HAL
open science

Le style de Proust dans la Recherche du temps perdu. Etude quantitative.

Étienne Brunet

To cite this version:

Étienne Brunet. Le style de Proust dans la Recherche du temps perdu. Etude quantitative.. VII International Symposium of the Association for Literary and Linguistic Computing, 1982, Pise, Italie. pp.51-76. ⟨hal-01574516⟩

HAL Id: hal-01574516

https://hal.science/hal-01574516v1

Submitted on 14 Aug 2017

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.

Autorisation HAL

HAL Authorization

Le style de Proust dans *La recherche du temps perdu*. Etude quantitative[1]

Etienne Brunet

Les grands écrivains sont ceux que l'ordinateur approche le plus volontiers. Il ne faut point voir là le signe de quelque discernement remarquable non plus que d'un regrettable sans-gêne, mais la marque d'une contrainte économique qui pèse sur ce genre d'entreprise et qui se justifie plus aisément dans le cas des grands textes comme la Bible, la Somme de Saint Thomas, l'oeuvre de Shakespeare ou celle de Proust. Aussi bien l'index que nous proposons n'est-il pas le seul qui ait été tenté sur la *Recherche du temps perdu* : depuis bien longtemps déjà, on dispose d'un *Dictionnaire des idées de Proust*, réalisé en 1968 par Pauline Newman Gordon[2]. Et plus récemment au Centre d'analyse des manuscrits modernes du C.N.R.S[3], un vaste projet a été engagé pour réaliser l'édition automatique des brouillons de Proust. Nul doute que cette dernière entreprise – dont on trouve peu d'exemples – n'apporte une aide précieuse à l'exégèse d'un texte variant dont la genèse, la composition et l'écriture posent des problèmes sans cesse renouvelés[4].

Notre index se situe entre ces deux entreprises : profitant des progrès accomplis depuis quinze ans par l'informatique, il ne se contente plus de proposer un choix de termes ou de champs sémantiques jugés proustiens et significatifs, mais fournit un relevé complet de tous les mots de la

---

1. Actes du colloque du VIIe International Symposium of the l'Association for Literary and Linguistic Computing (Pisa, 1982), parus dans *Linguistica Computazionale*, vol. 3, Supplément, 1983, Pisa, p. 51-76.

2. Standford University, Mouton. Préface d'Alphonse Juillard.

3. A l'École Normale Supérieure, rue d'Ulm, sous la direction de Louis Hay.

4. L'avancement des travaux est exposé dans un article de Louis Hay, « Passé et avenir : Le patrimoine des écrits », *Courrier du C.N.R.S.*, numéro 38, oct. 1980, p. 40-44. On y trouvera un exemple de l'édition automatique des variantes de Proust, réalisation de M. Hainsworth et J. L. Lebrave.

Recherche. En revanche, il s'en tient au texte de l'édition de la Pléiade et s'interdit toute incursion du côté des *Cahiers* et des manuscrits. Il n'est pas impossible qu'il puisse précisément aider au déchiffrage et au classement des manuscrits, la lecture associative des index permettant de rapprocher des passages qui ont la même inspiration, et peut-être la même date de composition, et qui ont été disloqués dans l'espace du livre par l'effet des ciseaux et de la colle dont Proust usait si souvent.

1. L'Index de la *Recherche du temps perdu*

Le tableau 1 restitue une des 1600 pages de notre index[5]. Les index sont devenus des produits courants qui n'appellent guère de commentaires. Nous nous contenterons de souligner quelques nouveautés.

1.1. Il s'agit d'abord d'un index lemmatisé et cette opération a été rendue malaisée par la taille du corpus. En recourant à un fichier modèle établi à Nancy[6], on a pu rattacher au lemme – à la vedette de regroupement – les formes ou graphies relevées dans le texte. Certes cette lemmatisation n'est pas sans reproche : les homographes ont reçu un traitement approximatif dans les deux fichiers et l'attention méfiante du lecteur est attirée sur ces cas douteux par un astérisque. De plus, le détail des 36.770 formes différentes et de leurs références est conservé et restitué dans l'index, tout en apparaissant subordonné au lemme grâce à l'artifice typographique de la minuscule initiale et du décrochement marginal. Bien entendu, l'ordre alphabétique gouverne d'abord les lemmes, puis les formes à l'intérieur des lemmes[7]. Le consultant alerté par le symbole de l'homographie peut donc contrôler lui-même les options de la lemmatisation et les rectifier au besoin pour son propre usage. Il peut en outre reconnaître si le lemme proposé se trouve réellement dans le texte sous la forme canonique : les lemmes

5. *Le Vocabulaire de Proust*, avec *l'Index complet et synoptique de A la recherche du temps perdu*, Slatkine-Champion, 1983, 3 vol., 1983 (préface de J.Y. Tadié).

6. Ce fichier, comme le fichier des données de Proust, est un produit de l'Institut National de la langue française (anciennement *Trésor de la langue française*).

7. C'est l'ordre alphabétique usuel, qui refuse toute valeur discriminante aux accents et qui a nécessité la traduction sous-jacente des mots en alphabet pauvre, même si la transcription utilise les caractères riches.

« introduits » sont ceux qu'aucune référence ne suit immédiatement et qui sont absents de la liste subséquente des formes regroupées[8].

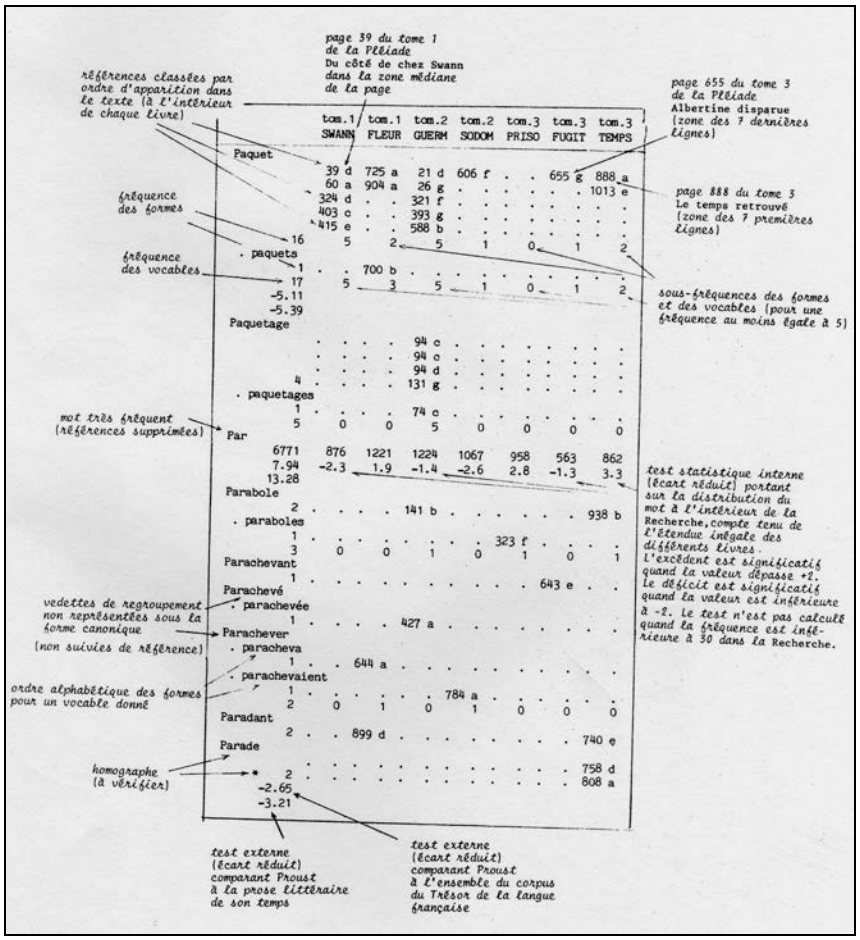


Tableau 1. Index de Proust (*A la recherche du temps perdu*)

1.2. L'index est exhaustif pour les formes, les lemmes et les fréquences des uns et des autres, et *quasi* complet pour les références. Ont été écartées, pour la seule raison qu'on devait gagner de la place, les références de 74 formes extrêmement fréquentes dont la localisation importait assez peu. Que pouvait-on gagner à connaître l'emplacement

8. Quand lemme et forme coïncident au même endroit de la liste alphabétique, on a évité un redoublement inutile; ainsi en est-il dans notre exemple pour *abaissement, abandon, abasourdi*, etc.

précis des 53.078 occurrences de la préposition *de* ou des 28.885 emplois de l'article *la* ? Ces 74 exclusions concernent les formes et non les lemmes, ce qui permet l'étude des temps et modes, même des verbes fréquents et des auxiliaires, puisque, parmi les formes verbales, on n'en a rejeté que 3 du verbe *être*, 3 du verbe *avoir*, 2 de *faire*, 1 de *dire* et 1 de *pouvoir*. Aucun substantif n'a été exclu, ni aucun adjectif ou participe. Le tableau ci-dessous donne la répartition des mots dans les sous-ensembles de la *Recherche* [9]:

	Swann	**Filles**	**Guer.**	**Sodom.**	**Prison.**	**Fugit.**	**Temps**	**Total**
Nbr. pages	428	528	592	534	410	272	360	3124
	Vocabulaire commun							
occurrences	176357	217281	237655	215179	164683	111162	144752	1267069
vocables	9029	9826	10265	9866	8633	6419	8240	18322
formes différ.	14677	16653	17221	16239	14162	10155	13033	36867
références								569415
	Noms propres							
occurrences	4718	5469	9138	9534	5301	3530	5017	42707
vocables	569	721	1040	1084	629	368	710	2976
références								32686

Au total notre Index comprend 569.415 références, à quoi il faut ajouter 32.686 références de noms propres réunis dans une liste spéciale[10].

1.3. Les références précisent le tome de l'édition de la Pléiade, la partie de l'ouvrage (de *Swann* au *Temps retrouvé*), la page et la zone dans la page. Comme dans les données de Nancy les lignes ont été recomposées tout en respectant le cadre de la page, il n'était plus possible de restituer le numéro exact de la ligne, lequel, au reste, ne figure pas dans l'édition de référence. Pour éviter au lecteur des lectures et des fatigues inutiles, un code ajouté au numéro de page le conduit directement à une zone de la page qui contient environ 6 lignes ou 70 mots. On a distingué 7 codes alphabétiques (de *a* à *g*), le code *d* par exemple renvoyant au milieu de la page, et le code *g* à la fin. En réalité pour constituer chaque référence, il a suffi de 3 ou 4 chiffres (indication de page) et d'une lettre (indication de zone), les précisions de tome et de partie figurant une fois pour toutes en haut de chaque page.

9. Il s'agit de l'édition de la Pléiade, réalisée par Pierre Clarac et André Ferré, en 1954. Cette belle édition est appelée à durer longtemps encore et l'étude approfondie des *Cahiers* ne semble pas devoir la remettre en cause.

10. Ont été écartées les abréviations *M.* (2.116 occ.) et *Mme* (3.089 occ.) et la particule nobiliaire de (4.316 occ.)

1.4. En effet, comme nous l'avons fait pour l'index de l'*Emile*[11], la présentation est synoptique et les 7 livres consécutifs de la *Recherche* apparaissent simultanément en 7 colonnes juxtaposées, ce qui facilite non seulement la lecture, mais aussi la mesure rapide de la distribution ou de l'évolution d'un mot à l'intérieur de l'oeuvre. L'impression visuelle que donne cette sorte d'histogramme renversé est confirmée – au moins chaque fois qu'une forme a au moins 10 occurrences – par la séquence chiffrée des fréquences. Ainsi la préposition *à* distribue 3.814 de ses occurrences dans *Du côté de chez Swann*, puis 4.716, dans *A l'ombre des jeunes filles en fleurs*, etc.

1.5. Ces fréquences absolues n'étant pas immédiatement comparables puisque l'étendue diffère d'un livre à l'autre, on les a doublées d'un test statistique qui permet de mesurer les variations d'emploi et d'apprécier les déficits et les excédents selon une mesure probabiliste. Il s'agit d'un écart réduit dont le calcul est classique : sachant que *Du côté de chez Swann* représente 176.357/1.267.069 = 0,1392 ou 14% de la *Recherche*, on peut s'attendre à trouver 27.468 x 0,1392 = 3.823 occurrences de la préposition *à*. Comme on en compte 3.814, le déficit est de 9 occurrences et cet écart est dit « réduit » lorsqu'on le pondère par l'écart-type théorique : racine carré de *npq*, soit racine carré de 27.468 x 0,1392 x 0,8608. On obtient z = -0,159. C'est cette valeur, tronquée à la première décimale, qui est reproduite dans la colonne réservée à *Swann* au dessous de la fréquence réellement observée, et qui est beaucoup trop faible pour être attribuée à une cause non aléatoire (pour que le seuil soit franchi qui autorise une conclusion, il faudrait que cette valeur atteigne 2 en valeur absolue – ce qui d'ailleurs se produit deux fois dans le cas de cette préposition *à*, dont l'excédent dans *Sodome* (+4,2) et le déficit dans le *Temps retrouvé* (-3,6) échappent pareillement au hasard)[12].

Cette suite d'écarts réduits est parfois révélatrice d'une évolution (croissance ou désaffection), comme celle qu'on remarque dans l'emploi grandissant du mot *vérité* (p. 1475 de l'index) :

Occurrences :	49	45	53	39	55	46	73	total 360
écarts réduits :	-0,1	-2,3	-1,9	-3,1	1,2	2,6	5,2	

11. *L'index-concordance d'Emile* in *Index des oeuvres de J.J. Rousseau*, Slatkine, 1980, 2 vol. (LX -) 585 p. (XXI -) 727 p.

12. Au lecteur que surprennent ces calculs, nous recommandons la lecture des manuels de Charles Muller, *Initiation aux méthodes de la statistique linguistique*, Hachette, 1973, et *Principes et méthodes de la statistique lexicale*, Hachette, 1977.

1.6. Les tests qui précèdent explicitent une comparaison interne des différents sous-ensembles de l'oeuvre de Proust. Notre index ménage aussi une comparaison externe et confronte pour chaque mot l'usage de Proust à celui de son temps et à celui de l'ensemble du corpus du *Trésor de la langue française*. Il s'agit là encore de l'écart réduit fondé sur les probabilités :

p1 = 1.267.069 / 70.273.852 = 0,01803
p2 = 1.267.069 / 12.216.571 = 0,10372[13]

Le même exemple de la préposition *à* montre un excédent tout à fait caractéristique chez Proust, qu'on établisse la comparaison avec le corpus entier (z = 14,48) ou avec la prose du début du XXe siècle (z = 19,60). On verra que cet excédent n'est pas propre à cette seule préposition.

2. Le Vocabulaire de Proust

La lecture d'un index comparatif et statistique est en effet plus riche d'informations et de suggestions que celle d'un simple relevé manuel, non point seulement parce que l'objectivité, l'exhaustivité et la fiabilité y sont garanties (et ce ne sont pas là de minces avantages), mais aussi parce qu'on y trouve à chaque page des aides à l'analyse, des amorces de synthèse et des clignotants qui sollicitent l'attention. Ainsi l'excédent surprenant de la préposition *à* engage le consultant sur la trace des autres prépositions, puis plus largement sur la voie des subordonnants et des mots grammaticaux dans leur ensemble.

2.1. Et il découvre alors un fait de style tout à singulier qui certes n'est pas insensible à la lecture, mais qui reçoit de la statistique une confirmation éclatante : l'abondance des mots grammaticaux quand ils sont des indicateurs de fonction et qu'ils servent à la structuration de la phrase et principalement à la subordination. Cela produit un déséquilibre dans la structure des fréquences, l'excédent des mots grammaticaux – qui se cantonnent dans les hautes fréquences – est compensé par un déficit dans les basses et moyennes fréquences, là où l'on rencontre ce qu'on appelle les mots sémantiques. La conséquence est que les tests de notre index, où les mots fréquents sont noyés dans la masse, apparaissent le

13. La probabilité *p2* est établie à partir des tranches 9, 10 et 11 du corpus du *Trésor de la langue française*, soit 231 textes complets échelonnés de 1893 à 1926. Quant à la probabilité p1, elle mesure le rapport de la *Recherche* à la totalité du même corpus, de 1789 à 1964, soit un millier de textes littéraires. Pour plus de précisions sur ce corpus, nous renvoyons le lecteur à notre ouvrage *Le vocabulaire français de 1789 à nos jours*, Genève, Slatkine, 1981, 3 vol. (XIV -) 852 p. (VIII -) 518 p. (XIV -) 454 p.

plus souvent négatifs, au point qu'on peut penser qu'il s'agit d'une erreur de calcul[14.] Ainsi dans notre exemple, on compte 13 valeurs négatives pour 7 positives et en poussant l'investigation jusqu'à la vingtième page, on obtient un rapport de 3 à 1 (238 valeurs négatives pour 81 positives). Jamais jusqu'ici nous n'avions observé de pareilles distorsions : dans les corpus de Chateaubriand, de Rousseau ou de Giraudoux que nous avons étudiés, les écarts de signe contraire s'équilibrent. Car ces auteurs font un choix dans le lexique alors que Proust affirme son originalité dans la syntaxe. Un simple pourcentage confirme ce trait stylistique : si l'on rassemble tous les mots très fréquents (dont la fréquence est supérieure à 100.000. dans le corpus du T.L.F.), leur masse représente 52,86% du total dans le T.L.F. et 59,01% chez Proust. Un écart aussi important nous a suggéré de parcourir la gamme des fréquences et d'étudier la position du corpus proustien à chaque palier. Le tableau 2 détaille 20 groupes de fréquences dans le corpus du T.L.F. : les hapax, les mots qui ont deux occurrences, ceux qui en ont 3 ou 4, 5 à 8, 9 à 16, etc.

Le déficit est constant dans les 16 premiers groupes, c'est-à-dire dans les fréquences basses et moyennes, mais un renversement de tendance s'opère dans le 17ème groupe quand la fréquence atteint 50.000 dans le corpus et 1.000 dans la *Recherche* ; au-delà de ce seuil les excédents sont très considérables chez Proust (écarts réduits de 14, 43 et 46) et les trois derniers groupes – qui ne rassemblent qu'une centaine de mots-outils – récupèrent les 51.772 occurrences perdues dans les 17 premiers groupes.

2.2. Ce déséquilibre du discours proustien au profit des très hautes fréquences explique que Proust ne spécule pas particulièrement sur la richesse ou la variété du vocabulaire. L'indice *w* qui mesure la richesse lexicale[15] est sensiblement le même (13,43) que celui qu'on observe dans le corpus du XXe siècle établi à Nancy (13,40). Une comparaison avec un romancier contemporain de Proust permet – sans indice – une

14. Le calcul n'est pourtant pas biaisé puisqu'en adoptant les probabilités fournies par le Centre de Nancy, on obtient les mêmes résultats : le *Dictionnaire des fréquences* et le *Répertoire des textes* donnent respectivement 85.161.918 et 1.507.681 pour l'étendue du corpus du *T.L.F.* et l'étendue du corpus de la *Recherche*, y compris les noms propres et les signes de ponctuation, soit une probabilité de p = 0,0177 qui est très voisine de celle dont nous nous servons (p 0,0176) et qui engendre des valeurs très proches pour l'écart réduit : pour *abandonnant*, 582 occurrences dans le corpus et 4 chez Proust, on obtient z = -1,98 au lieu de -2,02.

15. La formule, dans sa forme simplifiée, s'écrit ainsi: $w = {}_{N}V^{-a}$ pour *a* 0,172. Pour plus de détails sur cet indice, voir notre thèse *Le vocabulaire de Giraudoux*, Slatkine, 1978, p. 45 et *Le vocabulaire français de 1789 à nos jours*, Slatkine, 1981, p. 55.

comparaison directe. Les trois romans de Giraudoux *Suzanne*... (1921), *Siegfried* (1922), et *Juliette*... (1924) écrits par Giraudoux à la même époque que la *Prisonnière* représentent sensiblement la même étendue (164.004 occurrences contre 164.683). On y compte 10.108 vocables alors que le texte correspondant de Proust n'en a que 8.633. On saisit là l'effet de deux écritures différentes, voire opposées : la coquetterie stylistique de Giraudoux qui le pousse à s'affubler de mots rares ou surprenants est étrangère à Proust dont la simplicité lexicale (qui peut fort bien aller de pair avec la complexité syntaxique) ne cherche rien d'autre que la restitution la plus exacte possible du réel[16].

La comparaison avec le corpus de Chateaubriand – auquel Proust rend hommage, citant le fameux passage de la grive au gazouillement magique[17] – est plus facile car les deux corpus sont de taille voisine[18]. Or on trouve 1.300 vocables de plus chez Chateaubriand, alors même que les créations lexicales multipliées au cours du XIXe siècle offraient à Proust un choix beaucoup plus vaste dont il a moins profité que Giraudoux[19]. En prenant appui sur l'ensemble des données de l'Institut de la langue française et en utilisant la loi binomiale, on peut rapporter à une mesure commune ces trois écrivains. Proust apparaît comme le moins soucieux de variété lexicale, le déficit étant plus fort chez lui (z = -42,39) que chez Chateaubriand (z = -37,66) et chez Giraudoux (z = -29,50).

---

16. L'hostilité commune de Proust et Giraudoux au naturalisme ne doit pas faire illusion : quand l'un s'éloigne de Zola en créant un univers de fantaisie, l'autre au contraire contourne et dépasse Zola dans la voie du réalisme psychologique. L'emploi du mot *réel* et de ses dérivés est tout à fait révélateur à cet égard:

	réel	*réalité*	*réellement*	*réaliser*	*réalisation*	*réalisé*
Fréq. Giraudoux	30	26	4	7	1	5
Ecart réduit	-6,64	-9,51	-4,48	-5,21	-3,50	-3,10
Fréq. Proust	204	431	34	74	49	34
Ecart réduit	10,00	18,55	-0,86	2,93	9,19	3,46

Rappelons que le corpus de Proust représente deux fois celui de Giraudoux. Le rapport des occurrences devrait donc être du simple au double, alors qu'il est ici du simple au décuple. Jean Milly a donc raison quand il *constate que le mot « réalité » et ses parents « réel » et « réellement » sont employés chez Proust d'une façon exceptionnellement fréquente*. (*Proust et le style*, Minard, 1970, p. 39), mais il a tort quand il *marque peu d'intérêt pour le relevé d'écarts* – ce qu'il fait pourtant lui-même, à juste titre et contre son principe.

17. *Le temps retrouvé*, dans la Pléiade, p. 919.

18. 1.398.984 occurrences chez Chateaubriand contre 1.267.069 chez Proust.

19. La mesure de cette inflation lexicale a été tentée au chapitre II de notre ouvrage : *Le vocabulaire français de 1789 à nos jours*, Slatkine, 1981, p. 51-81.

Classe	Fréquence dans le T.L.F.	N chez Proust	N théorique	écart absolu	écart réduit
1	1	193	380	-187	-9,70
2	2	161	239	-78	-5,10
3	3 - 4	280	373	- 93	4,87
4	5 - 8	467	618	-151	-6,13
5	9-16	869	1076	-207	-6,38
6	17 - 32	1702	2016	-314	-7,05
7	33 - 64	3150	3746	-596	-9,83
8	65-128	6079	7103	-1024	-12,26
9	128 - 256	10848	12763	-1915	-17,11
10	257 - 512	16302	20309	4007	-28,37
11	513-1024	24248	29976	-5728	-33,39
12	1025-2048	34863	44184	-9321	44,75
13	2048 - 4096	50770	62496	-11726	-47,33
14	4097 - 8192	72768	76881	-4113	-14,97
15	8193 -16384	73064	80055	-6991	-24,93
16	16385 - 32768	74688	79734	-5046	-18,03
17	32769-65536	87412	87687	-275	-0,94
18	65537-131072	73057	69339	3718	14,25
19	131073-262144	91136	78990	12145	43,61
20	+ 262144	645012	609103	35909	46,43
		1.267.069		0	

Tableau 2 . Les groupes de fréquence

2.3. Certains songeront peut-être à expliquer cette relative pauvreté du vocabulaire proustien par le piétinement que la situation impose aux personnages enfermés dans le cercle étroit du Paris mondain. Et il est vrai que les enquêtes que Zola mène dans des milieux fort divers favorisent l'extension du vocabulaire et que l'enquête de Proust s'exerce plus en profondeur qu'en surface. Il semble toutefois que l'explication soit ailleurs : ni Chateaubriand, ni Giraudoux ne prétendent faire oeuvre documentaire. Mais ils ont un certain goût des mots rares ou recherchés, que Proust ne partage pas. Ce n'est pas que Proust soit timoré à l'égard du langage : il ne refuse pas les mots familiers, voire vulgaires[20], et il s'amuse fort des « cuirs » de Françoise ou du maître d'hôtel[21]. Il ne s'interdit pas de fabriquer des mots à l'occasion, en recourant aux ressources de la suffixation et de la préfixation et parfois aux deux

20. *Comme le gros pétard du baron de Charlus*, Pléiade, t. 2, p. 610.

21. Proust est très indulgent à l'égard du langage populaire et des déformations qu'il fait subir à l'orthographe, à la prononciation ou au sens des mots et après avoir corrigé un cuir de Françoise (*estoppeuse* pour *stoppeuse*), il se repent aussitôt de son intransigeance : *Ce reproche était particulièrement stupide car les mots français que nous sommes si fiers de prononcer exactement ne sont eux-mêmes que des « cuirs » faits par des bouches gauloises qui prononçaient de travers le latin ou le saxon (...). Le génie linguistique à l'état vivant, voilà ce qui eût dû m'intéresser dans les fautes de Françoise.* (*Sodome*, La Pléiade, p. 736).

procédés réunis dans la parasynthèse[22]. Les 200 exclusivités lexicales que nous avons relevées dans la *Recherche* appartiennent à ces deux catégories. Celles qui transcrivent des propos entendus sont teintées d'humour ou d'ironie à l'égard de Françoise (*alliancé, envahition, paperole, charlatante, copiateur, estoppeuse*), du maître d'hôtel (*enverjure, pistière* pour *pissotière*), de Bloch (*barbifiant*), de Norpois (*héraldiquement*), de Rachel (*vatique*) voire de Swann (*botticellien*) ou de Mallarmé (*effeuillaison* sur le modèle de *cueillaison*). Mais beaucoup d'autres ne sont que des outils que Proust a forgés pour la circonstance et qui n'ont d'autre intention que de circonscrire la réalité. Dans ces cas-là, Proust ne cherche guère le pittoresque ni l'archaïsme, ni même l'originalité et certaines de ses exclusivités[23] sont plus audacieuses que gracieuses (*inserviabilité, matérialistement, contagionnement, musculeusement, assouvissabie, annihilateur, complexement, désignatifs*, etc.). La liste ci-dessous rend compte des mots propres à Proust qui ont plus d'une occurrence :

alliancé	3	germanophobe	2	pistière	5
barbifiant	2	héraldiquement	2	plissage	2
botticellien	2	hétérostylé	2	présentateur	4
busquage	2	inserviabilité	2	schumannesque	2
caféterie	2	interchangé	2	stoppeuse	3
contagionné	2	matérialistement	2	toponymie	4
désintoxicant	2	musicographe	2	transplantant	2
effeuillaison	2	nasonnement	2	trional	7
envahition	2	onctueusement	2	vatique	2
enverjure	3	paperole	4	vicariant	2
germanophilie	5	pastellisé	2		

Pour être complète, la liste devrait incorporer les quelque 193 hapax qu'on rencontre sous la plume de Proust et qui représentent seulement la moitié de l'effectif attendu (380 : une simple règle de trois suffit pour ce calcul). C'est dire la discrétion de Proust dans la fabrication lexicale : il se contente des ressources que la langue offre dans ce domaine et sa marque est plus dans la combinaison des éléments du lexique que dans

22. Voici quelques exemples de créations parasynthétiques propres à Proust : *encaoutchouté, encauchemardé, enfarinement, engrillager.*

23. Ce sont des exclusivités relatives, si l'on peut dire, en égard au corpus de référence qui est loin de représenter la totalité de la langue écrite, même s'il est gros de 70 millions d'occurrences. On ne saurait sans contrôle les ranger parmi les créations absolues et les porter automatiquement au crédit de Proust

leur invention[24]. C'est dire que l'originalité de Proust est moins dans le lexique que dans la syntaxe, et moins dans la syntaxe que dans la pensée.

3. La Phrase

Peut-on, par le truchement des mots isolés, accéder à la syntaxe ou au rythme de la phrase ? Poser ainsi la question, c'est admettre que le traitement automatique est encore décevant lorsqu'il prend directement pour objet la construction et la signification du discours. Dès lors qu'on envisage des unités plus larges que le mot graphique, la lecture humaine et le découpage manuel restent les meilleures garanties. Et c'est ainsi qu'a procédé Conrad Bureau dans son étude de la phrase de Proust à travers *Combray*[25]. Ces garanties sont toutefois insuffisantes car la subjectivité introduite à l'entrée des données affaiblit la valeur des résultats : certes le découpage des syntagmes s'appuie sur des critères clairement définis, mais ces critères sont pour une part arbitraires et leur application particulière repose en bien des cas sur l'interprétation. Nous nous en tenons donc à la statistique proprement lexicale, quelles que soient ses faiblesses dans le domaine de la combinaison des mots, par où commence l'écriture. Ce que l'on perd ainsi en finesse, on le gagne en généralité et Proust nous donnerait peut-être raison qui préférait les *lois* aux *détails*[26]. Certes, comme le souligne Conrad Bureau, « un traitement statistique ne peut prétendre rendre compte de l'organisation proprement syntaxique des énoncés si le calcul ignore cette hiérarchisation pour ne

---

24. La position de Proust à l'égard du lexique, et plus généralement du langage, est dénuée de coquetterie. Et par là, il rejoint la doctrine classique qui prisait peu les joliesses et curiosités de plume. Comme le remarquait Jacques Bersani dans une récente émission télévisée (« Les vaches sacrées », avril 1981), l'adjectif proustien dans le sens de recherché est aussi éloigné de Proust que le marivaudage l'est de Marivaux. Comme Molière, Marivaux prétendait peindre le plus justement et le plus simplement la nature humaine et s'indignait qu'on lui reprochât de peser des « oeufs de mouches dans des balances de toile d'araignée », selon l'expression de Voltaire. Proust de la même façon s'indigne contre les épithètes de *fin*, *exquis*, *délicat* qu'on prodigue alors qu'il revendique le vivant et le vrai *cf.* L. Pierre-Quint, *Proust et la stratégie littéraire*, Buchet Chastel, 1954, p. 51. Sur la théorie du langage chez Proust, on consultera J. Milly, *Proust et le style*, Minard, 1970.

25. Conrad Bureau, *Linguistique fonctionnelle et stylistique objective*, Presses Universitaires de France, 1976, p. 264.

26. Conrad Bureau renonce lui-même à exploiter les détails dispersés et la variété infinie des structures de phrase. Après avoir recensé 335 constructions syntaxiques différentes, il conclut : « Il est inutile de vouloir établir une typologie des phrases de Proust », puisque le nombre de classes serait énorme et l'effectif de chacune le plus souvent nul (ouvrage cité, p. 217).

retenir que le nombre des éléments accumulés ». Mais précisément la hiérarchisation a des repères tangibles dans la phrase auxquels la statistique des mots donne accès.

3.1. Il s'agit des *mots grammaticaux* et nous abandonnons ici l'étude des basses fréquences pour celle des fréquences élevées. Conrad Bureau a beau ironiser sur le décompte « des *qui*, des *que*, des *parce que*, des *si*, des *comme* »[27], s'il avait fait ce dénombrement, il serait arrivé aux mêmes conclusions sur la complexité de la phrase proustienne, avec l'avantage d'une moindre intervention dans les données et d'une généralité plus grande puisque ce qu'on observe dans *Combray* est étendu à toute la *Recherche*. Si l'on s'en tient à la liste des indicateurs de fonction que cite Conrad Bureau, on voit aisément le goût de Proust pour la subordination dans les écarts considérables que l'on constate (en prenant pour référence l'ensemble du corpus du *Trésor*[28]):

	QUE	QU'	QUI	SI	COMME	POUR
Fréquence observée	25634	18459	15644	6590	7622	10357
Fréquence théorique	17804	11542	12903	4277	5187	8125
Ecart réduit	59,21	64,95	24,35	35,68	34,12	24,98

L'extrait du vocabulaire spécifique reproduit dans le tableau 3 montre que cette prédilection pour la phrase complexe ne s'exprime pas seulement par la surabondance des subordonnants (*que, qu', parce que, puisqu', puisque, quoique, tandis que, quand, comme, si*), mais aussi par celle des relatifs (*qui, que, dont, où, lequel, laquelle, lesquels, lesquelles, duquel, desquels, desquelles, auquel, auxquels, auxquelles*), par l'excédent des démonstratifs qui précèdent le relatif ou le substantif déterminatif (*celle* de ou *qui, celles, celui, ceux*), par l'excédent enfin de la plupart des prépositions (notamment les plus courantes *de, à, par, pour, sans, avec* ou celles qui concernent le temps *avant, après, dès, pendant, depuis, jusque*[29]).

27. Ouvrage cité p. 218.

28. Les écarts ne changent guère si l'on choisit pour « norme » la prose littéraire du temps de Proust, de 1900 à 1922.

29. Par contre, les prépositions relatives à l'espace sont le plus souvent déficitaires (*sur, sous, vers; derrière, autour, parmi, dans*). En fait la distinction n'est peut-être pas celle qui sépare le temps et l'espace, mais plutôt celle qui oppose la simple préposition à la locution conjonctive formée d'une préposition associée à *que*. Les prépositions de temps sont de ce dernier type et non les prépositions de l'espace.

VOCABULAIRE	NEGATIF		VOCABULAIRE	POSITIF	
-038.51	17963	il	+076.9.9	44093	que
-037.50	735	tu	+052.15	14221	elle
-031.75	21652	le	+037·91	11663	me
-029.14	11925	se	+036.54	6590	si
-029.22	25209	et	+033.67	5825	même
-029.19	400	te	+032.41	8581	mais
-027.20	16234	les	+032.33	2262	chez
-027.09	4267	sur	+030.58	14583	pas
-025.48	19471	1’	+030.49	7622	comme
-023.82	28885	la	+027.48	791	laquelle
-021.16	136	toi	+026.50	23620	ne
-020.73	284	donc	+026.28	1286	celle
-020.20	11221	des	+025.59	3218	ma
-019.98	27	lorsque	+024.97	4611	où
-016.48	258	pourquoi	+024.45	3165	quand
-016.41	197	voilà	+024.55	1830	car
-016.23	1334	deux	+024.71	10357	pour
-016.14	2631	ils	+023.78	1447	parce que
-016.01	68	voici	+022.58	15644	qui
-015.70	603	ça	+022.04	3584	ou
-014.77	359	trois	+021.92	1698	elles
-015.25	472	aujourd’hui	+021.54	493	celles
-013.69	533	contre	+021.54	1933	quelque
-013.62	872	sous	+020.92	499	lequel
-013.25	3036	ses	+020.64	356	certaines
-012.86	491	ni	+019.60	27468	à
-012.78	151	quatre	+018.66	875	pourtant
-012.61	20	tes	+017.74	178	duquel
-012.59	332	quoi	+017.74	232	lesquels
-012.50	94	ta	+017.43	2306	cela
-012.09	697	vers	+017.10	412	certains
-012.01	191	hé	+016.33	1230	celui
-011.71	214	ho	+016.14	213	lesquelles
-011.55	509	ha	+015.95	21720	je
-011.09	6184	vous	+015.81	6678	nous
-010.95	781	leurs	+014.84	85	desquelles
-010.68	7768	du	+014.62	2216	dont
-010.60	62	six	+014.54	97	desquels
-010.71	1486	rien	+014.44	3738	moi
-010.32	5003	son	+014.13	3704	autre
-011.90	557	votre	+013.47	798	pendant
-010.12	146	vos	+013.28	6771	par
-010.26	158	dix	+012.98	14729	une
-009.60	199	derrière	+012.48	604	longtemps
-009.24	968	toute	+011.96	559	puisque
-009.29	79	hier	+011.83	127	auxquels
-009.26	126	cent	+011.63	247	certaine
-008.82	73	ceci	+011.38	113	auxquelles
-008.78	340	quel	+011.18	444	malgré
-008.72	162	cinq	+010.71	158	quoique
-008.34	35	nul	+010.65	7094	lui
-008.32	526	nos	+010.50	867	maintenant
-008.20	35	trente	+010.47	3412	ces
-007.93	387	ton	+010.34	316	certain

Tableau 2. Les mots grammaticaux caractéristiques de Proust

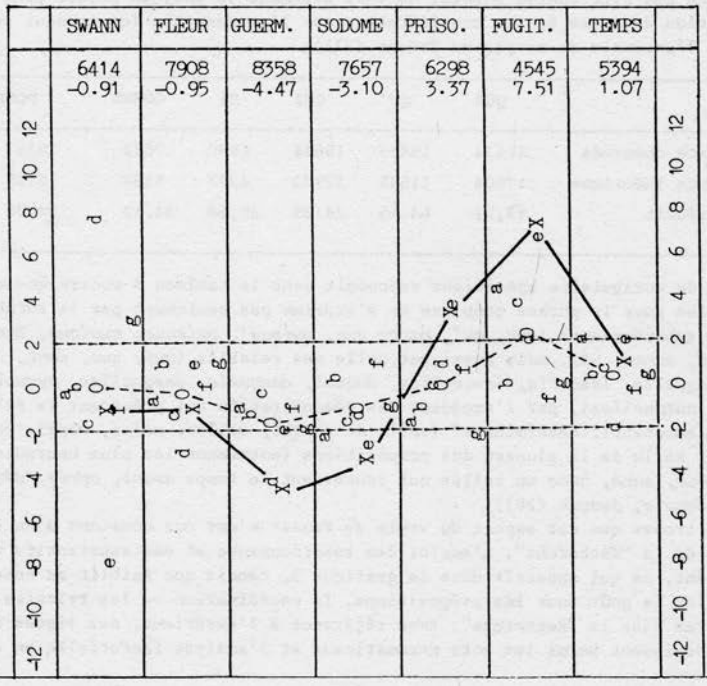
3.2. Il se trouve que cet aspect du style de Proust *n’est pas constant* d’un bout à l’autre de la *Recherche*. L’emploi des subordonnants et des substantifs va s’accentuant, ce qui apparaît dans le graphique 1,

tandis que faiblit au cours de la rédaction le goût pour les prépositions, la coordination ou les relatifs.

NOMBRE = 7
OCCURRENCES = 46574

CODE	MOT	FREQUENCE	ECARTS	REDUITS
a	parce	1447	23.78	31.61
b	puisqu'	158	3.32	3.14
c	puisque	401	12.42	12.14
d	qu'	18459	53.06	64.97
e	que	25634	56.00	59.21
f	quoique	123	10.80	0.38
g	tandis	352	3.43	5.16
	MOYENNE		23.25	25.23

		N-REEL	N-THEO	ECART	REDUIT	MOYENNE	TYPE
1	SWANN	6414	6482	-68	-0.91	0.094	4.36
2	FLEUR	7908	7986	-78	-0.95	-0.142	1.32
3	GUERM.	8358	8735	-377	-4.47	-1.511	1.25
4	SODOME	7657	7909	-252	-3.10	-0.921	1.26
5	PRISO.	6298	6053	245	3.37	0.272	1.88
6	FUGIT.	4545	4086	459	7.51	2.374	2.57
7	TEMPS	5394	5320	74	1.07	0.608	1.31



Graphique 1. Courbe des subordonnants

Si l'on enferme dans la *Recherche*, sans référence à l'extérieur, des lignes de force apparaissent parmi les mots grammaticaux et l'analyse factorielle en donne la traduction dans le graphique 2.

Graphique 2 : Analyse factorielle des mots grammaticaux

Les sept parties de la *Recherche* décrivent un arc de cercle autour du centre dans l'ordre chronologique[30], les trois premières occupant le haut du graphique, entourées de coordinations, de prépositions et d'articles, et les quatre dernières se situant au bas de la figure, avec les négations, les

30. Remarquons toutefois que le *Temps retrouvé* a tendance à se rapprocher de *Swann* et des *Jeunes filles en fleur*, et cette anomalie apparente, confirmée dans de nombreuses autres analyses de ce type, jette quelque lumière sur la genèse de l'oeuvre. On sait en effet qu'une large portion du *Temps retrouvé* a été rédigée dès le début de l'entreprise.

démonstratifs et les subordonnants. Si l'on ajoute aux mots grammaticaux les mots sémantiques, les catégories nominales se portent vers le haut, du côté des premiers textes, et le verbe vers le bas, du côté des derniers textes. Il s'agit là en fait d'un système dont la logique s'impose aussi à l'ensemble du corpus du T.L.F. de 1789 à 1964 et où l'on voit pareillement progresser les verbes avec les négations, les indéfinis, les interrogatifs et les subordonnants, et régresser les substantifs avec les articles, les prépositions, les coordinations et les relatifs. Est-ce à dire que Proust se laisse emporter par le courant de la langue? Ou bien s'agit-il de quelque loi de vieillissement individuel, que nous avons observée aussi chez Giraudoux et chez Chateaubriand, où le rapport verbes/substantifs va croissant avec l'âge, comme chez Proust ?[31]

3.3. La place nous manque pour décrire plus précisément la phrase de Proust. Une incursion rapide du côté des signes de ponctuation confirme un trait qui n'a échappé à personne, pas même au plus inattentif des lycéens : la phrase de Proust est longue. On y relève 29,21 mots en moyenne contre 14,60 dans l'ensemble du corpus du T.L.F., c'est-à-dire exactement deux fois plus[32]. Si l'on rapproche Proust de la prose de son temps, la disproportion est encore plus forte (12,80, soit un rapport de 2,25). Encore ne tenons-nous pas compte d'un artifice de Proust dont la ponctuation contrarie la syntaxe, et notamment quand un point s'interpose entre la principale et la subordonnée qui suit[33]. L'originalité de Proust est d'autant plus grande que le mouvement de la littérature depuis 1789 tend vers la phrase courte[34]. Et la comparaison avec Chateaubriand est révélatrice sur ce point : avec 17,8 mots en moyenne par phrase, la mesure rythmique de Chateaubriand n'a pas l'ampleur de celle de Proust. La phrase de Chateaubriand n'en a pas non plus la complexité syntaxique : les excédents qu'on vient de constater chez

31. Les variations du genre littéraire peuvent rendre compte en partie des faits observés, dans le cas de Giraudoux. Mais le témoignage de Zola sera plus difficile à récuser quand sera achevée l'étude que nous entreprenons présentement et qui porte sur l'intégralité des *Rougon-Macquart*.

32. On ne tient pas compte ici des noms propres ni des mots étrangers et l'ambiguïté du point n'a pas été dissipée. En réalité, si l'on raffine les données et si l'on n'accepte comme séparateurs de phrase que les ponctuations fortes (. ? et !), la moyenne s'élève à 15,82 dans le corpus du *T.L.F.* et dépasse 30 mots chez Proust.

33. C'est le cas des phrases introduites par *de sorte que*, dont le statut est assimilé à celui de l'adverbe de phrase *ainsi*. On trouve deux exemples de cette construction pages 554 et 555 *d'Albertine disparue*.

34. Le coefficient de corrélation chronologique est significatif (r =0,60)

Proust parmi les mots grammaticaux où s'articule la syntaxe, sont déficitaires chez Chateaubriand et notamment les relatifs et les subordonnants. Tandis que la phrase de Chateaubriand regorge de substantifs, d'adjectifs et de prépositions, enveloppés dans un mouvement souple et ondulatoire, celle de Proust a des vertèbres apparentes, des articulations serrées qui doivent moins à la rhétorique qu'à la logique. Le style de l'Enchanteur est celui de l'évocation quand le style de Proust est fait d'explication. Et l'explication est souvent laborieuse, elle s'accomplit par des précisions, des analogies, des distinctions, des approfondissements, des repentirs et mille parenthèses[35].

4. La Vision

Mais on aurait tort de s'appesantir trop longtemps sur les particularités lexicales, syntaxiques ou rythmiques d'un écrivain qui a une trop haute idée de l'art pour réduire son attention aux seuls procédés. « Le style pour l'écrivain, déclare-t-il dans le *Temps Retrouvé*, aussi bien que la couleur pour le peintre, est une question non de technique mais de vision »[36].

4.1. Nous ne donnerons de cette vision que quelques illustrations en renvoyant le lecteur à notre ouvrage pour de plus amples détails. Ainsi la liste que nous proposons dans le tableau 4 ne représente qu'un faible extrait du vocabulaire spécifique de Proust dont 4.733 éléments dépassent le seuil habituel de signification ($z > 2$ ou $z < -2$), soit 25% du vocabulaire. C'est dire le caractère original de l'univers de Proust qui puise relativement peu dans le fond commun de son époque. Parmi ces mots spécifiques, plus de deux sur trois représentent un déficit (3.309 négatifs pour 1.424 positifs). Nous avions déjà souligné ce fait en étudiant la structure lexicale et la richesse du vocabulaire.

35. Ou plus exactement 3886. Parenthèses, tirets et guillemets sont les seuls signes de ponctuation excédentaires chez Proust. Là-dessus, l'ordinateur apporte sa voix à l'archilecteur de Proust, aux côtés de Léo Spitzer, Curtius, Le Bidois, Léon Guichard, Ramon Fernandes et Léon Pierre-Quint. La répulsion que Proust éprouve pour les blancs, les alinéas et les signes de ponctuation est confirmée par les chiffres : les six signes principaux de ponctuation (.!?;:,) ont des déficits considérables chez Proust (écarts réduits respectivement de -129, -62, -42, -60, -33, -5).

36. *Le temps retrouvé* p. 895. Ce passage est cité par Gérald Antoine et par bien d'autres.

VOCABULAIRE NEGATIF			VOCABULAIRE POSITIF		
-022.79	531	main	+080.04	825	duchesse
-021.91	212	âme	+051.88	702	princesse
-021.09	321	enfant	+045.14	474	baron
-018.37	229	terre	+038.58	559	duc
-017.22	1523	homme	+034.83	1107	plaisir
-016.74	471	tête	+029.66	519	hôtel
-015.47	175	bras	+028.57	455	tante
-015.43	347	nuit	+027.18	391	genre
-014.74	353	an	+026.12	554	habitude
-013.20	40	justice	+026.06	1421	mère
-012.95	177	silence	+025.72	177	marquise
-012.63	386	porte	+025.54	1613	moment
-012.34	26	prêtre	+025.03	162	plage
-012.20	70	cri	+024.80	131	amabilité
-012.17	418	voix	+024.68	1486	air
-011.97	24	abbé	+023.91	227	jalousie
-011.77	75	dieu	+023.70	455	prince
-011.77	34	école	+022.66	183	valet
-011.70	586	coeur	+021.90	301	relation
-011.60	139	droit	+021.06	1091	nom
-011.55	2	juin	+020.22	80	altesse
-011.46	78	sang	+020.19	1498	ami
-011.28	154	travail	+020.00	462	salon
-011.20	42	soldat	+019.90	297	soirée
-011.01	21	religion	+019.75	336	maîtresse
-010.99	105	peuple	+019.55	106	patronne
-010.97	41	chien	+019.48	61	clan
-010.88	10	juillet	+018.65	260	charme
-010.83	7	octobre	+018.55	431	réalité
-010.80	14	mai	+018.48	350	souffrance
-010.77	269	eau	+018.37	685	désir
-010.75	5	novembre	+018.31	2035	femme
-010.64	154	ombre	+018.26	1063	gens
-010.63	3	avril	+018.04	740	cause
-010.60	2	décembre	+017.92	400	parent
-010.58	16	misère	+017.82	706	effet
-010.56	33	genou	+017.26	526	dame
-010.53	70	doigt	+017.25	288	cousin
-010.52	49	armée	+016.18	1736	fois
-010.46	116	mur	+015.00	229	midi
-010.37	6	août	+014.93	383	impression
-010.28	9	septembre	+014.85	149	revanche
-010.26	66	paix	+014.43	66	particularité
-010.23	3	février	+014.25	380	conversation
-010.21	189	bois	+014.18	242	promenade
-010.19	78	foi	+014.15	67	dîners
-010.14	15	chapitre	+014.09	236	égard
-010.10	219	pays	+013.71	315	mémoire
-009.77	61	ennemi	+013.63	106	invitation
-009.76	319	matin	+013.43	106	faubourg
-009.70	138	feu	+013.41	39	blanchisseur
-009.67	220	histoire	+013.38	109	élégance
-009.67	199	pièce	+013.35	522	façon
-009.64	31	vin	+013.16	88	chic
-009.57	50	dos	+012.91	81	ambassadeur
-009.56	65	papier	+012.91	344	voiture
-009.48	20	mars	+012.84	136	politesse
-009.43	115	épaule	+012.71	359	visite
-009.41	56	dimanche	+012.61	457	cas

Tableau 3. Le vocabulaire caractéristique de Proust (extrait)

SUBSTANTIFS					
+017.15	190	tante	+003.63	112	coeur
+012.07	32	pianiste	+003.59	72	rose
+011.17	41	clocher	+003.53	14	sol
+009.53	36	piano	+003.50	52	fenêtre
+009.45	191	père	+003.49	26	cocher
+009.23	71	docteur	+003.46	50	couleur
+008.27	101	rue	+003.44	49	lit
+008.18	71	maman	+003.42	16	bal
+007.84	110	parent	+003.36	18	délicatesse
+007.71	63	bois	+003.34	19	essence
+006.66	39	allée	+003.27	76	porte
+006.59	20	messe	+003.24	78	phrase
+006.58	17	samedi	+003.22	27	bord
+006.57	54	peintre	+003.18	20	paysage
+006.38	44	morceau	+003.12	19	village
+006.27	19	jardinier	+003.11	42	douceur
+006.26	63	église	+003.10	11	herbe
+006.23	48	arbre	+003.10	15	neige
+006.07	17	aubépine	+003.07	286	fois
+006.07	17	gothique	+003.07	10	notions
+005.87	23	dimanche	+003.05	14	pointe
+005.69	39	pierre	+003.05	20	renseignement
+005.58	56	jardin	+003.00	22	chic
+005.55	31	cuisine	+002.99	26	front
+005.55	69	eau	+002.96	27	épaule
+005.35	113	pensée	+002.93	21	domestique
+005.29	18	bouquet	+002.93	50	lumière
+005.29	18	mauve	+002.87	14	agitation
+005.23	15	feuillage	+002.87	14	branche
+005.23	37	odeur	+002.85	87	tête
+005.22	17	violette	+002.84	23	détail
+005.20	31	champ	+002.83	52	charme
+005.18	14	toit	+002.83	10	marché
+005.09	17	curé	+002.83	28	parties
+005.06	28	feuille	+002.77	16	horizon
+005.01	16	parc	+002.75	21	thé
+004.97	30	sonate	+002.75	130	côté
+004.97	85	fleur	+002.75	92	place
+004.88	59	ciel	+002.73	20	fruit
+004.87	29	rayon	+002.70	45	rêve
+004.82	23	pluie	+002.69	28	coin
+004.78	34	mur	+002.69	64	joie
+004.77	32	reflet	+002.65	50	chemin
+004.66	53	âme	+002.64	98	visage
+004.62	21	parfum	+002.61	10	syllabe
+004.56	73	soleil	+002.60	19	plaisanterie
+004.51	17	vitrail	+002.59	184	plaisir
+004.36	160	soir	+002.55	260	moment
+004.31	18	jaune	+002.55	12	notion
+004.30	16	soie	+002.54	9	rocher
+004.28	21	chaud	+002.54	9	tapisserie
+004.21	45	fête	+002.52	18	dieu
+003.85	12	sauvage	+002.50	17	doigt
+003.85	12	occupation	+002.50	17	hiver
+003.83	27	lune	+002.50	45	midi
+003.80	19	lampe	+002.50	20	passage
+003.72	12	volet	+002.48	11	boulevard
+003.71	61	bonheur	+002.47	47	promenade
+003.70	52	oncle	+002.43	9	flamme
+003.68	70	image	+002.42	16	nuage
+003.64	123	maison	+002.41	163	amour
+003.64	19	printemps	+002.41	32	campagne

Tableau 4. Le vocabulaire caractéristique de *Swann*

Mais le contenu nous important davantage présentement, jetons un oeil sur les premiers mots du tableau 4. Point de surprise : les éléments que nous trouvons en tête de liste sont ceux dont la relation aux thèmes proustiens est la plus étroite. Le milieu social dans lequel évolue le roman est circonscrit par les quatre premiers termes qui désignent les acteurs aristocratiques de la comédie proustienne : *duchesse, princesse, baron, duc.* Un peu plus loin dans la liste, on rencontre la *marquise,* le *prince,* les *valets,* les *altesses,* l'*ambassadeur,* le *marquis, etc.* Notons au passage que les dames passent devant les hommes à titre égal[37]. Les événements ou rites de la vie mondaine apparaissent dans la même liste (*salon, invitation, visite, visiteur, matinée, soirée, dîners, thé,* etc.) comme aussi les valeurs qu'on y prône : *élégance, chic, raffinement, goût, politesse, amabilité, beauté, prestige,* tandis que les relations de parenté *s'*affirment dans ce milieu mi-aristocratique, mi-bourgeois (*tante, mère, parent, cousin, nièce, fille, femme, mari, oncle, neveu, famille*). Seuls les personnages appartiennent ici au monde concret, en dehors de quelques rares objets (ou lieux) privilégiés : *hôtel, plage, mer, digue, falaise, clocher, rose, vitrail.* Tandis que le cadre spatial reste vague (*distance, intervalle, trajet, lieu, côté*), le champ psychologique s'ouvre largement à l'analyse (*signification, valeur, erreur, caractère, qualité, originalité, individualité, particularité, réalité, cause, effet, possibilité, altération, différence*) et aucun recoin de la Carte du Tendre n'est ici ignoré (*admiration, charme, agrément, bonté, tendresse, douceur, satisfaction, plaisir, désir, avances, amour, curiosité, soupçon, déception, mensonge, vice, jalousie, revanche, souffrance, maladie, dédain, froideur, indifférence, séparation*)[38]. Là est la dominante de l'univers proustien, et cela ne surprendra personne : le début de la liste ne pouvait manquer de souligner ces caractéristiques très évidentes[39]. Les révélations et les

37. On comparera dans le tableau 3 précédent les pronoms masculins *il* et *ils* qui se rangent à gauche parmi les déficits (-38,51 et -16,14), et les pronoms féminins *elle* et *elles* qu'on trouve à droite, dans la zone des excédents (52,15 et 21,92).

38. La liste de gauche qui restitue les spécificités négatives, c'est-à-dire les mots relativement peu employés par Proust, donne la contre-épreuve de la liste positive. Dès qu'il s'agit du monde extérieur ou de certaines réalités de la vie sociale, le désintérêt de Proust est manifeste. Ainsi pour les parties du corps (*main, tête, bras, genou, doigt, dos, épaule, bouche, jambe*), les éléments naturels (*terre, eau, feu, ciel, vent, sol, arbre, bois, route*), les repères du calendrier (*juin, juillet, octobre,* etc.), l'univers du travail (*justice, misère, pauvre, travail, lutte, droit*), de la religion (*prêtre, abbé, dieu, religion*) ou de la politique (*soldat, armée, paix, pays, peuple, ennemi, gouvernement*).

39. Les « évidences » du sentiment linguistique sont souvent rassurées par la ratification statistique, même si l'importance d'un mot ou d'un thème ne se mesure pas

surprises apparaissent plus loin dans la liste, dans la partie immergée de la masse lexicale. Nous n'en donnerons qu'un exemple : celui des couleurs. Certains passages colorés de la poésie proustienne peuvent faire illusion, la vision de Proust fait un usage discret de la couleur malgré les roses, les lilas et les aubépines. Tous les adjectifs de couleur sont déficitaires dans la *Recherche* par rapport à la littérature de la même époque (*blanc* -6,88, *jaune* -6,65, *rouge* -5,78, *vert* -8,63, *noir* -11,25, et même *bleu -0,12*). Deux nuances voisines cependant font exception : le *mauve* 7,78 et le *violet* 2,22 et aussi les tons imprécis qui ont recours au suffixe *âtre : blanchâtre* 2,73, *bleuâtre* 2,73, *rougeâtre* 2,34, *verdâtre* 1,13. Peut-être peut-on voir quelque héritage à peine conscient du symbolisme.

4.2. Bien entendu, la *vision* proustienne diffère d'un livre à l'autre et les couleurs sont plus vives dans *Swann* que dans la suite du roman, comme le montre la confrontation de *Swann* et de *Sodome* :

Le tableau 5 restitue l'atmosphère heureuse, colorée, champêtre, familiale et printanière du premier livre de la *Recherche*, où visiblement *Combray* a laissé des traces plus nettes qu'*Un amour de Swann*. L'explication réside dans le fait que le récit de l'enfance est unique, alors que l'expérience de l'amour et du monde est répétitive et que l'aventure de Swann annonçant celle de Charlus et du narrateur lui-même, le vocabulaire s'y trouve quelque peu banalisé. A l'opposé de *Swann*, le *Temps retrouvé* délaisse l'individu pour la condition humaine (ce qui suscite l'apparition du pluriel et du collectif: *ils, les, des, eux, ceux, leurs,*

---

seulement par la fréquence. Il est parfois plaisant que de bons auteurs dénoncent à l'avance le témoignage de la statistique – qui pourtant leur donnerait raison. Ainsi fait Conrad Bureau lorsqu'il évoque les surcodages « qui malgré la fréquence brute peu élevée de signes comme souvenir, se rappeler et retrouver, font de ces derniers les signes les plus privilégiés de l'œuvre » (ouvrage cité, p. 184). En fait ces mots appartiennent au vocabulaire hautement spécifique de Proust :

Souvenir	f 527	z 10,43,		*souvenir*	f 196	z 7,40
Rappelant	f 68	z 10,46,		*rappelé*	f 37	z 3,43
Rappeler	f 578	z 15,19,		*retrouver*	f 436	z 7,37

Il est vrai que Proust lui-même se fait du vocabulaire spécifique une idée paradoxale, puisque l'absence du mot ne décourage nullement son intuition : « je trouve *aberrant* extrêmement Renan. Je ne crois pas que Renan ait jamais employé ce mot. Si je le trouvais dans son oeuvre, cela diminuerait ma satisfaction de l'avoir trouvé ». *Correspondance Générale*, IV, p. 227. Il est vrai aussi que Proust n'a pas dédaigné des relevés statistiques, comptant chez Flaubert les *et*, les *tandis que*, les imparfaits et les participes présents.

mêmes, société, gens, êtres), le sentiment pour la réflexion (*vérité, réalité, essence, valeur*), le bonheur pour la mort (*mourir, mort, tué, guerre*), l'amour pour l'art (*livre, écrivain, art, oeuvre, littérature*), et l'espace pour le temps (*an, année, époque, durée, âge, temps, vie, génération, changement, mémoire, vieillesse, vieillard, jeunesse*). Des listes de spécificités analogues à l'extrait du tableau 5 ont été dressées pour chaque sous-ensemble de la *Recherche*, par catégorie grammaticale. Et de livre en livre, on voit la vision de Proust s'assombrir, comme le souligne Maurice Bardèche : « Est-ce une évolution, un approfondissement de son art, les fruits amers que portaient *Un amour de Swann* et *Les jeunes filles* ? La lumière heureuse qui baignait ces premières oeuvres a été remplacée par des couleurs d'orage, cuivrées, violacées, inquiétantes. »[40] Le point ultime de cette évolution n'est pas nécessairement le *Temps retrouvé*, dont beaucoup de passages ont été écrits dès 1914, mais assez souvent l'épisode d'Albertine (*Prisionnière* et *Albertine disparue*[41]), ce que l'on peut voir dans beaucoup de nos courbes constituées, à l'image des graphiques 3 et 4, autour d'un champ sémantique. Ainsi le déclin des termes qui évoquent la nature (graphique 3) est suspendu quand on aborde le *Temps retrouvé* (mais il est vrai qu'en privilégiant les entités comme *monde* et *matière*, Proust y parle moins en poète qu'en philosophe de la nature).

Inversement la pente statistique du temps monte avec la chronologie (graphique 4). Observons toutefois que l'unanimité ne règne pas parmi les constituants de cette classe sémantique et qu'en particulier le sentiment des rythmes courts (*minute, heure, jour, nuit, midi, soir, matin*), encore vif dans *Swann*, s'éteint au fil de l'oeuvre alors que s'impose de plus en plus l'obsession d'un temps large et enveloppant qui se mesure en années, en générations et conduit à la mort[42]. Le graphique 4 permet cette finesse d'analyse puisqu'elle individualise (par un code spécifique) l'évolution de chacun des constituants de la série.

---

40. *Marcel Proust romancier*, tome l, p. 249.

41. Comme notre index renvoie à l'édition de la Pléiade, nous avons gardé le titre de *Fugitive* qui s'y trouve, même si ce texte est plus connu sous le nom d'*Albertine disparue.*

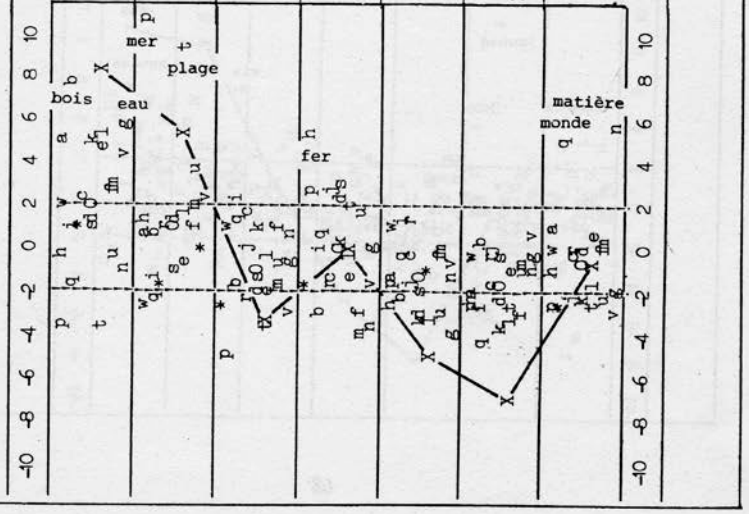
42. Bien entendu le thème voisin de la. mort suit la progression du temps. Voici la suite des 7 écarts réduits qu'on observe successivement :

– pour la *mort* : -3,6 ; -7,2 ; 4,4 ; 3,1 ; 0,7 ; 4, 7 ; 9,3.
– pour *mourir* : -2,8 ; -0,6 ; -0,9 ; -0,5 ; -1,4 ; 1,3 ; 6,1.

NOMBRE = 22
OCCURRENCES = 5357

CODE	MOT	FREQUENCE	ECARTS REDUITS EPOQUE	ECARTS REDUITS CORPUS
a	arbres	123	-4.55	-3.47
b	bois	189	-10.21	-7.47
c	campagne	151	-0.37	0.37
d	chemin	208	-2.14	-0.47
e	ciel	231	-8.61	-10.60
f	couleur	147	1.09	1.57
g	eau	221	-10.05	-8.89
h	fer	100	-3.10	-3.65
i	feu	118	-9.53	-9.83
j	fleur	100	1.92	1.47
k	fleurs	272	5.94	5.73
l	jardin	161	-4.40	-1.00
m	lumière	221	-5.92	-3.86
n	matière	127	4.51	0.67
p	mer	362	9.60	4.68
q	monde	1298	8.50	5.37
r	nature	310	-0.32	-9.48
s	pays	219	-10.09	-9.07
t	plage	153	25.56	29.06
u	route	118	-8.76	-8.77
v	soleil	319	-4.25	-2.98
w	terre	209	-18.17	-17.55
	TOTAL	5357		
	MOYENNE		-1.97	-2.18

SWANN	FLEUR	GUERM.	SODOME	PRISO.	FUGIT.	TEMPS
958	1066	910	916	578	328	601
8.41	5.36	-3.29	0.25	-4.79	-6.81	-0.47
2.088	1.283	-0.769	0.055	-1.166	-1.396	-0.429
2.94	3.18	1.89	2.38	1.46	1.35	2.26

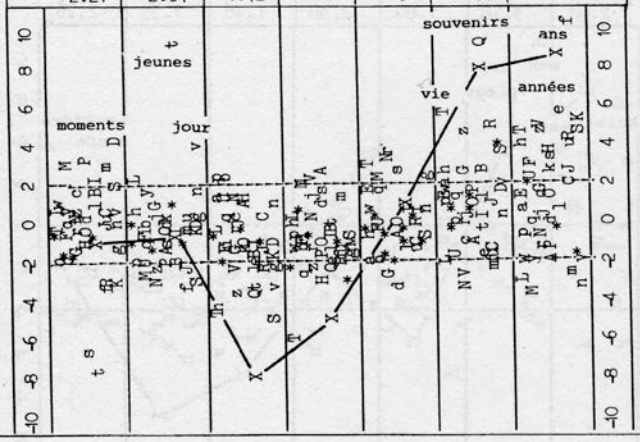


Graphique 3. Courbe du champ sémantique de la nature

NOMBRE = 45
OCCURRENCES = 18178

CODE	MOT	FREQUENCE	ECARTS REDUITS EPOQUE	ECARTS REDUITS CORPUS
a	ancien	199	5.66	6.30
b	ancienne	104	2.93	2.49
c	anciens	111	3.13	0.37
d	année	193	-0.79	1.06
e	années	359	8.20	9.34
f	ans	286	-14.15	-13.05
g	avenir	137	1.01	-2.68
h	changé	144	6.89	7.90
i	changement	107	9.30	6.62
j	changer	135	7.62	5.60
k	époque	211	8.95	3.77
l	habitude	458	25.37	25.82
m	heure	759	2.83	6.50
n	heures	481	-5.28	-5.76
p	histoire	175	-9.09	-10.34
q	instant	550	11.16	14.56
r	instants	115	3.62	6.86
s	jeune	889	7.04	6.85
t	jeunes	435	10.21	13.97
u	jeunesse	127	-2.10	-1.80
v	jour	1358	4.62	6.29
w	journée	150	-1.99	-1.03
y	jours	766	3.71	1.16
z	mémoire	269	13.81	11.76
A	minutes	155	2.24	2.13
B	mois	205	-7.07	-8.97
C	moment	1368	22.35	22.76
D	moments	245	13.09	14.22
E	nouveau	422	0.83	2.88
F	nouveaux	103	3.53	1.06
G	nouvelle	322	3.21	2.84
H	nouvelles	219	5.37	1.93
I	nuit	296	-15.48	-13.66
J	oublié	135	6.25	6.61
K	passé	454	5.26	6.01
L	semaine	103	-1.11	0.13
M	soir	720	-1.28	2.79
N	soirée	248	18.61	18.20
P	soirs	116	8.85	12.40
Q	souvenir	461	18.68	20.11
R	souvenirs	196	7.40	7.39
S	temps	1637	8.26	5.08
T	vie	1727	5.47	9.42
U	vieille	219	-4.00	-0.25
V	vieux	309	-10.91	-7.35
	TOTAL	18178		

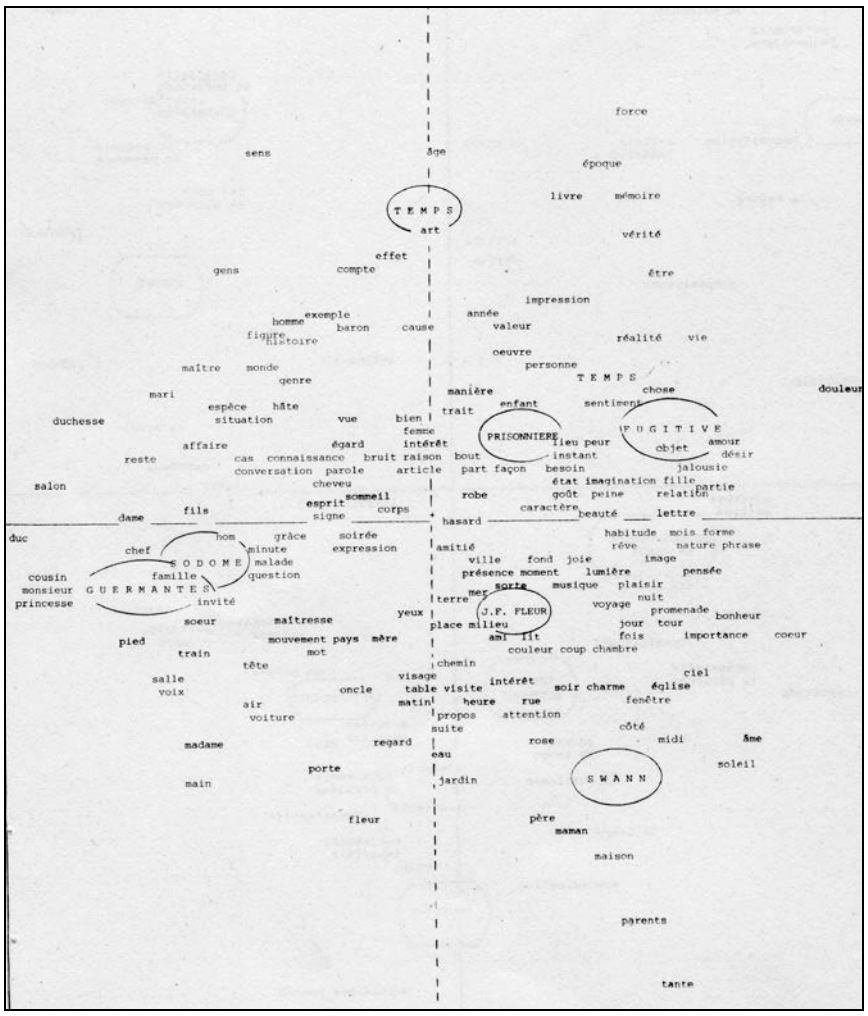
SWANN	FLEUR	GUERM.	SODOME	PRISO.	FUGIT.	TEMPS
2491	3085	2988	2844	2416	1906	2448
-0.83	-0.62	-7.99	-4.80	1.19	8.18	8.67
-0.194	-0.165	-0.951	-0.685	0.034	1.139	1.339
2.21	2.04	1.72	1.61	1.56	2.39	2.94



Graphique 4. Courbe du champ sémantique du temps

4.3. Cet assombrissement de la vision est visible d'un coup dans l'analyse factorielle que nous proposons (graphique 5) et qui concerne les 200 substantifs les plus fréquents de l'oeuvre. On voit se constituer des alliances et des oppositions : *Swann* et les *Jeunes filles* font cause

commune dans le quadrant inférieur droit. *Guermantes* et *Sodome* occupent la partie gauche tandis que la moitié supérieure du graphe est réservée aux derniers livres, de la *Prisonnière* au *Temps retrouvé*.



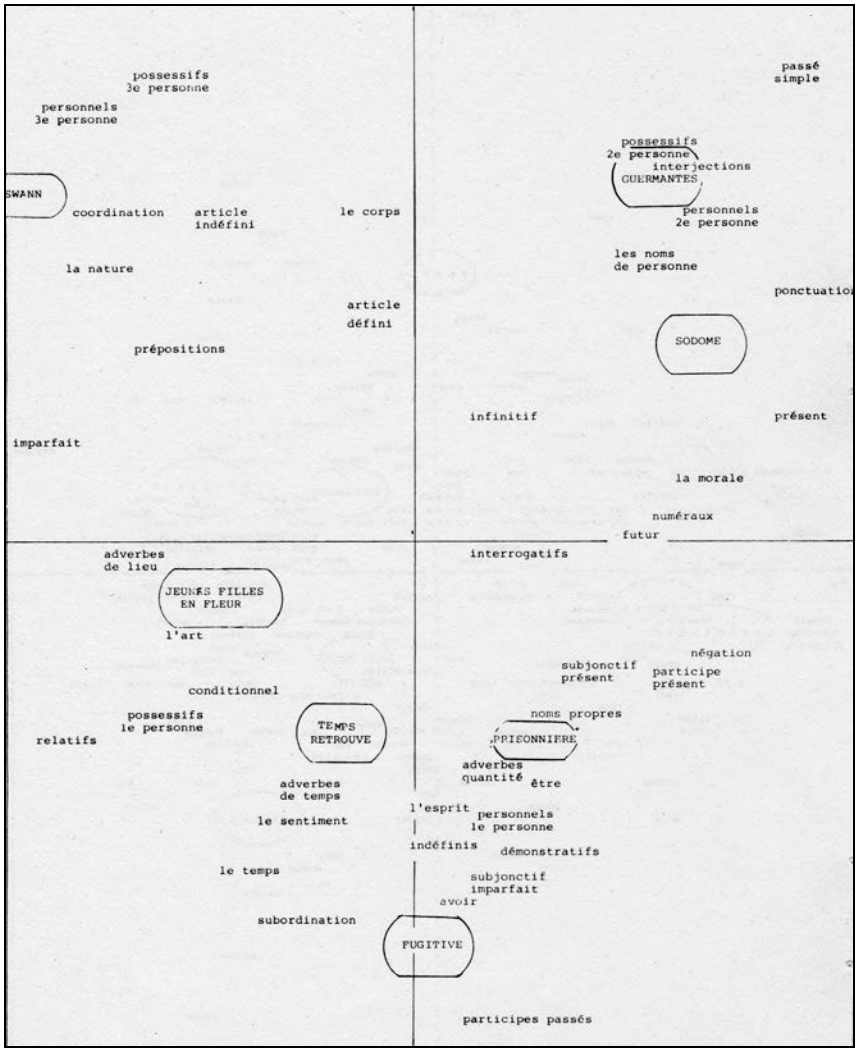
Graphique 5. Analyse factorielle des substantifs (facteurs 1 et 2)

Dans tous nos graphiques, nous observons ces trois pôles et la répartition des livres en trois paires, *Swann* et les *Jeunes filles*, *Guermantes* et *Sodome*, *Prisonnière* et *Fugitive*. Seule la place du *Temps retrouvé* est variable et incertaine – ce qui semble dénoncer une genèse complexe. Les indications des listes de spécificités se retrouvent ici, et il

est facile de superposer le tableau 5 et le bas du graphique, car les mêmes éléments apparaissent : *tante, parents, maman, père, maison, jardin, eau, ciel, âme, coeur, etc.* En négligeant la partie gauche (où se portent les acteurs de la comédie mondaine dans le voisinage de *Guermantes* et de *Sodome*), on touche du doigt l'évolution de l'univers proustien si l'on explore le graphique de bas en haut. Le paradis propre à l'enfance se maintient encore dans l'adolescence (l^er^ tiers inférieur) et dans les *Jeunes filles*, l'atmosphère reste heureuse et lumineuse (*charme, bonheur, plaisir, joie, rêve, coeur, lumière, couleur, ciel, mer terre, nature, jour, nuit*). Mais en passant la ligne médiane (axe des *x*), on bascule dans l'univers maudit de l'amour-maladie qui se déclare et se développe dans la *Prisonnière* et la *Fugitive*. Un halo sombre entoure ces deux constellations dans le graphique : *amour, désir, imagination, besoin, jalousie, peur, peine, douleur,* ce dernier terme étant le plus significatif, à l'extrême droite du graphique. L'analyse factorielle des verbes donne les mêmes indications (comme aussi celle des adjectifs et des participes). La zone d'influence d'Albertine y est pareillement fréquentée par le soupçon, la souffrance, la séparation et l'oubli : *savoir, comprendre, ignorer, expliquer, craindre, croire, supposer, cacher, apprendre, tromper, aimer, désirer, perdre, partir, quitter, revenir, souffrir, souvenir, oublier.*

4.4. Mais la vision proustienne n'est pas réduite aux éléments sémantiques du discours. Elle a un caractère global qui intègre la syntaxe, le rythme, le choix des catégories, des registres et des temps verbaux. Aussi bien avons-nous essayé de réunir en une seule analyse – ou synthèse – tous les éléments comptables qu'on a pu relever dans le discours proustien. Le résultat figure dans le graphique 6 où l'on reconnaît bien les trois pôles de la *Recherche* : *Swann* et les *Jeunes filles* dans la moitié gauche, *Guermantes* et *Sodome* dans le quadrant supérieur droit, la *Prisonnière* et la *Fugitive* dans le quadrant inférieur droit.

Une telle analyse est faite de compromis multiples et les distances sont parfois trompeuses quand on néglige les facteurs 3 et 4 qui complètent la représentation.



Graphique 6. Analyse factorielle globale de la *Recherche du temps perdu*

Néanmoins, on voit clairement comment se répartissent les grands champs sémantiques, les personnes, les catégories, les temps verbaux. Dans le premier groupe, règne l'imparfait nostalgique avec l'évocation de la nature, la prise en compte du corps et le recours aux catégories descriptives du substantif (articles et prépositions). L'épisode de « Swann » impose ici la troisième personne, qui cède la place à la seconde dans le deuxième groupe. Là l'univers de *Guermantes* et de *Sodome* multiplie en effet les situations de dialogue, ce qui entraîne

l'abondance des interjections et des signes de ponctuation et le recours privilégié au présent et au futur. Et quand les contraintes de la narration imposent le passé, c'est le passé simple qui est ici préféré. Du côté sémantique, les jugements moraux se développent alors que c'est le sentiment et la connaissance qui prévalent à partir de la *Prisonnière* (bas du graphique). Ici Proust utilise les temps composés (*être, avoir* et participes passés) et son discours multiplie les subordinations et les verbes, et parallèlement les négations et les subjonctifs. Ce n'est plus un discours descriptif, narratif ou dialogué, mais un soliloque intériorisé où s'impose la première personne. Enfin, la boucle se ferme avec le *Temps retrouvé* qui prend à son compte les choix de la *Fugitive*, mais en y ajoutant le thème du temps (substantifs et adverbes) et celui de l'art et en se rapprochant aussi des premiers livres de la *Recherche* par le recours à l'imparfait et au conditionnel.

Nous arrêterons là, sur cette image spectrale de la *Recherche*, notre survol rapide de l'oeuvre proustienne. Le lecteur peut être réticent devant ce raccourci synthétique dont les étapes intermédiaires semblent escamotées. Elles ne le sont point dans notre ouvrage auquel nous le renvoyons.