David ERL1CH
Équipe *Sémantique des textes*
Centre de linguistique française, Université de Paris IV

UNE MÉTHODE D'ANALYSE THÉMATIQUE EXEMPLES DE L'ENNUI ET DE L'AMBITION

1. Objectifs

Notre objectif est de proposer des méthodes lexicométriques, éventuellement systématisables, dans le cadre de l'étude thématique. Nous utilisons les ressources des systèmes FRANTEXT et DISCOTEXT développés par l'iNaLF. Dans un premier temps, nous esquisserons une méthode d'interrogation de ce corpus. Ainsi étudierons-nous comment localiser des *extraits* illustrant un thème à partir de formes[1]. De ces extraits nous tirerons des *corrélats* que nous étudierons de deux façons. Autour du thème de l'ennui, en nous aidant d'indicateurs statistiques, nous examinerons comment trouver de nouveaux extraits. Nous mettrons en perspective la méthode. Nous élargirons la problématique avec le thème de l'ambition en étudiant notamment les rapports qu'entretiennent les corrélats avec la trame narrative.

2. Présentation de la méthode

2.1. Commencer la recherche par des formes d'entrées

Le thème est représenté par une séquence linguistique (une phrase, un groupe nominal, un nom propre ou commun).

Le corpus littéraire d'étude n'est accessible que par extraits interrogeables par *mots-clés*. Il nous est donc nécessaire de traduire le thème en une liste de mots-clés que nous appellerons *formes d'entrée*. Nous nous placerons d'un point de vue pratique, en cherchant si une forme d'entrée peut permettre la sélection d'un grand nombre d'extraits qui illustrent le thème. Cela se vérifie si la fréquence de ces formes d'entrées dans le corpus d'étude n'est pas marginale ou si les extraits sélectionnés illustrent effectivement le thème.

Limitons-nous au cas où la forme d'entrée est un substantif, nom commun. La forme d'entrée qui se propose naturellement est le thème lui-même. Par exemple *ambition* pour l'ambition ; *ennui* pour l'ennui.

2.2.Difficultés

Le thème de l'ennui apparaît-il nécessairement partout où se trouve le vocable *ennui* ? Une fois la forme d'entrée choisie, nous devons éviter deux types de difficultés.

2.2.1. Le problème du bruit

Le bruit représente la quantité d'information parasite produite par le processus de recherche. En cas de bruit, les extraits obtenus sont nombreux, mais illustrent rarement le thème. Parmi les causes de ce phénomène, nous relevons :

1) l'entrée lexicale est polysémique ;

2) l'entrée lexicale possède plusieurs homographes ;

3) l'entrée lexicale possède une fréquence absolue extrêmement élevée. Des vocables courants se situent au sein d'extraits qu'il devient difficile d'exploiter.

2.2.2. La forme d'entrée est sous-reprêsentée dans le corpus d'étude

Parfois la fréquence absolue (en langue) de la forme d'entrée est faible. Certains intitulés thématiques ont été définis *a posteriori* par les critiques (l'auteur décrivait un thème sans le savoir). Un thème ne se « dit » pas, mais s'exprime au fil du récit romanesque. Par exemple *Les Choses* de Perec ne possèdent que trois occurrences de *ennui* alors que le thème de l'ennui y est largement présent.

2.3.Remèdes

Dans les deux cas, les mots-clés cernent insuffisamment le thème dans le discours, les formes d'entrée ne fournissent pas assez d'information. Ce supplément d'information nous est fourni par des *sources secondaires* : le dictionnaire général (qui peut fournir des formes dérivées) et le dictionnaire de synonymes. Ces outils procurent une *liste de formes supplémentaires* qui vont circonscrire ou élargir le spectre de résultats.

a) Pour restreindre le bruit, nous proposons d'introduire un critère exclusif en cherchant les extraits contenant deux mots-clés en cooccurrence (deux formes d'entrée).

Le problème de la distance (en terme de nombre de caractères) entre deux formes demeure ouvert. Il dépend des relations sémantiques qu'entretiennent les vocables dans le discours. Nous utilisons alors une liste de formes d'entrées pour obtenir les premiers extraits. Chaque extrait contiendra deux formes d'entrée. Ainsi, nous précisons le sens d'un terme en le restreignant à certains contextes d'emploi. Cette stratégie suppose que deux vocables auraient « une puissance thématique » supplémentaire : ils attirent les extraits pertinents.

b) Pour rompre le silence, nous proposons d'élargir la recherche à l'aide de formes supplémentaires qui accèdent tout simplement au titre de nouvelles formes d'entrée. Dans ce cas, nous élargissons le sens d'un terme à l'aide de termes voisins.
Ces solutions accroissent l'ordre de complexité du problème, car les mêmes questions de polysémie et d'homographie se reposent avec les nouvelles formes employées.

2.4. Extraits et listes de corrélats

A la suite de l'élargissement des requêtes, nous avons obtenu un nombre appréciable d'extraits. Chaque *extrait* relevé à partir d'une forme d'entrée précise le thème. Il draine un vocabulaire caractéristique. Ceci laisse penser que les formes les plus fréquemment en cooccurrence avec une forme d'entrée apportent un supplément d'information. Ces *corrélats* autorisent la recherche d'extraits nouveaux dans lesquels le thème reste présent.

3. Etude expérimentale autour de deux thèmes : l'ambition et l'ennui

3.1. Introduction

3.1.1. Corpus

Nous avons travaillé autour de deux thèmes :

—L'ennui dans un corpus de romans et de poésies : Maxime Du Camp :
Mémoires d'un suicidé ; Gustave Flaubert : *Madame Bovary* ; Charles Baudelaire : *Les Fleurs du mal* ; Paul Verlaine : *Poèmes saturniens* ; J.K.
Huysmans : À *rebours* ; Emile Zola : *La Joie de vivre* ; Jules Laforgue : *Les*
Complaintes ; Mallarmé : *Poésies.*

—L'ambition dans un corpus de romans : Honoré de Balzac : *Le Père Goriot,*
César Birotteau, Le Lys dans la vallée, Stendhal : *Le Rouge et le noir,* Gustave
Flaubert : *L'Éducation sentimentale.*

Le premier corpus nous a été suggéré par É. Martin (1991), le second fut établi d'après un dictionnaire thématique. Ces corpus seront désignés par le nom de

corpus d'étude. D'autre part nous avons défini *un corpus de référence* à partir des romans et poésies de la base DISCOTEXT. D nous sert à établir les fréquences de référence des différentes formes pour les tests statistiques[2].

3.1.2. Statistiques

Nous utiliserons l'outil statistique de deux façons.

—*Classification* : Les différentes listes de formes gagnent à être ordonnées suivant le résultats de tests statistiques (souvent on mentionnera le X2) qui fournissent des ordres de grandeur. Ils permettent d'apprécier en particulier les irrégularités de distribution de certaines formes suivant les corpus-thèmes ou les différentes oeuvres.

—*Sélection* : Devant l'abondance de résultats, il nous sera nécessaire d'éliminer certaines formes jugées non significatives. Ces critères d'exclusion revêtent essentiellement un intérêt pratique ; ils fixent arbitrairement un seuil à partir duquel les données véhiculent un bruit trop élevé.

Nous avons abusé de la loi normale : « On s'aperçoit très vite que seuls les vocables les plus fréquents se prêtent à son application » (DUGAST [1980]). Mais signalons que pour les faibles fréquences la loi de Poisson modélise bien mieux le comportement d'une forme. En fait, nous ne pouvons discuter ici des méthodes qu'il eût fallu employer dans les différents cas.

3.2. L'ennui

Nous examinerons comment les corrélats (formes extraites du corpus d'étude) et formes extraites de sources secondaires (formes extraites de dictionnaires) permettent la découverte d'extraits originaux. Nous comparerons les résultats obtenus à partir de ces deux types de formes. Puis nous envisageons les rapports des corrélats avec le thème et le corpus d'étude.

La forme d'entrée *ennui* compte 86 occurrences dans le corpus, soit 45 occurrences de plus que « le hasard » ne le permettrait.

3. 2.1. Sélection de nouvelles formes d'entrée

Établissons une première liste de formes qui apparaissent dans *la même phrase* que *ennui.* Ces formes sont recueillies manuellement : celui qui étudie le thème recueille les formes qu'il juge intéressantes[3]. Les plus fréquentes sont : *abattement (2), bâillant (2), dégoût (4), dégoût de la vie (1), désespérance (2), désœuvrement (2), engourdissement (2), existence (3), incuriosité (2), inaction (2), inutile (2), las (2), morne (2), mort (2), néant (3), regret (2), vide (3).*

Nous pouvons comparer cette liste à celle que nous aurait directement donné la consultation d'une source secondaire, par exemple le *Petit Robert* : abattement, *absurde,* dégoût, désœuvrement, inaction, *langueur, lassitude, mélancolie, monotone, oisiveté, spleen, vague,* vide. Remarquons qu'une grande partie de ces formes n'a pas été détectée parmi les extraits (en *italiques* les formes non relevées).

Dans le tableau statistique ci-dessous, nous présentons en gras les formes retenues : elles ont une fréquence faible dans le corpus ou elles ont une significativité élevée et une fréquence moyenne[4] ; en italique, les formes rejetées : elles ont une signification trop faible pour une fréquence trop forte[5].

Formes	Occurrences dans le contexte de *ennui*	fréquence dans le corpus d'étude	(f-fth)^ 2 / (fth)
FORMES OBTENUES DANS LES EXTRAITS			
abattement	2	**5**	*
dégoût	4	**32**	**13**
désespérance	2	**4**	*
désœuvrement	2	**10**	**43**
engourdissement	2	**4**	*
ennui	x	86	**129**
incuriosité	2	**3**	*
inaction	2	**3**	*
inutile	2	*47*	*6*
las	*2*	*30*	*9*
morne	2	*33*	*4*
mort	2	*299*	*13*
néant	3	*37*	*18*
regret	*2*	*46*	*33*
vide	*3*	*18*	*1*
FORMES FOURNIES PAR DES SOURCES SECONDAIRES			
absurde	source II	**9**	(0)
langueur	source II	**17**	**22**
lassitude	source II	**8**	**72**
mélancolie	source II	17	(0)
monotone	source II	27	**36**
oisiveté	source II	**10**	(14)
spleen	source II	11	**82**
vague	source II	44	(20)

♦ Test non applicable (fréquence trop faible)

Dans un deuxième temps, recherchons tous les extraits contenant les corrélats sélectionnés. Ces corrélats deviennent de nouvelles formes d'entrée : *abattement ; dégoût ; désespérance ; désœuvrement ; engourdissement ; ennui ; incuriosité.*

Ajoutons les formes données par les sources secondaires qui satisfont aux critères de sélection : *absurde ; langueur ; lassitude ; monotone ; oisiveté ; spleen.*

La recherche avec de nouvelles formes d'entrée nous fournit un certain nombre d'extraits. Nous examinons chaque extrait pour obtenir de nouveaux corrélats. Le tableau suivant synthétise les résultats.

Forme	Fréquence	%d'extraits pertinents	X2 (voir note S)
FORMES OBTENUES DANS LES EXTRAITS			
abattement	**S**	**100%**	*
dégoût	**32**	**41%**	**13,5**
désespérance	**4**	**100%**	**9,6**
désoeuvrement	**10**	**100%**	**43**
engourdissement	**4**	**0%**	
ennui	**86**	**58%**	**129**
inaction	**3**	**100%**	*
FORMES FOURNIES PAR DES SOURCES SECONDAIRES			
absurde	**9**	**11%**	**0**
langueur	**17**	**71%**	**22**
lassitude	**8**	**75%**	**72**
monotone	**27**	**59%**	**36**
oisiveté	**10**	**90%**	**15**
spleen	**11**	**82%**	**82**

Test non applicable (fréquence trop faible)

Les formes les plus significatives statistiquement sélectionnent les extraits pertinents les plus nombreux. Ce phénomène s'estompe pour des formes de fréquence élevée.

3.2.2. Des formes significatives non relevées dans le contexte de ennui : *spleen et oisiveté*

Les extraits relevés autour de ces deux formes drainent un vocabulaire caractéristique, comme nous le prouve le nombre de formes relevées. *Spleen* apparaît comme un bon équivalent d'*ennui.* Il n'a qu'un seul sens et s'emploie surtout littérairement. Mais on le trouve surtout chez Huysmans et Laforgue. Ainsi *spleen* n'apparaît-il pas dans *Les Fleurs du mal* (honnis « Spleen et Idéal », titre qui n'a pas été relevé) ! Ce type de forme se comporte comme si elle était
substituable à la forme d'entrée.

***spleen* (9 extraits pertinents sur 11 dépouillés) et *oisiveté* (9 pertinents extraits sur 10 dépouillés) sont tous deux significatifs (x^2 de 82 et 15 respectivement).**

« Justement, Lazare était la, assis près du fourneau, dans cette *oisiveté* fiévreuse *qui le dévorait.* » (Zola)

« Malade *d'oisiveté,* n'ayant goût à rien, il trouvait trop rude même de lire, et passait ses jours à se *dévorer* ». (Zola)

« et il n'était pas plutôt retombé dans l'oisiveté, qu'il *s'y dévorait* de honte et de malaise » (Zola)

« la solitude profonde et l'*oisiveté* absolue auxquelles j'étais condamné, les inquiétudes qui me *dévoraient...* » (Du Camp)

Nous voyons que *oisiveté* est souvent accompagnée de *dévorer.*

« Quand le *spleen le pressait,* quand les temps pluvieux d'automne, l'aversion de la rue, du chez soi, du ciel en boue, jaune, des nuages en macadam, l'assaillait » (Huysmans)

« ...H *tomba,* désorienté dans le *spleen* » (Huysmans)

« ...La dégoûtation de l'existence s'accentue et le *spleen écrase* » (Huysmans)

Chaque forme est employée de préférence par un auteur *(l'oisiveté* chez Zola, le *spleen* chez Huysmans). Le registre de sensations correspondant change également, d'un côté l'oisiveté ronge et dévore, de l'autre le spleen écrase et presse.

3.2.3. Des formes toujours relevées dans le contexte de ennui : incuriosité, inaction, engourdissement

Ces formes sont pratiquement toujours présentes avec la forme *ennui.* Elles appartiennent à des classes de fréquence faibles (moins de 10 occurrences dans le corpus d'étude). En voici des exemples :

Chez Baudelaire *Y ennui* est « le fruit de la mome *incuriosité* »

Chez Huysmans la langue latine « dégageait une telle *incuriosité,* un tel *ennui* qu'il fallait dans les études de linguistique au style français du siècle de Louis XIV pour en rencontrer une aussi volontairement débilitée, aussi solennellement harassante et grise ». Plus loin on retrouve les deux vocables en apposition : « après l'*ennui* et *l'incuriosité* du premier empire... »

Chez Du Camp : « — Pourquoi ne pas vivre ici, me disais-je, dans la nonchalance orientale, *tuant l'ennui par l'engourdissement* et les chagrins du bien-être ? »

Chez Huysmans : « Dans cet *engourdissement,* dans cet *ennui* désœuvré où il plongeait... »

Chez Zola : « Lui, malgré son indifférence, trouvait dans les querelles une secousse à *l'engourdissement de son ennui,* s'y entêtait souvent par cette distraction de se donner la fièvre »

3.2.4. Une forme non significative qui fournit peu de résultats : absurde

Absurde (2 extraits pertinents sur 9 extraits dépouillés), décrit incorrectement le thème (au XIX^{e} siècle). Cela se vérifie par le fait *qu'absurde* n'est absolument pas significatif dans le corpus (le test du x^ fournit un nombre 0). En tant qu'adjectif, il s'applique surtout à la réalité sociale : « absurde tyrannie » (Du Camp), « absurde prison » (Du Camp), « absurde législation » (Flaubert), « code théologique absurde » (Huysmans). La révolte et la violence de la prise de conscience de l'absurde s'oppose à l'apathie intellectuelle, l'absence de volonté.

3.2.5. L'emploi de monotone

Chez les poètes le monotone est associé aux mouvements naturels de l'eau sur une rive, le bruit du ressac, associé à une musique[6]. Il est étonnant de constater que le soleil lui aussi engendre la monotonie. Il rappelle le scintillement de l'eau

calme sous le soleil, qui vient se déverser régulièrement aux pieds du poète. Du Camp détaille cela : « La Seine noire et rapide se brisait aux piliers du pont, coulait en se rayant de *reflets de lumière* et semblait nous appeler par son murmure *monotone* et plaintif. »

Les extraits suivants, moins riches en corrélats, font appel une image mécanique de la monotonie : l'horloge, le bélier : « un coucou accroché au mur battait régulièrement son tic-tac *monotone* » (Du Camp) ; « bercé par le bruit *monotone* du balancier » (Du Camp) ; « *monotone* comme le ronflement d'une toupie » (Flaubert) ; « Mon esprit est pareil à la tour qui succombe / Sous les coups du bélier infatigable et lourd. / II me semble, bercé par ce choc *monotone, I* Qu'on cloue en grand hâte un cercueil » (Baudelaire).

3.3. Le thème et la liste de corrélats

En dépouillant l'ensemble des extraits, nous obtenons une nébuleuse de formes qui gravitent autour du thème. Les nouveaux corrélats relevés sont :

inîîni(s), lassitude(s)

tristesse(s)

douleur(s), **triste(s),** vague/vaguement

crépuscule, dévor(er), écras(er), immense, lourd(es), malad(e, -if), misère(s), octobre, profond(es), solitude

abîme, **automne,** curiosité, désert, désespérance, désespoir, désolé(e), insurmontable, oisi(f, ve), seul.

A partir de la liste de corrélats, on peut relever différents traits du thème. Dans le temps, par exemple, l'ennui apparaît de préférence au *crépuscule* d'un *dimanche d'octobre.* On pourrait aussi établir un bestiaire de l'ennui : des *larves répugnantes de regrets* chez Huysmans au *chien déshérité* de Zola.

Nous pouvons alors tenter de retracer le parcours de l'ennuyé en classant les formes recueillies.

—La monotonie et l'inaction conduisent à un ennui pathologique.

—Le décor est désespérant. Verticalement tout n'est que poids, chute et
abîme. Horizontalement les âmes errent dans des plaines infinies et désolées,
« une amertume plus vaste qu'un désert » (Du Camp).

—Les souffrances morales s'expriment sous forme de « regrets du passé »
(Du Camp), tristesse, « d'infini de souffrances oubliées »
(Huysmans) et laissent
un sentiment d'impuissance et de défaite.

—Ces douleurs mènent au sommeil vers la mort, le néant.

3.4. Retour au corpus

3.4.1. Les différentes œuvres face à la méthode

Nous avions considéré notre corpus d'étude comme un ensemble homogène, Il devient maintenant instructif d'étudier la manière avec laquelle chaque œuvre réagit au stimulus, la forme d'entrée. Pour cela, étudions :

— la significativité statistique des formes dans chacune des œuvres par
rapport au corpus thème et notamment le comportement des formes de fréquence
relativement faible,

— le comportement des œuvres de chaque auteur en évaluant le nombre
d'extraits et le nombre de formes obtenues pour chacun.

Les formes rares sont relativement plus employées par Verlaine ou par Laforgue chez qui les formes de faible fréquence sont fortement caractéristiques.

Quand le nombre de formes prélevées dans un extrait est élevé (5 chez Verlaine ; 4,2 chez Laforgue ; 3,6 chez Du Camp), les œuvres ont une propension à contenir un vocabulaire caractéristique du thème.

Même si le nombre d'extraits par pages[7] est identique chez Flaubert et chez Verlaine (0,1), le nombre de formes par extraits est fort différent (5 contre 2,9). D'autre part, les formes sont statistiquement très saillantes chez Verlaine, alors qu'elles ne le sont pas du tout chez Flaubert qui se refuse en effet à l'interrogation statistique.

3.4.2. L'ensemble du corpus

Pour confirmer que le thème est bien présent dans le corpus d'étude, nous terminerons notre examen en considérant l'ensemble. Pour l'ensemble du corpus d'étude, soit 553 853 occurrences représentant 2,7 % du corpus de référence, le test du x^ indique que l'ensemble des formes est significatif du corpus.

	corpus d'étude	*corpus de référence*	*X2*	*formes d'entrée*
Nbre d'occurrences f<100 Nbre d'occurrences f<30	1301 838	30505 17357	290,4 304,4	37 formes 35 formes

Nous examinons le comportement statistique de tous les corrélats pris en bloc. Le vocabulaire constitué de formes de fréquences inférieures à 100 est plus significatif que celui des formes inférieures à 30. Cela résulte de l'élimination des deux formes *triste* et *mort.* Ainsi, le fait d'éliminer ces deux formes nous décharge-t-il le devoir de dépouiller 463 extraits (= 1301-838).

L'élimination trop drastique des corrélats les plus fréquents nous prive de formes intéressantes, toutefois un seuil (ici une fréquence de 100 dans le corpus d'étude) nous permet de concentrer les recherches autour des formes les plus spécifiques.

4. Ambition

Pour ce thème, nous nous attacherons à tenter une nouvelle approche. Nous nous interrogions sur la densité d'extraits dans une œuvre, et le nombre de formes qu'on pouvait en extraire. Cela nous amène à nous poser la question de la répartition de ces extraits dans une œuvre particulière.

4.1. Généralités

Nous avons pris pour mots-clés *ambition* et *ambitieux.* Comme précédemment, nous dressons la liste des corrélats et en extrayons les formes les plus fréquentes.

Le thème ne se caractérise pas par un vocabulaire très spécifique. L'ambition est une maladie de l'âme qui prend appui sur une intrigue. On relève ainsi des formes très générales et assez fréquentes : *but, action, succès, moyens.* Ainsi les formes les plus significatives du corpus possèdent-elles des fréquences très élevées : *fortune* : 307, *amour :* 585, *intérêts* : 76, *succès* : 104.

Le tableau suivant synthétise les résultats.

formes	fréquence dans corpus thème	X2	formes	fréquence dans corpus thème	X2
fortune	307	313,82	limite à P=0,001		
ambition	69	92,321	parvenir	21	8,188
intérêts	76	83,583	parvenu	20	7,02
ambitieux	28	59,291	moyens	55	6,778
succès	104	51,79	dominer	11	4,009
prétentions	24	46,876	limite à P=0,05		
vanité	55	38,564	carrière	21	3,734
intérêt	112	31,658	envie	95	2,507
orgueil	97	16,347	reconnaissance	49	2,839
privilège	16	16,014	estime	31	1,479
pouvoir	160	12,485	fierté	22	1,301
triomphes	9	11,111	convoitise	6	1,029
limite à P=0,001			triomphe	38	0,23
parvenir	21	8,188	privilèges	5	0,086
parvenu	20	7,02	réussite	2	1,044
moyens	55	6,778	gloire	45	6,404

L'Éducation sentimentale semble présenter des résultats très en deçà des autres œuvres. Cela suppose que le thème de l'ambition n'y laisserait pas de traces au niveau du lexique. Peut-on parler à ce propos d'un cas Flaubert ?

En revanche, les autres œuvres du corpus se partagent les différentes formes selon des logiques qu'il est intéressant d'analyser. *Ambition* et *ambitieux* sont répartis harmonieusement dans le corpus. Toutes les irrégularités de distribution des corrélats pourront être analysées sans réserve dans l'ensemble des œuvres du corpus.

Deux formes s'avèrent plus significatives que la forme d'entrée *ambition* : *fortune* et *amour.* La répartition de *amour* est très irrégulière. *Le Lys dans la vallée* totalise de nombreuses occurrences. Dès lors, les résultats globaux s'en trouvent faussés. Cela s'expliquerait par le fait que ce roman illustre aussi le thème de l'amour pur et désintéressé, non plus comme moyen supplémentaire d'ascension sociale ou comme faiblesse mortelle pour l'ambitieux. La présence de ces deux thèmes dans un même roman engage à séparer les termes se rapportant à l'un et à l'autre. Nous pouvons décider alors d'éliminer *amour,* ou de s'en servir comme filtre d'exclusion. Les formes qui se trouveront dans le voisinage *d'amour* seront suspectes.

La fortune, dans son sens le plus général, étant l'objet principal de l'ambitieux il n'est pas surprenant de la trouver en abondance dans le corpus. *Pauvre* obtient une saillance particulière dans *Le Père Goriot. Orgueil* et *vanité* se détachent dans *Le Rouge et le Noir. Jeune* n'a pas de saillance particulière. Mais Eugène de Rastignac et Julien Sorel, grâce à leur jeune âge, font émerger — relativement — cette forme dans leurs corpus respectifs.

4.2. *Le Père Goriot*

Considérons la répartition des formes *ambition* ou *ambitieux* dans le déroulement du *Père Goriot.*

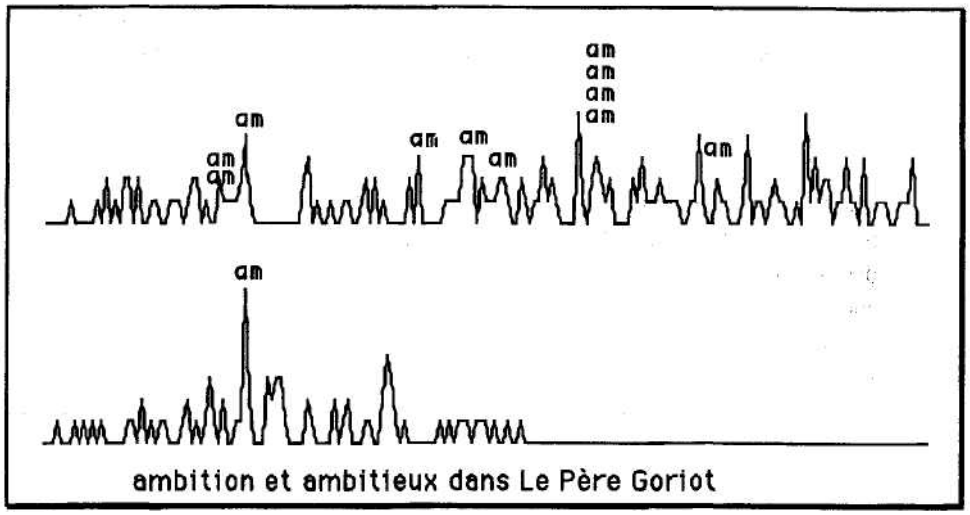

Fig.2 : répartition des vocables *ambitieux* et *ambition* dans *Le Père Goriot*
En abscisse les numéros de page ; en ordonnée le nombre de corrélats dans une page :
y = nbre_corrélats(page x) ; « am » indique la présence de « ambitieux » ou « ambition » dans une page

On peut lisser la courbe précédente en effectuant : y = [nbre corrélats (page x) + nbre corrélats (page x-l)+ nbre corrélats (page x+l)]-l. On indique les pages où y est supérieur à un seuil donné.

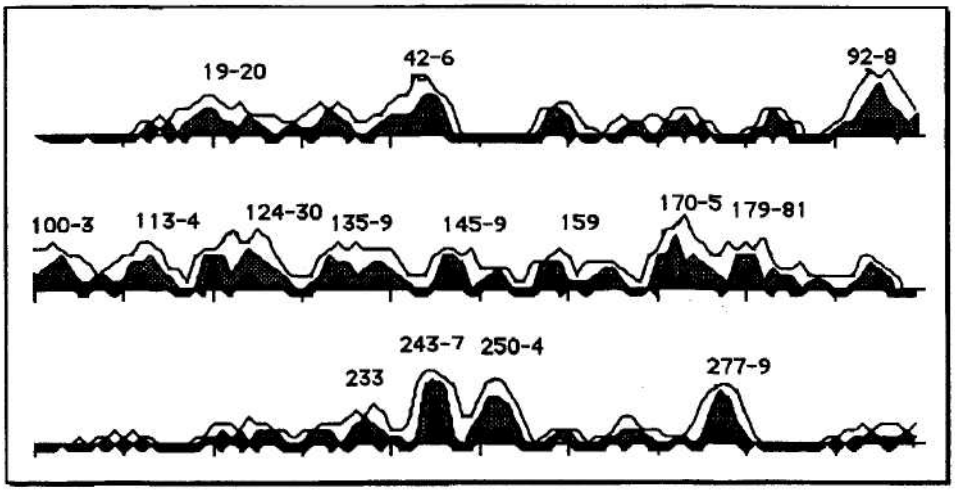

Fig.3 : répartition des vocables *ambitieux* et *ambition* dans_L« *Père Goriot*
En abscisse les numéros de page. En ordonnée le nombre de corrélats dans une page :
y = [nbre_corrélats(page x)+ nbre_corrélats(page x-l)+ nbre_corrélats(page x+l)]-l

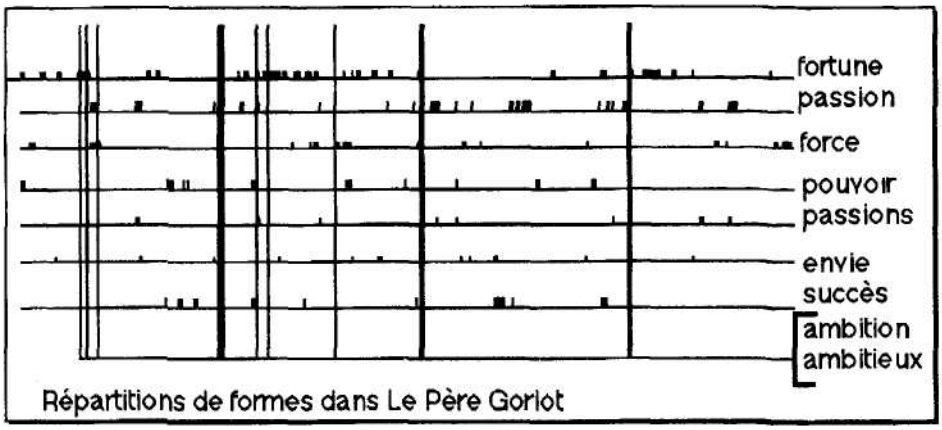

Fig. 4 : répartition de différents vocables dans *Le Père Goriot.* En abscisse, les numéros de pages. En ordonnée, la présence des occurrences. Les traits verticaux repèrent la présence de *ambition* ou *ambitieux*

On remarque très nettement les distributions « en rafale » dont fait mention Lafon f 1984]. On peut, en rapportant ce schéma aux deux précédents, indiquer quel est la part de telle ou telle forme au cours du déroulement romanesque. *Fortune* n'apparaît qu'en trois endroits distincts qui ne correspondent pas aux lieux de *passion.* Ainsi les deux vocables participent-ils différemment à l'élaboration thématique. Le lien entre ces différents graphiques et la composante dialectique du récit s'établit aisément. On connaît l'intrigue du roman. Eugène de Rastignac, jeune provincial arrive à Paris et loge à la pension Vauquer. Parmi les pensionnaires se distinguent deux personnages mystérieux : Le Père Goriot, un vieil homme, et Vautrin, un solide gaillard. Nous apprendrons que le premier est le père méprisé de deux des femmes les plus en vue de Paris : la duchesse de Restaud et la comtesse de Nucingen. Quant au second, il n'est rien d'autre que Jacques Colin, dit *Trompe-la-mort,* le légendaire trésorier du bagne.

Les moments de tension des pages 46-8 et 90-2 correspondent à deux articulations du roman : la présentation de Rastignac et son entrée dans le monde ; la décision du *méridional* de se donner les moyens de son ambition et sa demande d'argent à sa famille exsangue. On explique ainsi les principaux « pics » :

124-30 : Vautrin donne ses premiers conseils (diaboliques) à Rastignac pour réussir.

135-9 : Rastignac réussit à s'introduire dans la loge de Mme de Nucingen dont il désire faire la conquête pour s'ouvrir les portes du tout-Paris.

170-5 : Première conquête : Delphine de Nucingen.

179-81 : Deuxième entretien avec un Vautrin plus diabolique que jamais.

Le pic pp. 240-250 marque la victoire définitive de Rastignac sur Nucingen, celui des pages 277-80 décrit le dernier bal de Mme de Beauséant, protectrice d'Eugène qui quitte la haute société (figurant ainsi l'émancipation finale du héros).

Les périodes de « plat » correspondent aux moments où la vie de la pension Vauquer revient à l'avant-scène : pp. 0-30 présentation générale de la pension, pp. 50-60 discussion à la pension, pp. 190-220 l'arrestation de Vautrin, pp. 260-270 et 275 jusqu'à la fin, la mort tragique du père Goriot.

La répartition du vocabulaire suit une logique propre. Comme nous l'avions vu, l'amour entretient des rapports avec l'ambition. *Passion* se groupe essentiellement autour des scènes où Rastignac conquiert Mme de Nucingen (ce qui annonce une conquête sociale). *Fortune* se groupe autour de deux moments charnière : la première entrevue avec Vautrin qui va lancer Rastignac ; la conquête de Nucingen de qui il obtient une invitation au bal de Mme de Beauséant.

4.3. Réseau d'associations

Afin de regrouper les principaux corrélats dans une structure de type réseau sémantique, le problème se posait dans le choix et la nature des liens. Nous avons exploité la présence de plusieurs corrélats au sein des mêmes pages. Il convient d'éliminer les termes de trop haute fréquence qui faussent les résultats (surtout force et fortune). On peut voir alors se découper plusieurs réseaux d'associations.

Le schéma suivant présente l'exemple des corrélats d'*ambition* dans *Le Lys dans la vallée* et *Le Père Goriot.* Ici ne figurent que les liens correspondant à au moins cinq pages concordantes. *Succès, pouvoir, intérêt et passion* ont les fréquences les plus importantes, ce qui explique leur rôle d'attracteurs. On voit se regrouper *jouissances, envie et passions* autour de *passion.* En outre, il est intéressant de noter que *ambition* n'occupe qu'une position périphérique.

D'autre part, on observe certaines associations prévisibles : *intérêt* et *intérêts, passion et passions, ambitieux* et *ambition.*

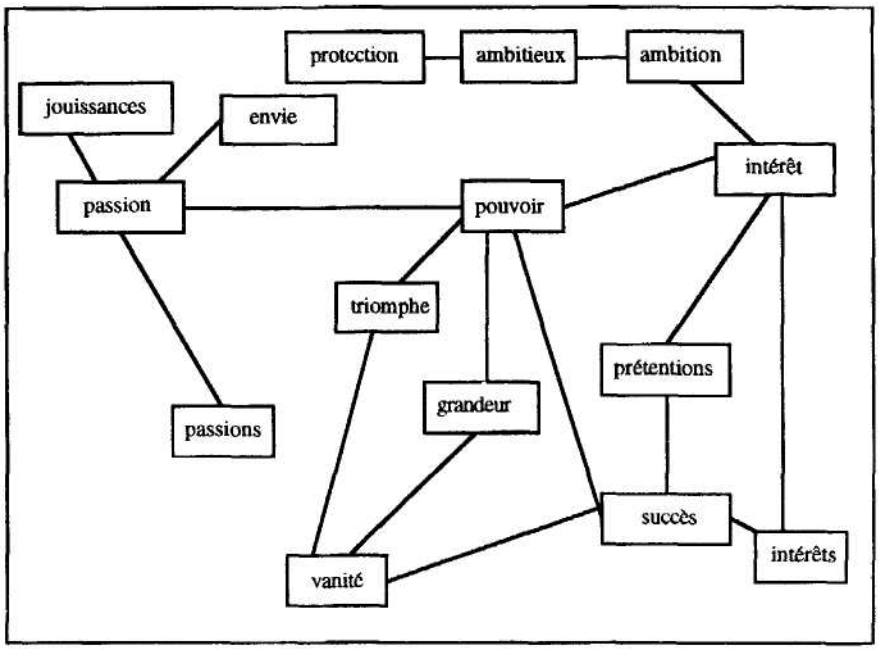

Fig. 5 : Réseau d'associations entre formes dans *Le Père Goriot* et *Le Lys dans La vallée*

Nous pouvons rapporter les différentes régions de ce réseau aux différents emplois *d'ambition.* On distingue un pôle sensuel autour de la passion. Il s'agit vraisemblablement des corrélats de *amour.* Mais gardons nous de généraliser de tels graphes.

5. Conclusions

5.1. L'étude expérimentale

Nous supposions de prime abord qu'un thème puisse être représenté par une suite *de formes d'entrées.* Cela supposait également que le corpus thématique contînt ces formes de façon *privilégiée. Ennui* et *ambition* satisfont ces hypothèses puisqu'ils sont présents avec des fréquences acceptables et possèdent un degré de signification élevé.

Les statistiques sont extrêmement fécondes. Elles assurent un filtre qui évite la dispersion des recherches. Les évaluations montrent que certaines formes moins significatives du corpus fournissent de faibles résultats. *Absurde* pour le thème de l'ennui n'a ainsi procuré que deux extraits.

Elles mettent en lumière les irrégularités du corpus. Par exemple, les œuvres de Haubert sont déficitaires en formes significatives et fournissent peu d'extraits en regard de leur taille.

Les statistiques apportent enfin une information intrinsèque sur le corpus et ses différentes composantes. La répartition des formes significatives dans *Le Père Goriot* nous paraît éclairante à cet égard.

Nous avons pris le parti de tenir compte, de manière prioritaire, des formes de *fréquence inférieure* à un seuil arbitraire. Pour le corpus *ennui* par exemple, ôter les formes de fréquence inférieure à 30 (soit 27 formes) fait chuter la significativité de la liste de corrélats dans le corpus.

6.2. Comparaison et interprétation des deux recherches

Le thème de l'ennui se présente comme un cas idéal. D'une part, la forme choisie était la plus significative de la liste. D'autre part, le vocabulaire dès formes relevées avait une fréquence moins importante en langue. Beaucoup de formes possédaient des fréquences moyennes et une signification (statistique) élevée. L'ennui se décompose en une faisceau de traits sémantiques qui permet la classification de presque tous ses corrélats. Les principaux décrivent un moment particulier de l'âme. Chacun renvoie à une modalité particulière selon que l'on invoque le temps, le goût, où que l'on privilégie l'écrasement, l'anéantissement.

L'ambition présente un caractère tout autre. Il est peut-être fortuit que les formes d'entrée de ce thème soient aussi nombreuses (même ordre de fréquences, de significativité), mais les corrélats obtenus requièrent une autre interprétation. L'ambition résume un processus. En cela, l'intrigue révèle les différents traits de l'ambitieux en soumettant ses aspirations au contact avec le monde réel. Dès lors, la description des états d'âme du héros n'occupe qu'une partie du champ thématique. Les calculs et l'élaboration d'une stratégie sociale sont primordiaux. L'objet de l'ambition reste relativement secondaire, peu importe qui devenir, il faut devenir quelqu'un.

Le thème apparaît dans des passages déterminés (notamment dans le genre romanesque). Les corrélats possèdent surtout une puissance de localisation des moments charnières où l'ambitieux se voit décrit intérieurement, dans ses pensées plutôt que dans ses actions.

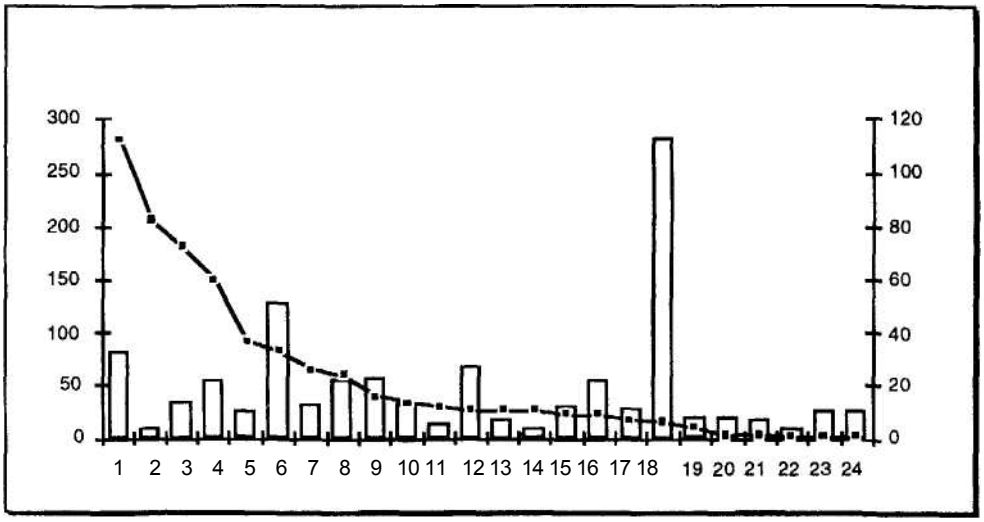

Fig.6 : répartition des formes suivant leur fréquence et leur significativité dans le corpus « ennui »
axe vertical gauche : test du X^ (points)
axe vertical droit : fréquence (colonnes)
axe horizontal : formes classées par signification décroissante

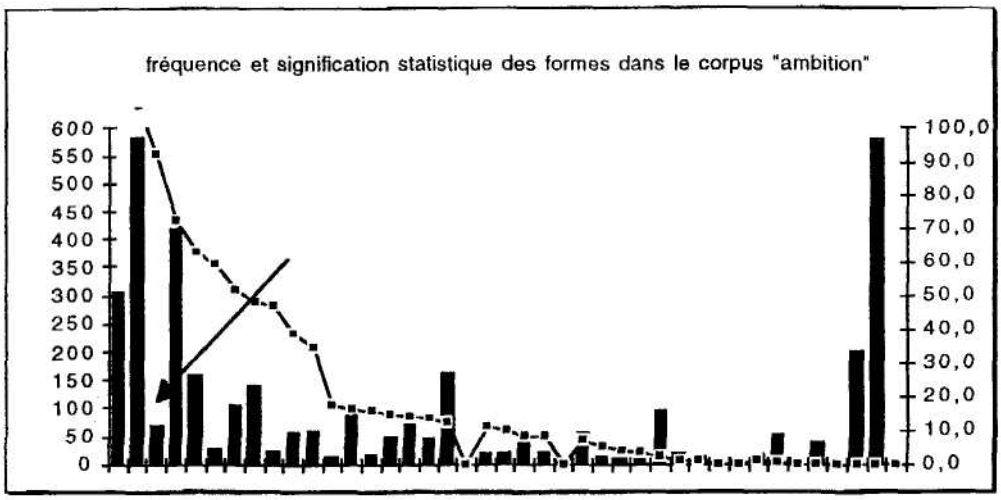

Fig.6 : répartition des formes suivant leur fréquence et leur significativité dans le corpus « ambition »
la flèche indique la forme *ambition*
axe vertical gauche : test du X^ (points)
axe vertical droit : fréquence (colonnes)
axe horizontal : formes classées par signification décroissante

Ces deux graphiques illustrent bien les différences entre les deux thèmes pour la répartition des formes les plus significatives.

6.3. Les corrélats, le corpus et le thème

Que peut-on affirmer sur la notion de thème à la lumière de la méthode proposée ? Nous assimilons un thème littéraire à un vocable où à un ensemble de vocables à partir desquels nous formons un vocabulaire caractéristique constitué de ces vocables et de leurs corrélats. Ce vocabulaire caractéristique est significatif du corpus étudié. La plupart des seuils statistiques se fondent sur le corpus thématique dans son ensemble. Si ce dernier est inhomogène, on ne pourra obtenir de résultats précis. Cela sous-entend que le *même thème* (le même vocabulaire caractéristique) se trouve dans les différentes œuvres du corpus. A contrario pourra-t-on éliminer certaines œuvres. L'étude du *Père Goriot* nous a montré que l'on pouvait rattacher les groupements de vocables caractéristiques issus d'un corpus thématique au déroulement de l'intrigue d'un seul roman.

Le relevé des corrélats demande une intervention du chercheur. On ne saurait attribuer au thème une existence totalement intrinsèque au sein d'une œuvre. Pour mettre en place une méthode objective, il faudrait en extraire les vocables les plus caractéristiques, puis les classer par thèmes (voir Guiraud [1954]). Mais cela ne supprimerait pas la subjectivité du chercheur. Comme le souligne Lafon [1984] : « Le modèle statistique est de nature totalement étrangère à la réalité linguistique. D n'est pour nous qu'un instrument de mesure permettant de détecter les formes qui justement s'éloignent le plus de lui afin de donner une description précise de la réalité. »

NOTES

1. Mot au sens graphique du terme : suite de caractères délimitée par un espace ou un signe de ponctuation. Ce n'est pas une notion de linguistique.
2. Muller [1977] écrit justement que «une fréquence observée ne saurait devenir *caractéristique* que comparée à une fréquence théorique, donc par référence à un texte ou à un corpus plus étendu que celui qui est en question. »
3. Nous aurions pu envisager de récolter systématiquement toutes les formes (éventuellement autres que des grammèmes). Mais nous n'avions pas les moyens informatiques d'effectuer d'aussi lourdes opérations.
4. Nous considérerons les formes admissibles de faible fréquence : les vocables peu fréquents sont souvent d'un emploi plus précis, donc s'attachent mieux à une description. Mais cette remarque n'est pas une justification.
5. Nous abordons par la suite le problème du seuil à partir duquel l'on peut accepter ou rejeter une forme. Il nous semble dépendre de nombreux facteurs (notamment des fréquences du vocabulaire caractéristique du thème).

«Les sanglots longs/Des violons /De l'*automne I* Blessent mon cœur/ D'une *langueur / Monotone »* (Verlaine) rappellent « Et non comme pleure la rive / Quand son jeu monotone ment » (Mallarmé). *Cf.* aussi : « L'automne faisait voler la grive à travers l'air atone / Et le *soleil dardait un rayon monotone I* Sur le bois jaunissant où la brise détone » (Verlaine), et : « Je vois se dérouler des rivages heureux / Qu'éblouissent les feux d'un *soleil monotone* » (Baudelaire).

Le nombre moyen d'extraits pertinents par page permet de juger la patience d'une œuvre relativement au thème étudié.

BIBLIOGRAPHIE

DUGAST (D.), *Statistique lexicale,* Genève, Slatkine, 1980.

DUGAST (D.), Définition des notions de répartition et de localisation des vocables dans le discours. Pour une réhabilitation de la Loi de Poisson. Essai de sémantique quantitative, *Cahiers de Lexicologie,* vol. XLII-1983,1983. ERLICH (D.), *Thématique assistée par ordinateur,* mémoire de DEA à l'Université de Paris XI, 1991.

GUIRAUD (P.), *Les Caractères statistiques du vocabulaire,* Paris, PUF, 1954. LAFON (P.), *Dépouillement et statistiques en lexicométrie,* Genève, Slatkine, 1984. MARTIN (É.), *Reconnaissance de contextes thématiques dans un corpus textuel. Éléments de lexico-sémantique,* Paris, CNRS-INaLF ; Didier Érudition, 1993 (Études de sémantique lexicale).

MULLER (Ch.), *Initiation aux méthodes de statistique linguistique,* Paris, Hachette, 1973.

MULLER (Ch.), *Principes et méthodes de statistique lexicales,* Paris, Hachette, 1977. *La Recherche française par ordinateur en langue et littérature. Actes du colloque organisé par l'université de Metz,* Genève, Slatkine, juin 1983. SAGNES (Guy), *L'Ennui dans la littérature française de Flaubert à Laforgue,* Paris, Armand Colin, 1969. TROUSSON (R.), *Thèmes et mythes : questions de méthodes,* Bruxelles, Éditions de l'Université de Bruxelles, 1981.