



**HAL**  
open science

## Flaubert traité par Hyperbase

Etienne Brunet

► **To cite this version:**

| Etienne Brunet. Flaubert traité par Hyperbase. Revue Flaubert, 2003, 3, pp.1-19. hal-01362745

**HAL Id: hal-01362745**

**<https://hal.univ-cotedazur.fr/hal-01362745>**

Submitted on 9 Sep 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Flaubert traité par HYPERBASE<sup>1</sup>

Etienne Brunet

Il y a bien longtemps, à la fin des années 70, nous avons fait le projet d'une monographie flaubertienne à l'image de celles que nous avons réalisées sur Proust, Zola et Hugo. Stendhal, Chateaubriand et Claudel étaient aussi en bonne place dans nos cartons. Tous les matériaux nous étaient fournis par la mine de Nancy qui était alors peu exploitée en dehors des besoins propres des rédacteurs du T.L.F. Bien des choses ont changé depuis, sauf les données. Avec la microinformatique et Internet le transfert et le traitement des textes ont gagné en puissance et en souplesse. Mais les données de Nancy sont restées celles que des bataillons de dactylos avaient dépouillées il y a trente ans, sur des bandes perforées. Le carton a fait place à la bande magnétique, puis au disque dur et enfin à la technologie du laser, tandis que la B.N.F. et FRANTEXT prenaient le relais du T.L.F., par l'entremise d'Internet. Mais les données initiales sont restées disponibles et elles demeurent à peu près fiables malgré l'injure du temps<sup>2</sup>. Toutefois s'agissant de Flaubert elles ne sont pas complètes et nous avons eu recours à d'autres serveurs pour constituer le corpus : principalement le site ABU et un site dévoué à Flaubert et créé par J.B. Guinot. Nous laissons à ce chercheur le soin de préciser et de justifier les éditions choisies, que nous avons reprises sans autre forme de procès. Notre intervention dans les données s'est bornée à une harmonisation minimale, afin de disposer d'un corpus homogène<sup>3</sup>.

---

1. Article paru dans *Revue Flaubert* 3, 2003, Les apports de l'informatique et de l'Internet aux nouvelles formes d'édition, de lecture et d'interprétation des textes flaubertiens (<http://flaubert.univ-rouen.fr/revue/revue3/>). Seule la mise en page a été revue et corrigée.

2. En passant d'un support à l'autre, les textes ont reçu parfois des égratignures. Un toilettage a été nécessaire dont ont bénéficié les textes de Nancy communiqués au site Gallica de la B.N.F.

3. Cela concerne surtout la ponctuation et la mise en page : par exemple un double trait d'union tient lieu de tiret de dialogue dans les textes d'ABU.

En réalité nous avons constitué deux corpus, le second étant consacré uniquement à la correspondance de Flaubert, que D. Girard et Y. Leclerc viennent de publier sur Internet à partir de l'édition Conard. Certes une partie de cette correspondance était intégrée dans la première monographie, afin de donner une image plus riche de l'écrivain. Mais cette correspondance est si volumineuse qu'elle dépasse en taille l'ensemble des autres écrits de Flaubert et l'étude comparative que nous envisagions aurait souffert de cette disproportion, où le genre épistolaire aurait fait la loi. Il valait mieux que les genres et les œuvres où Flaubert s'est illustré soient représentés de façon équilibrée, afin de les comparer plus aisément. Et pareillement, en isolant la correspondance dans un corpus particulier, on se donnait le moyen de neutraliser le genre et le thème, pour mieux cerner la chronologie et l'évolution de l'écrivain.

On trouvera ci-dessous (figure 1) le menu principal de la première des deux monographies. Comme ces deux bases ont été mises sur Internet à la libre disposition des chercheurs<sup>4</sup>, notre ambition est moins d'exploiter et de commenter les résultats auxquels elles conduisent que d'expliquer leurs fonctions et leur mode d'emploi, afin que les spécialistes de Flaubert en fassent un usage mieux qualifié.

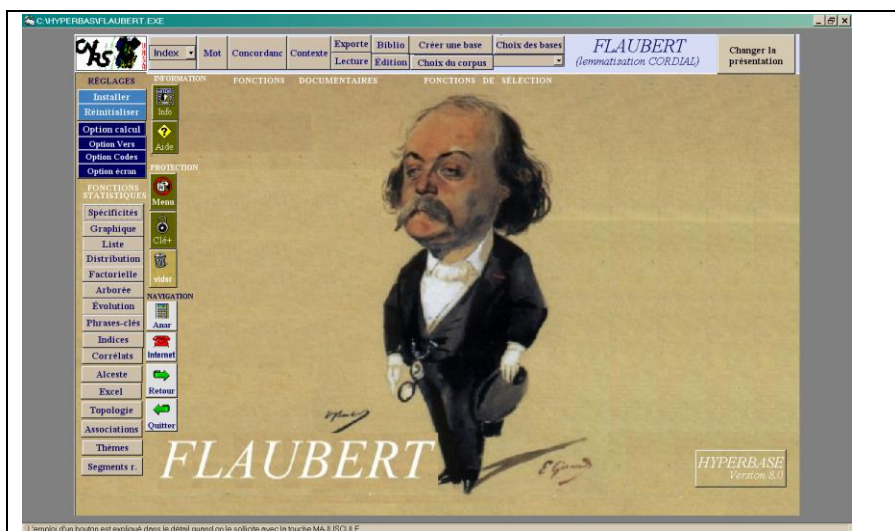


Figure 1. Le menu principal de la base Flaubert

4. Voici l'adresse de ces deux bases (respecter la casse) :

<http://www.unice.fr/ILF-CNRS/HYPERBAS/flaubert.exe>

<http://www.unice.fr/ILF-CNRS/HYPERBAS/flaucorr.exe>

On distinguera deux séries de fonctions : les unes, documentaires, sont groupées horizontalement au haut de l'écran, les autres, vouées à la statistique, occupent la marge gauche. On n'insistera guère sur les premières, car leur utilité s'accorde avec leur facilité. Il serait oiseux de s'appesantir sur les programmes de concordance ou de recherche de contextes. Un exemple, relatif au mot de Cambronne, suffit à illustrer cette fonctionnalité (figure 2). Flaubert n'est pas bégueule, mais soucieux des genres et des convenances, il n'emploie ce mot que dans sa correspondance.

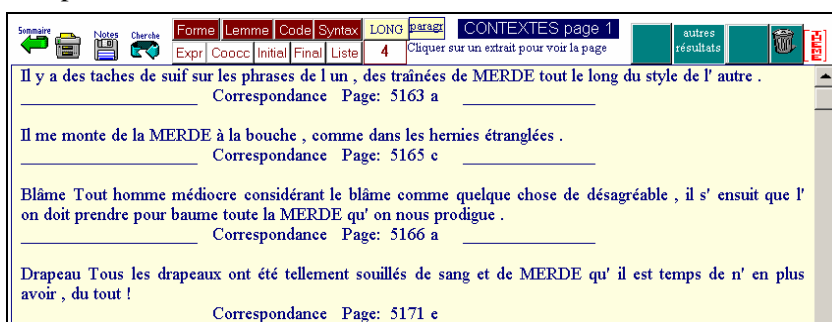


Figure 2. Les contextes du mot de Cambronne

Mais même dans cette fonction documentaire très traditionnelle, la statistique pointe son nez. Quand un mot (ou un ensemble de mots) produit une moisson suffisante de contextes (par exemple plus d'un millier pour le mot *femme*), la fonction THEME observe tous les mots présents dans l'entourage immédiat du mot choisi pour pôle et compare la fréquence de ces corrélats dans ce sous-ensemble à celle qui est la leur dans le corpus entier. Dans la liste qui en résulte on découvre une constellation thématique qui circonscrit l'univers féminin, avec ses composantes dans l'imaginaire flaubertien : l'*amante*, l'*épouse*, la *mère*. Le tableau 1 restitue deux de ces constellations : la première, consacrée à la forme au singulier, est plus narrative, la seconde rend compte du pluriel et paraît appartenir à l'ordre de la description et à l'évocation de tableaux collectifs<sup>5</sup>.

5. On s'attend évidemment à ce que le singulier appelle le singulier, et le pluriel le pluriel, là où les règles imposent l'accord, ce qui est le cas des déterminants, adjectifs et verbes en relation directe avec le mot-pôle. Mais la concordance du nombre – comme on dit la concordance des temps – dépasse largement la portée de la grammaire (en particulier les règles n'exigent pas que le pluriel du sujet doive nécessairement entraîner le pluriel dans les compléments). Dans la liste de gauche tous les termes sont au singulier (sauf

écart	corpus	texte	mot	HIERARCHI	écart	corpus	texte	mot	HIERARC
199.01	658	644	FEMME		185.94	472	471	FEMMES	
21.52	8126	368	UNE		26.41	15301	581	DES	
15.28	4047	184	SA		22.82	44	18	NUES	
13.90	156	24	VIEILLE		21.28	13	9	MOITIÉS	
12.78	29	9	VÊTUE		18.49	226	35	ENFANTS	
10.42	298	27	JEUNE		16.12	7	5	AIMÉES	
9.93	55	10	NUE		15.50	566	50	HOMMES	
8.68	251	21	BONNE		15.03	20	8	ASSISES	
8.30	29	6	SAUVAGE		14.70	20706	521	LES	
8.10	1246	55	CETTE		12.61	1877	88	LEURS	
8.00	22	5	MARIÉE		11.34	26	7	VIEILLARDS	
7.65	209	17	BELLE		9.30	85	11	ROBES	
7.40	815	39	MA		8.14	105	11	NUS	
6.94	28	5	JOLIE		8.14	105	11	BELLES	
6.83	53	7	SEINS		7.76	27	5	LUXE	
6.52	57	7	AMANT		7.39	3538	98	AVEC	
6.47	235	16	PAUVRE		7.33	1072	42	CES	
6.36	110	10	MARI		7.26	56	7	VOILES	
6.28	295	18	ENFANT		6.57	83	8	PARFUMS	
6.02	49	6	MONTRE		6.55	123	10	ALLAIENT	
5.73	87	8	AIMER		6.48	67	7	JAUNES	
5.66	332	18	MAISON		6.41	932	35	TOUTES	
5.45	57	6	ASSISE		6.24	54	6	HABITS	
5.44	183	12	PREMIÈRE		6.19	181	12	CHEVAUX	
5.26	98	8	MÉDECIN		6.14	40	5	NÈGRES	
5.11	280	15	PETITE		6.09	1534	48	AUX	
5.09	63	6	RICHE		5.78	44	5	CORBEILLES	
5.00	950	34	HOMME		5.47	48	5	VIEILLES	
4.87	21774	430	LA		5.39	49	5	FLEURS	
4.87	49	5	L		5.31	90	7	VILLES	
4.80	50	5	POUSSÉ		5.24	115	8	TÊTES	
4.69	114	8	PRÊTRES		5.00	268	13	FLEURS	
4.63	250	13	AIME		4.90	1897	50	LEUR	
4.53	226	12	ENFANTS		4.89	689	24	DONT	
4.48	10426	220	D'		4.88	153	9	NOIRS	

Tableau 1. L'environnement lexical de *femme* et *femmes*

Même la lecture est assistée par la statistique. La page qu'on a sous les yeux met en relief les mots qui sont spécifiques du texte considéré, comme les termes *objet* et *connaître* dans l'exemple de la figure 3, emprunté à *Bouvard et Pécuchet*. Dans d'autres pages du même texte se trouvent soulignés les mots qui marquent un progrès, et plus souvent un échec, dans la quête du savoir à laquelle se livrent les deux héros du roman : *apprendre, observer, indiquer, bibliothèque, scientifique, découverte, réflexion, méthode, expérience*, etc.

On remarquera que le corpus de Flaubert a été lemmatisé (grâce au logiciel *Cordial*), le texte de la figure 3 apparaissant dans deux séquences alignées, à gauche les formes, à droite les lemmes. En réalité, deux autres champs, pareillement alignés, contiennent les codes grammaticaux et les structures syntaxiques. On peut ainsi choisir une forme et repérer immédiatement tous les passages où elle apparaît, mais le choix peut se faire aussi sur les trois autres objets. Ainsi en cliquant sur le code *Vmis3sv* (V=verbe, m=principal, i=indicatif, s=passé simple, 3 = troisième personne, s=singulier), qui correspond au premier verbe (*laissa*)

*seins, enfants* et *prêtres*) et dans celle de droite le pluriel est systématique (seule exception : *luxe*).

du précédent extrait, on peut faire défiler les 13.606 contextes où un passé simple est employé à la troisième personne. Mais leur nombre est tel qu'on préférera les compter, observer leur distribution dans les textes du corpus et faire la comparaison avec d'autres catégories.

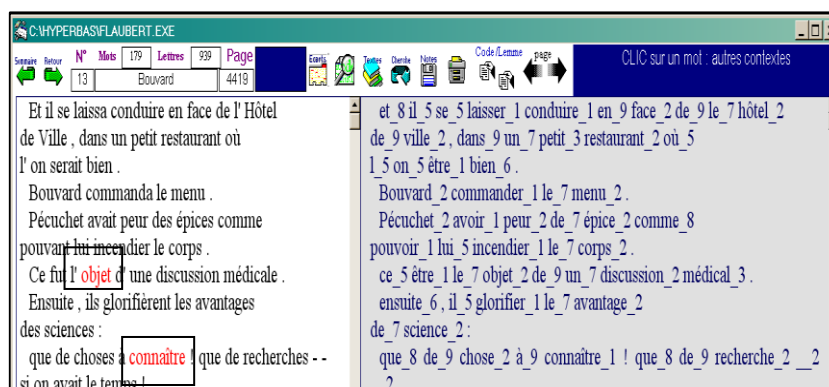


Figure 3. Mise en relief des mots significatifs

Dans la phase de préparation et d'indexation, le logiciel s'emploie à de tels décomptes, en relevant dans le corpus 976.382 occurrences (ou mots), 36.575 formes différentes et 19.405 lemmes. Naturellement ces relevés sont faits aussi pour chaque texte, chaque classe de fréquence, chaque catégorie, etc. En prenant appui sur les lemmes (le calcul peut se faire aussi sur les graphies ou sur les codes), il est possible de calculer la distance qui sépare un texte de tous les autres respectivement. Pour chaque couple de textes, on prend en compte tous les mots rencontrés et leur répartition, partagée ou exclusive, dans les deux textes. La figure 4 dessine la carte typologique qui résume l'ensemble de ces mesures de proximité ou d'éloignement. Ainsi pourrait-on représenter la carte géographique d'un pays quand on connaît le tableau des distances de ville à ville.

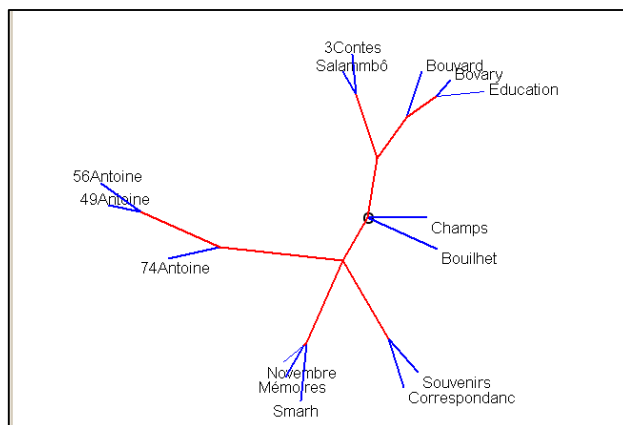


Figure 4. Analyse arborée de la distance intertextuelle (calculée sur les lemmes)

Observons toutefois que seule compte la longueur des segments qu'il faut suivre pour aller d'un point à un autre ; les angles, les directions et l'orientation sont arbitraires et indifférents. Cela ne gêne en rien la lisibilité du graphe, qui met en relief les textes qui partagent soit le même thème (les trois versions de la *Tentation de Saint Antoine*, au haut du graphique), soit la même époque (les écrits de jeunesse *Mémoires d'un fou*, *Smarh* et *Novembre*, à gauche), soit le même genre (les trois romans à sujet moderne à droite, *Bovary*, *l'Éducation* et *Bouvard*, sont séparés de ceux qui traitent une époque plus ancienne, *Contes* et *Salammbô*). C'est aussi le lien du genre qui rapproche l'écriture personnelle des *Souvenirs* et de la *Correspondance*, en bas à droite. Quant au récit de voyage *Par les champs et par les grèves* et à l'évocation de son ami *Bouilhet*, ces deux textes sont isolés, au centre du graphique, sans liaison avec le reste de l'œuvre.

Prenant appui sur le même tableau de distances, l'analyse factorielle propose une typologie tout aussi claire (figure 5). Cette fois l'espace est orienté et les points cardinaux ont une signification : le premier facteur oppose la droite et la gauche et met en relief le déchirement de Flaubert, « suspendu entre le double abîme du lyrisme et du vulgaire »<sup>6</sup>. La tentation du lyrisme est à droite, associée surtout aux œuvres de la première période (mais avec des résurgences tardives dont témoigne la version 1874 de la *Tentation*), l'écriture plus sèche et moins personnelle vers laquelle tend Flaubert est à gauche, atteignant son point extrême avec la trilogie *Bovary-Éducation-Bouvard*. Une opposition verticale

6. Lettre du 20 mars 1851 à Louise Colet.

vient corriger et compléter cette première dichotomie : elle semble tenir au genre, les œuvres de fiction s'installant au haut du tableau, quand les autres occupent le bas.

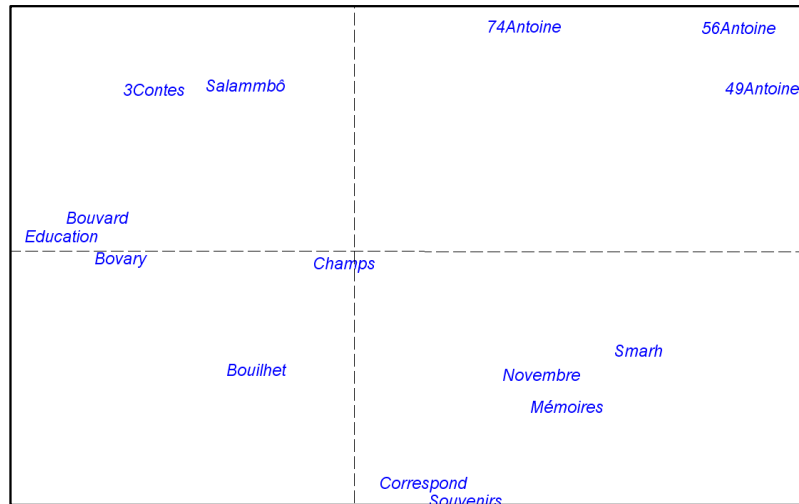


Figure 5. Analyse factorielle de la distance intertextuelle

D'aucuns penseront que le choix des sujets gouverne partiellement les alliances et les oppositions. Aussi convient-il d'éliminer l'influence du thème en évacuant le sens des mots et en ne retenant du texte que l'aspect grammatical. Les données sont fournies par le logiciel sous la forme d'un tableau à deux dimensions, dont chaque case indique l'effectif d'une partie du discours dans un texte particulier. L'aimantation que produit ce filtre grammatical dans la figure 6 n'est pas sans rappeler celle de la figure 5 – ce qui semble indiquer que les lignes de force appartiennent au style plutôt qu'au thème, puisqu'elles se maintiennent quand le sens des mots est écarté. La disposition des textes est en effet semblable, au moins pour le premier facteur : c'est comme précédemment la dérive chronologique qui parcourt l'espace de droite à gauche et dont l'influence n'est plus masquée par d'autres facteurs. Les textes que le thème ou le genre avaient agglutinés se disloquent en tronçons séparés qu'emporte le courant : ainsi en est-il des trois versions de la *Tentation* qui s'échelonnent de 1849 à 1874, et des trois romans de la trilogie qui suivent à distance le même chemin.



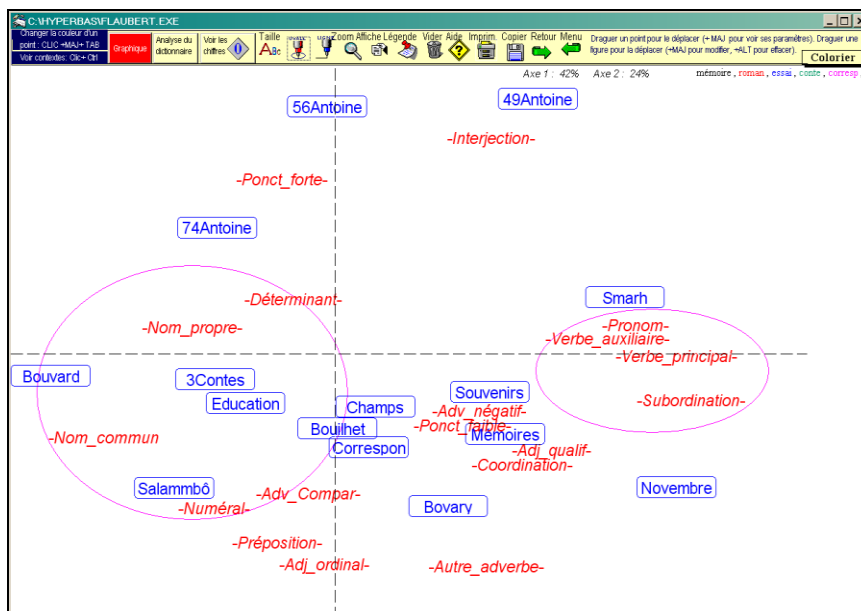


Figure 6. Les parties du discours. Analyse factorielle

L'analyse factorielle est ici plus qu'une observation : c'est aussi une explication. Considérons en effet les éléments qui figurent en rouge sur le graphe et qui représentent les lignes du tableau. Comment ne pas voir qu'ils s'ordonnent pareillement en deux camps, dont la rivalité a été maintes fois constatée dans d'autres corpus. Le verbe campe solidement à droite, en compagnie de ses acolytes habituels : pronoms, adverbes et subordinants. Le nom règne à gauche, qu'il s'agisse du nom propre ou du nom commun. Les déterminants l'accompagnent, les numéraux aussi et les prépositions<sup>7</sup>. On peut ainsi caractériser l'évolution stylistique de Flaubert par un passage progressif des catégories du verbe à celle du substantif, tendance que l'on a observée aussi à plus grande échelle dans la littérature française, du XVI<sup>e</sup> siècle à nos jours<sup>8</sup>.

7. La syntaxe explique la liaison forte qui s'établit entre le substantif et les déterminants. Elle ne justifie qu'en partie la relation substantif-préposition, car la préposition peut introduire aussi bien un pronom et un infinitif. De plus, beaucoup de prépositions entrent dans la composition des subordinants et annoncent une proposition subséquente, et donc un verbe.

8. Sur ce point nous renvoyons le lecteur à *Nouvelle Histoire de la langue française* et à notre contribution : « Ce que disent les chiffres », Le Seuil, 1999, p.675-727.

Un fait pourtant échappe aux prévisions : c'est le sort réservé à l'adjectif qualificatif. Sa place habituelle est dans la sphère d'influence du substantif. Or il refuse de suivre le chef de file dans la section gauche et se tient à l'écart dans la zone droite, progressivement désertée. Il faudrait voir de plus près les manuscrits de Flaubert pour examiner si les adjectifs ont fait l'objet d'un ostracisme grandissant, comme nous l'avons observé pour Zola. Flaubert, en s'éloignant du romantisme a-t-il raturé le signe le plus visible de l'écriture romantique, dénoncé avec humour par Musset dans *Dupuis et Cottonnet* ? On ne trouve pas sous sa plume de déclaration de guerre, comme on en a trouvé chez Giraudoux : « Déteste l'adjectif et chéris la raison »<sup>9</sup>. Tout au plus Flaubert avoue-t-il son embarras pour manier cet outil : « Oh ! pauvre amie, si tu pouvais assister à ce qui se passe en moi, tu aurais pitié de moi, à voir les humiliations que me font subir les adjectifs et les outrages dont m'accablent les *que* relatifs. »<sup>10</sup> Mais les chiffres ne souffrent guère la contestation : la courbe de l'adjectif suit une pente descendante dans la figure 7, qui rend compte d'un effectif de 41000 adjectifs. La pente serait plus forte encore si le « bâton » dévolu à la correspondance n'avait pas été placé arbitrairement à la fin de la série (en réalité toutes les époques sont représentées dans ce recueil épistolaire). La diagonale descendante rend compte d'une désaffection croissante à l'égard d'un colifichet suranné, que Flaubert assimile aux rubans et qui allait faire fureur dans l'écriture artiste et pittoresque des Goncourt et de Daudet : « Ce sont toujours des singes, des perroquets, des adjectifs et des rubans ! »<sup>11</sup>

---

9. *Juliette au pays des hommes*, Paris, Grasset, 1967, p. 138.

10. *Correspondance*, 1847. Remarquons que la distribution des pronoms relatifs est parallèle à celle des adjectifs (et des subordonnants). Les coefficients de corrélation chronologique sont respectivement de -0,60 (relatifs), -0,43 (adjectifs) et -0,38 (subordonnants).

11. *Tentation de Saint Antoine*, 1849.

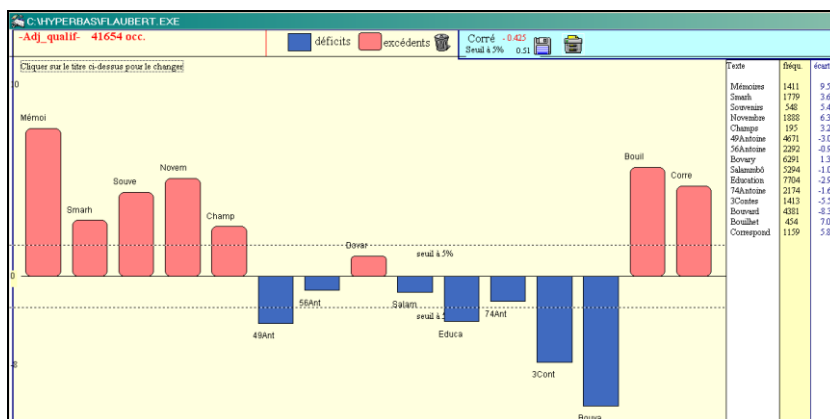


Figure 7. Le déclin de l'adjectif chez Flaubert

Mots en régression			
Coeff.	Fréq.	Mot	
- 0.693	135	vide	- 0.648
- 0.692	32	rêverie	- 0.644
- 0.795	57	frais	- 0.643
- 0.789	113	sourire	- 0.642
- 0.774	107	volupté	- 0.642
- 0.768	204	rêver	- 0.642
- 0.766	52	crépuscul	- 0.641
- 0.748	70	humide	- 0.638
- 0.747	264	joie	- 0.638
- 0.746	43	depuis	- 0.638
- 0.744	1846	voir	- 0.638
- 0.738	699	entendre	- 0.637
- 0.736	228	feuille	- 0.636
- 0.736	34	sécher	- 0.635
- 0.733	224	herbe	- 0.634
- 0.727	800	regarder	- 0.634
- 0.725	127	baiser	- 0.632
- 0.719	1404	si	- 0.632
- 0.709	31	trace	- 0.631
- 0.708	39	bercer	- 0.629
- 0.708	20	enivrement	- 0.629
- 0.701	51	vanité	- 0.628
- 0.700	41	délice	- 0.627
- 0.698	148	larme	- 0.622
- 0.697	49	misérable	- 0.622
- 0.696	237	regard	- 0.621
- 0.695	753	encore	- 0.621
- 0.648	702	aimer	
- 0.644	35	tombe	
- 0.643	37	éclair	
- 0.642	512	nuit	
- 0.642	38	amertume	
- 0.642	28	lasser	
- 0.641	370	appeler	
- 0.638	649	sentir	
- 0.638	150	pluie	
- 0.638	62	gonfler	
- 0.638	23	contour	
- 0.637	252	coucher	
- 0.636	315	battre	
- 0.635	24	fatigué	
- 0.634	53	peser	
- 0.634	51	vide	
- 0.632	698	coeur	
- 0.632	132	triste	
- 0.631	1375	venir	
- 0.631	21	fourmi	
- 0.629	35	rosée	
- 0.628	104	libre	
- 0.627	73	avenir	
- 0.622	187	déjà	
- 0.622	57	boue	
- 0.621	55	ruisseau	
- 0.621	32	ardent	

Tableau 2. Les vocables en régression

Les adjectifs fournissent un lot important à la liste des vocables en régression qui disparaissent sous la plume de Flaubert et qu'on a représentés dans le tableau 2. Derrière la tête de liste *frais*, suivent les notations vaguement descriptives dont beaucoup évoquent un sentiment ou une couleur (*humide, misérable, vide, pauvre, vain, triste, vague, confus, infini, inconnu, fou, fier, doux, suave, ardent, libre, tendre, amoureux, blanc, rose, doré.*) Mais les substantifs (le premier est *volupté*)

et les verbes (derrière *sourire* et *rêver*) remplissent pareillement ce magasin des antiquités romantiques que Flaubert, sans doute à regret, s'est résigné à remiser au grenier. Pour la plupart ce n'est pas leur étiquette grammaticale qui est en cause, mais les traits sémantiques dont ils sont porteurs, dont la mode s'est détachée et dont Flaubert a perdu le goût.

On pourrait suivre de texte en texte l'évolution de Flaubert en parcourant la galerie des portraits lexicaux que dessine le calcul des spécificités. Mais la place nous manque et nous nous contenterons d'un portrait robot où les caractéristiques de Flaubert lui-même, tous textes confondus, se détachent sur la toile de fond de la littérature française (tableau 3).

Vocab. Spécif. POSITIF		29.01	915	160	logique	21.44	791	114	prêtres
Ecart Corpus Texte Mot		28.91	7307	569	cependant	21.30	1508	167	aperçut
		27.96	182948	6449	s'	20.80	2158	205	poussière
80.94	323745 15301	27.25	7793	571	or	20.75	3975	304	reprit
76.29	492769 20705	26.80	1790	222	louise	20.68	10087	577	oh
74.28	243 187	26.71	562	113	j	20.45	258086	8085	dans
74.00	75 102	26.33	3009	301	poitrine	20.13	1633	168	ailes
70.58	134 131	25.76	19317	1043	donc	20.07	1233	141	larges
61.92	5755 878	25.58	610	114	répliqua	19.58	793	106	ça
47.45	1930 374	25.09	1767	209	dieux	19.41	5029	341	figure
47.12	26818 1877	24.08	1012	145	infini	18.82	4910	329	bord
45.31	1589 322	23.70	8428	551	milieu	18.72	34224	1398	sous
41.20	300 119	23.65	1124	152	montagnes	18.71	2938	232	quelquefo
38.09	282 107	23.63	2562	251	orgueil	18.57	2021	181	chevaux
37.48	405 128	22.85	14451	791	bras	18.40	1919	174	taille
37.33	287 106	22.33	136680	4712	comme	18.20	701026	19933	et
36.95	202774 7677	22.04	628	102	miraille	18.19	2086	182	étoiles
34.90	574 145	22.03	1829	193	c				
31.59	89300 3711	21.90	3187	273	robe				
30.17	412 106	21.80	638	102	chevelure				
30.09	90585 3683	21.76	1555	173	écria				
29.91	418 106	21.65	115651	4046	sa				
		21.44	36888	1567	puis				
Vocab. Spécif. NEGATIF		-24.10	40921	263	)	-16.03	35823	423	jamais
N° Ecart Corpus Texte Mot		-23.41	40245	274	(	-15.79	55831	815	ma
		-22.49	92423	1246	cette	-15.75	97700	1677	bien
		-22.12	157050	2563	j'	-15.70	22175	190	déjà
-46.63	431159 6011	-22.02	92788	1275	m'	-15.57	92692	1580	ai
-45.24	396110 5469	-20.81	157506	2653	mais	-15.41	44793	612	peut
-39.10	342841 5007	-19.75	79885	1128	mon	-15.40	51993	753	encore
-36.13	305816 4535	-19.21	613321	13052	à	-15.16	1279768	29364	de
-33.25	229220 3252	-18.82	166682	2973	a	-14.68	49641	792	peu
-32.99	58150 213	-18.42	205106	3832	n'	-14.34	63189	1019	fait
-32.31	267396 4085	-18.38	34354	328	très	-14.29	19588	178	moment
-29.88	300597 4967	-18.34	42317	470	aussi	-14.09	15502	114	dis
-28.49	171164 2444	-17.49	73938	1108	être	-13.88	126263	2391	c'
-28.06	141768 1899	-17.41	25852	210	notre	-13.75	27426	331	avais
-26.12	1652388 36127	-16.75	36997	423	dire	-12.84	26769	342	petit
-25.88	136975 1933	-16.69	19555	125	ici	-12.81	31993	443	été

Tableau 3. Flaubert comparé à l'ensemble des écrivains français  
Excédents (en haut) et déficits (en bas)

On reconnaît quelques accidents thématiques qui tiennent à certains sujets : c'est à *Salammô* que les *barbares* et les *esclaves* doivent leur promotion, à *Madame Bovary* le *pharmacien* et le *diable* à *Saint Antoine*. Quant à l'humble *Félicité* des *Trois Contes*, profitant de l'homographie, elle s'efface et se cache derrière la minuscule. Nul besoin de tant de calculs pour ces observations triviales. La présence insistante des parties du corps n'est pas surprenante, mais c'est plutôt chez Zola qu'on attendait – et qu'on trouve – la *poitrine*, le *bras*, le *ventre*, la *tête* et la *robe*. Mais la singularité qu'on pouvait difficilement soupçonner et qui apparaît dès les premiers mots de la liste, c'est le goût du pluriel : les écarts réduits sont énormes pour *des*, *les*, *ils*, *leurs*, *ses*, *leur*, *elles*, *eux*. Il est vrai que ces formes appartiennent presque toutes à la troisième personne et que leur usage est naturel chez un romancier. Mais les formes correspondantes du singulier (*il*, *elle*, *la*, *le*, *un*, *sa*) sont nettement moins prisées. Le rapport pluriel/singulier n'est d'ailleurs pas constant chez Flaubert : modéré lorsqu'un héros est au devant de la scène (*Bovary*, *Éducation*), il grossit démesurément dans les grandes fresques collectives et descriptives comme *Salammô* et la *Tentation*.

L'extrait présenté ici est trop court pour admettre beaucoup de verbes. On en trouve pourtant dans la liste, presque toujours à l'imparfait (successivement *étaient*, *apercevait*, *marchait*, *faisaient*, *entendait*, *portait*, *allaient*, *montait*, *avaient*, etc.), ce qui confirme l'analyse de Proust, à propos de l'« éternel imparfait » flaubertien : « J'avoue que certain emploi de l'imparfait de l'indicatif – de ce temps cruel qui nous présente la vie comme quelque chose d'éphémère à la fois et de passif qui, au moment où il retrace nos actions, les frappe d'illusion [...] – est resté pour moi une source inépuisable de mystérieuses tristesses. »<sup>12</sup> Mais là encore la manière de Flaubert est variée et ce qu'en dit Proust s'applique surtout à l'*Éducation sentimentale*, ce « grand trottoir roulant », au « défilement continu, monotone, morne, indéfini »<sup>13</sup>. La narration se fait au présent dans la *Tentation*.

---

12. À la recherche du temps perdu, I, p. 43.

13. *Chroniques*, p. 194.

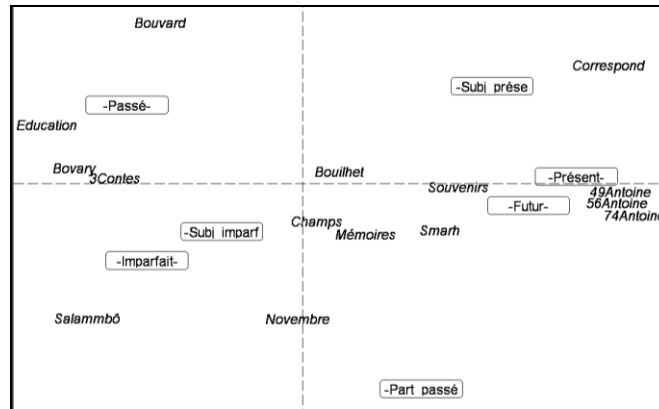


Figure 8. Analyse factorielle des temps verbaux

On s'en convaincra en considérant l'analyse factorielle des temps verbaux réalisée dans la figure 8. L'interprétation est d'une grande simplicité : à droite des temps du présent, où évoluent les premiers écrits de Flaubert, à gauche les temps du passé, et les dernières publications de Flaubert. Là encore on observe une dichotomie et une évolution d'un style à l'autre.

On laissera au lecteur le soin d'interroger la base et d'y exploiter beaucoup de fonctions qu'on ne peut détailler ici. L'accès aux codes grammaticaux et aux structures syntaxiques permet des investigations jusque là impraticables. La phraséologie, le rythme du discours, voire les sonorités, tout est sujet à mesure, et parfois à découverte, y compris même la thématique. Il ne s'agit pas seulement de circonscrire une constellation de corrélats autour d'un mot, comme indiqué dans le tableau 1. Ni de caractériser un texte par un ensemble d'extraits spécifiques<sup>14</sup>. L'ambition, permise par un traitement sémantique de *Cordial*, vise à rendre compte des idées, des sentiments, des actions, bref des thèmes exprimés dans un texte. En réalité *Cordial* fait appel à un thésaurus de référence, où sont cataloguées les disciplines, les concepts et les connaissances. Tout un jeu d'étiquettes hiérarchisées est mis en place, parmi lesquelles chaque mot du texte doit faire son choix. Sans doute ces

14. Le calcul des spécificités est appliqué non seulement aux formes et aux lemmes, mais aussi aux codes grammaticaux et aux phrases. Ainsi l'extrait retenu par le filtre comme étant le plus typique de la première *Tentation* est le suivant :

*Ni mon CORPS ni mon ESPRIT ne sont plus, mon CORPS est de la MATIÈRE de toute MATIÈRE, mon ESPRIT de l'essence de tout ESPRIT, mon âme est toute l'âme !*

étiquettes sont-elle parfois trop proches des représentations modernes, et s'appliquent –elles malaisément aux sociétés du passé, sans compter les bévues auxquelles l'homographie et même la polysémie peuvent donner lieu. Derrière « cinétique » on peut comprendre « mouvement » ; mais que recouvrent les termes d' « interdépendance », de « production » et de « grandes notions » ? Pourtant, malgré les faiblesses et les incertitudes du codage sémantique, les résultats auxquels il conduit ne sont pas dénués d'intérêt. On les a reproduits dans la figure 9. Cette analyse montre que la même aimantation des textes, déjà observée au niveau lexical et syntaxique, se retrouve au niveau thématique : les trois versions de la *Tentation* occupent un même quadrant, lieu de débat où la volonté est aux prises avec l'idéal (*spiritualité, homme, droit, éthique*). La trilogie, de *Bovary* à *Bouvard*, s'installe dans le quadrant opposé, là où sont en cause les faits de société (*pouvoir, conflits, quotidien, société, action, mouvement*). Les écrits personnels, ceux des premiers textes autobiographiques et de la correspondance, se groupent dans le quadrant inférieur droit, où il questionne de l'art, de la santé, du corps et de l'espace.

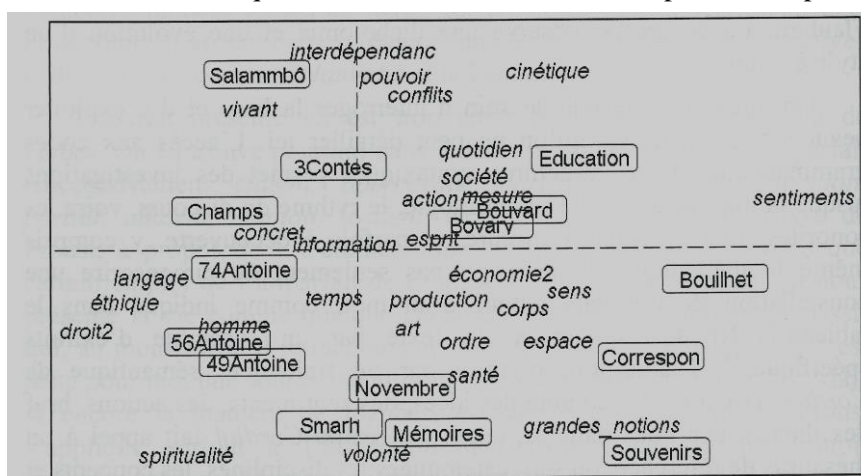


Figure 9. Analyse factorielle des thèmes

L'importance du sujet semble moins considérable dans la base réservée à la correspondance (Flaucorr.exe), que nous nous proposons d'examiner maintenant. Car Flaubert écrit au jour le jour au gré des circonstances, au hasard des rencontres. Quand il est en voyage, les lettres qu'il envoie sont de vrais messages personnels, sans idée de publication, comme c'est le cas de la fausse correspondance que Hugo développe dans le *Rhin*. Les thèmes abordés sont ceux du moment. Nul souci de différer, d'orienter, de taire ou d'embellir. L'importance du

genre est aussi abolie. La correspondance est un genre à part entière, dont la définition ne fait pas difficulté. Le genre est ainsi neutralisé, puisqu'on l'a isolé et que tous les textes qu'on va comparer sont « à genre constant » et ne varient pas sous ce rapport. Reste une seule variable à étudier : la chronologie. Le corpus fourni par D. Girard et Y. Leclerc est divisé par années, soit une quarantaine d'années entre 1830 et 1880 (certaines années sont absentes). Ces cinquante années ne s'écoulent pas comme un long fleuve tranquille.

Il y a des événements marquants comme la liaison avec Louise Collet, le voyage en Orient, la guerre de 1870, ou la ruine du mari de sa nièce, sans parler des événements tout aussi considérables que sont pour Flaubert le succès ou l'échec de ses livres. Ces facteurs extérieurs fournissent la matière de la correspondance, mais la manière leur échappe en grande partie. Les déceptions, les joies, les voyages, les travaux, les amours et les peines ont accompagné Flaubert tout au long de son existence. Mais il ne les accueille pas de la même façon, suivant qu'il est jeune ou qu'il est vieux. Et c'est ce que montre le coefficient de corrélation, appliqué à tous les vocables du corpus. Ce coefficient met en relief les mots que Flaubert emploie de plus en plus et ceux qu'il emploie de moins en moins. Nous n'avons retenu que ces derniers dans le tableau 4.

Coeff.	Fréqu.	Mot	- 0.694	87	courir	- 0.655	74	travers
			- 0.694	79	étrange	- 0.653	50	odeur
- 0.834	4131	y	- 0.691	259	déjà	- 0.649	218	chaque
- 0.787	3608	comme	- 0.689	47	doigt	- 0.648	77	épaule
- 0.786	7451	qui	- 0.682	7931	que	- 0.648	55	sec
- 0.754	1206	quelque	- 0.681	3397	tout	- 0.645	63	pitié
- 0.749	1099	beau	- 0.680	136	ouvrir	- 0.644	703	adieu
- 0.749	100	quelque	- 0.679	6368	dans	- 0.639	5701	en
- 0.736	160	quoique	- 0.678	69	vert	- 0.638	107	lorsque
- 0.734	301	surtout	- 0.677	517	sous	- 0.637	5518	tu
- 0.733	167	feu	- 0.673	1681	toi	- 0.633	880	tout
- 0.702	51	visage	- 0.669	278	soleil	- 0.632	33	rose
- 0.696	9221	ce	- 0.666	5367	;	- 0.630	110	route
- 0.696	1653	où	- 0.661	46	singulier			

Tableau 4. Les vocables en régression dans la correspondance de Flaubert

Les premiers mots de la liste (*qui, que, quoique, comme, lorsque*) soulignent un changement dans la syntaxe : l'abandon progressif des constructions complexes, complétives ou relatives, ce qu'on avait déjà observé pour l'ensemble de l'œuvre. Mais l'abandon est surtout thématique : c'est un renoncement à la vie (*vivre, vie, bonheur, beau, pur, rire, sourire, danser, rêver, passion, ardent, amour, sens, gloire, jeunesse*), au monde et à la nature (*soleil, mer, terre, fleur, odeur, herbe*).



*sable, montagne, ombre, vent, étoile*), aux couleurs (*vert, rose, pâle, bleu, couleur*) et au corps (*visage, doigt, épaule, peau, poitrine, pied, corps, chair, cou, oreille, lèvres, voix, bouche, nez, regard, tête, gorge*, etc.) La liste des mots en progression est le négatif de la précédente : les soucis, les contraintes, les échéances et l'amertume emplissent tout l'espace. Tout y est désenchanté et prosaïque, mis à part quelques éclats de révolte. Et le temps y pèse comme un couvercle.

On peut mesurer cet appauvrissement du vocabulaire qui s'accorde avec l'assombrissement du sentiment. Le calcul de l'accroissement lexical se fait en notant année après année tous les mots qui n'ont pas été encore rencontrés. Mais on peut s'y prendre de deux façons, soit qu'on adopte, comme c'est naturel, l'ordre chronologique, soit qu'on procède en sens inverse en commençant par la dernière tranche et en remontant le temps. Quelle que soit la démarche, la figure 10 conclut à l'épuisement du vocabulaire.

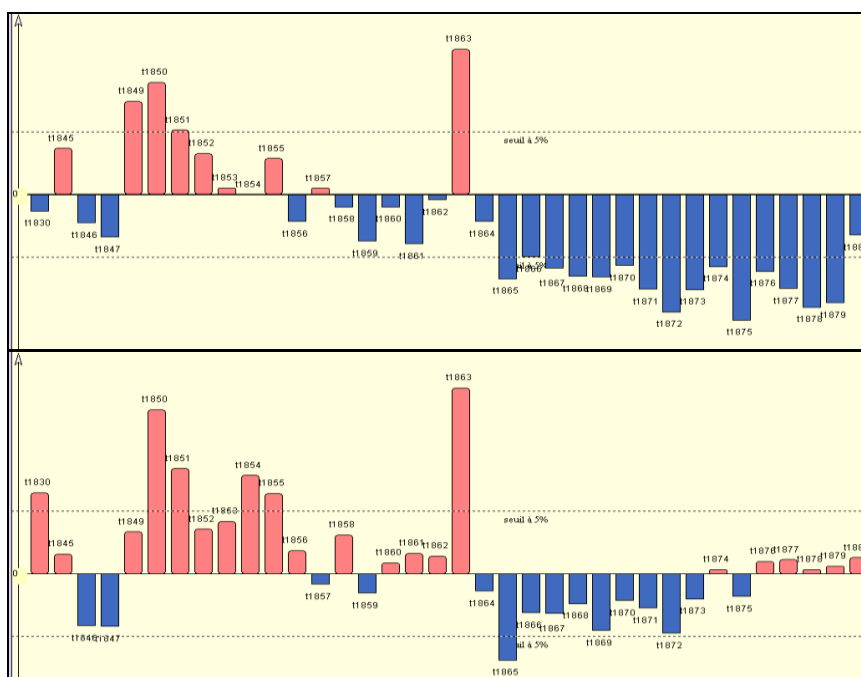


Figure 10. L'accroissement du vocabulaire (ordre chronologique et inverse)

Quant à la syntaxe, on retrouve dans la figure 11 les phénomènes observés dans la figure 6. Le passage progressif du verbe au substantif ne doit donc pas être attribué à l'inégale exploitation des genres, puisqu'il se

maintient quand le genre épistolaire est seul en cause. Et cela n'est pas dû à quelque relâchement de la plume, qu'on pourrait admettre dans la correspondance lorsqu'on règle des questions matérielles avec un éditeur, un parent proche ou un créancier (ce qui se produit plus souvent dans les dernières années). Car le même mouvement stylistique s'exerce au même moment dans les romans que les affres du style ont le plus tourmentés. Il s'agit donc d'une tendance sourde et profonde, dont ni l'auteur, ni la critique ne semble avoir eu une conscience nette.

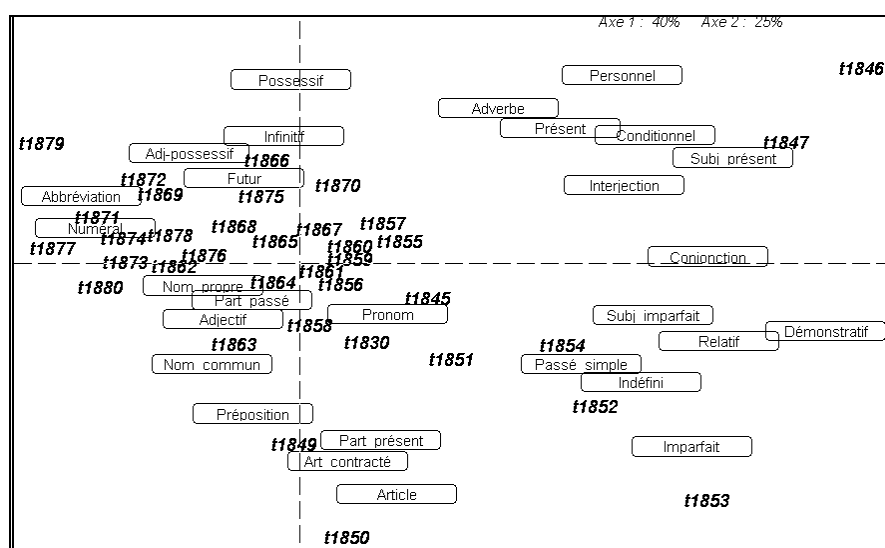
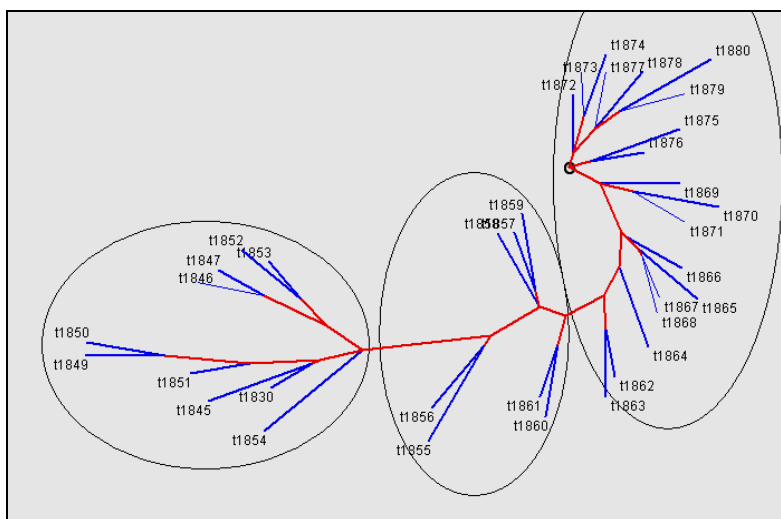


Figure 11. Parties du discours et temps verbaux dans la correspondance de Flaubert

Naturellement le calcul de la distance lexicale (figure 12) met en relief cette dérive du temps, où trois périodes se distinguent, la première s'étendant jusqu'au premiers mois de 1855 (c'est à ce moment-là que Flaubert rompt définitivement avec Louise Colet), et la dernière commençant en 1863-64 (c'est le début de la correspondance avec Georges Sand). Ce changement de destinataire explique en partie le changement de ton. Les lettres à la seconde sont fraternelles et sereines, quand la passion remplit celles qui s'adressent à la première.



**Figure 12. La distance lexicale**  
(ici établie sur les lemmes et la méthode Jaccard)

La figure 13 met en œuvre un objet dont nous n'avons pas parlé encore : les structures syntaxiques. Or le logiciel utilisé ne se contente pas d'accoler un lemme et un code grammatical à chaque graphie rencontrée. Il mémorise les séquences observées, c'est-à-dire les combinaisons de codes, en s'arrêtant aux ponctuations. Ainsi la formule rituelle qui clôt certaines lettres : « je vous embrasse tendrement » suit un moule structurel ainsi codé : ppvr (pronom+pronom+verbe+adverbe). L'expression exacte ne se rencontre que 8 fois. Mais le schéma est productif et en variant les pronoms, les temps et modes du verbe et les adverbes, on recueille 108 exemples si on maintient une contrainte : la présence du verbe *embrasser*. Si cette contrainte disparaît, la structure est reconnue plus de 3000 fois. Elle est d'ailleurs en constante diminution, ce qui ne saurait surprendre, puisque tous ses constituants sont individuellement en déclin, comme l'indique la figure 11. Ces structures, comme aussi les codes grammaticaux, servent à affiner les recherches, en imposant des filtres précis à la sélection proposée. Mais elles peuvent aussi être traitées en tant que telles, indépendamment des mots et des codes à travers lesquels elles sont instanciées dans le discours. C'est le cas dans la figure 13, qui, après repérage et décompte de toutes les structures du corpus, en étudie la distribution dans les textes et établit à partir de là un calcul de proximité. Les trois périodes précédemment distinguées se reconnaissent sans peine dans le graphe obtenu.

L'évolution constatée chez Flaubert n'est donc pas seulement dans le choix des mots. Elle est dans la syntaxe, dans le rythme de la phrase, dans l'oreille et le gueuloir de l'écrivain.

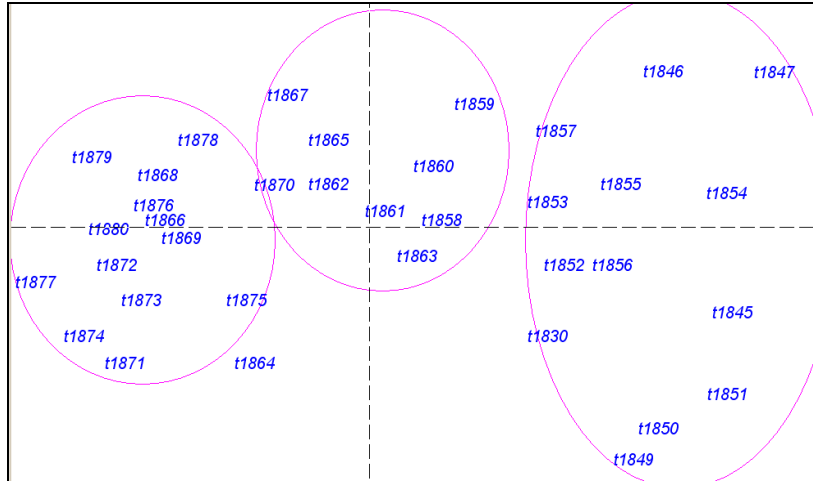


Figure 13. La distance « syntaxique ». Analyse factorielle

S'agissant de rythme, un dernier point de vue peut être abordé : celui de la ponctuation. On croit sur ce point devoir faire confiance à l'édition Conard. Car il serait vain de porter son attention sur la ponctuation de l'éditeur, si ce n'est pas celle du manuscrit. Déjà la figure 6 indiquait la tendance, les ponctuations faibles se situant du côté des premières années et les ponctuations fortes dans les dernières. Le détail de la distribution est délivré dans la figure 14.

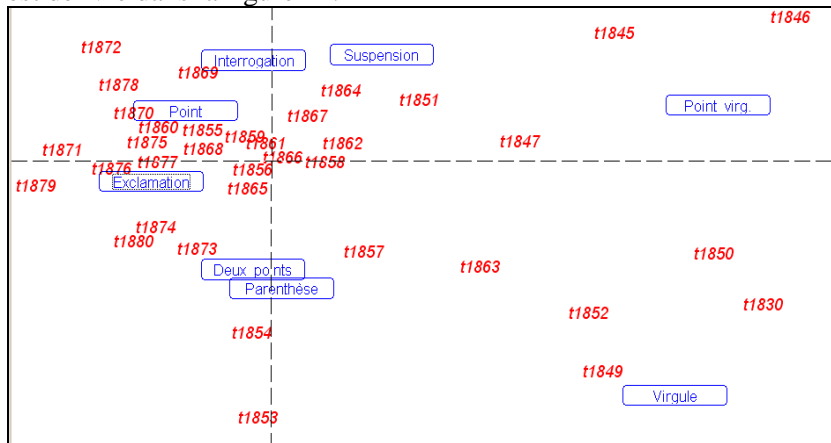


Figure 14. La ponctuation dans la correspondance de Flaubert

Virgules et points-virgules sont effectivement à l'extrême droite, au début de la chronologie, alors que les points, les interrogations et les exclamations sont à l'opposé, dans une zone où les dernières années se concentrent. La phrase de Flaubert se raccourcit donc dans sa correspondance (cela est vrai aussi de l'ensemble de l'œuvre), mais elle ne s'apaise pas. Si les points sont en augmentation, les points d'interrogation et surtout d'exclamation croissent plus vite encore. Flaubert a une nature éruptive. Et plus le temps passe, plus il trouve d'occasions de s'indigner. Il est vrai que les épreuves, nationales, professionnelles, familiales ou personnelles se multiplient à la fin de sa vie.

En conclusion, il ne nous échappe pas que notre parcours autour de Flaubert n'est guère qu'un papillonnage virevoltant. Notre intention était de montrer la voie en nous posant temporairement sur certains points qu'il faudrait approfondir. Bien d'autres ressources restent encore inexploitées. Mais plutôt que de saccager le champ de recherche, par une exploitation intempestive, nous le transmettons, dans sa virginité première, aux spécialistes de Flaubert. Ils en feront un meilleur usage.