diversité lexicale d'un texte. Il y a pourtant quelques distorsions et le problème est de savoir à laquelle des deux méthodes attribuer la déviance. Il ne sert à rien de solliciter le témoignage des autres indices (W, Uber ou Rubet) puisque, s'appuyant sur *N* et *V* comme la loi binomiale, ils ne peuvent que souscrire au classement de cette dernière. Tous ces indices luttent à l'envi contre l'influence de la longueur et sur ce point l'entropie n'est pas sans défaut. Observons en effet la progression de *H* sur une même colonne, c'est-à-dire dans un même texte. Partout l'entropie est croissante d'un mouvement régulier[8] qui ne souffre aucune exception. Et cela tient à sa définition, l'entropie partant nécessairement de zéro (lorsque le texte n'a qu'une occurrence et qu'un vocable) et s'élevant jusqu'à un maximum de log *N*, qu'elle est loin d'atteindre (dans l'ordre de grandeur de nos textes, l'entropie se situe à 3 environ, alors que la limite est fixée près de 4). Et cette caractéristique nuit à l'utilisation de l'entropie quand les textes comparés sont de longueur différente. Ainsi, à son terme, la *Débâcle*, parvient à une entropie de 3,243 qui dépasse celle du *Rêve* (3,221) dont le vocabulaire est plus riche mais dont l'étendue est trois fois moindre. Mais si dans la *Débâcle*, on prend la valeur de *H* au niveau du *Rêve*, c'est-à-dire au bout de 200 pages, l'entropie n'est plus que de 3,186. Cette influence de la longueur fait ainsi gagner 8 places à la *Débâcle* dans le classement de l'entropie, et 4 à *Germinal*, et dans le même temps elle pénalise le *Rêve* de 4 places.

[8] Il ne s'agit pas d'un mouvement uniforme mais d'un mouvement uniformément ralenti – ce qui invite à en rendre compte par une équation du second degré.

Tableau 6 : Classements selon l'entropie

PAGE	RAQ	PER	POR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEB	PAS
10	19	14	3	9	18	22	10	15	13	21	12	20	11	6	17	1	16	2	5	4	8	7
20	19	20	7	3	18	22	5	11	17	21	13	16	12	9	8	2	10	1	14	4	6	15
30	19	14	8	3	18	22	9	11	20	21	10	17	15	13	7	2	16	1	12	4	6	5
40	19	15	7	4	12	22	14	11	20	21	16	17	13	9	10	2	8	1	18	3	6	5
50	20	11	5	3	10	22	15	14	19	21	17	18	12	9	13	4	8	1	16	2	6	7
60	21	12	5	3	10	22	15	14	18	20	17	19	13	8	11	4	9	1	16	2	6	7
70	20	14	5	3	10	22	16	12	17	21	19	18	13	8	11	4	6	1	15	2	7	9
80	18	15	4	1	10	22	13	12	16	21	20	19	14	7	9	5	6	2	17	3	8	11
90	18	16	4	1	9	22	8	11	17	21	20	19	14	7	12	5	6	3	15	2	10	13
100	18	15	3	1	7	22	10	12	17	20	21	19	13	9	14	5	6	4	16	2	11	8
110	17	15	3	1	7	22	10	11	18	19	21	20	13	9	14	5	6	4	16	2	12	8
120	17	13	2	1	6	22	10	11	18	19	20	21	12	9	14	4	7	5	16	3	15	8
130	17	14	2	1	6	22	11	9	18	19	20	21	12	10	13	4	7	5	16	3	15	8
140	18	16	2	1	7	22	9	10	17	20	19	21	12	11	13	4	6	5	14	3	15	8
150	18	16	2	1	6	22	8	10	17	21	19	20	11	12	13	4	7	5	14	3	15	9
160	18	16	3	1	7	22	8	9	17	21	19	20	11	13	12	4	5	6	14	2	15	10
170	19	16	3	1	7	22	10	8	17	21	18	20	11	14	12	4	5	6	13	2	15	9
180	18	16	3	1	5	22	10	8	19	20	17	21	12	14	11	4	7	6	13	2	15	9
190	19	16	3	1	5	22	11	9	18	21	17	20	12	14	10	4	7	6	13	2	15	8
200	18	17	2	1	5	22	12	9	19	20	16	21	11	13	10	4	7	6	14	3	15	8
210	18	17	3	1	5	22	12	8	19	21	16	20	11	13	10	4	7	6	14	2	15	9
220	18	19	3	2	5	22	13	9	17	21	16	20	11	12	8	4	6	7	15	1	14	10
230	17	19	3	1	5	22	12	9	18	21	16	20	11	13	8	4	6	7	15	2	14	10
240	18	19	4	1	5	22	12	9	17	21	16	20	10	14	8	2	6	7	15	3	13	11
250	19	17	3	1	5	22	13	9	18	21	16	20	10	14	6	4	7	8	15	2	12	11
260	19	17	2	1	5	22	15	9	18	21	16	20	10	12	6	3	7	8	14	4	13	11
270	19	17	3	1	5	22	15	8	18	21	16	20	11	12	6	2	7	9	13	4	14	10
280	19	17	3	2	5	22	14	7	18	21	16	20	10	15	6	1	8	9	13	4	12	11
290	19	17	3	2	5	22	15	6	18	21	16	20	10	13	7	1	8	9	14	4	12	11
300	19	17	3	2	5	22	14	7	18	21	16	20	11	13	6	1	8	9	15	4	12	10
310	19	17	3	2	5	22	13	7	18	21	16	20	11	14	6	1	8	9	15	4	12	10
320	19	17	3	2	5	22	13	7	18	21	16	20	11	15	6	1	8	9	14	4	12	10
330	19	18	4	2	5	22	14	7	17	21	16	20	10	15	6	1	8	9	13	3	12	11
340	19	18	3	2	5	22	14	7	17	21	15	20	9	16	6	1	8	10	13	4	12	11
350	19	18	4	2	6	22	14	5	17	21	15	20	9	16	7	1	8	10	13	3	12	11
360	19	18	4	3	6	22	15	5	17	21	14	20	9	16	7	1	8	10	13	2	11	12
370	20	18	4	3	7	22	15	5	17	21	13	19	9	16	6	1	8	10	14	2	11	12
380	20	18	4	3	7	22	15	5	17	21	13	19	8	16	6	1	9	10	14	2	11	12
390	20	18	4	3	7	22	15	5	17	21	13	19	8	16	6	1	9	10	14	2	11	12
400	20	18	4	3	7	22	15	5	17	21	13	19	8	16	6	1	9	10	14	2	11	12
410	20	18	4	3	7	22	15	5	17	21	13	19	8	16	6	1	9	11	14	2	10	12
420	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	9	11	14	2	10	12
430	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	9	11	14	2	10	12
440	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	9	11	14	2	10	12
450	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	9	11	14	2	10	12
460	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	10	11	14	2	9	12
470	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	10	11	14	2	9	12
480	20	18	4	3	7	22	15	6	17	21	13	19	8	16	5	1	10	11	14	2	9	12
490	20	18	4	3	7	22	15	6	17	21	13	19	9	16	5	1	10	11	14	2	8	12
500	20	18	4	3	7	22	15	6	17	21	13	19	9	16	5	1	10	11	14	2	8	12
510	20	18	4	3	7	22	15	6	17	21	13	19	9	16	5	1	10	11	14	2	8	12
520	20	18	4	3	7	22	15	6	17	21	13	19	9	16	5	1	10	11	14	2	8	12

C'est pourquoi les classements n'ont de valeur qu'au même niveau et nous les avons calculés tranche après tranche dans le tableau 6. On les lira sur les lignes de cette matrice. La première montre que dès la dixième page l'*Œuvre* a pris la tête du corpus au regard de la diversité lexicale. Elle la cède ensuite au *Rêve*, puis à la *Curée* avant de la retrouver définitivement. En queue de classement au contraire, le même texte (la *Conquête*) se maintient d'un bout à l'autre. Si l'on veut à toute force fixer un classement, on peut choisir un niveau qui représente le meilleur compromis entre les textes longs et courts. Ce pourrait être par exemple la 30^{e} ligne qui indique pour chaque texte l'entropie atteinte à la page 300. On se rapproche un peu alors du classement de la loi binomiale

(0,91). Mais le tableau 6 est aussi intéressant par ses colonnes qui montrent l'évolution interne de chaque texte. Certains, mal classés au départ, améliorent leur position au fil des pages. Certains autres au contraire dépensent, sans compter et sans attendre, leurs richesses lexicales et s'épuisent par la suite. Le *Ventre de Paris* a un profil caractéristique du premier groupe, comme aussi *Germinal*, alors que le *Rêve* représente le second.

Ce qui sépare l'entropie des autres indices tient aussi au fait que la mesure ne porte pas exactement sur le même objet. Tout indice qui met en rapport *N* et *V* est nécessairement très sensible aux basses fréquences qui influent grandement sur *V* (les seuls hapax, suivant les cas, représentent 30 à 50 pour cent des vocables). Le calcul de l'entropie fait au contraire la part belle aux hautes fréquences, au point que nous avons dû filtrer les fréquences extrêmes. Il tend à négliger par contre les basses fréquences et en particulier se montre totalement indifférent aux hapax car le terme $n_i\ x_i\ \log(x_i)$ est nul lorsque x_i vaut 1 (log(1) valant zéro). Les textes où les hapax foisonnent, principalement l'*Assommoir* et le *Ventre de Paris* reculent donc dans le classement de l'entropie.

Mais Évrard ne destine pas l'entropie à ce rôle de classement. Car ce n'est pas pour lui la valeur finale de l'entropie qui importe (puisqu'elle dépend de la longueur) mais bien plutôt le profil de cette entropie, la courbe qui rend compte d'une progression plus ou moins accentuée. Une fois de plus, comme dans le calcul de l'indice Uber, nous sommes ramenés à un problème d'ajustement. Et comme Dugast, Évrard songe à une fonction du second degré du type $y = ax^2 + bx$ (il n'y a pas de terme *c* parce que la courbe commence nécessairement à 0, quand le texte n'a qu'un mot). Comme dans le cas de l'indice Uber, on aurait pu recourir à un autre modèle d'ajustement, par exemple à la fonction logarithmique $y = a + b\ log(x)$, *y* représentant l'entropie et *x* l'étendue du texte. Les résultats eussent été acceptables avec un coefficient de détermination proche de 0,90 (soit une corrélation de 0,95). Voir la partie supérieure du tableau 7. L'ajustement de la fonction du second degré est meilleur encore et les valeurs des paramètres *a* et *b* s'accordent avec les observations faites par Évrard sur son corpus latin : *a* est négatif (et *b* positif) et la concavité de la parabole est tournée vers le bas. Les valeurs relevées dans les 22 textes sont du même ordre et autorisent, semble-t-il, des classements auxquels Évrard donne une signification précise. Le paramètre *b* rendrait compte de la volonté plus ou moins accusée de diversification lexicale, alors que le paramètre *a* marquerait la limite de l'« essoufflement », la pesanteur qui infléchit la courbe vers le bas.

Tableau 7 : Entropie

APPROXIMATION LOGARITHMIQUE

		b	a	r^2	r	rang b	rang a
1	RAQ	0.10637	2.07363	0.94242	0.97079	3	3
2	FER	0.08703	2.25783	0.88180	0.93904	21	18
3	FOR	0.10810	2.09573	0.91887	0.95858	2	4
4	CUR	0.10293	2.16360	0.85874	0.92668	7	10
5	VEN	0.11852	1.96605	0.94410	0.97165	1	1
6	CON	0.09995	2.03671	0.97900	0.98944	9	2
7	FAU	0.09260	2.20938	0.89745	0.94734	15	13
8	EXC	0.10487	2.09704	0.94706	0.97317	5	5
9	ASS	0.09224	2.18196	0.93381	0.96634	16	12
10	PAG	0.09824	2.11220	0.91461	0.95635	10	8
11	NAN	0.09032	2.20949	0.94676	0.97302	17	14
12	POT	0.08883	2.21502	0.90481	0.95121	19	16
13	BON	0.10018	2.13530	0.94659	0.97293	8	9
14	JOI	0.08771	2.26430	0.87405	0.93491	20	19
15	GER	0.10411	2.10131	0.94568	0.97246	6	6
16	OEU	0.08897	2.29717	0.91455	0.95632	18	21
17	TER	0.10492	2.10743	0.88828	0.94249	4	7
18	REV	0.09302	2.26865	0.79653	0.89249	14	20
19	BET	0.09593	2.16477	0.95784	0.97869	12	11
20	ARG	0.09652	2.22369	0.86762	0.93146	11	17
21	DEB	0.08367	2.30317	0.88176	0.93902	22	22
22	PAS	0.09361	2.21120	0.89373	0.94538	13	15

$y = a + b \log(x)$ pour y = entropie et x = N

r= 0,67 r= 0,88 r= 0,86

CALCUL DES PARAMETRES A ET B

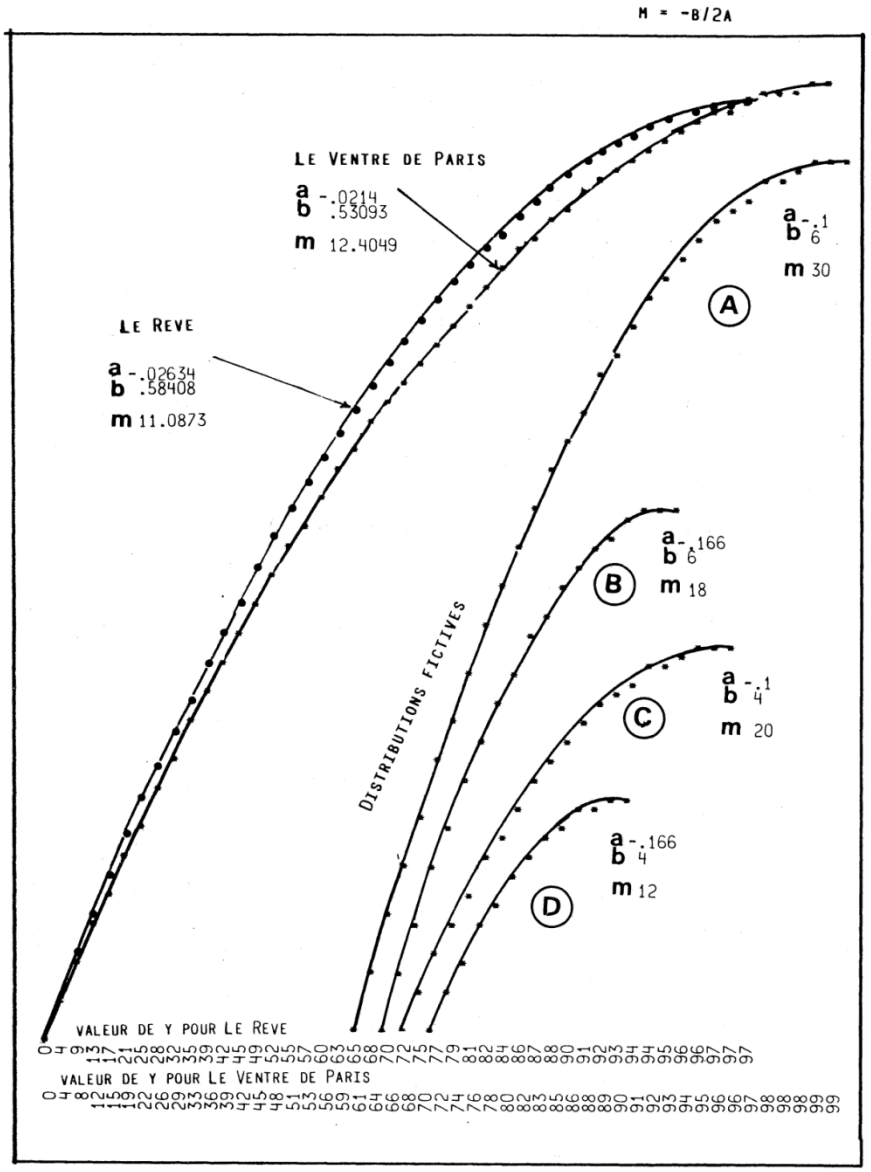
	N	log(N)	x^{max}	rang x max	b	a	rang b	rang a
RAQ	37866	10.5418	11.3528	18	0.56151	-0.02473	16	17
FER	55619	10.9263	11.1375	21	0.57113	-0.02564	20	21
FOR	67902	11.1258	11.9886	2	0.54812	-0.02286	5	3
CUR	60778	11.0150	11.8222	5	0.55659	-0.02354	13	9
VEN	63921	11.0654	12.4049	1	0.53093	-0.02140	1	1
CON	64869	11.0801	11.7725	7	0.53094	-0.02255	2	2
FAU	68551	11.1353	11.5519	13	0.55634	-0.02408	11	12
EXC	73468	11.2046	11.8124	6	0.54857	-0.02322	6	6
ASS	92591	11.4359	11.5190	14	0.55222	-0.02397	10	11
PAG	59546	10.9945	11.6419	10	0.54554	-0.02343	3	7
NAN	82310	11.3182	11.4343	15	0.55662	-0.02434	14	14
POT	78642	11.2727	11.3903	16	0.55653	-0.02443	12	16
BON	86645	11.3696	11.7281	8	0.55005	-0.02345	9	8
JOI	68722	11.1378	11.3525	19	0.56490	-0.02488	18	18
GER	96246	11.4747	11.8460	4	0.54681	-0.02308	4	4
OEU	77252	11.2548	11.3883	17	0.57192	-0.02511	21	19
TER	94299	11.4542	11.8676	3	0.54876	-0.02312	7	5
REV	38554	10.5598	11.0873	22	0.58408	-0.02634	22	22
BET	73002	11.1982	11.6680	9	0.55003	-0.02357	8	10
ARG	83015	11.3268	11.5756	11	0.56489	-0.02440	17	15
DEB	108657	11.5960	11.2305	20	0.56961	-0.02536	19	20
PAS	65620	11.0916	11.5576	12	0.55823	-0.02415	15	13

$y = ax^2 + bx$ r= 0,96

pour y= entropie et x= log(N)

En théorie, les paramètres *a* et *b* ont effectivement cet effet sur une fonction du second degré. Et on le comprendra avec 4 exemples fictifs que nous avons projetés à droite de la figure 8. La pente initiale est surtout déterminée par le facteur *b* et elle est plus forte en *A* et *B* (*b* = 6) qu'en *C* et *D* (*b* = 4). La courbure, au contraire, est d'autant plus accusée que *a* est plus grand, et c'est le cas en *B* et *D* (où *a* vaut -0,166, contre -0,1 dans *A* et *C*). La diversité maximale du vocabulaire suppose donc une pente raide (*b* fort) et une courbure peu prononcée (*a* faible). Une entropie restreinte (c'est-à-dire la sobriété lexicale) devrait au contraire gonfler le paramètre *a* et diminuer *b*. Or qu'observe-t-on ? Des situations intermédiaires du type *B* ou *C*, où *a* et *b* sont à la fois faibles ou forts. La partie gauche de la figure 8 représente les deux textes qui s'opposent le plus par la courbe de leur entropie : le *Rêve* et le *Ventre de Paris*. Le *Rêve* a les plus fortes valeurs de la série pour les deux paramètres *a* et *b* et le *Ventre de Paris* les deux plus faibles. Les deux courbes se différencient nettement sur le graphique : celle du *Rêve* s'élève et retombe plus rapidement, en une trajectoire plus verticale que la pesanteur épuise plus vite. Celle du *Ventre de Paris* est plus tendue et elle va plus loin. Or on observe les mêmes effets balistiques dans les 22 textes de notre corpus. Partout une relation constante est établie entre les facteurs *b* et *a*, entre l'angle d'attaque et l'épuisement de l'énergie. Si on fait un classement sur *a* et sur *b*, la corrélation est extrêmement forte (r = 0,96). Dès lors il devient difficile de donner un sens particulier et différent à l'un et à l'autre paramètres, puisqu'ils vont de pair et qu'apparemment ils se rattachent à une même cause et rendent compte du même phénomène. D'ailleurs dans les données étudiées par Évrard, la liaison entre *a* et *b* parait également établie. Précisons d'ailleurs que les paramètres *a* et *b* qu'on obtient par une approximation logarithmique sont eux aussi corrélés entre eux et qu'ils entretiennent des liens étroits avec ceux de la fonction du second degré. Dans le premier cas (*a* et *b*) le coefficient est de 0,88, dans le second (al et a2, bl et b2) la corrélation s'élève à 0,86 et 0,67 respectivement.

Tableau 8 : Courbe de l'entropie ($y = ax^2 + bx$)



Que conclure de notre coûteuse expérience ? Qu'il faut sans doute renoncer à distinguer les paramètres *a* et *b* puisque dans les faits observés ils donnent le même enseignement. Mais cette information n'est pas indifférente : la forme de la courbe de l'entropie renseigne sur la façon

dont un auteur gère le stock lexical, en prodigue ou en économe, en visant le court terme ou les longues échéances. Et de ce point de vue, les deux textes représentés dans le graphique 8, qui sont tous les deux parmi les plus riches des *Rougon-Macquart*, n'usent pas semblablement de leurs richesses. Dans l'univers de sacristie du *Rêve*, on sent que Zola est entré un peu par effraction, armé de dictionnaires et d'ouvrages spécialisés[9]. Il est pressé d'écouler son butin, presque à la sauvette, au lieu que dans le *Ventre de Paris*, Zola se sent chez lui, dans un monde familier dont il connaît depuis longtemps les inépuisables ressources. Il fait faire à son lecteur le tour du propriétaire mais il garde des réserves, des secrets et l'on sent qu'il pourrait offrir une seconde visite totalement différente de la première[10].

Il existe cependant un moyen d'amplifier les minces différences qui peuvent séparer *a* et *b* afin de souligner le dessin de la courbe. Il s'agit tout simplement de projeter le point ultime que la courbe peut atteindre sur l'axe des *x* avant de rebrousser chemin, c'est-à-dire avant que la seconde branche de la parabole n'entame son mouvement de chute[11].

---

[9] Le dossier préparatoire du *Rêve* est détaillé par H. Mitterand, pp. 1610-1615, la Pléiade tome IV. On y trouve un bric-à-brac de lectures sur la *Légende dorée*, l'architecture des cathédrales, les cérémonies religieuses, les règles administratives de l'adoption, les armoiries les vitraux, les broderies. Beaucoup de dictionnaires et d'encyclopédies dans cette documentation : l'*Art du brodeur* de Saint Aubain, l'*Almanach du Commerce*, le *Dictionnaire* de Savary, le *Dictionnaire de l'Industrie*, la *Légende dorée* de Jacques de Voragine, le *Grand Dictionnaire* de Pierre Larousse, le *Dictionnaire d'Architecture*, le *Dictionnaire des cérémonies et des rites sacrés*.

[10] La préparation livresque du *Ventre de Paris* est beaucoup plus légère : quelques ouvrages d'histoire sur le Second Empire, quelques notices sur les déportations en Guyane et quelques rapport sur l'administration des Halles, voir H. Mitterand tome 1, pp. 1622-1623. Les descriptions du *Ventre de Paris* sont faites *de visu* comme en témoigne Paul Alexis qui a souvent accompagné Zola dans la visite des Halles : « Un crayon à la main. Zola venait par tous les temps, par la pluie, le soleil, le brouillard, la neige, et à toutes les heures, le matin, l'après-midi, le soir, afin de noter les différents aspects. Puis, une fois, il y passa la nuit entière, pour assister au grand arrivage de la nourriture de Paris », cité par H. Mitterand, tome 1, p. 1617.

[11] Cela ne s'observe pas dans la réalité où l'entropie croît sans cesse, du moins dans l'ordre de grandeur où nous avons placé nos observations. C'est donc la portion ascendante de la courbe qui est utile, mais presque toute la portion ascendante, car on s'approche très près du sommet. Ainsi dans le *Rêve*, la limite log *N* a pour valeur 10,5598 alors que le sommet théorique de la courbe est de : $-b / 2a = 11{,}0873$. Dans le cas du texte le plus long, la *Débâcle*, il semble qu'on ait franchi la limite et qu'on amorce le mouvement de descente : l'axe de symétrie (11,2305) se situe en deçà de la dernière valeur de *x* (11,5960). Comme en réalité l'entropie n'a cessé de monter (mais très faiblement), il s'agit plutôt d'une imperfection de l'ajustement.

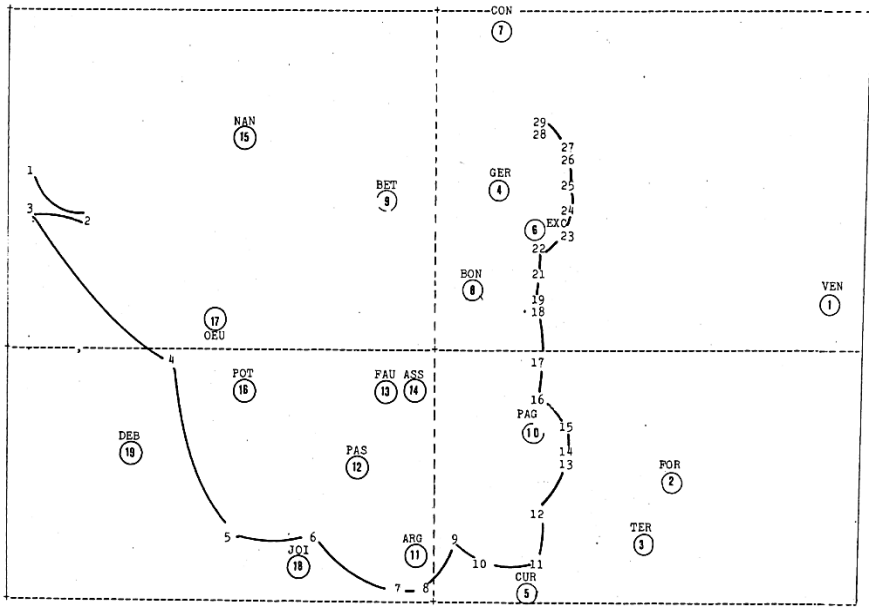
Dans une fonction du second degré, ce point qui marque l'axe de symétrie est le quotient $-b/2a$ qui dans nos données prend une valeur proche de 11. Ce point virtuel se situe plus ou moins loin (de 11,01 à 12,40) et cela suffit à déterminer si la trajectoire est plus raide ou plus molle, c'est-à-dire si le discours garde une énergie potentielle ou s'il a épuisé ses réserves. Le classement obtenu sous ce rapport est indiqué dans la quatrième colonne inférieure du tableau 7. Il est d'ailleurs très semblable à celui que permettent les paramètres *a* et *b*, en se rangeant toujours du côté de *a* lorsqu'il y a quelque désaccord entre *a* et *b* (et même en ajoutant une surenchère à la tendance marquée par *a*). Étienne Évrard avait donc raison de privilégier le paramètre *a* et de l'interpréter comme le signe de l'essoufflement lexical. Mais l'axe de symétrie, qui tient compte à la fois de *a* et de *b*, nous paraît un meilleur indice. Et surtout, il ne nous semble pas que le paramètre *b* ait quelque rapport direct avec la diversité lexicale. L'indice de la diversité (ou richesse) lexicale, c'est la valeur elle-même de l'entropie[12], à tel moment d'un texte que l'on voudra et particulièrement à la fin, quand les jeux sont faits. La forme de la courbe est une autre notion, qui met en action le mouvement du texte et indique une tendance. Ainsi, un homme peut être riche ou pauvre (c'est l'entropie) et, étant ce qu'il est, s'employer à l'être plus ou à l'être moins (c'est la tendance).

Mais le tableau à deux dimensions comme le tableau 5 ne peut pas ne pas donner l'idée d'une analyse factorielle. De si nombreux exemples de ce puissant outil ont été proposés dans le domaine linguistique qu'il semble oiseux d'en exposer les principes. Nous ne parlerons ici que des contraintes, auxquelles il a fallu souscrire en écartant les lignes et les colonnes trop courtes afin d'avoir une matrice exempte de trous. On n'a donc retenu que les 29 premières lignes, en considérant que tous les textes étaient suspendus à la 290e page, ce qui est suffisant pour l'indication d'une tendance. Les trois textes qui n'atteignaient pas cette limite (*Thérèse Raquin*, *Madeleine Férat*, et le *Rêve*) n'ont pas été pris en compte. C'est donc un tableau de 29 lignes (29 tranches de 10 pages) et de 19 colonnes (19 textes) qui a été soumis à une analyse factorielle des correspondances. Le programme reconnaît sans peine l'existence de données sérielles et sur le graphique 9, il dispose dans l'ordre les 29 maillons d'une chaîne qui décrit un croissant, comme c'est l'habitude : dans le quadrant supérieur gauche, les 3 premières tranches, puis les 5 suivantes dans le quadrant inférieur gauche, les 9 suivantes de l'autre côté

[12] Avec les réserves que nous avons dites.

de l'axe des *y* et les 12 dernières en haut et à droite. Or, sur un tel plan les textes se répartissent suivant que leur entropie est relativement plus forte dans les premières ou dans les dernières tranches. Ceux qui dépensent dès les premières pages leurs ressources lexicales sont à gauche, ceux qui les ménagent et les réservent pour la fin, à droite. La répartition selon le 1er facteur (ce qui sépare la gauche de la droite sur le graphique 9) reprend clairement le classement auquel nous a conduit l'axe de symétrie dans la méthode précédente. Les deux termes extrêmes sont les mêmes et sur la marge droite, le *Ventre de Paris* s'oppose à la *Débâcle* à l'extrême gauche. Nous avons fait figurer en médaillon le classement de la méthode précédente. On voit qu'il est partout respecté sauf une entorse mineure dans sa partie centrale, la *Bête Humaine* et une *Page d'Amour* (rangs 9 et 10) ayant passé la frontière.

Figure 9 : Analyse factorielle de l'entropie (19 textes et 29 tranches)



Observons ici encore qu'il s'agit de tendances et que l'analyse porte sur des profils, la valeur absolue de l'entropie n'intervenant pas. C'est pourquoi l'analyse ignore les riches et les pauvres et ne veut considérer que ceux qui s'enrichissent et ceux qui s'appauvrissent. Et de fait, on trouve des riches et des pauvres des deux côtés. Notons que l'« essoufflement » lexical – le tassement de l'entropie – se produit plus

souvent à la fin des *Rougon-Macquart* et que les derniers romans de la série chronologique se trouvent à gauche, là où les ressources s'épuisent, tandis que les quatre premiers titres se portent à droite[13]. Est-ce assez pour suggérer que Zola aurait en vieillissant le souffle plus court ? Non pas que ses romans tendent à devenir moins longs – ils gagnent au contraire en étendue – mais le rythme du renouvellement lexical semble baisser et une fois le thème posé, Zola paraît se laisser aller à la répétition. On aimerait disposer des données des *Trois Villes* et des *Quatre Évangiles* pour contrôler cette évolution.

Au reste, l'évolution du flux lexical, ainsi mesurée par le calcul de l'entropie, pourrait être saisie par des moyens plus simples. Quand on a établi l'effectif des occurrences et celui des vocables pour chaque moment du discours (tableau 1 et 2), toutes les méthodes sont bonnes pour mettre en relief l'évolution de chaque texte. La plus élémentaire est le rapport $V / \sqrt{N}$ qu'on a calculé dans le tableau 10 et qui donne lieu à l'analyse factorielle de la figure 11.

Tableau 10 : Rapport $V/\sqrt{N}$ de 10 en 10 pages (multiplié par 100)

	RAQ	FER	FOR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEB	PAS
10	2179	2202	2309	2283	2136	1935	2231	2179	2231	2027	2198	2139	2226	2291	2139	2451	2151	2401	2295	2354	2315	2293
20	2372	2286	2493	2574	2347	2124	2550	2447	2394	2250	2462	2409	2485	2512	2454	2630	2452	2847	2453	2656	2575	2441
30	2501	2502	2575	2753	2539	2213	2598	2629	2447	2351	2654	2569	2580	2595	2554	2918	2580	2986	2565	2846	2678	2697
40	2523	2561	2715	2841	2702	2276	2611	2696	2500	2415	2678	2630	2662	2698	2626	2982	2788	3093	2579	2962	2748	2753
50	2541	2615	2812	2977	2787	2238	2658	2704	2593	2507	2675	2677	2732	2759	2643	3035	2874	3103	2616	3064	2868	2812
60	2546	2616	2921	3026	2836	2317	2667	2699	2630	2582	2687	2675	2730	2827	2683	3093	2893	3084	2632	3112	2849	2850
70	2570	2552	2969	3055	2829	2307	2644	2736	2632	2612	2675	2660	2732	2829	2787	3094	2975	3065	2638	3138	2864	2855
80	2577	2531	3010	3089	2826	2356	2722	2805	2663	2576	2653	2638	2734	2861	2796	3092	3005	3025	2636	3111	2860	2821
90	2591	2558	3009	3103	2826	2378	2786	2825	2678	2560	2652	2667	2748	2848	2765	3055	3021	2980	2707	3106	2835	2807
100	2585	2562	2983	3133	2880	2391	2799	2808	2691	2602	2633	2657	2754	2840	2743	3049	3023	2935	2697	3077	2800	2901
110	2573	2533	2965	3129	2880	2374	2772	2814	2685	2631	2623	2639	2735	2820	2725	3025	3009	2902	2673	3062	2765	2884
120	2545	2580	2961	3101	2883	2353	2741	2837	2662	2626	2625	2620	2749	2807	2718	3032	2985	2856	2676	3023	2737	2862
130	2533	2549	2953	3064	2882	2336	2717	2840	2662	2603	2628	2600	2735	2760	2735	3016	3001	2826	2666	3009	2718	2819
140	2486	2518	2956	3046	2866	2339	2773	2822	2654	2572	2631	2588	2748	2742	2728	2995	2988	2780	2674	3017	2697	2802
150	2462	2486	2936	3016	2866	2337	2770	2803	2653	2549	2615	2592	2754	2720	2739	2978	2970	2747	2655	3003	2691	2767
160	2441	2473	2914	2987	2834	2341	2759	2814	2642	2518	2611	2572	2743	2691	2721	2964	2971	2708	2649	2992	2685	2753
170	2439	2451	2877	2967	2830	2341	2720	2803	2638	2504	2626	2555	2721	2667	2704	2956	2964	2705	2641	2983	2675	2731
180	2430	2435	2851	2953	2852	2346	2714	2790	2625	2497	2615	2552	2706	2648	2716	2945	2928	2694	2630	2950	2652	2721
190	2409	2433	2833	2920	2837	2343	2684	2773	2611	2487	2614	2560	2693	2633	2721	2932	2904	2694	2601	2938	2639	2701
200	2391	2405	2813	2909	2820	2343	2660	2755	2592	2483	2606	2543	2688	2632	2715	2934	2901	0	2584	2916	2620	2692
210	2377	2382	2791	2883	2817	2341	2637	2742	2576	2465	2595	2522	2675	2615	2699	2913	2881	0	2547	2906	2595	2674
220	2361	2359	2773	2847	2820	2346	2613	2721	2580	2449	2591	2522	2646	2598	2708	2903	2859	0	2538	2901	2599	2653
230	2349	2342	2763	2838	2836	2342	2601	2704	2562	2433	2585	2511	2636	2573	2693	2893	2854	0	2523	2886	2580	2621
240	0	2328	2757	2829	2818	2340	2588	2688	2557	2417	2576	2511	2644	2548	2701	2890	2841	0	2515	2868	2564	2600
250	0	2321	2740	2817	2822	2344	2567	2689	2541	2402	2559	2491	2634	2530	2711	2876	2825	0	2506	2852	2545	2578
260	0	2329	2738	2788	2798	2353	2544	2677	2530	2387	2559	2477	2617	2523	2705	2858	2812	0	2480	2829	2531	2564
270	0	2311	2724	2780	2796	2358	2541	2667	2507	2377	2547	2468	2603	2506	2685	2863	2798	0	2470	2808	2509	2552
280	0	2296	2698	2764	2772	2347	2536	2671	2505	2365	2534	2448	2604	2486	2665	2865	2792	0	2456	2791	2509	2535
290	0	2273	2683	2761	2749	2329	2514	2664	2493	2362	2534	2427	2581	2478	2652	2861	2778	0	2432	2781	2500	2527
300	0	0	2668	0	2744	2331	2498	2649	2481	2365	2523	2408	2565	2467	2647	2847	2769	0	2409	2773	2485	2515
310	0	0	2653	0	0	2332	2480	2637	2476	0	2519	2399	2546	2456	2640	2849	2764	0	2401	2759	2470	2511
320	0	0	2649	0	0	2323	2476	2643	2472	0	2508	2382	2541	2434	2628	2849	2750	0	2404	2745	2458	0
330	0	0	0	0	0	0	0	2636	2472	0	2504	2366	2533	2428	2625	2843	2730	0	2403	2728	2450	0
340	0	0	0	0	0	0	0	2636	2466	0	2499	2350	2525	0	2610	2820	2714	0	2399	2709	2436	0
350	0	0	0	0	0	0	0	2639	2456	0	2492	2343	2522	0	2598	2816	2705	0	0	2695	2432	0
360	0	0	0	0	0	0	0	2647	2453	0	2488	2332	2517	0	2586	2816	2686	0	0	2688	2423	0
370	0	0	0	0	0	0	0	0	2444	0	2489	2337	2503	0	2575	0	2674	0	0	2676	2408	0
380	0	0	0	0	0	0	0	0	2442	0	2484	2341	2506	0	2569	0	2657	0	0	2658	2401	0
390	0	0	0	0	0	0	0	0	2436	0	2471	2337	2495	0	2557	0	2644	0	0	2658	2389	0
400	0	0	0	0	0	0	0	0	2438	0	2474	0	2504	0	2554	0	2632	0	0	0	2380	0
410	0	0	0	0	0	0	0	0	2425	0	0	0	2497	0	2540	0	2622	0	0	0	2386	0
420	0	0	0	0	0	0	0	0	2419	0	0	0	2491	0	2540	0	2611	0	0	0	2383	0
430	0	0	0	0	0	0	0	0	2413	0	0	0	0	0	2532	0	2596	0	0	0	2373	0
440	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2518	0	2584	0	0	0	2372	0
450	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2504	0	2578	0	0	0	2381	0
460	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2494	0	0	0	0	0	2376	0
470																					2377	
480																					2382	
490																					2374	
500																					2371	
510																					2361	
520																					2363	

[13] Mais on trouve aussi de ce côté deux grandes fresques tardives, *Germinal* et la *Terre*.

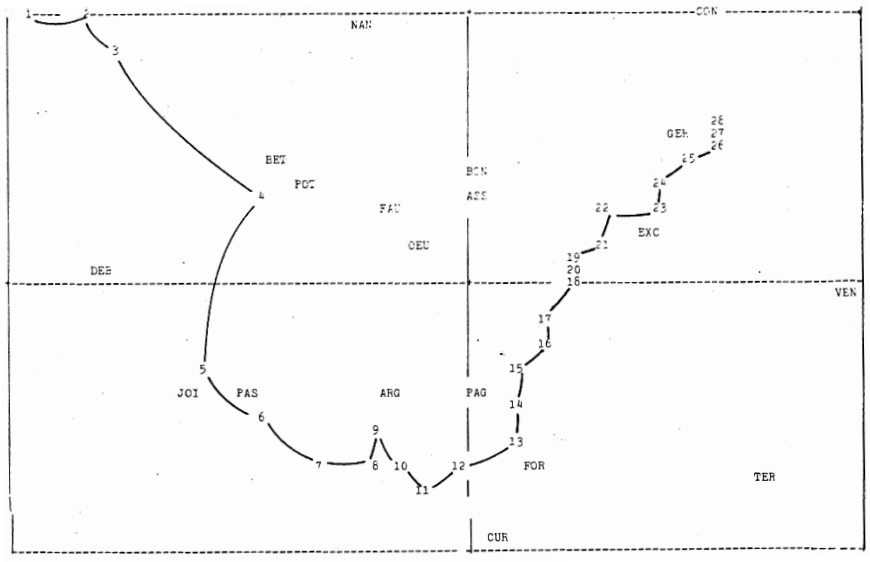


Figure 11 : Analyse factorielle du rapport $V/\sqrt{N}$ (19 textes et 29 tranches)

Comment ne pas voir dans ce graphique la pure superposition de celle de la figure 9, qui avait été réalisée à partir de l'entropie ? Même forme en croissant ordonnant les tranches de la 1ère à la 29e. Même disposition des textes : à gauche ceux dont la variété lexicale s'épuise plus vite, à droite ceux qui renouvellent davantage leur vocabulaire.

Ainsi retrouvons-nous Pierre Guiraud au terme de notre parcours. La formule si simple qu'il a proposée jadis et qu'on a parfois décriée donne au bout du compte les mêmes indications que les calculs les plus sophistiqués – que Guiraud avait aussi suggérés. N'a-t-on donc pas fait de progrès depuis vingt ans dans une discipline que Guiraud a fondée pour l'abandonner une fois qu'elle fut devenue grande ? Certes on a gagné en puissance et les dénombrements qu'il a faits, lui, à la main, se font maintenant à grande échelle et à grande vitesse avec des machines. On a gagné aussi en précision et en fiabilité avec la facilité des contrôles et des regroupements. Mais n'a-t-on pas perdu en puissance de raisonnement, en fertilité d'imagination, en rapidité d'anticipation ? La longueur et l'inconfort des calculs manuels ont précipité Guiraud sur la voie des raccourcis, des audaces, des hypothèses et des découvertes.

Mais hélas, quand sont venus les ordinateurs, Guiraud s'en est allé.