INTRODUCTION

La recherche présentée ici a été effectuée[1] dans le cadre d'un projet pluridisciplinaire, intitulé *Accès sémantique aux banques textuelles*, qui réunit des informaticiens et des linguistes de deux laboratoires du CNRS, le CRIN (Centre de recherches en informatique de Nancy) et l'INALF (Institut National de la Langue Française).

L'INALF a mis à la disposition du public, dès 1986, une importante banque de textes du XVI^e^ au XX^e^ siècle, à dominante littéraire, nommée FRANTEXT[2], et les études menées dans le cadre du projet *Accès sémantique aux banques textuelles* portent sur le corpus ROMAN (de 1830 à 1970[3]) de cet ensemble.

Les banques textuelles sont en développement constant et les méthodes d'investigation de leur contenu sont bien souvent insatisfaisantes parce qu'elles négligent de prendre en compte les caractéristiques du texte, et tout d'abord le fait que celui-ci ne se réduit pas à "une suite de phrases". L'action de recherche que nous menons se fonde sur l'évidence que les structures textuelles sont essentiellement sémantiques, et que, pour l'analyse sémantique, le problème fondamental est celui de la sélection de contextes pertinents. Nous mettons à profit les connaissances acquises en sémantique du contexte et nous utilisons la pondération par un test statistique pour élaborer des outils d'aide à l'analyse de corpus électroniques[4], permettant à l'usager de constituer des sous-corpus sémantiquement riches, par rapport aux objectifs de sa recherche.

Les structures textuelles variant avec les genres et les discours, on fait l'hypothèse que la ponctuation est employée de façon particulière selon les textes en fonction de ces paramètres, et on l'étudie comme les autres signes, en corpus, pour repérer les régularités et les contrastes liés aux *composantes textuelles*, thématique, dialectique, dialogique et tactique, les instances systématiques qui règlent la production et l'interprétation des suites linguistiques. On cherche donc à intégrer la ponctuation dans le cadre théorique d'une sémantique textuelle[5].

---

[1] Ce texte a paru en 1998 (Bourion E., 1998) : nous avons supprimé ici les annexes, le glossaire et la bibliographie qui font double emploi avec ceux de la thèse (des redites cependant sont inévitables).

[2] Ces textes ont été saisis sur support mécanographique à partir de 1963 pour constituer le corpus de référence du dictionnaire *Trésor de la langue française des XIX^e^ et XX^e^s.* ; la proportion était fixée à l'origine à 80% de textes littéraires et 20% de textes techniques pour le corpus servant à l'élaboration du dictionnaire, et donc pour la période XIX-XX s. Mais au fil des années, le corpus s'est constamment enrichi, avec des textes de la période mais aussi des siècles précédents, et ce développement se poursuit toujours (actuellement, le corpus représente 2650 œuvres et environ 180 millions d'occurrences -*occurrence* étant défini comme signe entre deux blancs-). D'autre part, l'extension des moyens techniques a permis de mettre à la disposition des chercheurs et du public ce corpus informatisé, qui est interrogeable en ligne, grâce à un logiciel élaboré par J. Dendien, depuis 1986 (actuellement sur 64 stations à travers le monde) ; depuis 1996, il est également consultable sur Internet.

[3] Le corpus ROMAN comporte 345 œuvres et représente 40 millions d'occurrences environ.

[4] Ces outils sont élaborés par l'auteur et J. Maucourt, informaticien à l'Inalf.

[5] Il n'est pas possible, dans le cadre de cet article d'exposer les concepts théoriques sur lesquels la recherche s'appuie, mais nous signalerons rapidement les notions utiles à la présentation des résultats (en italiques) et nous renvoyons au glossaire en annexe et aux ouvrages signalés en bibliographie.

I. LA PONCTUATION, LA SEMANTIQUE INTERPRETATIVE ET LES TEXTES ELECTRONIQUES

"La problématique du texte, fondée sur l'analyse différentielle [...] définit le sens par l'interaction paradigmatique et syntagmatique des signes linguistiques, non seulement entre eux, mais avec le texte dans sa globalité"[6]. Pour la sémantique interprétative, l'accès au sens est considéré comme un processus de type "reconnaissance de formes" (et non comme un "calcul"). Son approche du texte se situe dans la tradition herméneutique : le global détermine le local (cercle philologique ou herméneutique). Les concepts de la sémantique interprétative permettent de décrire les "parcours interprétatifs" et d'expliciter comment le linguistique impose des contraintes sur les représentations mentales.

La cohésion textuelle est engendrée par la récurrence de traits sémantiques et l'effet de ces récurrences produit *des isotopies sémantiques.* Les isotopies génériques à l'œuvre dans un texte constituent des fonds sémantiques sur lesquels de détachent des *molécules sémiques*, induisant l'impression référentielle, qui donne l'illusion de "voir tel personnage, tel décor, telle action en train de se dérouler". Puisqu'en fait le sens d'un mot est indécidable si l'on ne connaît pas son contexte, nous utilisons le test statistique pour "filtrer le contexte", pour en faire apparaître un ensemble de signifiants dont la présence simultanée dans l'entourage du *mot-pivot* de recherche s'explique par des raisons sémantiques : les signifiés de ces mots partagent avec le mot-pivot des traits sémantiques (ou sèmes) récurrents, qui, organisés en structures, sont le support linguistique des représentations mentales. En étudiant les relations, marquées par le système linguistique, entre le mot-pivot et les *corrélats* (sélectionnés par le test statistique[7]), on peut interpréter les sèmes et leurs interactions en structures, les formes sémantiques (dont les thèmes).

Les signes de ponctuation sont traités comme les autres signes, puisqu'ils participent pleinement et d'emblée à l'interprétation du sens. Dans le signifié de certains ponctèmes, on retrouve des traits attestés dans les lexèmes et les grammèmes[8]. Les ponctèmes sont étudiés par la sémantique interprétative selon différentes approches :

a) En règle générale, ponctuation faible et ponctuation forte peuvent servir à sélectionner des zones de localité où la propagation est plus ou moins facilitée : ces signes peuvent permettre de filtrer, parmi les *cooccurrents* sélectionnés par le test de l'écart réduit, ceux qui ont des relations sémantiques privilégiées avec le mot-pivot, eu égard à leur distance à celui-ci (distance "faible, moyenne, forte"). Ils sont utilisés également, dans les programmes d'aide à l'analyse sémantique que nous élaborons, pour définir les fenêtres de sélection, et constituer les corpus de travail, en fonction de l'influence qu'ils ont sur la diffusion des traits sémantiques.

---

[6] F. Rastier, Pour une sémantique des textes - questions d'épistémologie, *in Textes et sens,* F. Rastier (éd.) Paris, Didier, 1996, pp. 9-35

[7] Cf. E. Bourion, Le réseau associatif de la peur, in *L'analyse thématique des données textuelles, l'exemple des sentiments*, F. Rastier (éd.), Paris, Didier, 1995, pp. 107-145 ; pour la formule du test, v. annexe.

[8] Ici on convient de nommer *ponctèmes* les signes de ponctuation, *lexèmes* les "mots sémantiques" et *grammèmes* les "mots-outils".

b) Dans une analyse thématique de corpus, le signifié des ponctèmes sélectionnés par le test statistique[9] s'interprète, comme celui des autres signes, à l'aide des composantes textuelles : par exemple on fait l'hypothèse qu'un sentiment dont la molécule sémique comporte le trait /imperfectif/, comme *l'ennui*, sera fréquemment associé aux points de suspension, et qu'en revanche, *la peur*, qui comporte les traits /intensif/ et /perfectif/, comptera le point d'exclamation parmi ses corrélats. Les résultats observés dans FRANTEXT ont en partie seulement validé ces hypothèses (*cf.* II).

c) Certains ponctèmes, comme les points, peuvent servir de barrière à la diffusion des traits sémantiques en constituant une démarcation forte pour le sens avec une autre marque (de paragraphe, par exemple, pour la période) ; tandis qu'un ensemble de virgules peut rassembler (tout en les séparant, dans une énumération par exemple) des mots dont le signifié a des traits sémantiques communs. Un tiret ou une parenthèse qui suivent un mot peuvent introduire un contexte qui signale que le contenu sémantique doit être remanié (ce qui s'interprète à l'aide des instructions d'*assimilation*, *dissimilation*, *virtualisation* ou *actualisation* de sèmes, apportées par le contenu de la parenthèse ou de l'énoncé qui suit le tiret[10]).

II. LA PONCTUATION DES QUATRE CORPUS "SENTIMENTS"

Selon la méthode détaillée dans Bourion 95, on a voulu étudier la répartition et la valeur des ponctèmes dans des corpus représentatifs à la fois d'un genre, le roman et d'une thématique générique, celle des sentiments. En leur appliquant les programmes mis au point pour l'analyse sémantique, on a comparé les corpus de deux sentiments thymiques, la peur et la colère, un sentiment relationnel, la pitié et un sentiment existentiel, l'ennui.

II. 1. La pondération statistique

Elle opère en calculant une fréquence théorique d'apparition d'un signe dans le voisinage du mot-pivot dans un *corpus de travail* par rapport à la fréquence de ces signes dans le corpus de référence, le corpus ROMAN. Le test de l'écart réduit permet de constater la déviation d'une fréquence observée (Fobs.) par rapport à une fréquence théorique (Fthéor.) : plus le score d'écart réduit est élevé, moins la cooccurrence s'explique par le hasard. On a retenu, à partir des travaux de statistique lexicale, le seuil r = 3, au-delà duquel on considère que la cooccurrence est non aléatoire[11].

[9] Pour un exposé de la méthode suivie pour étudier la composante thématique d'un corpus, cf. Bourion 95 et *Le thème de la parure*, à paraître.

[10] Quand on lit, dans R. Sabatier, *Le Chinois d'Afrique*, (éd. Albin Michel, Paris, 1966, p. 316) *ma bonté-pitié-charité*, on doit interpréter ce contenu complexe, à partir des trois sémies-types (sur ce concept et celui de sémie-occurrence, proposés pour traiter du remaniement apporté en contexte à la "signification en langue", répertoriée par les dictionnaires, v. F. Rastier, *Sémantique pour l'analyse*, Masson, Paris, Didier, 1994, pp. 52-58) et des composantes textuelles (par exemple, il faut prendre en compte les traits sémantiques de l'acteur qui énonce ce propos - le foyer énonciatif - pour juger de la distance de l'énonciateur à son discours).

[11] Ce seuil a été déterminé pour les "mots" et, à notre connaissance, il n'y a pas eu d'études de ce type sur les ponctèmes, auxquels nous avons, dans cette première expérience, appliqué le même seuil : on pourra voir que ce qui est signifiant, c'est la variation des ponctèmes sélectionnés et les contrastes entre les scores. Comme nous

Pour cette expérience, le corpus PEUR comporte donc toutes les attestations du mot-pivot *peur*, rencontrées dans le corpus ROMAN, accompagnées d'un contexte de 10 mots avant et 10 mots après, le point arrêtant la sélection ; il en est de même pour les corpus PITIE, ENNUI et COLERE[12].

II. 2. Contraste entre les quatre corpus "sentiments"

Les quatre corpus sont assez homogènes en ce qui concerne la proportion des lexèmes, grammèmes et ponctèmes, comme le montrent le tableau 1 et le diagramme 1.

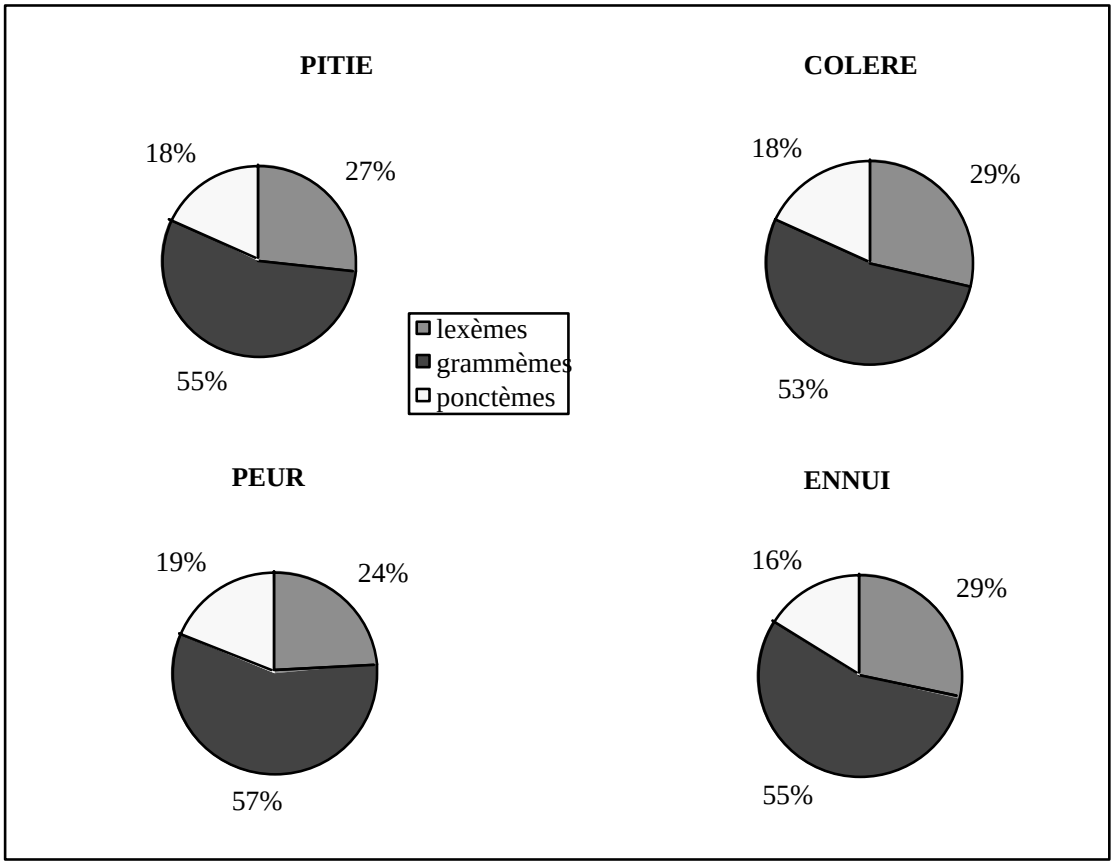


Diagramme 1

	PEUR	COLERE	PITIÉ	ENNUI
nb signes	138594	67798	50177	26792
lexèmes	33695	19331	13564	7713
grammèmes	78748	36566	27452	14885

souhaitons contribuer à l'établissement d'une typologie des genres et des discours fondée sur des critères linguistiques, issus d'observation de corpus attestés, les travaux en cours permettront de faire des propositions pour le genre du roman (v. IV pour le contraste "roman traditionnel", "nouveau roman").

[12] Il est bien clair que tous les passages de textes centrés sur la thématique de la peur (de la colère, etc.) ne figurent pas dans le corpus ainsi sélectionné : il s'agissait ici de contraster les quatre corpus de sentiments au plan des signes de ponctuation et non de faire une recherche thématique pour déterminer les contenus associés ou pour préciser comment on lexicalise, dans le corpus ROMAN, la peur, l'ennui ou les autres sentiments.

ponctèmes	26151	11901	9161	4194

Tableau 1 : Composition des quatre " corpus sentiments "

II. 3. La pondération des ponctèmes par le test statistique

La comparaison porte sur des corpus de taille variable, mais ils sont pondérés par le test de l'écart réduit, et contrastés, par rapport au même corpus (corpus de référence, ROMAN). La variable "genre textuel" est contrôlée, autant que faire se peut, par le fait qu'on travaille à l'intérieur d'un corpus appartenant à un seul genre textuel. On fait l'hypothèse que les différences de score, de corpus à corpus, réfèrent à des aspects sémantiques, liés aux thématiques différentes (peur, colère, pitié, ennui), et qu'on doit donc pouvoir les interpréter grâce aux traits sémantiques stables des molécules sémiques PEUR, COLERE, PITIE et ENNUI.

Le tableau 2 donne les résultats de la sélection par le test de l'écart réduit pour chaque corpus et le tableau 3 précise l'ensemble des scores des ponctèmes ; le graphique 2 permet de visualiser les contrastes dans la répartition des ponctèmes sélectionnés par le test[13].

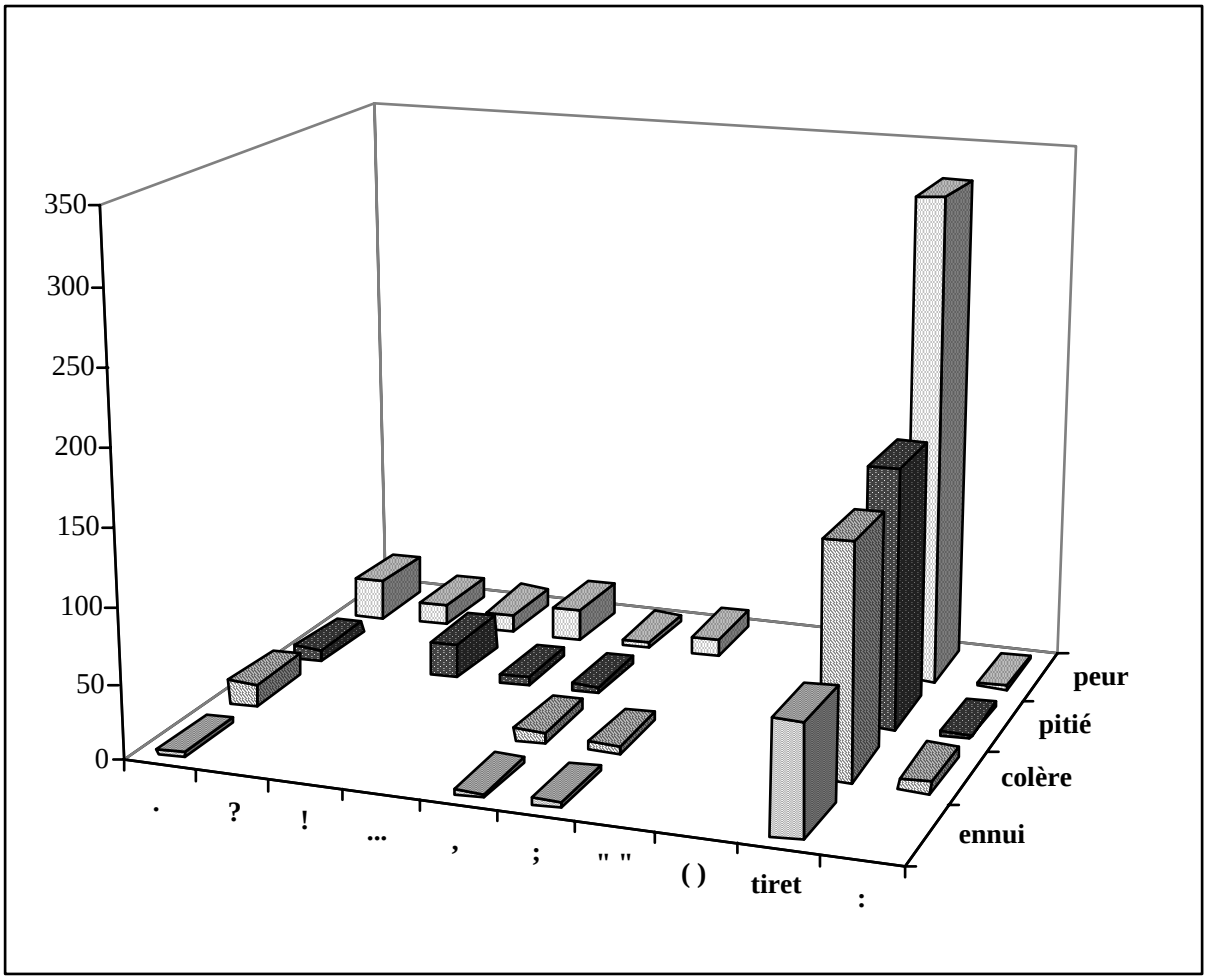
	PEUR	**PITIÉ**	**COLÈRE**	**ENNUI**
.	26	8	16	4
?	15	ns	ns	ns
!	13	22	ns	ns
...	21	5	ns	ns
,	4	5	8	3
;	12	ns	5	4
" "	ns	ns	ns	ns
( )	ns	ns	ns	ns
:	3	3	7	ns
-	326	172	154	73

Tableau 2 : ponctèmes sélectionnés par le test, dans les quatre corpus "sentiments"

[13] Dans le graphique ne figurent pas les valeurs au-dessous du seuil de sélection.

	PEUR	**PITIÉ**	**COLÈRE**	**ENNUI**
.	26	8	16	4
?	15	-0.9	-3	-2
!	13	22	1	-2
...	21	5	-5	-5
,	4	5	8	3
;	12	2	5	4
" "	-2	-2	-4	-6
( )	-2	-1	-1	-2
:	3	3	7	-0.9
-	326	172	154	73

Tableau 3 : scores des ponctèmes dans les quatre corpus "sentiments"



Graphique 2 : les ponctèmes dans les quatre corpus "sentiments"

II. 4. Commentaires des résultats

On remarque que la ponctuation forte, l'ensemble des points, est plus représenté dans le corpus PEUR que dans les autres ; cela est vrai du point "simple"[14], mais encore plus des points dits "expressifs", point d'exclamation, d'interrogation et aussi de ceux de suspension. Le point d'exclamation, sélectionné avec un score d'écart réduit très important dans le corpus PEUR, a un score encore plus élevé dans le corpus PITIE, mais n'est pas sélectionné dans les deux autres. Ces observations entraînent à calculer la longueur moyenne de l'unité "phrase"[15], définie comme "unité textuelle comprise entre deux signes de ponctuation forte" (tableaux 4 et 5) : même si ce calcul est grossier, les résultats montrent que cette unité peut signaler des différences entre corpus, qui doivent s'interpréter par les composantes textuelles.

[14] Malgré la sous-évaluation du point qui vient du fait que la sélection automatique des énoncés du corpus de travail s'arrêtait au point dans ce programme. Dans le corpus PEUR ce critère a limité des énoncés constitués de phrases elliptiques, cf. les exemples cités.

[15] Dans l'avenir, et au fur et à mesure que les textes du corpus seront balisés avec les codes SGML et TEI, les programmes que nous élaborons offriront une image plus réaliste de l'unité "phrase" puisque le code "s" permet d'évaluer leur nombre exact (v. conclusion).

	PEUR		**COLÈRE**		**PITIÉ**		**ENNUI**	
	nb	%	nb	%	nb	%	nb	%
ponctèmes forts (. ? ! ...)	10381	39,6	4202	35,3	3461	37,7	1424	33,9
ponctèmes faibles (, ;)	11650	44,5	6052	50,8	4004	43,6	2312	55,1
autres signes (- ( ) " :)[16]	3906	14,9	2092	17,5	1380	15,1	458	10,9

Tableau 4 : ponctuation forte vs *ponctuation faible*

	PEUR	**COLÈRE**	**PITIÉ**	**ENNUI**
nb moyen de signes entre 2 ponctèmes forts	13	16	14	18
nb moyen de grammèmes entre 2 ponctèmes forts	7	8	7	10
nb moyen de lexèmes entre 2 ponctèmes forts	3	4	3	5

Tableau 5 : longueur moyenne de la phrase dans les 4 corpus

a) Le corpus PEUR

C'est le corpus qui connaît le plus fort pourcentage de ponctèmes forts (tableau 4), et ces ponctèmes sont sélectionnés tous les quatre par le test statistique (tableau 2) alors que pour les autres corpus, on passe de trois pour PITIE, à un, le point, pour COLERE et ENNUI, mais avec un score faible dans ce dernier corpus. Ce fait s'accompagne pour PEUR d'un score faible de la virgule, et d'un score particulièrement élevé des tirets, que l'on a interprété par la lecture des contextes[17] : dans ce corpus on rencontre beaucoup plus de phases courtes, et de dialogues, qui justifient le score important du tiret, signe démarcatif de prise de parole d'un acteur, et bien présent aussi à cause de l'inversion du pronom personnel sujet du verbe support (*dit-il, répéta-t-elle*, etc.[18]). Ce fait est à mettre en relation avec les régularités dans la lexicalisation de ce sentiment que nous avions notées dans le précédent travail (Bourion 1995) : pour les narrateurs, "faire vrai", c'est montrer que la peur empêche la parole, et que l'expression est

[16] Les "ponctèmes forts" sont ici : . ! ? ... On ne peut tenir compte des tirets, à cause de leurs valeurs différentes (trait d'union ou tiret de dialogue), mais l'analyse des contextes montre que si l'on pouvait les faire entrer dans le calcul après désambiguïsation, ils renforceraient la tendance. Les "ponctèmes faibles" sont la virgule et le point-virgule. Les deux-points ne sont pas comptabilisés, pour la même raison de valeur, puisqu'ils peuvent être considérés tantôt comme signes forts, tantôt comme signes faibles. Pour ces calculs, on a tenu compte de tous les ponctèmes, qu'ils soient sélectionnés ou non par le test de l'écart réduit, puisqu'ils sont pris en compte ici au titre de leur fréquence absolue dans chaque corpus.

[17] Cf. les énoncés donnés à titre d'exemple.

[18] Au moment où cette expérience a été menée, le programme ne pouvait pas différencier le tiret du trait d'union : actuellement, un dictionnaire complet des formes composées et un catégorisateur morpho-syntaxique (élaborés à l'Inalf par J. Maucourt et M. Papin), vont permettre de reconnaître les mots-composés et d'avoir des calculs statistiques plus fiables. Cependant, avec la loi des grands nombres, l'incidence du trait d'union ne semble pas très importante (cf. IV, note 36), et pour des comparaisons de corpus où le tiret a un score d'écart réduit important, on peut invoquer l'argument du tiret de dialogue, comme la lecture des contextes l'a montré pour les corpus étudiés.

impossible ou en reste au plan infra-langagier, en utilisant des phrases courtes, voire elliptiques, des interjections, et aussi en alternant des dialogues, où le point d'exclamation marque l'intensité (où le point d'interrogation, sélectionné seulement dans ce corpus, est bien représenté aussi - cf. son score) et des parties narratives. Quelques énoncés tirés des corpus illustrent les différences entre ces deux modes narratifs.

> *"Alors souriant d'un sourire étrange et la prunelle fixe, les dents serrées, il s'avança en écartant les bras. Elle se recula tremblante. Elle balbutiait :*
>
> *- Oh ! Vous me faites peur ! Vous me faites mal ! Partons.*
>
> *- Puisqu'il le faut, reprit-il en changeant de visage. Et il redevint aussitôt respectueux, caressant, timide." Flaubert G.,* Madame Bovary, *1857, p. 183*

> *"Elle s'agite : "Parce qu'il regarde les gens au fond des yeux ?*
>
> *- Tu as peur, Laura ?*
>
> *- Peur ? Tu es fou. Je n'ai peur d'aucun homme, aucun."*
>
> *Et elle serre nerveusement son sac contre son fragile ventre de femme. "Paysan C.,* Les feux de la Chandeleur, *1966, p. 75*

> *"Alors ! ... oui. Alors ! ... Eh bien ! j'ai peur de moi ! j'ai peur de la peur ; peur des spasmes de mon esprit qui s'affole, peur de cette horrible sensation de la terreur incompréhensible.*
>
> *Ris si tu veux. Cela est affreux, inguérissable." Maupassant G. De,* Contes et nouvelles, *1883, t. 2, p. 853*[19]

> *"J'ai peur", dit-elle. Je lui demandai, de quoi elle avait peur et elle me répondit : "Vous n'avez pas peur, vous ?" Alors, je gardai le silence. C'était vrai, j'avais peur, moi aussi. Elle dit encore : "Vous ne sentez pas qu'il se passe quelque chose ? - Où ça ? - Où ça ! où ça ! Autour de nous !" Elle haussa les épaules : "Ah ! je suis toute seule ! toute seule ! et j'ai peur !" Leroux G.,* Le parfum de la dame en noir, *1908, p. 134*

> *"- Je n'ai pas de vieux... et si tu dois me parler de choses qui ne te regardent pas, je pars*
>
> *- Essaye un peu*
>
> *- Mon pauvre petit bonhomme, tu crois que tu me fais peur*
>
> *- Je te fais peur*
>
> *- Non*
>
> *- Si*
>
> *- La preuve*
>
> *- Quelle preuve ?" Rivoyre Ch. De,* Les Sultans, *1964, p. 190*[20]

> *"Brusquement, le sol manqua sous eux, ils faillirent rouler dans un canal profond, dont ils avaient atteint la berge sans s'en apercevoir. Ils s'arrêtèrent, soufflèrent un moment, se regardèrent sans oser avouer leur peur. Alain tordit son mouchoir trempé et s'essuya le visage. Ils eurent tout à coup un tressaillement épouvantable, se jetèrent l'un contre l'autre." Van Der Meersch M.,* Invasion 14, *1935, p. 57*

[19] Le corpus comporte actuellement quelques textes intitulés "contes" ou "nouvelles" : la question d'en faire on non un corpus séparé sera étudiée à l'aide des critères linguistiques de genre, sur lesquels nous travaillons.

[20] On observera cet exemple "atypique" sans signes de ponctuation en dehors du tiret ; il est de la deuxième moitié du XX^e^s., et avec les textes traités en partie IV, on peut juger de l'intérêt d'études par tranches chronologiques pour l'évolution de l'emploi des ponctèmes.

"On crut à une feinte, à une ruse, à un mauvais vouloir de malfaiteur, et les deux hommes armés, le rudoyant, l'empoignèrent et le plantèrent de force sur ses béquilles.

La peur l'avait saisi, cette peur native des baudriers jaunes, cette peur du gibier devant le chasseur, de la souris devant le chat. Et, par des efforts surhumains, il réussit à rester debout.

"En route !" dit le brigadier." Maupassant G. De, Contes et nouvelles, *1884, t. 2, p. 443*

"Pour l'intelligence de ce qu'allait tenter Jacques Collin, il est nécessaire de faire observer que tous les assassins, les voleurs, que tous ceux qui peuplent les bagnes ne sont pas aussi redoutables qu'on le croit. A quelques exceptions très rares, ces gens-là sont tous lâches, sans doute à cause de la peur perpétuelle qui leur comprime le cœur." Balzac H. De, Splendeurs et misères.., *1847, p. 548*

"La peur le galope. Il voudrait crier ; il ne peut pas. Il se sent emporté comme un fétu dans une avalanche : impossible de s'accrocher à rien : tout a chaviré, tout sombre avec lui... Enfin la gorge se desserre, la peur s'y fait un passage, jaillit en un cri d'horreur, qui s'étrangle aussitôt." Martin Du Gard R., Les Thibault, *1929, t. 1, p. 1253*

Le sème /intensité/, très représenté dans ce corpus (dans les lexèmes, grammèmes, ponctèmes, mais il est exprimé aussi dans le procédé de la répétition), est rapporté au point de vue de la composante thématique, et des critères issus des autres composantes textuelles (*dialogique*, *dialectique* ou *tactique*) : récit *vs* énonciation représentée, paroles d'acteurs différents, rapportées ou non en discours direct[21], dans un style "haché", où dominent des phrases courtes.

b) Le corpus PITIE

Le point d'exclamation, non sélectionné dans les corpus COLERE et ENNUI, a un score plus important avec *pitié* qu'avec *peur* : il rend compte des traits /intensité/ et /imploration/. On remarque également que les corpus PITIE et PEUR ont en commun un score plus important du tiret, par rapport aux corpus COLERE et ENNUI, ce qui marque une alternance de dialogues et de parties narratives dans ces corpus. Dans les corpus PEUR et PITIE, dominent les traits sémantiques /intensité/ et /perfectif/, ainsi que /ponctuel/[22], et ces traits expliquent les analogies de sélection des ponctèmes (comme de certains lexèmes).

"- O Stephen ! mon ange, grâce ! grâce ! aie pitié d'une pauvre femme qui n'a plus de force pour te résister !

Oh ! Ce serait lâche d'abuser de ma faiblesse ! Je te haïrais, je te mépriserais... laissez-moi, laissez-moi, homme vil ! Je vous hais, je vous méprise ! ... " Karr A., Sous les tilleuls, *1832, p. 302*

"Que je suis malheureuse si je vous afflige encore !

- Lélia, n'auras-tu pas pitié de moi ?

- Pitié de toi ! Et que puis-je faire de plus ? Je t'ai soumis toutes les puissances rebelles de mon âme." Sand G., Lélia, *1833, p. 216*

[21] Sur les différentes parties de ce corpus, v. ci-dessous III.
[22] Le test a sélectionné le verbe *prendre* uniquement pour la forme *pris(e) de peur* ou *de pitié* ; en revanche le trait /imperfectif / et /itératif/ de la molécule de *ennui* explique d'autres lexicalisations (v. ci-dessous).

"J'ai toujours ouï dire qu'il était périlleux de traverser la nuit les juiveries, qu'il y pleuvait des chaudières et des matras, des chats noirs, des mandragores, des chauves-souris, des feux grégeois...

- Pouvez-vous bien, à votre âge, croire pareilles balivernes ? Un homme de loi ! un docteur ! vous faites pitié !

Maître Bonaventure, par mon honneur ! Je puis vous attester que si la nuit il pleut en ce quartier, à coup sûr, ce ne sont ni des mandragores ni des chats noirs."

Borel P., Champavert, *1833, p. 130*

"M. Buré passa la bride de son cheval dans son bras, et lui et son frère s'éloignèrent lentement.

- Ainsi, disait le capitaine, tu me le jures ! point de grâce ! point de pitié !

- Fie-toi à ma haine.

- Il faut qu'il meure aux galères !

- J'ai de quoi l'y envoyer." Soulié F., Les mémoires du diable, *1837, p. 208*

"Oh ! Mon Dieu ! Mon Dieu ! Mon Dieu ! Est-il un Dieu ? S'il en est un, délivrez-moi, sauvez-moi ! Secourez-moi ! Pardon ! Pitié ! Grâce ! Sauvez-moi ! Oh ! Quelle souffrance ! Quelle torture ! Quelle horreur !" Maupassant G. De, Contes et nouvelles, *1886, t. 2, p. 1115*

"Oh ! - pensait-il parfois, -au moins, si ma victime m' a échappé... si je n' ai pu me venger en détail... que je me venge bien sur cette société tout entière ! Oh ! Que c'est pitié... pitié de voir ces savants, ces philanthropes, cette élite de Paris, de leur Paris... du monde... être joués par un misérable esclave, un pauvre nègre, qui a encore le dos tout meurtri des coups de fouet du commandeur... " Sue E., Atar-Gull, *1831, p. 38*

"- Mon père ! père chéri ! Regarde-moi ! ... Regarde-moi ! ... aie pitié de moi ! et ne demande pas que s'ouvre ma bouche qui doit rester close à jamais... et crois-moi. Ne crois pas ces hommes !" Leroux G., Rouletabille chez le Tsar, *1912, p. 108*

"Il ne songea pas une seule fois, distinctement du moins, au grand changement qui venait de s'opérer dans son sort. Quel regard ! se disait-il ; que de choses il exprimait ! quelle profonde pitié ! Elle avait l'air de dire : la vie est un tel tissu de malheurs ! Ne vous affligez point trop de ce qui vous arrive !" Stendhal, La Charteuse de Parme, *1839, p. 253*

c) Le corpus COLERE

Si, comme on l'a vu, la peur empêche la parole, au contraire, la colère s'exprime par des mots, et quand *cri* est sélectionné avec *colère*, ce mot lexicalise le trait /paroles/, comme *accent*. Ni le point d'exclamation, ni le point d'interrogation ne sont sélectionnés, mais les scores importants des deux points (le plus élevé des quatre corpus) et du tiret montrent que les parties dialoguées sont bien représentées dans ce corpus également.

"Albert l'écouta en frémissant tantôt d'espoir, tantôt de colère, parfois de honte ; car, par la confidence de Beauchamp, il savait que son père était coupable, et il se demandait comment, puisqu'il était coupable, il pourrait en arriver à prouver son innocence." Dumas A. Père, Le comte de Monte-Christo, *1846, p. 392*

"Anne-Marie ne vivait plus. Elle sentait au-dessus d'eux tous quelque chose d'affreux en suspens. La journée ne se passerait pas sans un malheur. Comment dire maintenant ce qu'elle avait vu cette

nuit ? Elle tremblait comme la feuille à l'idée de la colère où le père entrerait. Ne pas l'avoir réveillé ? Alors que l'autre était là ! Qu'on tenait pareille occasion de lui laver la tête avec du plomb de calibre ! " Pourrat H., Gaspard des montagnes*, 1922, p. 143*

"- Ces saintes mains ! murmura-t-elle en s'efforçant bravement de sourire, bien que ses yeux fussent pleins de larmes.

Tout à coup, elle rougit fortement de nouveau, et un sentiment qui ressemblait sans doute autant à la colère qu'à la honte gonfla ses lèvres." Bernanos G., La joie*, 1929, p. 547*

"En effet, le cri de la mère change ; maintenant c'est un cri de colère : "Ah ! tu ne veux pas me rendre mes petits !" C'est un cri de colère terrible, irrésistible ; il révolte l'air tout autour." Frapié L., La Maternelle*, 1904, p. 135*

"Février s'achevait, il ne lui restait que quelques jours pour l'envoi au Salon, c'était un désastre.

Un soir, devant Christine, il jura, il lâcha ce cri de colère :

- Aussi, tonnerre de Dieu ! est-ce qu'on plante la tête d'une femme sur le corps d'une autre ! ... Je devrais me couper la main." Zola E., L'œuvre*, 1886, p. 119*

"Toute sa carcasse tremblait d'une colère si terrible qu'elle faisait peur à regarder. Voilà : c'était un homme qui ne pouvait pas avoir tort, qui ne pouvait pas céder." Genevoix M., Raboliot*, 1925, p. 222*

"Il justifiait tous ses abandons passés et futurs ; il semblait heureux de n'être plus rien, il ne voulait plus être rien.

Je souhaitais éclater de colère et l'avertir de ma déception, de mon mépris naissant. Mais il faisait preuve d'une maîtrise étonnante dans l'analyse de la situation et des caractères." Drieu La Rochelle P., Rêveuse bourgeoisie*, 1937, p. 289*

"Je sortis indigné, le coeur gros de colère et de haine. A compter de ce jour, ce fut entre Navarin et moi la guerre sans merci. A chaque rencontre, je l'insultais et le provoquais, et il se mettait en fureur : c'est une satisfaction qu'il ne me refusait jamais." France A., Le petit Pierre*, 1918, p. 45*

d) Le corpus ENNUI

C'est également le critère "parties narratives, descriptives *vs* passages de dialogues" qui rend compte des différences observées entre le corpus ENNUI, et les autres : la longueur moyenne des phrases est plus importante, ce qui s'accompagne d'un "déficit" en ponctèmes du corpus ENNUI par rapport aux autres, et du fait que la composition moyenne des phrases comporte à la fois plus de lexèmes, de grammèmes et de ponctèmes que les autres corpus (v. tableau 5). C'est le corpus qui compte le moins de ponctèmes sélectionnés par le test, 4 seulement (5 pour COLERE, 6 pour PITIE et 8 pour PEUR).

Pour les points de suspension, il y a lieu de réviser l'hypothèse initiale : la sémantique de l'ennui, sentiment existentiel, se lexicalise en fait dans des unités plus longues que les autres sentiments, avec moins de ponctèmes, ce qui s'explique par de moindres variations modales ; de même, on y trouve une proportion plus élevée des ponctèmes faibles par rapport aux ponctèmes forts (tableau 4). En fait, les points de suspension ne sont sélectionnés dans aucun des corpus et c'est aussi le cas des parenthèses. É. Brunet[23] a noté que ces signes sont en progression constante, de 1789 à nos jours, si on les observe dans la perspective d'analyse par tranches du corpus de FRANTEXT, qu'il a adoptée dans cet article : il rattache ce fait à l'importance croissante du discours direct (les chiffres sont plus importants dans le genre "prose littéraire") et aussi, pour la parenthèse, aux liens avec le discours à la première personne.

Dans ces phrases plus longues[24], les traits sémantiques de la molécule *ennui* sont lexicalisés dans des mots comme *fatigue* (qui a un score bien plus élevé près de *ennui* [16] que près de *colère* [3] ou *peur* [4]) : ce mot lexicalise les traits /imperfectif/, /intensité/ et /dysphorie/ propres à la molécule sémique de *ennui*[25]. De même, si *mortel* et *mourir* sont sélectionnés dans ce corpus, c'est pour lexicaliser les traits /intensité/ et /itératif/[26].

> *"Je désertais de plus en plus mon propre bureau, et c'était chez tante Claire, à côté de l'ours aux pralines, que je subissais avec plus de résignation la torture des devoirs ; sur le mur, dans un recoin caché de la boiserie de cette chambre, un portrait à la plume du Grand-Singe subsiste encore, avec d'autres bonshommes de fantaisie ; l'encre a pâli, jauni, mais on les a respectés et, quand je les regarde, je retrouve encore du mortel ennui, de l'étouffement glacé, - des impressions de collège, enfin." Loti P.,* Le roman d'un enfant, *1890, p. 216*

> *"La terre fut oubliée, le symbole du grand amour fécond. Mes squales patients et attentifs, requis par l'odeur de la proie mûre, émergèrent de mes sillages. Je sombrai plus irréparablement aux impénitences, je reniai la beauté un instant reconquise. Les stupeurs, les lassitudes, un mortel et léthargique ennui de nouveau furent la litière de mes apostasies." Lemonnier C.,* L'homme en amour, *1897, p. 222*

---

[23] *La ponctuation et le rythme du discours (d'après les données du TLF), C.U.M.F.I.D.*, n° 13, 1981, p.1-28.
[24] *Cf.* la citation de Senancour, pour un cas extrême.
[25] Sur l'étude du thème de l'ennui, v. F. Rastier, *La sémantique des thèmes ou le Voyage sentimental*, in Rastier 95, pp. 223-249.
[26] Alors que *mourir* sélectionné avec *peur* lexicalise les traits /intensité/ et /perfectif/, cf. "mourir de peur". En revanche *mort*, qui est sélectionné près de de *peur*, ne l'est pas près de *ennui*, ce qui attire l'attention sur ce que Coseriu nommait les *solidarités lexicales*, qui sont issues d'autres systèmes de normes que le système de la langue.

Mais ces caprices, d'abord si fréquents et si impétueux, sont devenus rares et tièdes ; car l'enthousiasme aussi s'est refroidi et c'est après de longs jours d'assoupissement et de dégoût que je retrouve parfois de courtes heures de jeunesse et d'activité. L'ennui désole ma vie, Pulchérie, l'ennui me tue. Tout s'épuise pour moi, tout s'en va. J'ai vu à peu près la vie dans toutes ses phases, la société sous toutes ses faces, la nature dans toutes ses splendeurs." Sand G., Lélia, *1833, p. 203*

"Elle pouvait se passer de lui, et il retombait dans l'ennui de son existence vide, un ennui qui le laissait les mains ballantes, changeant de siège, se promenant avec des regards désespérés aux quatre murs, s'oubliant devant la fenêtre, sans rien voir." Zola E., La joie de vivre, *1884, p. 933*

"Vos instincts ne vous portent point au crime ; ils repoussent l'infamie. Vous fûtes un type de candeur et de grâce, vous n'êtes aujourd'hui le type de rien : vous vous ennuyez ! L'ennui n'avilit ni ne dégrade, mais il efface, il détruit !

- Vous le savez sans doute, madame l'abbesse, répondit Sténio avec aigreur, car j'ai surpris le secret de vos nuits, et je sais que vous ne lisez pas, que vous ne dormez pas, que vous ne priez pas ; je sais que, vous aussi, l'ennui vous dévore !" Sand G., Lélia, *1839, p. 524*

"Enfin, on partit, et ce jour-là, en effet, l'armée pivota sur sa gauche (...) De Contreuve à la vallée de l'Aisne, les plaines recommençaient, se dénudaient encore ; la route, en approchant de Vouziers, tournait parmi des terres grises, des mamelons désolés, sans un arbre, sans une maison, d'une mélancolie de désert ; et l'étape, si courte, fut franchie d'un pas de fatigue et d'ennui, qui sembla l'allonger terriblement." Zola E., La Débâcle, *1892, p. 98*

"- En vérité, ce serait peut-être un bonheur pour moi, répondait Lucien ; ou, pour parler avec l'exactitude mathématique que nous aimons, rien ne peut être pour moi aggravation de peine ; je crois, sans trop présumer, être parvenu au comble de l'ennui." Stendhal, Lucien Leuwen, *1835, t. 1, p. 128*

"Elle restait assise, dans sa cuisine, ou dans sa chambre, d'où par-dessus les cheminées elle apercevait le sommet d'un arbre, dans un jardin d'hôpital. Elle ne lisait pas, elle essayait de travailler, elle s'engourdissait, elle s'ennuyait, elle pleurait d'ennui ; elle avait un pouvoir singulier de pleurer, indéfiniment : c'était son plaisir. Mais quand elle s'ennuyait trop, elle ne pouvait même plus pleurer, elle était comme gelée, le coeur mort." Rolland R., Jean-Christophe, *1908, p. 810*

"Je me demande quelquefois où me conduira cette contrainte qui m' enchaîne à l'ennui, cette apathie d'où je ne puis jamais sortir ; cet ordre de choses nul et insipide dont je ne saurais me débarrasser, où tout manque, diffère, s'éloigne ; où toute probabilité s'évanouit ; où l'effort est détourné ; où tout changement avorte ; où l'attente est toujours trompée, même celle d'un malheur du moins énergique ; où l'on dirait qu'une volonté ennemie s'attache à me retenir dans un état de suspension et d'entraves, à me leurrer par des choses vagues et des espérances évasives, afin de consumer ma durée entière sans qu'elle n'ait rien atteint, rien produit, rien possédé." Senancour E.-P. De, Obermann, *1840, t. 1, p. 157*

Cette première expérience d'analyse des ponctèmes à la lumière des traits sémantiques stables et structurés définis par les autres signes, a permis de mettre en évidence que les différences entre les scores des ponctèmes au test statistique sont bien liées à celles des différentes thématiques et à leurs modes de lexicalisation : il est donc nécessaire de rapporter les éléments observés aux composantes textuelles puisqu'ils sont inséparables du thème et de l'ensemble du contenu textuel. Ces résultats nous ont confortée dans l'objectif de concevoir des programmes d'aide à l'analyse sémantique qui puissent faire apparaître les liens entre

lexèmes et ponctèmes, qui concourent à l'interprétation du sens et définissent ensemble les zones de localité dans lesquelles l'interprétation est pertinente.

III. LE CORPUS PEUR ET SES SOUS-CORPUS : PONCTUATION ET CRITERES ENONCIATIFS

III. 1. Les hypothèses

Le corpus *peur* a fait l'objet de trois examens différents :

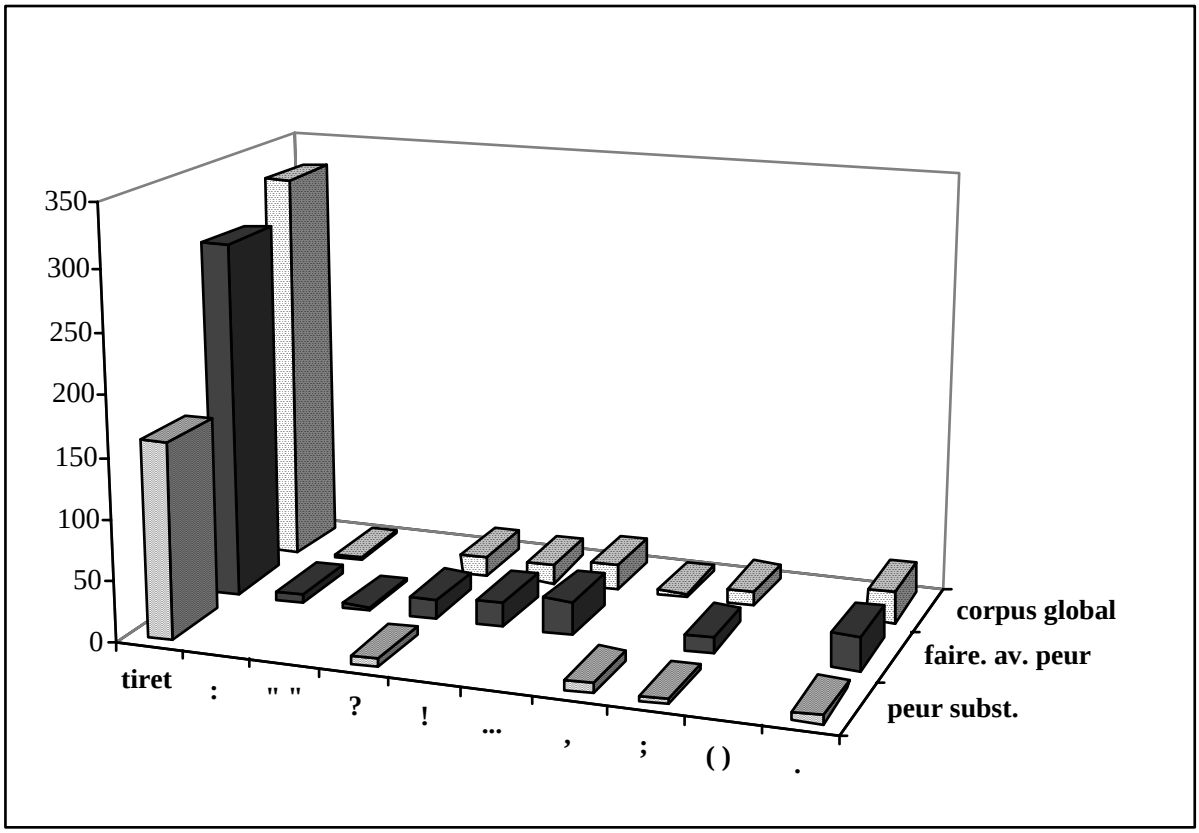
- **corpus 1** : corpus global (ensembles des contextes de 10 mots avant + 10 mots après *peur*).
- **corpus 2** : de ce corpus global, on ne garde que les contextes où *peur* figure avec les verbes *faire* et *avoir* (sans unité graphique intercalée).
- **corpus 3 :** corpus 1 - corpus 2, c'est-à-dire (à peu de choses près) le corpus de *peur* comme substantif.

La recherche sur le thème de la peur, effectuée dans le corpus ROMAN autour de 26 mots-pivots, avait montré en effet que la thématique globale des mots de la peur excluait de façon régulière la "parole représentée". Mais la comparaison entre les quatre corpus "sentiments" avait mis en évidence qu'autour du mot *peur* on pouvait rencontrer tantôt des séquences de dialogues (autour de *faire / avoir peur*, principalement, cf. les énoncés ci-dessus), tantôt des parties de récit. Nous avons fait l'hypothèse que le paradigme des ponctèmes devait être représenté différemment selon ces critères énonciatifs et qu'il devait y avoir une variation nette, entre les ponctèmes sélectionnés par le test statistique, dans les corpus 2 et 3.

Le tableau 6 et le graphique 3 présentent les résultats du test statistique.

	1 : corpus global	**2 : *faire/avoir peur***	**3 : *peur* (subst.)**
.	26	28	7
?	15	15	6
( )	-2	-1.9	-0.8
"	-2	3	-8
!	15	21	-3
...	21	28	-0.5
,	4	-3	9
;	12	12	4
:	3	7	-2
-	326	294	161

Tableau 6 : les ponctèmes dans les différents corpus PEUR



Graphique 3 : les sous-corpus PEUR

III. 2. Commentaires des résultats statistiques

Les différences observées, à propos du statut "sélectionné ou non" des ponctèmes, comme au plan de leur score d'écart réduit, s'expliquent bien par la répartition entre passages descriptifs et passages de dialogues. Les parties narratives dominent largement dans "peur substantif", ce qui explique les scores inférieurs du point, du point d'interrogation, d'exclamation et des points de suspension, alors que la virgule y a un score d'écart réduit plus élevé. En revanche, les deux-points, marque de dialogue, ont un score plus élevé dans le corpus 2 ("faire / avoir peur"), alors qu'ils ne sont pas sélectionnés dans le corpus "substantif" : de même, les guillemets ne sont sélectionnés que dans ce corpus, qui contraste nettement avec le corpus global, comme avec le corpus "substantif"[27].

On a pu remarquer d'après différentes études en cours que la valeur des points de suspension peut aussi être très variable : elle semble, en tout état de cause, servir autant, et peut-être plus[28] à noter qu'un contenu sémantique est chargé d'intensité pour un acteur que la représentation, par le narrateur, d'une interruption dans un échange de paroles[29], ces deux valeurs ne s'excluant d'ailleurs pas dans de nombreux énoncés du corpus PEUR. Il semble aussi

[27] V. les énoncés du corpus donnés en exemples dans II.
[28] Dans le corpus ROMAN, en tous cas : est-ce une particularité due au genre ? Il faudra aussi vérifier si cette valeur est un fait d'évolution, car elle semble plutôt se développer depuis la fin du XIXème siècle. Sur l'importance de prendre en compte la dimension historique, v. É. Brunet, *loc. cit.*, et N. Catach, *Liaisons -HESO,* 1989, n° 16-17, p. 24-26.
[29] Cf. ce qui a été dit de l'hypothèse sur les points de suspension dans le corpus ENNUI.

qu'on puisse y repérer une valeur "dialectique" de changement d'acteur : c'est l'intérêt des études de corpus de faire ressortir ce genre de régularités et de permettre de préciser la valeur différentielle des ponctèmes.

Si l'on compare les résultats du corpus global et ceux du corpus 2 ("faire / avoir peur"), on observe qu'ils sont très proches, les parties de dialogues étant très riches en points, points d'exclamation, d'interrogation, de suspension, aussi bien que de marques de dialogues (tiret et deux points). Les résultats globaux "lissent" donc des distinctions importantes, et l'on voit comment l'interprétation de résultats, aussi bien que la délimitation des corpus de travail, ou, à plus forte raison, la mise en œuvre de calculs statistiques, nécessitent une certaine connaissance du corpus[30] et une attention aux différentes variables. A présent que le progrès technique permet de disposer d'un volume important de textes électroniques, il faut bien être conscient du fait que ces corpus ne sont pas utilisables pour des études de contenu sans certaines précautions et sans outils spécifiques, fondés sur les connaissances des différentes disciplines qui ont étudié les textes[31].

IV. LA PONCTUATION DE TROIS ROMANS, CONTRASTEE SUR LE CORPUS DE REFERENCE

La recherche précédente sur la ponctuation s'appliquait à un corpus multi-auteurs, dans un genre précis, et dans un corpus de travail sélectionné selon des contraintes sémantiques, la thématique des sentiments.

Dans une autre étude, on a appliqué le test statistique aux signes de trois romans du corpus, en les contrastant un par un sur le corpus de référence. On voulait vérifier :

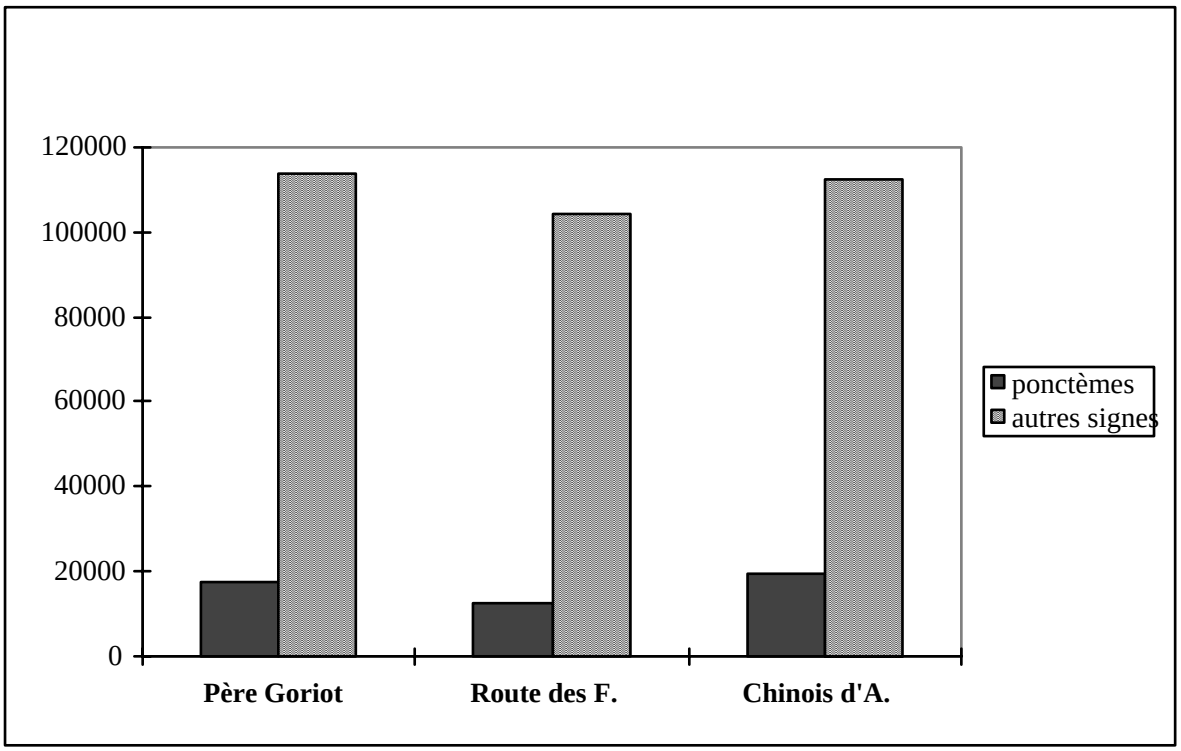
- si ce genre de contraste pouvait permettre d'offrir à l'utilisateur d'une banque de données textuelles un "aperçu" sur les mots "importants" du texte qui lui permette de formuler ou conforter des hypothèses de recherche[32]

- si dans des textes appartenant au genre du roman la répartition des ponctèmes se modifiait en diachronie, en fonction de l'évolution de ce genre dans la période contemporaine. On a comparé un roman "traditionnel" du XIX^{e}s., *Le Père Goriot* (de H. de Balzac, 1835), un texte appartenant à l'école du "nouveau roman", *La Route des Flandres* (de C. Simon, 1960 [33])

[30] Il n'est pas inutile de rappeler aussi que ce corpus doit être aussi homogène que possible au plan des contraintes de genre, pour qu'il soit possible d'interpréter les résultats.
[31] V. IV et la conclusion à propos d'autres problèmes d'interprétation.
[32] La méthode des "mots-clés", utilisée aussi bien en analyse littéraire que dans les systèmes documentaires s'est révélée bien pauvre, parce que trop liée à la subjectivité, ou à un thésaurus établi sans connaissances préalables du corpus : dans tous les cas, les limites de ces méthodes viennent de ce qu'elles ne s'appuient ni sur une sémantique de la langue des textes pris en compte, ni sur la reconnaissance du statut sémiotique de la langue et des textes (et donc aussi du thésaurus). Puisque pour nous un mot n'est pas interprétable hors contexte, nous ne considérons pas que les mots sélectionnés par le test statistique dans un roman (contrasté sur le corpus global, pourtant très important) puissent être considérés comme des mots-clés donnant *de facto* accès aux thématiques, mais les expériences menées montrent que ces listes sorties par la machine permettent d'appréhender le texte et ses structures avec des éclairages intéressants.
[33] Ce roman est le seul représentant de cette école littéraire dans le corpus (qui comporte des œuvres de 1830 à 1970) : nous avons utilisé l'édition de Minuit, Paris, 1960. Pour *Le père Goriot*, l'

et un autre texte qui porte la marque des recherches contemporaines sur le plan diégétique, *Le Chinois d'Afrique* (de R. Sabatier, 1966) [34] .

Ces textes ont été choisis en fonction des hypothèses énoncées ci-dessus, et aussi parce que leurs volumes sont très proches, comme le montre le graphique [35] : en effet, on ne peut comparer des œuvres trop différentes en volume, car les structures textuelles varient selon le "projet esthétique" et le "contrat énonciatif", même implicite, passé avec le lecteur.



Graphique 4 : répartition des signes

Le tableau 7 récapitule les résultats du test statistique, qui montrent que le "système de ponctèmes" utilisé dans chaque texte est très particulier : ces résultats sont présentés en format "3 D" dans le graphique 5.

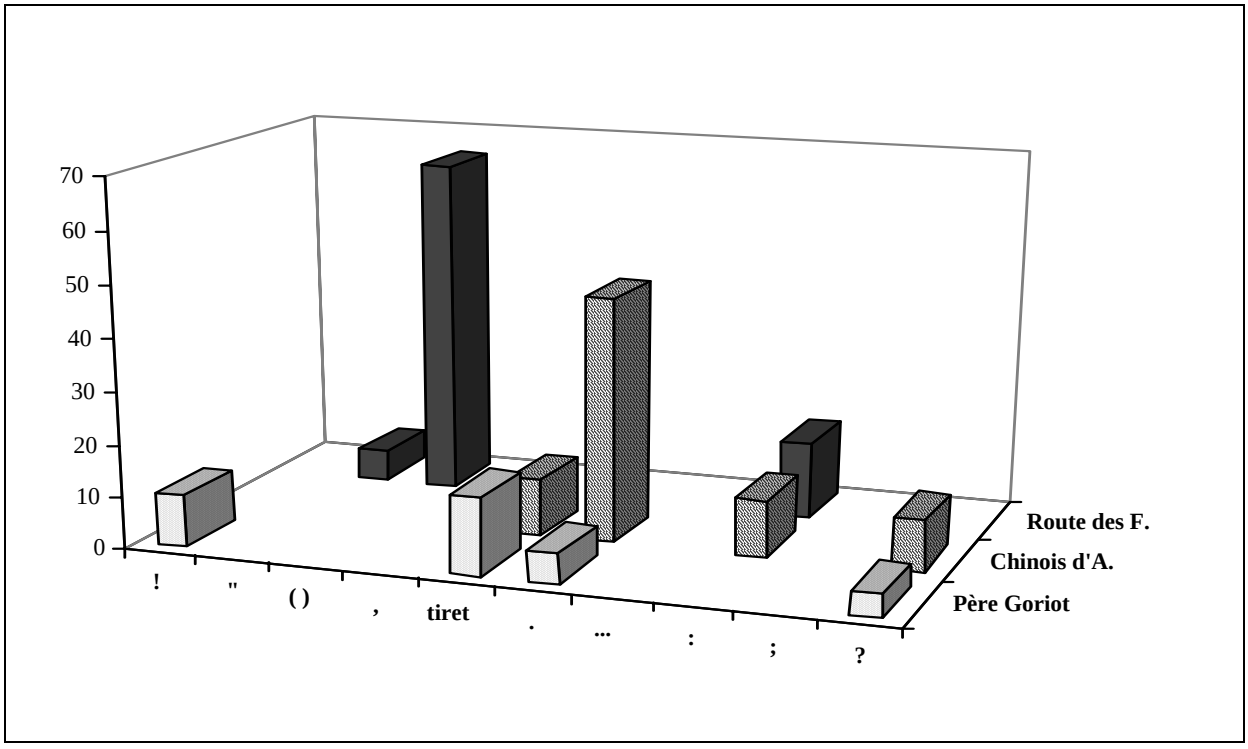
édition utilisée est celle de P.-G. Castex, Paris, Garnier Frères, 1960 et pour *Le Chinois d'Afrique* l'édition Albin Michel, Paris, 1966.

[34] Il ne semble pas que R. Sabatier soit explicitement reconnu comme appartenant à ce mouvement. Plusieurs participants au Colloque de Cerisy en 1971 ont proposé de nommer "Nouveau Nouveau roman" le stade postérieur au "premier nouveau roman" des années cinquante et les nouvelles tentatives, cf. J. Ricardou, *Le nouveau roman,* Paris, Seuil, Points, 1973, 1990, p. 151.

[35] Une des fonctionnalités du menu "sélection bibliographique" du logiciel de FRANTEXT permet de visualiser rapidement le volume des textes, et on peut ensuite connaître très facilement le nombre d'occurrences d'une liste de formes, comme par exemple le paradigme des signes de ponctuation, dans chaque texte.

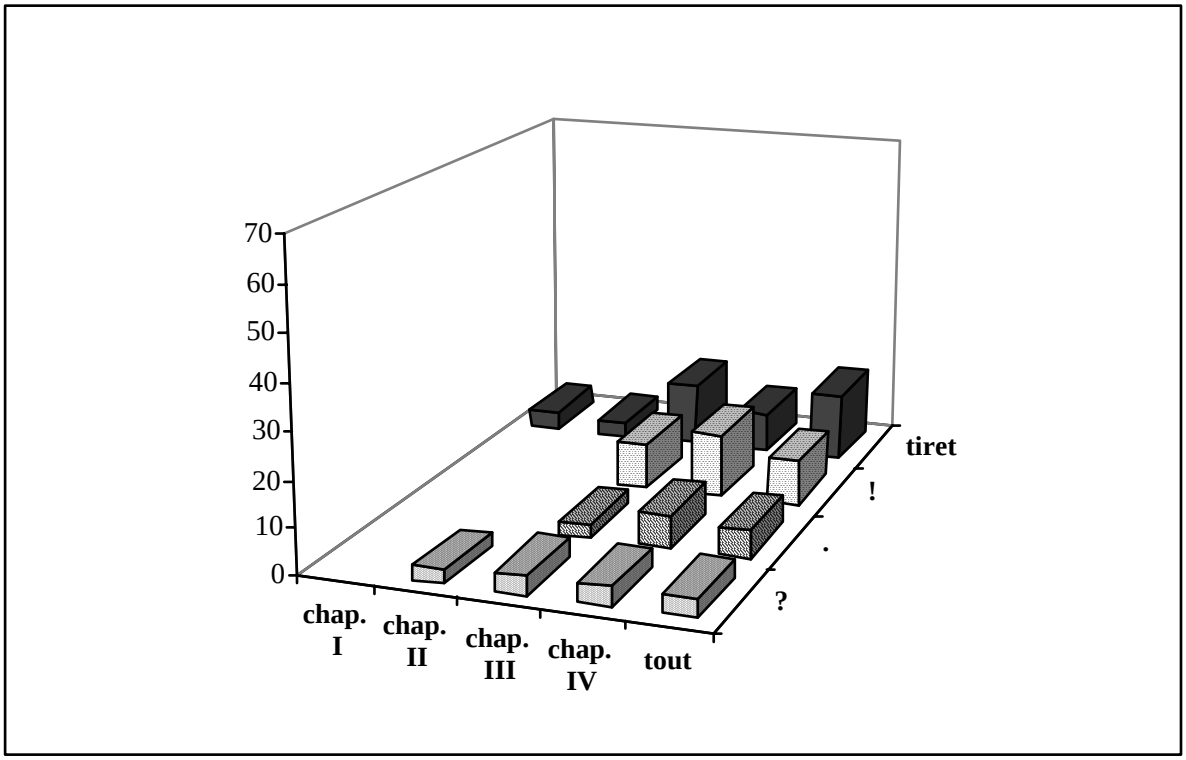
PONCTEME	PÈRE GORIOT	CHINOIS D'AFRIQUE	ROUTE DES FLANDRES
!	**10**	ns	ns
"	ns	ns	**6**
( )	ns	ns	**65**
,	ns	ns	ns
-	**15**	**11**	ns
.	**6**	**47**	ns
...	ns	ns	ns
:	ns	**11**	**15**
;	ns	ns	ns
?	**4.5**	**10**	ns

Tableau 7 : comparaison des ponctèmes de trois romans



Graphique 5 : scores des ponctèmes dans les trois romans

Le test ne sélectionne que quatre ponctèmes dans *Le Père Goriot* et *Le Chinois d'Afrique*, et trois seulement dans *La Route des Flandres*. On a ensuite contrasté chacune des subdivisions dont les auteurs ont structuré leur texte sur le corpus de référence, pour observer si le "système des ponctèmes" mis en place par chaque écrivain était stable dans l'ensemble du texte : les graphiques 6, 7 et 8 montrent la répartition des ponctèmes dans ces subdivisions .

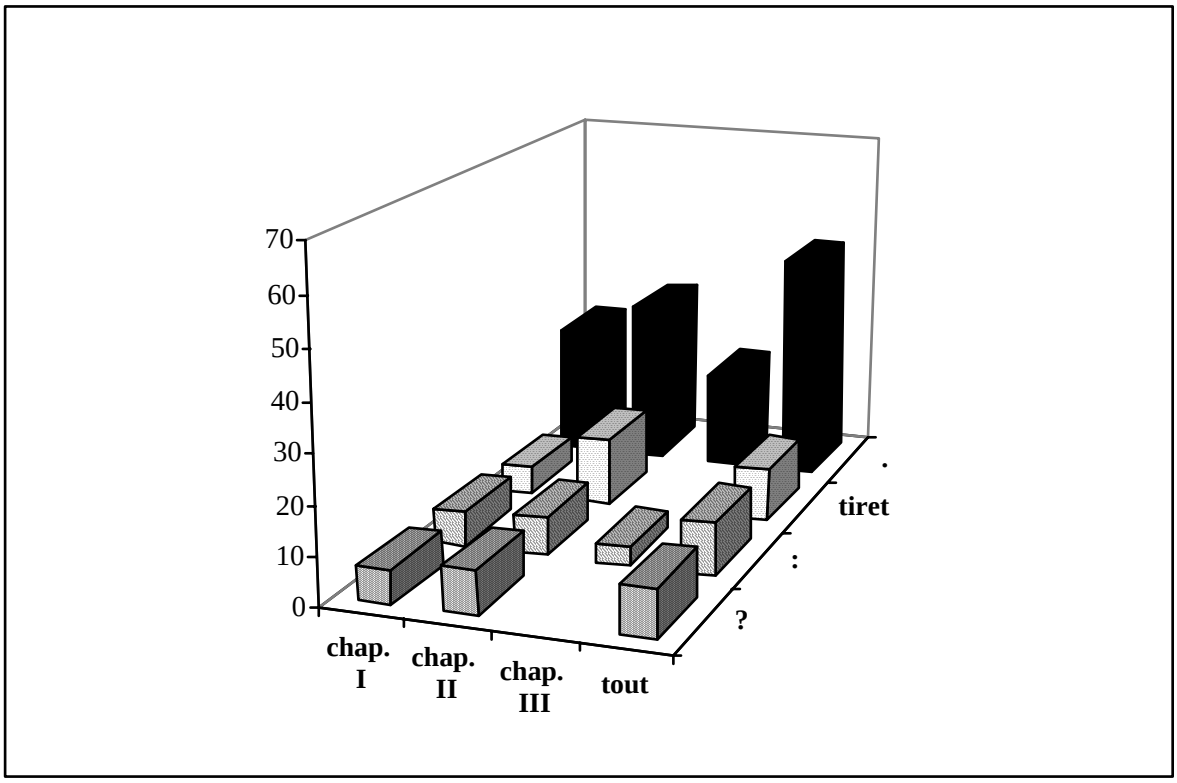


Graphique 6 : les ponctèmes dans Le père Goriot

Le Père Goriot est un texte où aucun score de ponctème n'atteint des extrêmes[36], si on le compare avec les deux romans modernes, et aussi avec les quatre corpus "sentiments" (cf. II). Le score de l'ensemble du texte "lisse" les différences entre les chapitres : le chapitre un, qui décrit la pension Vauquer et ses occupants dans de longues parties narratives se caractérise par un déficit de ponctèmes sélectionnés, et le tiret[37] seul ressort, en raison des dialogues de la fin du chapitre, où la parole représentée des différents pensionnaires complète la description que le narrateur en a faite. En revanche, dans les chapitres trois et quatre, qui sont les plus "dramatisés" le test a sélectionné le point d'exclamation, le point d'interrogation et le tiret ; ces signes et le point ont d'ailleurs un score encore plus marqué dans le chapitre quatre, intitulé *La mort du père*, où quelques épisodes revêtent une grande importance pour plusieurs acteurs de *La Comédie Humaine* et sont donc particulièrement riches aux plans thématique, dialogique et tactique.

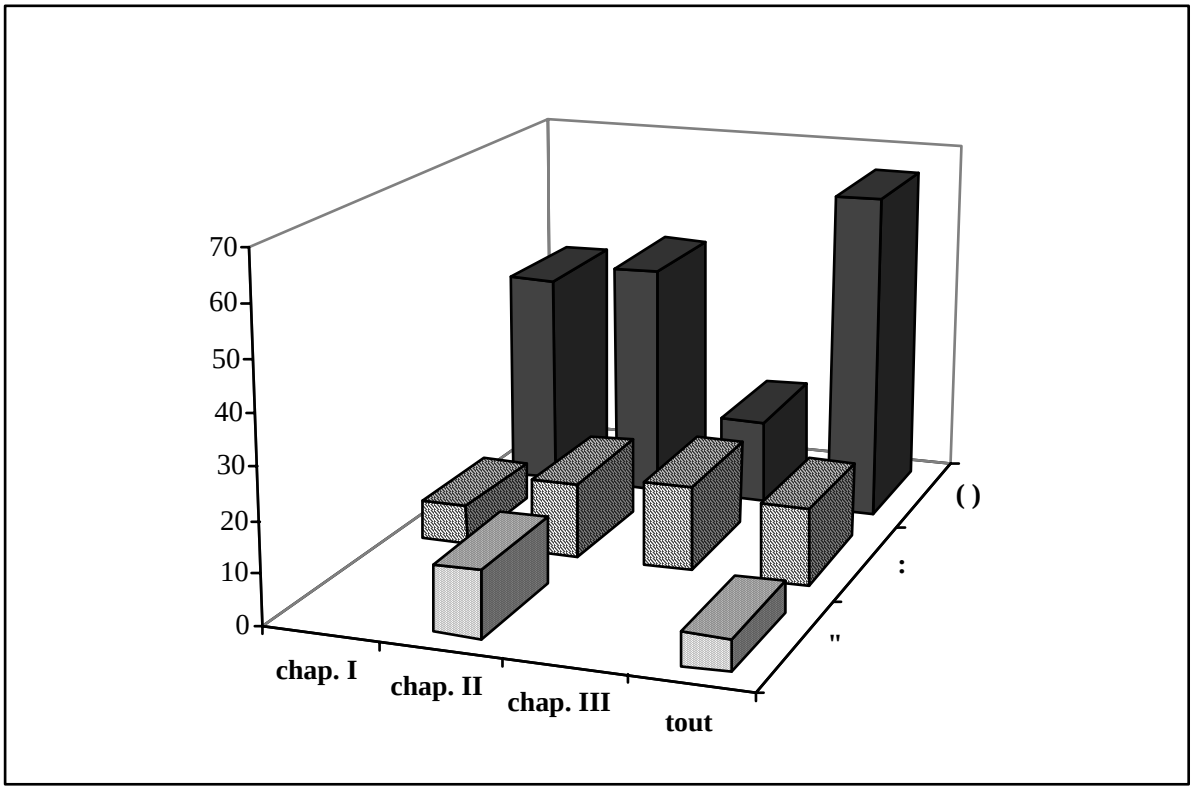
[36] Dans cette étude, on a voulu comparer les systèmes de ponctèmes de trois œuvres : pour une étude détaillée d'un auteur, surtout du XIX^{e}s., il y aurait lieu de vérifier la ponctuation de l'auteur, et les éventuels changements opérés par les éditeurs.

[37] On remarque aussi la différence très nette de score du tiret entre les trois romans et par rapport aux scores des quatre corpus "sentiments" : les calculs de scores seront plus fiables à l'avenir quand la valeur "trait d'union" sera éliminée par le rattachement au dictionnaire de formes composées (cf. note 17), mais, avec toutes les précautions qui s'imposent, on peut néanmoins observer des différences d'échelle significatives. De plus, le cas de *Le Chinois d'Afrique*, est tout à fait intéressant, car il y a accord total entre le score du tiret et des deux points, qui sont dans ce texte vraiment employés ensemble comme marque de dialogue, alors que ce texte comporte bien, comme les autres, son lot de mots composés.



Graphique 7 : les ponctèmes dans Le Chinois d'Afrique

Dans *Le Chinois d'Afrique*, le score égal des deux points et des tirets rend compte de l'importance de la parole représentée en discours direct, toujours introduit par ces deux marques, tandis que les guillemets servent à marquer le discours intérieur, ou une distance de l'énonciateur (par exemple quand il reprend les termes d'un autre acteur). Le score important du point, atypique aussi bien par rapport aux deux autres textes, que par rapport aux scores de ce signe dans les quatre corpus "sentiments" (cf. II) s'explique par une caractéristique de ce roman : le mode narratif se caractérise par des phrases très courtes et un texte "découpé" à l'extrême. En effet, structuré en trois chapitres (nommés "parties"), il est encore articulé par des subdivisions (huit dans la première partie, sept dans les deux autres), elles-mêmes fractionnées encore (avec le signe que les typographes appellent "astérisque", consistant en un motif de trois étoiles - marquant les ruptures thématiques et dialogiques), puis en paragraphes et enfin en phrases (souvent des prises de paroles marquées par les deux points, le retour à la ligne et le tiret). Ce dispositif sert un projet esthétique où le narrateur se situe clairement dans un mode "objectiviste", dans un texte où la prise en charge du nom de l'acteur principal et sa description sont l'objet d'un dispositif polyphonique très particulier. Le graphique 7 permet de constater une différence nette entre le troisième chapitre et les deux premiers, au plan des ponctèmes, qui doit s'étudier à l'aide des composantes textuelles.



Graphique 8 : les ponctèmes dans La Route des Flandres

Pour La Route des Flandres, on retrouve en tête, sans grande surprise, les parenthèses chères à C. Simon, avec un score vraiment atypique, par rapport aux deux autres romans, aux corpus "sentiments" et à d'autres études menées parallèlement ; le point n'est pas sélectionné et ces deux éléments renvoient à la longue phrase simonienne. La pondération sur un ensemble de textes important comme le corpus ROMAN, corpus de référence, a bien mis en évidence cette particularité, ainsi que l'emploi des deux points, qui servent à introduire la parole représentée (mais sans l'aide du tiret, comme c'est le cas chez Balzac ou Sabatier). On constate également un score atypique des guillemets, qui sont rarement sélectionnés, on l'a vu, dans les autres expériences rapportées ici (seulement dans le sous-corpus "faire / avoir peur"), et aussi dans les études en cours[38]. Cependant le test met en lumière que même ces signes d'emploi particulier à C. Simon connaissent un score différent selon les parties du texte. Dans le chapitre trois on constate une diminution de score des parenthèses ; de même, on observe que la sélection des guillemets dans l'ensemble du texte est imputable en fait uniquement au chapitre deux tandis que le score des deux points est inférieur dans le chapitre un aux deux autres chapitres. Nous pensons que ces sortes de "photographies" d'une œuvre sont intéressantes pour l'usager d'une banque textuelle, et qu'avec les listes de lexèmes et grammèmes sélectionnés par le test de l'écart réduit, ainsi que les programmes de sortie des contextes riches en signes sélectionnés, le chercheur peut émettre et vérifier des hypothèses

[38] On a observé (II. 4, corpus ENNUI) que É. Brunet les rattache à l'évolution moderne des modes narratifs et aux parties de texte à la première personne.

sur les structures textuelles, mettre en évidence des contrastes qui interrogent son approche du texte et stimulent l'interprétation.

CONCLUSION ET PERSPECTIVES

Ces premiers éléments d'études menées sur des corpus littéraires importants, en vue d'intégrer les signes de ponctuation à une sémantique textuelle se sont révélés encourageants. Toutes sortes de problèmes techniques subsistent mais on a vu que pour permettre à l'usager d'une banque textuelle d'interpréter certains éléments difficiles, il est nécessaire de pouvoir multiplier les calculs, de faire évoluer les corpus de travail en fonction d'hypothèses successives et d'avoir des documents de présentation des résultats qui soient clairs à interpréter. Il est nécessaire d'améliorer encore les outils que nous élaborons, mais cependant, la loi des grands nombres joue sur un corpus de référence important et l'interprétation des contrastes est toujours un moyen d'approcher au plus près les structures textuelles.

Dans la nouvelle phase de travail, nous mettons en œuvre un chantier de codage des textes avec les balises SGML et TEI, qui permettent de fournir des possibilités nouvelles de structuration des corpus de travail, sur des critères philologiques et herméneutiques : information sur la langue du texte (pour contraster les emplois ponctuels de formes étrangères), codages des différentes subdivisions, des noms (de lieux, de personnes) importants pour "situer" l'action historiquement et géographiquement, mais aussi des traits sémantiques (plus ou moins d'acteurs ayant le trait /masculin/ ou /féminin/, /noble/, etc. ?), des citations et de leur source (intertextualité), etc. Mais on pourra aussi constituer des corpus de travail sur la base des critères énonciatifs (comme la connaissance du corpus nous a permis de le faire pour PEUR) grâce aux balises de dialogues (plus sûres que les seules marques de ponctuation, car, comme on l'a vu, les systèmes varient suivant les auteurs). Les signes de ponctuation serviront, dans ce cadre, à mieux cerner les limites dans lesquelles appliquer les calculs statistiques et/ou les programmes de concordances "riches sémantiquement". On pourra ainsi utiliser les richesses des banques de données textuelles et mener des études particulières sur certains signes de ponctuation, des études contrastives sur les ponctèmes en fonction des thématiques et des genres, mieux appréhender les systèmes de normes, les contraintes liées aux genres textuels et l'évolution diachronique des genres.