Chapitre 7

Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus

7.1. Introduction

Le lecteur sera peut-être d'emblée frappé par le paradoxe de ce titre : lire des corpus ? quelle étrange idée aujourd'hui ! C'est pourtant bien ce dont il s'agit et ce que cette contribution voudrait apporter d'original à l'état de l'art.

En effet, le développement de la/des *linguistique(s) de corpus*, tant comme aire de recherche que comme objet d'enseignement, oblige à connaître et à problématiser une équivoque, comme souvent en français liée à l'interprétation de la préposition de. Au premier chef, *linguistique de corpus* désigne une option théorique et méthodologique en sciences du langage, de Bloomfield à Biber, en passant par Harris[1], pour ne citer qu'eux. En substance : tenir un discours scientifique sur une langue en tant que système signifiant, c'est avant tout, sinon exclusivement, décrire les régularités observées dans le corpus de cette langue, c'est-à-dire une collection d'énoncés effectivement produits dans une synchronie socioculturelle circonscrite avec rigueur, et destinés à représenter le meilleur échantillon possible de cette langue. Cette classe de corpus, sous cette acception, recouvre des ensembles dont l'accroissement est par définition facteur de qualité et de précision. Lire de tels corpus n'aurait aucun sens.

Chapitre rédigé par Jean-Marie VIPREY.

1. Notamment (Bloomfield, 1930 ; Biber, 1999 ; Harris, 1969).

Mais si la/les *linguistique(s) de corpus* sont l'objet d'une si vive demande aujourd'hui, ce n'est pas seulement parce qu'entendue(s) comme *supra* elle(s) détermine(nt) le développement des « *industries de la langue* ». C'est aussi et surtout que les méthodes qui les constituent sont appelées à se transposer dans toutes les sphères des sciences humaines, dès lors que leurs *corpus* comportent un composant verbal. Ainsi, de la linguistique *par* les corpus, passe-t-on à une/des linguistique(s) *pour* les corpus, ce qui est à la fois la même chose… et tout autre chose !

Il n'est pas trivial aujourd'hui d'insister sur cette distinction définitionnelle. Nous entendrons ici par *corpus* les collections de données textuelles que constituent les sciences humaines pour les explorer systématiquement, dans le cadre d'une étude *a priori* plus vaste. Ces corpus ne sont plus nécessairement des échantillons : ils peuvent être supposés exhaustifs (quitte à admettre que certaines pièces puissent manquer) : collection de tracts, d'interviews, de romans, etc.

Avant de restreindre plus encore notre propos, considérons la posture du chercheur en sciences humaines face à son corpus. Quelle qu'en soit la taille, celui-ci est constitué de textes qui renvoient en fait au *discours*[2], à la formation discursive qui est l'objet d'étude en dernière analyse. Le chercheur ne peut en aucun cas se dispenser de lire extensivement sinon tout son corpus, du moins de substantielles parties de celui-ci. Si le corpus est de taille modérée (centaines de milliers de mots), on admettra mal qu'il n'ait pas fait l'objet d'une lecture « à l'œil et à la main », linéaire ou non, et plus ou moins exhaustive. Mais même s'il est plus vaste (millions, voire centaines de millions de mots), et s'il n'est alors plus du tout question d'une lecture exhaustive, on n'admettra toujours pas que l'étude ne comporte pas des citations substantielles. Et celles-ci ne seront pas convoquées comme des échantillons de la « langue », mais comme des éléments de discours, face auxquels l'activité interprétative est plénière, puisqu'ils actualisent l'instance discursive étudiée.

Nous sommes donc à l'exact opposé du linguiste « bloomfieldien » (que l'on prendra – non certes sans un certain « utile » abus – comme anti-modèle), qui doit écarter autant que possible l'intuition qu'il a du sens des énoncés recueillis.

2. Il est fait référence ici, sans pouvoir la développer, à la sphère de l'Analyse de Discours. Voir notamment (Pécheux, 1990 ; Maingueneau, 1991 ; Bakhtine, 1977).

Il se trouve par ailleurs que, trop souvent, l'activité de *citation* s'expose naïvement au soupçon de partialité. Dans la période récente, liée à la généralisation du micro-ordinateur, il est devenu très commode de trouver et filtrer, dans la masse d'un corpus, les énoncés susceptibles de venir à l'appui d'une thèse. L'abus de cette facilité, aggravé par le vernis de scientificité qu'a pu fournir la technologie nouvelle, a même pu alimenter certaines réticences, pas toujours de très bonne foi elles non plus. Pour parer à ce risque, l'application de méthodes statistiques rigoureuses est de première utilité, qui permettent de régler *ad optimum* la fiabilité des extractions, leur impartialité et/ou leur « pluralisme » (la diversité des hypothèses au service desquelles elles sont susceptibles de venir). Cette impartialité, ce pluralisme, ce sont d'autres termes pour désigner la plus ou moins grande conformité de la panoplie d'extraits invoqués, à la vision synthétique du corpus qu'en donne la statistique.

Mais le propos est plus vaste : l'état de l'art permet – et exige – aujourd'hui d'envisager de *lire* effectivement les corpus, quelle qu'en soit la taille.

Nous ne visons bien sûr pas le simple fait que les micro-ordinateurs (et surtout les réseaux auxquels ceux-ci font référence) peuvent stocker et afficher des ressources gigantesques, ni même ceci, encore bien trivial, que les recherches d'occurrences, concordances et index dynamiques, facilitent l'accès aux données utiles. Nous visons moins encore le domaine des bases de données structurées. Il s'agit de faire monter du corpus, c'est-à-dire par définition de son étude, les éléments de sa propre *viabilisation*[3], au premier rang desquels ceux d'une cartographie statistique hypertextuelle. Ainsi le corpus devient-il plus accessible au fur et à mesure qu'il est exploré, comme il en va de la lecture « naturelle » ; ainsi peut-on espérer circonscrire l'effet d'hyperinterprétation dit *boîte noire*, et remettre en selle un critère élémentaire de la recherche scientifique : la connaissance critique, toujours approfondie, de ses propres données.

On dessine donc, dans le domaine restreint des corpus autres que strictement linguistiques, l'opposition entre deux perspectives sémantiques : l'une *exogène*, reposant sur des thesaurus *ad hoc* et/ou *a priori*, l'autre *endogène*, interposant le moindre appareil possible entre le corpus et son expert, et construisant l'appareil

3. *Viabilisation* s'entend ici : mise en place d'une « voirie », c'est-à-dire des éléments de repérage permettant l'accès, l'exploration, l'acquisition. Il s'agit d'un *processus* instruit tout au long de l'exploration, et non d'un appareil préalable et *ready made ;* en ce sens, il est *endogène*.

sémantique de l'intérieur, comme une part constitutive du corpus, ou du moins comme un élément de son infrastructure.

7.2. Sens et corpus littéraire : brève histoire d'un long malentendu.

Parmi les corpus de sciences humaines, tels que circonscrits *supra*, ceux qui relèvent de l'analyse littéraire occupent une place assez particulière[4].

Tout d'abord, ils sont en droit incontournables dans leur champ. Il semble assez hasardeux d'étudier la littérature sans étudier les textes, d'étudier par exemple Baudelaire sans réunir au moins le corpus de ses écrits, le théâtre classique sans recueillir les textes dont cette notion d'histoire littéraire délimite l'ensemble.

Ensuite, il est largement admis que ce qui définit littérairement un auteur, une école, une œuvre, c'est au premier chef une *langue* singulière (quoi que l'on suppose ensuite sous ce vocable). On peut même y voir un point de rencontre tout à fait privilégié avec la linguistique de corpus dans son histoire même : qu'est-ce d'autre que Baudelaire en tant que langue, voire que langage (englobant les autres niveaux sémiotiques envisageables), que le(s) systèmes(s) descriptibles dans son corpus ?

Quel que soit le degré d'accord sur ce second point (le premier étant difficilement négociable), on conviendra en outre que la densité et l'hétérogénéité du discours littéraire en fait un terrain d'expérimentation commune passablement probant et que les problèmes de *lecture* et de *sens* s'y configurent d'une façon plutôt emblématique. C'est pourquoi (outre le fait que l'auteur de ce chapitre est un spécialiste de littérature française) nous centrons maintenant le propos sur ce domaine[5].

Car en matière de corpus d'analyse littéraire, nous devons tout d'abord rappeler que la question du *sens* se pose d'emblée au niveau méthodologique, comme une question historique, en France en tout cas.

---

4. Voir aussi *infra*, la *mise en garde* finale, qui rediscute cette place.

5. Sans perdre de vue le risque majeur de cette option : on fera ici référence à un certain état de l'art, à certaines discussions, qui sont partiellement propres à un champ disciplinaire particulier. Il faut se souvenir qu'une part notable des pionniers des sciences du langage, et aussi de la statistique lexicale, puis textuelle, étaient des « littéraires ». On ne peut sans cela comprendre la structure même des ouvrages de référence de Muller (Muller, 1997a et b).

Les premiers travaux explicites de collection et d'étude de corpus littéraires en France furent des index-concordances[6], à la fois héritiers de toute la tradition philologique et précurseurs de l'hypertexte informatisé. Les auteurs étaient des spécialistes reconnus des œuvres concernées, et leurs productions furent reçues comme de nouvelles modalités d'accès à la textualité, permettant de passer un cran déterminant dans la quête de scientificité.

Or il se trouva que les mêmes courants universitaires durent endosser en même temps d'autres responsabilités, jouant notamment un rôle-clé dans la rénovation des sciences du langage, sur le fond partiellement conflictuel de l'introduction du distributionalisme, et de l'apogée du structuralisme. L'entreprise conjointe du *Trésor de la Langue Française* (TLF), puis de *Frantext*, est emblématique de l'obscurcissement graduel de la notion de *corpus*.

Avant même le lancement du TLF, Guiraud (Guiraud, 1954) utilisa la *Liste Vander Becke*[7] comme *système de référence* pour déterminer les *caractères arithmo-sémantiques* du vocabulaire d'un texte donné. En résumé, les fréquences dans la compilation Vander Becke servent à établir les valeurs de référence (la *norme*) pour les effectifs réels des items lexicaux de ce texte (*Les Fleurs du mal*, par exemple), et les termes significativement suremployés accèdent au statut de *mots-clés*, caractéristiques de l'œuvre. Guiraud ne semble pas avoir, en 1954, employé le terme de *corpus*. Significativement, la *compilation* Vander Becke est en outre réduite au statut de *liste*. Dans Guiraud (Guiraud, 1960)[8], on note cependant (p. 79) le terme de *corps* pour désigner à nouveau *Vander Becke* (et la « liste » d'*Eldridge*), terme qui semble être la plus approchante des allusions à cette notion.

Ce n'est qu'avec l'entreprise du TLF que s'impose inévitablement la notion de *corpus*. Après 1960, la *norme* de référence pour déterminer les spécificités d'un texte ou d'un auteur, est de plus en plus régulièrement désignée comme le corpus du TLF, chaque fois qu'est convoquée une *stylométrie* de l'écart. Le *corpus* du TLF (collection d'échantillons de « la langue littéraire » aux fins d'une étude diachronique du lexique français), fut en l'occurrence détourné de son emploi, c'est-

6. Par Paul Imbs, Bernard Quemada, Georges Matoré, Pierre Guiraud, notamment. Index publiés en polycopiés, dont rend compte notamment Muller (Muller, 1993).

7. Guiraud (Guiraud, 1954 : 63 ssqq) : *compilation de 88 textes de 13 000 mots chacun*, en référence à Vander Becke (Vander Becke, 1931), *French Word Book*, Mac Millan (New York).

8. Alors même que le terme reste exclu lorsque sont évoqués les *dépouillements* Henmon, Vander Becke et Gougenheim (Guiraud, 1954 : 93) pour les *vocabulaires de base.*

à-dire par définition, dénaturé (la *nature* d'un corpus ne pouvant être rien d'autre que sa *fonction*). Et *Frantext*, base de textes et documents, connaît désormais deux emplois-types dominants : mine d'exemples attestés et assignés ou norme lexicométrique, avec des bonheurs très divers.

L'explosion de l'offre numérique a fait le reste, de sorte qu'aujourd'hui, dans le champ des études littéraires, il reste d'une nécessité élémentaire de stipuler qu'un *corpus* n'est pas une collection de données que l'on « trouve » (sur Internet ou ailleurs), mais que l'on *élabore* dans une intention particulière de recherche, et qui reçoit de cette élaboration son homogénéité singulière. Tandis qu'on nommera de préférence *base(s)* les ressources numérisées éventuellement hétérogènes d'où l'on peut extraire certains des matériaux d'un *corpus*.

Cette élaboration comporte divers aspects (choix éditoriaux, recueil des variantes, philologie, balisage, reconnaissance linguistique) dont la mise en œuvre est continuellement déterminée par les partis pris de la recherche et remise en jeu par la dynamique de ceux-ci, dynamique permanente où nous situerons d'abord le *sens*.

Le travail sur (dans) le corpus, que permet notamment sa viabilisation informatisée, consiste donc essentiellement à former un *sens* à partir des manipulations qui font alterner vues analytiques et synthétiques sur les *significations*, multipliant ainsi les aspects de ces dernières tout en leur assignant des bornes définies.

S'il existe un champ de l'analyse littéraire, de la lecture critique, il ne peut devoir son autonomie qu'au dépassement du choix classiquement proposé quant au *sens : stable* ou *insaisissable*, entre les pôles duquel la discipline oscille toujours trop violemment, sans normes propres, notamment sous le vocable de *stylistique*. Lorsque Riffaterre (Riffaterre, 1971 : 64-94) propose le recours au *contexte immanent*, il est abusivement (et significativement…) assimilé au Cercle de Prague et à Mukarovski[9], au détriment de la voie qu'il indique pourtant avec vigueur dans un contexte historique contraignant : il n'y a pas de ligne de partage objective entre un *soubassement* et des *faits marqués*, pas plus qu'entre faits de *signification* et faits de *sens*. Autrement dit : le corpus textuel n'a pas une topologie donnée *a priori*, mais c'est la manipulation qui en est faite, qui lui donne forme et sens, qui le *viabilise*.

9. Voir notamment (Ducrot et Schaeffer, 1995 : 158).

En d'autres termes encore, le *sens* peut être réductible à l'*intention critique*[10] si on prend de cette dernière une acception dynamique : frottement aux reliefs qu'elle-même constitue dans le parcours du corpus[11].

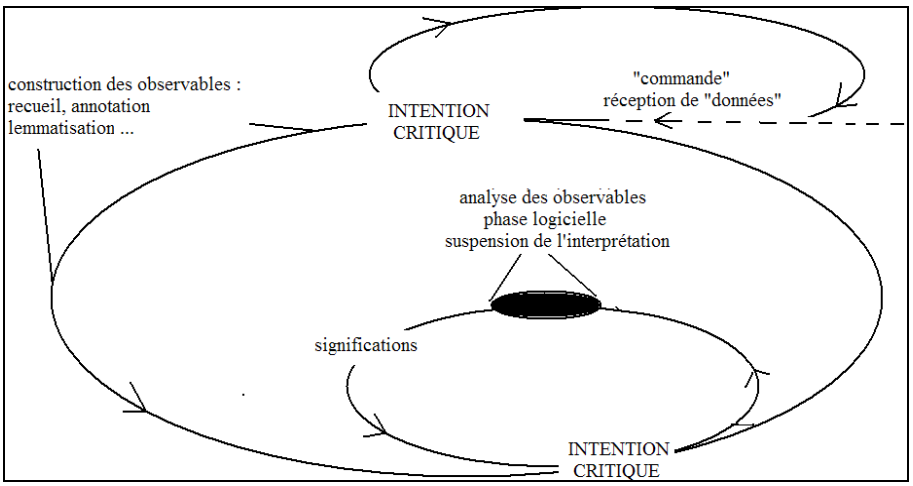

Figure 7.1. *voir note 10.*

10. Cette – trop – brève conclusion renvoie à la – très – vaste discussion, commune aux sciences du langage et de la littérature, sur l'*intentionnalité*. Nous proposons ici que *l'intention critique*, régime haut de l'activité de lecture, dépasse au moins provisoirement la dichotomie searlienne, à notre sens très douteuse en matière verbale, de *l'intention en action* (que Roussin et Schaeffer in Ducrot et Schaeffer (1995 : 85-89) glosent en *celle que [l'auteur] a donnée effectivement*) et de l'*intention préalable* (que les mêmes glosent en *celle que l'auteur a voulu lui donner*). Cette problématique rejoint celle, sans doute plus vaste, de l'*ethos* telle que développée autour notamment des travaux de Jean-Michel Adam (1999 et 2002). Pour ce qui nous concerne immédiatement, le modeste schéma de la figure 7.1 tente de situer l' *intention critique* au lieu de bifurcations entre les différentes tâches de la recherche sur corpus, présentées comme un ensemble potentiellement ininterrompu de boucles récursives. On rappelle ici, très simplement, que l'objectivité des données et de leur traitement est toute relative, qu'elle est en fait *objectivation* : processus doté d'un sujet dont il est intéressant de formaliser la place. Un développement plus substantiel sera accessible dans la publication du colloque *L'Analyse textuelle et comparée des discours* (Université de Lausanne, 6-8 Mai 2004) à paraître en 2005 aux éditions Slatkine sous le titre : *Sciences du discours en dialogue. Textualité & comparaison.*

11. Rastier (1989 : 18) *le texte apparaît comme une série de contraintes qui dessinent des parcours interprétatifs*.

7.3. Profils macrodistributionnels.

Pour illustrer ce propos modestement empirique, nous centrons maintenant l'intérêt sur quelques premières ouvertures offertes par le recours aux statistiques probabilistes[12], sans jamais occulter les problèmes que soulève cette pratique. A commencer par ceci : en tant que telles, les opérations statistiques ne nous donnent pas accès au *sens* plénier[13] (que nous venons de définir), mais à des *significations* de plus ou moins grande portée selon la portée même de la statistique employée. C'est l'oubli de cette distinction qui explique l'improductivité décourageante d'un grand nombre d'études littéraires « assistées par ordinateur ». En outre, toute recherche de signification, la plus atomique qui soit, ne peut se comprendre que dans la saisie simultanée d'une signification globale, l'une nourrissant l'autre.

Clarifions déjà que la signification d'un lexème est le système de ses emplois. Nous pouvons l'asserter quel que soit le corpus considéré et indépendamment de tout autre corpus de *référence*. Prenons l'adjectif JEUNE[14] dans La *Comédie humaine* de Balzac. Voici comment nous pouvons (sans prétendre à l'exhaustivité) amorcer le déploiement de sa signification.

Il présente 6 068 occurrences, dont 4 851 au singulier (dont 32 avec une initiale majuscule) et 1 217 au pluriel (2 avec majuscule). Il est le plus employé des adjectifs, devant même GRAND et PETIT (dans la terminologie de Guiraud, c'est un *mot-thème* de l'œuvre).

Le tableau 7.1 indique comment les occurrences se distribuent ainsi dans les 86 parties (nouvelles et romans).

---

12. Nous restreignons volontairement la présentation qui suit à deux techniques de l'ordre des statistiques probabilistes : l'écart-réduit comme expression valuée de la distribution des items lexicaux aux divers niveaux du texte, et l'Analyse Factorielle des Correspondances (AFC) comme moyen de généraliser cette expression. Chacune de ces techniques est convoquée au titre de représentant d'une classe de techniques équivalentes, parmi lesquelles seule une pratique approfondie des utilisateurs experts pourra faire un choix. L'écart-réduit est défini et développé au mieux par (Muller, 1997b : 69ssqq). Pour une introduction à l'AFC, on pourra consulter (Cibois, 1994) et surtout (Lebart et Salem, 1994).

13. Contrairement à ce qui est suggéré par le fonctionnement des « boîtes noires » d'analyse de données « textuelles ».

14. Le corpus n'est pas à ce jour entièrement lemmatisé. Pour l'adjectif JEUNE, les occurrences des deux formes ont été regroupées. Plus loin, lorsqu'on étudiera les contextes de deux autres vocables (REGARD et REGARDER), ces contextes auront été lemmatisés. Par convention, nous présentons un vocable en PETITES CAPITALES, et des formes occurrentes, en *italiques*.

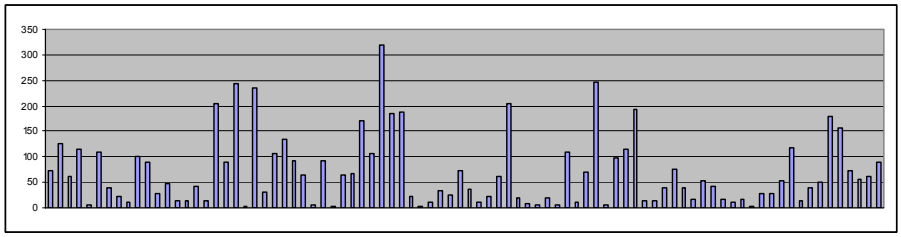

Tableau 7.1. *Distribution (brute) de* JEUNE *dans les 86 parties de La Comédie humaine.*

C'est ce que nous nommerons macro*distribution*. Elle est ici représentée par les *effectifs* bruts (par exemple il y a 318 occurrences de JEUNE dans *Illusions perdues*). Cette présentation est bien sûr très insuffisante, puisqu'elle ne tient pas compte de la longueur, très diverse, des parties concernées : ainsi, *Illusions perdues* étant l'un des romans les plus longs de l'ensemble, il est trivial d'observer qu'un item quelconque y a plus d'occurrences qu'ailleurs. On établira donc une norme, dite d'*équidistribution* : *Illusions perdues* « couvrant » 5,75 % de l'ensemble, l'effectif de référence de JEUNE y est de 6 034 (occurrence totale) x 5,75 %, soit 346,68.

On pourra dès lors employer ces valeurs de référence, pour calculer l'*écart-réduit* qui en sépare les valeurs constatées (« réelles »), et obtenir ainsi (tableau 7.2) une vision plus rationnelle de la distribution.

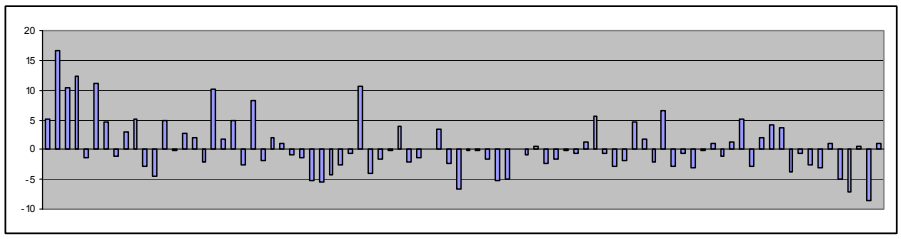

Tableau 7.2. *Distribution de* JEUNE *dans La Comédie humaine (écarts-réduits).*

L'écart-réduit à l'équidistribution est de –1,59 (déficit) pour *Illusions perdues*, qui vient sous cet angle loin derrière d'autres textes, marqués au contraire par le suremploi (excédent) de JEUNE, et dont nous donnons une « tête de liste » dans le tableau 7.3.

Titre	écart-réduit	effectif	référence	(longueur)
Le Bal de Sceaux	16,71	127	32,30	(25774)
La Vendetta	12,38	114	37,96	(30294)
Une Double famille	11,14	109	39,36	(31414)
Le Cabinet des Antiques	10,74	170	76,60	(61130)
La Bourse	10,44	61	17,45	(13924)

Tableau 7.3. *Les textes de CH employant le plus (en écart-réduit) le vocable* JEUNE.

Voilà donc le *profil*[15] *macrodistributionnel* de JEUNE dans *La Comédie humaine*. Ce profil est l'un des constituants de base de la *signification* de cet item, et ce quel que soit le principe qui prévaut pour le partitionnement : si le corpus n'est pas divisé en textes distincts, il peut l'être par tranches de même longueur ; si les parties peuvent être, en rompant ou non l'ordre linéaire de leur disposition première dans le corpus, regroupées en classes selon des principes pertinents (genre, longueur, datation, etc.), cette classification donnera lieu à un profil correspondant. Par exemple, on a regroupé dans le tableau 7.4 les 86 textes en 6 classes de longueurs, avec pour seuils intermédiaires 10 000, 20 000, 30 000, 50 000, 100 000 mots[16].

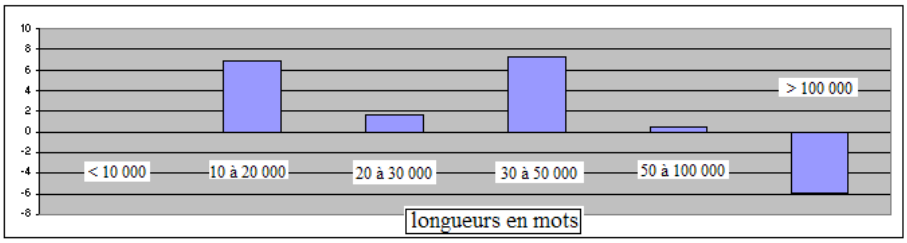

Tableau 7.4. *Distribution de* JEUNE *dans CH (classes de longueurs).*

Le sous-emploi de JEUNE dans les 18 romans de plus de 100 000 mots est *significatif* (écart-réduit –5,98).

15. Nous entendrons par *profil* le système des saillances et le « relief » ainsi formé, tel qu'il peut s'ordonner selon divers critères, pertinents par eux-mêmes (comme la chronologie) ou dans le but d'une comparaison plus ou moins généralisée.
16. Pour la classe des textes les plus courts, la barre est invisible, car l'écart-réduit est très proche de 0 (0,11).

Si nous regroupons (tableau 7.5) les 86 textes en 8 ensembles successifs de dimensions comparables (environ 600 000 occurrences), reflétant grossièrement la diachronie de publication (édition de Furne), le profil est hautement significatif[17].

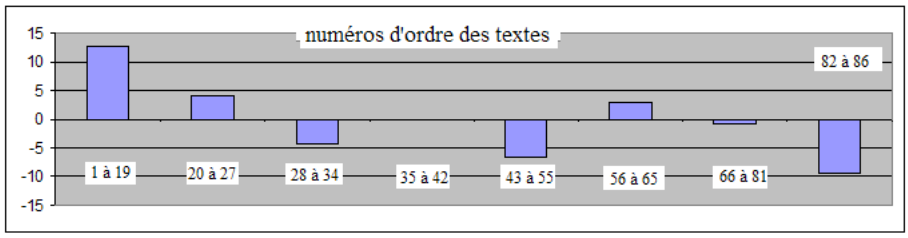

Tableau 7.5. *Distribution de* JEUNE *dans CH (tranches diachroniques).*

Les écarts-réduits sont considérables : de +12,5 à -9,3.

Un certain nombre d'items très importants (par l'effectif) connaissent ainsi une décroissance de leur emploi, qui est la propriété majeure de leurs macroprofils. C'est notamment le cas de *lettre*, de *mère*, d'*amour*, de *salon*, de *fortune*, de *mariage*, de *plaisir*, de *bonheur*, de *cœur*[18]…

A l'inverse, voir (tableau 7.6) le macroprofil « diachronique » de CORPS.

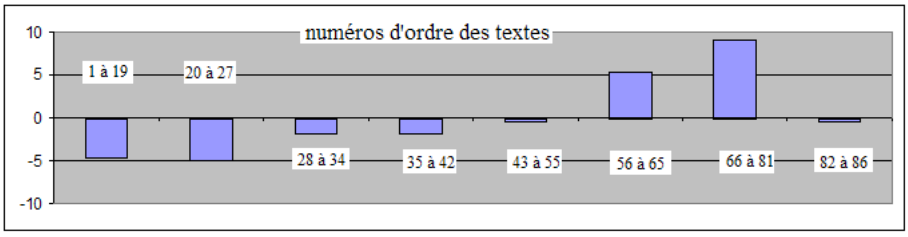

Tableau 7.6. *Distribution de* CORPS *dans CH (tranches diachroniques).*

17. *Significatif* doit être pris dans un sens précis. Pour être rigoureux, la probabilité que cette distribution soit due au hasard est infinitésimale. Un ou plusieurs *facteurs* méritent très probablement d'être recherchés. Mais ces facteurs peuvent néanmoins être triviaux. Il est parfaitement possible par exemple que le sous-emploi de JEUNE dans les 18 romans de plus de 100 000 mots s'explique par le fait que les textes longs sont de plus en plus nombreux sur l'axe diachronique.

18. Ici, on a décompté les occurrences des formes de singulier seulement.

Parmi les items d'effectif comparable, *mort,* TEMPS*, lit,* HOMME*, bois, nature, ciel* présentent un profil analogue.

Insistons sur le fait que ces propriétés font pleinement partie de la signification.

L'analyse statistique multidimensionnelle, et notamment l'Analyse Factorielle des Correspondances (AFC), permet de généraliser ce que nous venons de faire de manière assez intuitive.

Classiquement, l'AFC est employée pour offrir une expression synthétique de la distribution des faits linguistiques discrets (et notamment des items lexicaux) dans des ensembles textuels, diachroniques et/ou contrastifs.

Les méthodes multidimensionnelles comparent et classent les profils distributionnels[19] : elles extraient les parentés et contrastes les plus saillants, par passes successives, qui correspondent aux *facteurs*[20] de contingence.

Ces passes, pour épuiser l'information du tableau, sont aussi nombreuses que ce tableau comporte de dimensions (nombre de colonnes).

Les *facteurs* apparaissent par degrés décroissants de pertinence, de sorte que les résultats peuvent être examinés selon des degrés croissants de finesse (et corrélativement, de « coût »[21]).

19. Les profils par écarts-réduits, qui sont ici développés pour donner la meilleure visualisation de la distribution d'un individu, ne sont pas calculés par l'AFC, qui compare les profils proportionnés à la marge correspondante. Voir notamment (Lebart et Salem, 1994 : 79-87).

20. Cette acception du terme renvoie, sans la recouper exactement, à celle décrite dans la note 17.

21. Moyennant certaines précautions, qui concernent avant tout la dynamique distributive du tableau (considération du chi-2 total et de la décroissance de l'information extraite pour les facteurs successifs, prise en compte des contributions des lignes et des colonnes à chaque facteur), on considérera l'examen du graphe des deux premiers facteurs comme étant de première intention et sa représentation comme la meilleure synthèse de la distribution d'ensemble étudiée. Voir (Lebart et Salem, 1994 ; Cibois, 1994).

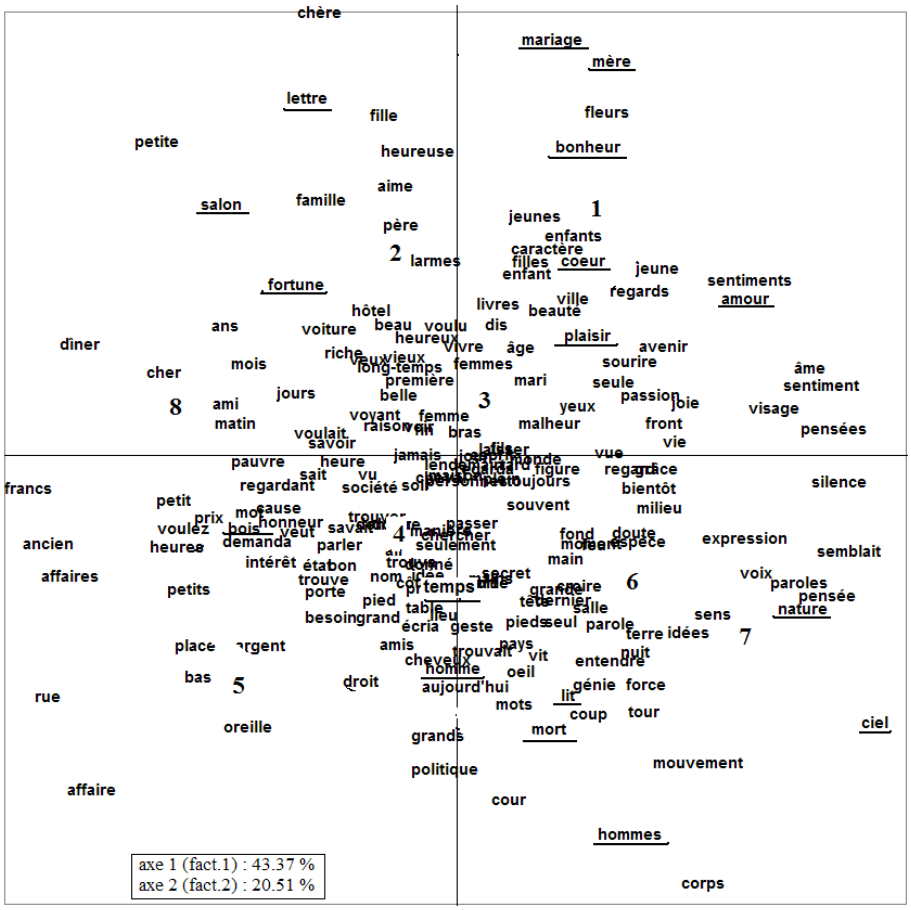

Tableau 7.7. *Macrodistribution de 200 items fréquents dans CH (AFC).*

La projection graphique des nuages de points (lignes et colonnes[22]) offre une synthèse spatiale en 2 dimensions de la distribution d'ensemble telle qu'elle aura été paramétrée. Ce qui est à l'évidence apprécié dans l'AFC par les chercheurs en sciences humaines, c'est d'abord sa capacité à offrir des synthèses, et disons-le : des *vues* synthétiques[23].

22. Les points colonnes seulement s'ils sont distincts des points lignes, ce qui n'est pas le cas des matrices « carrées » de cooccurrence, par exemple – ou de contiguïté (Lebart et Salem, 1994 : 200-207) – qui présentent un axe diagonal de symétrie.

23. Ce qui ne nous dispense, ni de la précaution d'examiner sérieusement les résultats, ni du scrupule de publier l'ensemble de ceux-ci, en annexe de nos travaux, notamment en ligne.

Ainsi retrouvons-nous, sur le graphique présenté en 7.7, certaines des observations déjà effectuées de manière plus manuelle, précédemment.

Sur ce graphe, les proximités entre points expriment, dans la dynamique de l'ensemble des contraintes, très complexes, des distributions particulières, les parentés de profils les plus significatives ; les points numérotés figurent les 8 tranches diachroniques du corpus, et leurs profils (qui croisent bien sûr les profils des items) les amènent à des positions de voisinage qui peuvent leur permettre de rendre compte, pour des items en position saillante, de la tendance générale de leur profil qui les a amenés là. On repère ici trois zones assez distinctes, correspondant aux 2 premières séries (textes 1 à 22), aux 3 suivantes (textes 23 à 55), et aux trois dernières (textes 56 à 86). Il s'agit probablement d'une évolution thématique, et probablement reconnaissable par un lecteur assidu de Balzac.

La position d'un item lexical sur ce graphique, sous réserve d'un examen plus approfondi et détaillé, est déjà une indication utile sur sa macrodistribution, donc sur un aspect de sa signification. Celle-ci étant replacée dans une vue aussi synthétique que possible, elle en acquiert une portée accrue. Surtout, vu sous un autre angle, c'est de la signification globale du système que ce graphique nous offre un aperçu commode.

7.4. Profils microdistributionnels

A côté des ces macroprofils, et sans doute plus universellement significatif encore, le profilage *microdistributionnel*. Nous ne considérons plus des segments continus, relativement longs, et représentatifs d'une macrostructure, mais des segments finement discontinus. Mesurons la distribution de *fortune* dans les 200 ensembles formés par les contextes étroits de 200 items parmi les plus fréquents du corpus[24]. En d'autres termes, recensons les cooccurrences de *fortune* avec ces 200 items, mais conservons à l'esprit l'idée fondamentale de *contexte discontinu*.

Cette problématique rejoint ce que nous développons dans la dernière partie de ce chapitre. Les travaux évoqués ici sur le corpus de Balzac sont en ligne, accessibles à partir du site de la Maison des Sciences de l'Homme de Franche-Comté, http : //msh.univ-fcomte.fr

24. Pour toute la suite, sauf mention contraire, les paramètres sont : 15 mots à gauche et à droite de chaque occurrence du pivot, dans les limites d'une même phrase (définie par la ponctuation).

Comme ces contextes sont constitués de phrases, ou d'empans verbaux autour des pivots que sont les 200 items *colloqués*, ils ont une longueur, comme les macrosegments. On peut donc y inférer une valeur de référence à laquelle comparer l'occurrence effective de *fortune*, et ainsi établir un profil *microdistributionnel* en termes d'écarts-réduits. Ces écarts-réduits sont les indices de cooccurrence, ici (tableau 7.8) dans les limites de la phrase, de *fortune* avec les 200 items visés.

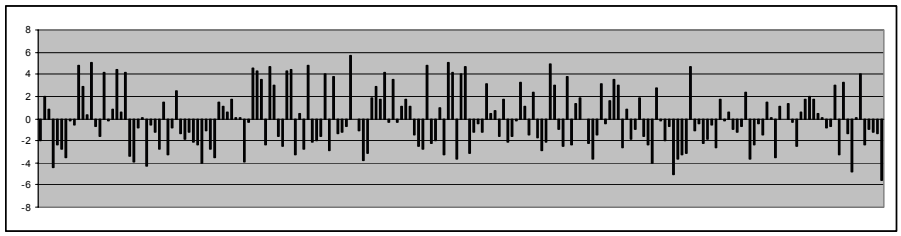

Tableau 7.8. *Cooccurrence de fortune avec 200 items fréquents dans CH (écarts-réduits, classé par ordre alphabétique).*

Il peut sembler absurde de « présenter » un tel profil s'il est classé, comme c'est le cas dans nos matrices de cooccurrence, par ordre alphabétique. Les éléments de pertinence (barres saillantes en excédent ou en déficit) n'y sont repérables qu'individuellement, en désordre, et on préférera disposer d'abord de la liste des items triés par indice décroissant (tableau 7.9).

cooccurrent	francs	enfants	nom	grande	livres	famille	avenir	honneur	laisser	riche	fils
écart-réduit	12,48	11,28	9,41	9,28	8,69	6,56	6,25	5,34	5,08	5,04	4,64
cooc. Brute	101	65	51	57	30	44	26	24	26	27	42

Tableau 7.9. *Les plus forts cooccurrents de fortune parmi 200 items fréquents dans CH.*

On y verra cependant beaucoup plus clair si l'on envisage de nouveau la véritable utilité de ces profils, qui consiste à les comparer entre eux (pourvu qu'ils soient établis selon les mêmes critères). Il ne sert à rien de mettre vis-à-vis du précédent un second profil microdistributionnel (classé alphabétiquement), qui concerne cette fois *visage* (tableau 7.10).

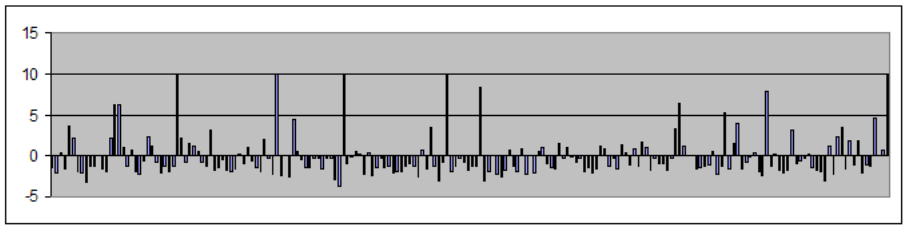

Tableau 7.10. *Cooccurrence de visage avec 200 items fréquents dans CH (écarts-réduits, classé par ordre alphabétique).*

Cependant, si nous n'ordonnons plus la série dans le trivial ordre alphabétique, mais d'après l'indice de cooccurrence (croissant) avec *fortune*, alors la comparaison peut s'exercer mieux : tout d'abord, ce qu'est devenu par construction le profil de *fortune* (tableau 7.11).

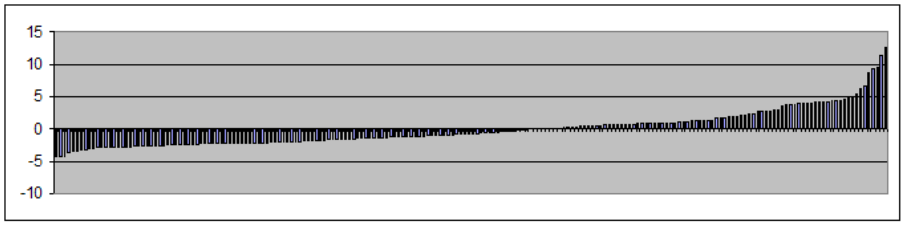

Tableau 7.11. *Cooccurrence de fortune avec 200 items fréquents dans CH (écarts-réduits, croissants).*

Et dans ces nouvelles conditions, celui de *visage* (tableau 7.12).

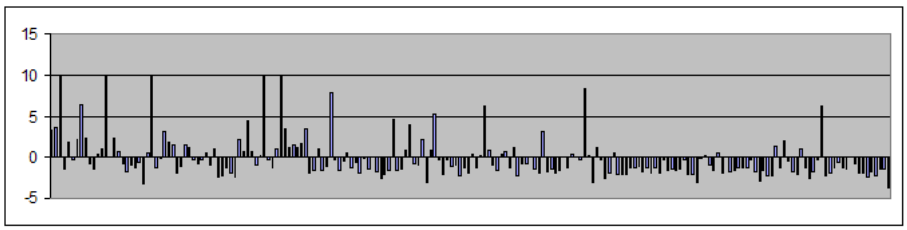

Tableau 7.12. *Cooccurrence de visage avec 200 items fréquents dans CH (classés dans le même ordre que supra).*

Le profil de *visage* est grossièrement à l'envers de celui de *fortune*.

En d'autres termes, ses collocations excédentaires tendent à être les collocations déficitaires de *fortune*.

Il ne s'agit pas d'un fait d'*antonymie*, bien au contraire : deux *antonymes* s'il en existe auront des profils collocatifs voisins, manifestant l'existence de leur invariant sémantique.

Ces deux-ci sont au contraire des items qui relèvent de secteurs sémantiques distincts dans la structure du vocabulaire du corpus à l'étude.

Ces « secteurs », que nous nommerons des *isotropies*[25] (*trepein* : incliner à), ne peuvent pas être mieux mis en évidence que par l'AFC.

Il serait en effet absurde d'inviter le lecteur à procéder aux multiples comparaisons intuitives qu'exigerait l'exploration d'un tel corpus.

Il s'agit au contraire de donner corps à la notion de *profil microdistributionnel*, comme instant décisif de la *signification*, et ceci dans le cadre nécessaire de la confrontation des profils, qui donne accès à la signification d'ensemble que, rappelons-le, nous visons pour l'essentiel dans cet exposé.

On trouvera donc en 7.13 le graphique qui découle de la confrontation généralisée des 200 microprofils évoqués *supra*, par l'AFC.

25. (Viprey, 1997 : 154-157).

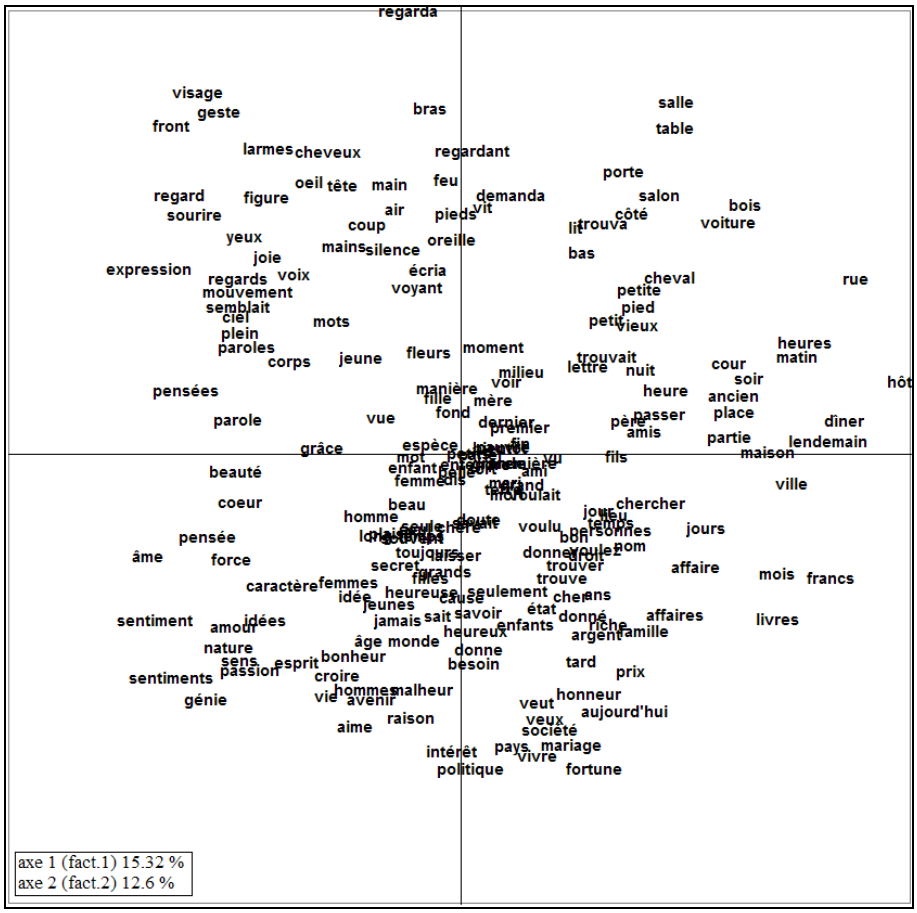

Tableau 7.13. *Microdistribution de 200 items fréquents dans CH (AFC).*

On y repère *fortune* (en bas) et *visage* (en haut et à gauche[26]), dont les positions respectives sont dues (massivement, bien sûr) à leur opposition de profil. On y repère surtout, dans un continuum qui permet de (et oblige à) respecter la nature très fine et polydimensionnelle de l'organisation textuelle, des aspects difficiles à percevoir synthétiquement sans cela, de la structuration propre du vocabulaire de *ce* corpus.

26. On aura compris que le choix de ces deux items (comme *supra* de JEUNE, etc.) n'était pas fortuit.

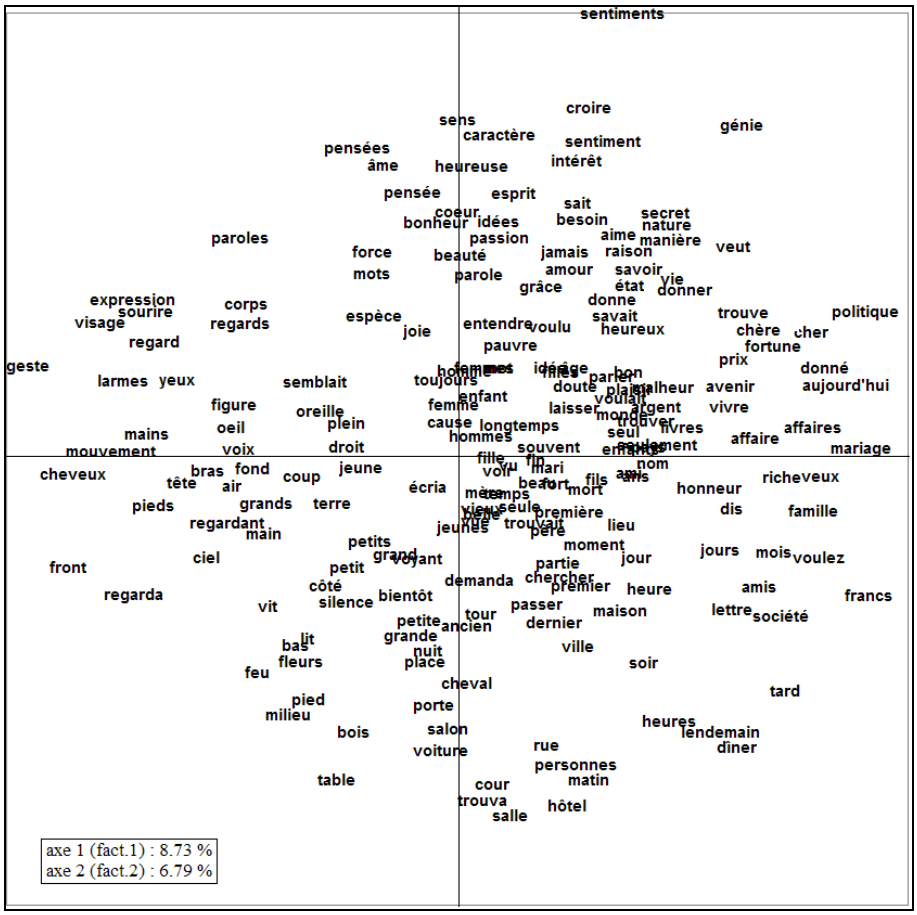

Tableau 7.14. *Microdistribution dans l'œuvre narratif de Maupassant, des 200 items les plus fréquents dans CH (AFC).*

Aussi importantes sont les proximités que les saillances. Ainsi, si l'on repère en bas à droite des rapprochements jugés hâtivement « triviaux », on aura intérêt à se demander si les mêmes proximités seraient extraites par cette procédure d'un autre corpus, d'une part ; et surtout d'autre part si elles occuperaient une position saillante (à la périphérie du graphe, là où en AFC les positions sont *significatives*).

Si l'on applique exactement la même procédure, concernant les mêmes items[27], au corpus narratif de Maupassant, on y observe facilement un très significatif invariant (tableau 7.14).

Pour rendre compte de cet invariant, nous aurons « en pratique » recours à une technique qui ne peut être présentée intégralement ici, impossibilité qui va d'ailleurs nous amener vers un autre aspect de notre problématique (voir section 7.6). Dans un environnement d'hypertexte expert, sur ordinateur, nous proposerions de colorier de couleurs distinctes les items du graphe Maupassant, selon le secteur où ils se trouvent dans le graphe Balzac. Nous pouvons tout juste ici (tableau 7.15) utiliser un code : les noms des points (les items) sont remplacés par des signes conventionnels. C'est la distribution dans Maupassant qui sert de référence.

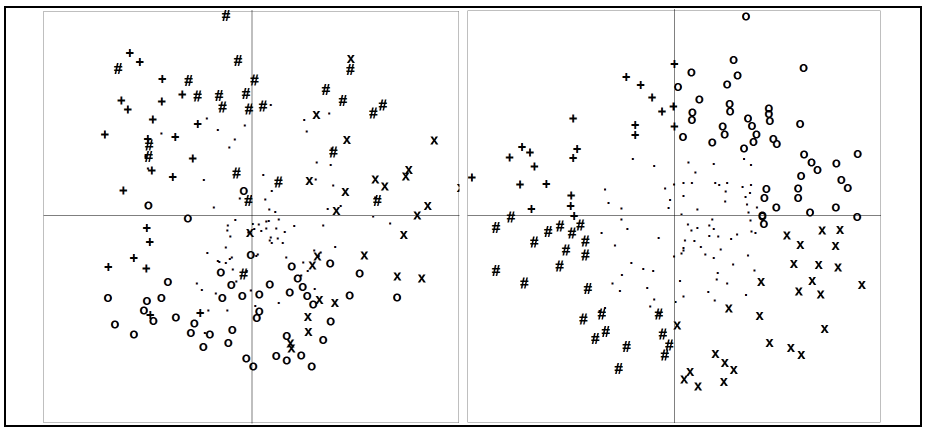

Tableau 7.15. *Microdistribution comparées de 200 items dans CH (à gauche) et dans l'œuvre narratif de Maupassant (à droite) (AFC).*

Un fort invariant s'impose[28], qui signale à la fois une parenté lexico-thématique forte, liée à des contraintes génériques (discursives), et la robustesse du modèle

27. Ce qui implique déjà une sensible différence, puisque les items qui occupent les 200 premiers rangs dans *La Comédie humaine* n'occupent pas ces mêmes rangs dans l'œuvre de Maupassant…

28. L'orientation sur les axes est moins importante ici que le degré de conservation des groupements en périphérie, même s'ils changent de place. Il ne s'agit bien sûr que d'une grossière observation. Des techniques plus fines restent à mettre au point.

probabiliste pour cette classe d'observations[29]. Mais les éléments éventuels de la variation, qui sont bien aussi importants du point de vue de la signification, ne peuvent être récupérés qu'en un second temps ; dans l'hypertexte expert, ils seront marqués et cliquables.

7.5. Structuration thématique endogène

Voyons maintenant ce que suggère la cartographie AFC en rapport avec la pratique, longuement établie, de l'approche thématique en littérature, constituant massif de la sémantique dans ce domaine.

On pourrait se contenter de noter que cette pratique est *grosso modo* naïve dans son ensemble, d'après les arguments suivants : (1) le choix du « thème » à étudier est généralement arbitraire et intuitif ; (2) la configuration de ce « thème » l'est plus encore, reposant sur des associations lexicales *a priori*, non critiquées ; (3) le relevé du « thème » présente à nouveau les mêmes caractères de naïveté, exacerbés par la dimension des corpus.

Cette critique, indispensable, doit néanmoins être modulée, si l'on veut éviter d'entrer dans une perspective purement descriptive et techniciste, qui n'a aucune chance de saisir ou de créer du *sens* dans la pratique critique des textes. La modulation ne peut être apportée que par l'*intention critique* que nous avons cernée *supra*. Dans la pratique critique exercée, l'intuition du spécialiste devient un constituant de son expertise. Ainsi doit-il savoir reconnaître, prudemment, la pertinence de certaines saillances d'une carte AFC, qui montent du corpus, avec lesquelles l'hypertexte expert lui permet d'instaurer un dialogue fécond, car encadré.

Construire un thème ne saurait en tout cas consister à appliquer une liste toute prête d'items. Par exemple, le thème du REGARD dans *La Comédie humaine*. Si l'on admet que le REGARD n'est pas la VUE, pas plus que les YEUX[30], etc., le plus sûr semblerait de commencer par recueillir les contextes du seul vocable REGARD, qui présente au total 2 790 occurrences (1 945 au singulier, 845 au pluriel). On aura bien sûr repéré, *supra* (tableau 7.13), que les deux formes *regard* et *regards* pointent

29. On repère là l'une des frontières méthodologiques de ce mode d'exploration : faute de connaître, et d'estimer, la pondération de ces deux facteurs d'invariance, on oubliera les risques d'artefact liés au modèle, et l'on tombera dans le plus plat positivisme.
30. Ce qui n'est rien moins qu'assuré : tant d'analyses thématiques considèrent de tels abus comme des évidences premières !

clairement dans un groupe isotropique, en haut à gauche, assez suggestif. Mais comment obtenir une vue plus précise de la manière dont s'organise l'amorce de ce thème particulier ?

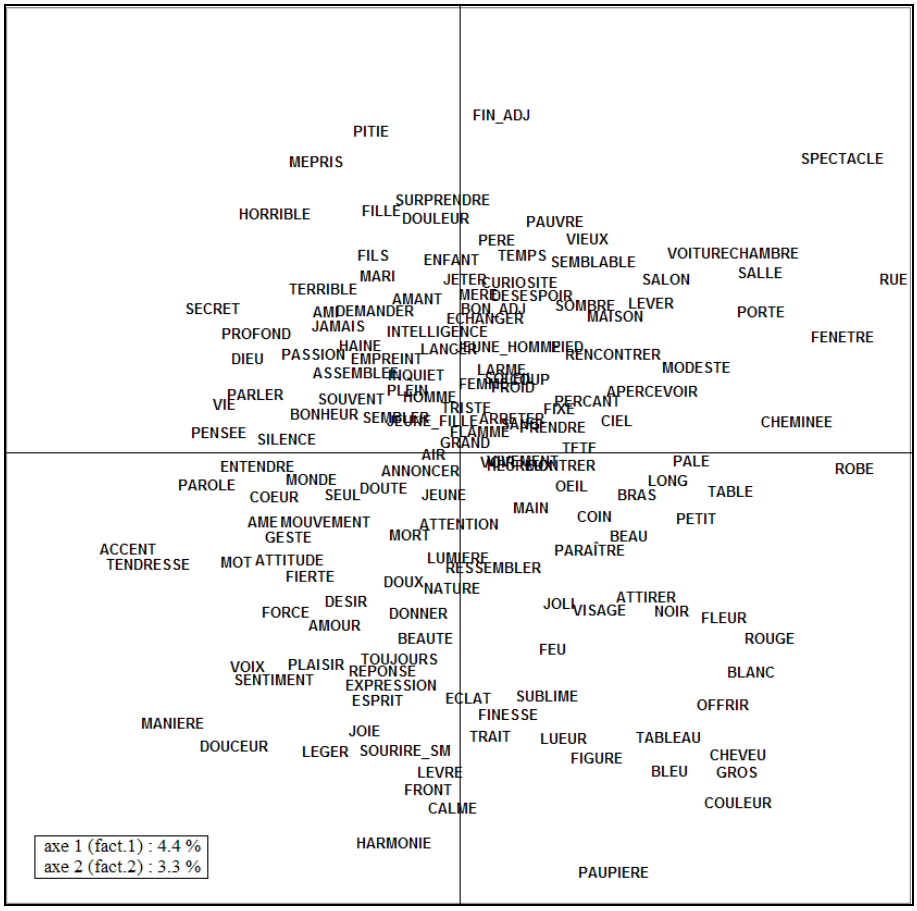

Tableau 7.16. *Microdistribution de 159 vocables dans les contextes du vocable* REGARD.

Le principe de recueil des données sera le suivant : relever les 2 790 contextes suivant un paramètre déterminé et constant (par exemple les 2753 phrases[31], ou encore 2 790 fois *n* mots autour du pivot, à gauche et à droite, ou à gauche, ou à droite). Opérons une expérimentation sur 10 mots à droite et à gauche. Le sous-

31. Et non 2 790 en raison d'occurrences multiples dans une même phrase.

corpus a été lemmatisé dans l'environnement DiaTag[32]. On établit une liste des plus fréquents vocables (159 items) de ce sous-corpus, dans les classes des substantifs, adjectifs, verbes et adverbes circonstants. Dans chacun des 2 790 segments ainsi constitués, on en relève toutes les cooccurrences internes

Le tableau 7.16 présente le graphe des deux premiers facteurs de cette matrice de cooccurrence par l'AFC.

Il s'agit d'une vue inédite sur le thème, non triviale : l'intuition peut s'y reconnaître en partie, mais s'y voir aussi prise en défaut et surtout corrigée. Cette synthèse polarise assez nettement, sur l'axe 1 – horizontal, l'opposition habituellement posée comme évidence première entre *concret* (à droite) et *abstrait*. Sur l'axe 2 – vertical, on repère au moins, à droite, deux ordres contrastés de « visibles », et à gauche deux modalités de « climats ». Interpréter ces polarisations (ces *isotropies*) ne saurait être l'objet de ce chapitre : au contraire, ce qu'elles suggèrent amènera le lecteur expert à explorer plus avant les cotextes pertinents (notamment, grâce à l'interactivité de la carte, en collectant vocables et listes de vocables) et en se procurant d'autres synthèses complémentaires. Nommer aussitôt ces zones du vocabulaire, par des catégories exogènes, du type de celles que l'on emploie dans les thesaurus, ce serait effectuer de manière très prématurée l'opération de *réduction* propre à tout discours critique et, en ce sens, certes nécessaire à terme,

Voyons maintenant ce qu'il résulte d'une éventuelle extension du thème, par exemple en associant REGARDER à REGARD. Ces deux termes font après tout l'objet d'une entrée commune dans la plupart des dictionnaires papier, et REGARD y a pour première définition *action de regarder*. Notons cependant qu'une simple connaissance intuitive du lexique permet de savoir (au moins) que :

(1) le *regard* n'est pas toujours *l'action de regarder*, mais très souvent *l'expression des yeux*, c'est-à-dire *du visage* (jusque dans l'étymologie de ce terme)

(2) *regard* peut selon les cas (2a) être paraphrasé par *expression*, par *sourire*, par *figure*, dans d'autres cas (2b) pas du tout ; dans d'autres encore, c'est un verbe qui bloque plus (2c) ou moins (2d, 2e) la paraphrase :

2a : Anastasie comprit le regard de Monsieur de Trailles

32. Développé à l'Université de Franche-Comté pour l'étiquetage lexico-morpho-flexionnel assisté des corpus (Viprey, 2004).

2b : Il voyait, à travers le cachemire, les teintes rosées du corsage que le peignoir, légèrement entr'ouvert, laissait parfois à nu, et sur lequel son regard s'étalait.

2c : Mademoiselle Taillefer coula timidement un regard sur le jeune étudiant.

2d : […] Il jeta sur les deux interlocuteurs un regard lumineux et plein d'inquiétude

2e : La jeune femme […] lui lança un de ces regards froidement interrogatifs qui disent si bien […]

La cartographie AFC peut-elle aider le lecteur de corpus à distinguer les grandes disjonctions sémantiques sous-jacentes ?

Nous supposons que, si le texte a été lemmatisé, les 6 115 occurrences de REGARDER et de REGARD n'ont pas été désambiguïsées selon les diverses disjonctions envisageables (à l'exception de composés connus de notre dictionnaire, comme *en regard de*). Faire monter les disjonctions du corpus, plutôt que de lui appliquer une catégorisation *a priori*, aussi savante soit-elle, c'est précisément ce que nous attendons de la statistique multidimensionnelle.

Le couplage REGARD/REGARDER nous fournit justement l'occasion d'un test exigeant. Le tableau de cooccurrence constitué suivant les mêmes règles que ci-dessus pour REGARD, livre le graphe 7.17, sur lequel nous avons repéré en outre les vocables se trouvant significativement plus en contexte avec REGARDER qu'avec l'ensemble REGARDER/REGARD (écarts-réduits supérieurs à 1.5) – en gras.

Le test paraît assez probant : un important invariant s'affirme, et une disjonction assez claire se confirme à la combinaison des deux premiers facteurs (en haut et à droite). Certes elle n'est pas dichotomique ; des items non spécifiques de REGARDER viennent se mêler aux autres. C'est précisément une vertu de l'AFC de ne pas dichotomiser, mais d'offrir un continuum suggestif.

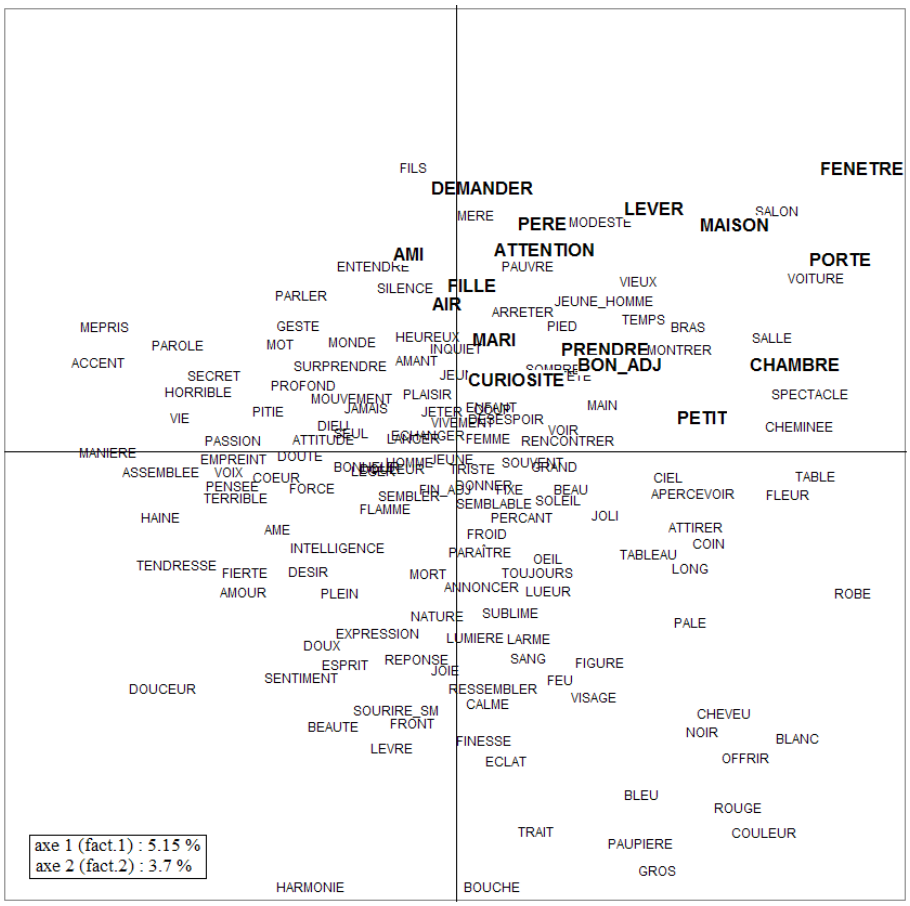

Tableau 7.17. *Microdistribution de 159 vocables dans les contextes des vocables* REGARD *et* REGARDER.

Si nous revenons brièvement à la carte antérieure (7.16), nous pouvons à la lumière de ce test passer un cran du travail d'interprétation. Parmi les items du pôle inférieur droit, le verbe OFFRIR est particulier. Sa concordance (70 occurrences en contexte avec REGARD dans un empan de 10 mots à droite et à gauche) indique 51 emplois du type OFFRIR au(x) REGARD(S), la plupart des 19 restants ayant la même signification, mais plus ou moins implicitement construite. Son voisinage avec TABLEAU est intéressant ; il indique la sollicitation d'un *regard* esthétisé ou artistique. Or, en 1.17, la zone *isotropique* d'OFFRIR partage la position des cooccurrents forts de REGARDER sur le 1er facteur, puis s'y oppose sur le second.

Une fois encore, on ne peut présenter sur papier qu'une vue schématisée, alors que l'hypertexte expert offre des items cliquables. La disposition spatiale sur le graphe répond des parentés et contrastes de profils (micro) distributionnels saisis dans la globalité de l'espace des réseaux tel qu'il est comptabilisé dans le tableau de contingence, tandis que l'éclairement rend compte d'une hiérarchie de parentés.

L'important est ici de noter que l'approche est susceptible de concerner non seulement les cas où des disjonctions sont attendues en raison des connaissances préalables, mais tous les items et groupes d'items de tout niveau (pourvu que ce niveau ait été exploré et codé). Ces graphes dynamiques (constituables en temps réel) pourraient devenir de véritables rubriques pour des dictionnaires de corpus, fonctionnant comme synthèses du *champ sémantique* tel que tentent de le décrire les *exemples* des dictionnaires papier : chaque item du graphe étant cliquable, il donnera accès à une liste structurée (concordance et contextes élargis) des collocations qu'il recouvre.

Un autre aspect essentiel de la sémantique de corpus en littérature ne peut être ici qu'évoqué : il concerne l'étude de la signification dont peuvent se charger les unités infra-morphématiques, notamment la syllabation, les phonèmes et les mannequins phonétiques[33], dans le cadre de la prosodie et/ou des allitérations. Nous renvoyons pour la théorisation à un article et à un ouvrage de l'auteur[34]. En résumé, la statistique textuelle permet d'aborder rationnellement de telles configurations, dont l'intuition et le relevé manuels ne sont envisageables qu'à des échelles très réduites.

7.6. Corpus et résultats d'analyse.

Le problème « technique » soulevé à plusieurs reprises à propos de la présentation des graphiques reflète celui, fondamental, du statut des résultats. L'exposition de cette problématique devrait permettre une claire compréhension de ce qui est en jeu sous le concept « lire les corpus ».

Jusqu'à une date récente, la plupart des environnements d'analyse statistique pour les corpus présentaient leurs résultats comme des sortes de culs-de-sac. Le *retour au texte*, condition *sine qua non* du sens de la mise en œuvre de telles procédures, était d'une pratique plus que fastidieuse, et demeura longtemps plus un

33. ou *patrons*. Groupements significativement récurrents de phonèmes, notamment consonantiques (Viprey, 2002).
34. Viprey (Viprey, 2000 et 2002).

slogan qu'une réalité ; en tout cas, il n'était réglé que par la mémoire évanescente de certaines vues. En outre, faire des comparaisons entre graphes, y repérer l'impact d'un constituant, y intégrer des éléments d'une recherche antérieure, n'était pas même envisageable.

De manière générale, tous ces travaux s'effectuaient d'une manière brutalement discontinue. C'est, à notre sens, ce qui a donné parfois aux recherches sur corpus « assistées par ordinateur » l'allure d'un marteau-pilon destiné à disséquer une mouche, et aussi ce qui a découragé de très nombreux chercheurs, qui n'y voyaient plus un rapport de *sens* (rendement) suffisant. On comprend aussi comment, à la réception de résultats auxquels il était difficile de venir articuler d'autres calculs, mais aussi de confronter synthétiquement des contextes, ont pu se développer des tendances à l'interprétation « sauvage », d'une part, à la démission critique face aux boîtes noires, de l'autre.

Une réponse positive à ces difficultés consiste au contraire à intégrer les techniques de recherche au corpus lui-même, à faire des résultats en cours l'appareil privilégié de navigation au sein du corpus de manière à y dessiner des parcours critiques seuls à même d'en développer le *sens*.

Une des notions centrales d'un tel dispositif est d'associer au corpus numérisé un thesaurus « sémantique » dynamique, *endogène*, fait de listes et de tableaux, bien sûr (de *dictionnaires* en un sens assez classique), mais aussi de graphes qui fonctionnent comme un atlas dans un hypertexte, c'est-à-dire cliquables selon des modalités aussi diverses que l'exige la logique d'exploration. Rien, ou le moins possible, de préconstitué, sauf s'il s'agit d'importer, aux fins des comparaisons et des convergences les plus sensées, des éléments de thesaurus eux-mêmes endogènes construits dans l'exploration de corpus dont on cherche à établir la convergence et la divergence.

Ainsi se justifie le programme de développement de l' *hypertexte critique* (ou *expert*), qui systématise ces propositions à partir de l'état de l'art. Il doit s'agir d'un nouveau mode d'exposition des discours critiques, qui offre tous les moyens au lecteur de reproduire les expériences décrites, de les intégrer, et de les généraliser. Quand nous disons : reproduire les expériences décrites, cela ne signifie en aucune manière que toute l'expérience critique soit reproductible (c'est-à-dire réductible, c'est-à-dire encore transmissible) ; nous visons ce qui, de l'expérience, aura été formalisé dans le parcours objectif du corpus, notamment grâce à une sémantique libérée des *a priori* des catégories de pensée.

L'*hypertexte critique*[35] est un environnement, dédié aux corpus de taille moyenne (de quelques milliers à quelques dizaines de millions de mots), qui les prend en charge à l'état « brut » (au format .txt[36]), et autorise leur enrichissement processuel, à commencer par leur étiquetage formel (forme graphique, lemme, classe grammaticale, flexion) dans un cadre fortement conventionnel et échangeable. C'est sur cette base que le moteur d'exploration vient s'accoupler, à l'état embryonnaire (un logiciel de fonctions : affichage, calculs, dialogues), et se développer dans le cours même de l'activité de recherche : annotation multidimensionnelle (constituée de commentaires discursifs et de liens), cartographie, dictionnaire dynamique ; l'historique de cette navigation est l'un des modes privilégiés de lecture du discours critique, sa « trace » la plus cohérente (mais jamais la seule ni surtout sans jamais avoir à être suivie linéairement). Autre avantage : une formalisation ainsi entendue limite le risque d'abus de formalisme, de descriptivisme gratuit, puisqu'elle expose en permanence la gestuelle de l'expert.

Elle propose une alternative à la connaissance *a priori* des rapports sémantiques « en langue » (thesaurus préconstitués), dont une linguistique textuelle proprement dite ne peut que révoquer en doute la pertinence. Elle peut libérer en outre par là l'*isotopie* de ses mésinterprétations dans la vulgate stylistique et lexicologique[37].

---

35. *Hyperbase*, développé par Etienne Brunet à Nice, a été présenté à de nombreuses reprises lors des sessions successives des *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT)*, dont on trouvera les références dans la bibliographie. Depuis 1997, nous développons à Besançon un environnement tel que celui décrit ici, dans l'équipe *Recherches Informatisées en Sciences des Textes (RIST)* du LASELDI (Laboratoire de Sémiolinguistique, Didactique, Informatique, EA 2281 du MENRT). Les deux environnements prototypés, ®DiaTag et ®Astartex, ont été présentés à travers certaines de leurs applications, lors des *JADT* 1998, 2000, 2002 et 2004, et lors des colloques annuels du réseau international du logiciel ®Intex (Silberztein, 1993), les *Journées Intex* 2000, 2002 et 2004 (voir en bibliographie). Les environnements ®SATO (Daoust : http : //www.ling.uqam.ca/sato), développé par le Centre ATO de l'UQAM et ®NEURONAV (Lelu, 1998), diffusé par la firme Diatopie, comportent plusieurs fonctionnalités typiques de ce que nous entendons ici, pour n'évoquer que le domaine francophone. Le développement de ces méthodes et des environnements adéquats (®Intex, ®DiaTag, ®Astartex) constitue le cœur du programme affecté au pôle *Archive, Bases, Corpus* (coordonnateurs : Pr. Claude Condé, Jean-Marie Viprey) lors de l'affiliation de la Maison des Sciences de l'Homme de Franche-Comté.

36. Ou encore, dans les formats HTML, XML, etc., c'est-à-dire des formats où l'énoncé est disjoint de tous les commentaires, qu'ils concernent ou non la mise en forme graphique.

37. Voir la critique de cette notion, et surtout de ses mésemplois, dans (Viprey, 1997). En substance, la notion d'*isotopie* (Greimas, 1986) pose en tant que telle de redoutables problèmes de formalisation. Vulgarisée dans le domaine des études de textes, elle devient trop souvent le synonyme confus de *champ notionnel*, lieu de la projection la plus incontrôlable

La pratique critique des corpus devrait y gagner l'autonomie qui lui est nécessaire vis-à-vis du TALN/IA. Aux impératifs de la traduction automatique, qui gravitent autour de la *réduction* du *sens* (par éviction de tout ce qui relève de l'*ambiguïté*) à une *signification* univoque extralinguistique, et de la négation de la surface textuelle, s'opposent diamétralement ceux de l'interprétation critique. Ce n'est pas pour autant que ces derniers seraient indifférents à la formalisation, et non explicitables, au contraire[38]. Mais la formalisation herméneutique est un incessant déploiement, une recherche *de et parmi* l'ambiguïté.

MISE EN GARDE.– Les limites de cette présentation interdisent qu'y soient traités à fond trois problèmes essentiels, liés à la problématique du *sens en corpus*, dont il ne faut pas pour autant déduire qu'ils passent inaperçus. Le premier est celui de l'interdiscours et des sources exogènes de la signification. L'hypertexte expert n'a pas pour vocation de clore le corpus sur lui-même. Le geste de clôture relève de l'*intention critique ;* à ce titre il est prudent, provisoire et constamment alternant avec le geste comparatiste. La mise en regard de 2 ou N corpus, qui appartient de droit à la lecture scientifique de chacun de ces corpus, est certainement nécessaire à l'approfondissement de leur signification. Non seulement les mêmes calculs doivent être menés sur des corpus à comparer, mais les résultats de l'un doivent éclairer, orienter l'exploration de l'autre (nous pensons notamment ici à l'éclairement des graphiques d'AFC microdistributionnelle, sur l'exemple donné de Balzac à Maupassant). De façon plus générale, la signification d'un item ne se réduit pas à son emploi dans un corpus circonscrit.

Cela soulève le second problème, celui de l'articulation du lexique et du vocabulaire (et au-delà, de la langue et du discours). Que se passe-t-il quand on collationne des dizaines de grands corpus d'auteurs littéraires français du XIX^e^ siècle ? Que se passe-t-il, d'autre et/ou de semblable, si ce sont les centaines de millions de mots de vingt ans du journal *Le Monde* ? Que se passe-t-il enfin lorsque D. Biber[39] réunit le corpus qui lui est nécessaire pour produire une grammaire de l'anglais parlé ? Nous avons tenté, pour circonscrire notre sujet et le faire entrer dans ce livre, d'esquisser une distinction entre *corpus de langue* et *corpus spécialisé*. On peut d'abord noter qu'en réalité, les mêmes collections de données peuvent entrer dans l'une ou l'autre catégorie, selon le point de vue qui préside à leur analyse :

---

des représentations sémantiques *a priori*, non explicitées, non formalisées, non contrôlées, de chaque analyste.

38. Ni d'ailleurs qu'ils ne soient conscients d'opérer des *réductions*.

39. On trouvera l'ensemble de la bibliographie de D. Biber à la fin de (Biber, 2004).

classification opératoire. Plus au fond, non seulement comme l'écrit Saussure, très utilement cité et commenté par Adam (Adam, 1999 : 23-27), *la langue n'est créée qu'en vue du discours*, mais encore nous n'entendons radicalement par *langue* rien d'autre que le système verbal qui se manifeste dans *un discours* tel qu'il s'engrène dans *le discours* universel (c'est l'option de Bakhtine[40]). Sous cet angle, les corpus *littéraires* sont emblématiques. La littérature est un *discours* non pas « comme un autre », mais parmi d'autres, constitué comme tel socialement (et non par un décret divin). Le même corpus (Balzac), relève de plusieurs sphères de discours et, comme tel, de plusieurs disciplines plus ou moins convergentes. L'analyse littéraire, ou plus exactement textuelle (mieux que *stylistique* ou *poétique*), rejoint l'histoire littéraire moderne pour considérer qu'une *langue* particulière, individualisée, y est à l'œuvre. Le recours à de tels corpus dans le cadre de tentatives pour cerner « le » français d'une synchronie, et sa sémantique, en est-il pour autant inutile ? Oui, si et seulement si l'on oublie la place qu'occupe l'activité littéraire dans l'activité langagière.

La très grande récurrence des énoncés littéraires dans l'activité du locutorat, explique certainement bien autant la « polysémie » que l'on prête à la littérature, que tout paramètre interne, structurel, de ces énoncés. L'analyse sémantique de ces discours, qui ne se réduit certes pas aux formalismes statistiques, peut aider à la modélisation d'une herméneutique ouverte, dialogique. Nous présentons le programme scientifique et technique d'hypertexte expert comme une alternative à la description lexicographique des disjonctions sémantiques, nécessairement dichotomique et réductrice. La description dynamique du vocabulaire (lexique du corpus) compte parmi les moins vaines opérations en quoi consisterait la *lecture des corpus*.

Ce qui nous amène à la troisième question : quelles sont les unités sémantiques dont nous parlons et sur lesquelles nous travaillons ? Existe-t-il aujourd'hui une théorie de la segmentation qui, dépassant le fonctionnalisme, puisse venir dialoguer avec les contraintes amont et aval de l'étiquetage des corpus précieux[41] ? De ce point de vue, les recherches probabilistes sont loin d'avoir atteint leur rendement maximal. Sur le socle conceptuel des *segments répétés* (Salem), elles devraient

40. Voir notamment (Bakhtine, 1977 ; Adam, 1999 : 24 et *passim*).

41. On entend ici par précieux tous corpus dont il est prévisible qu'ils feront l'objet d'explorations récurrentes, selon divers points de vue. Ces corpus méritent un étiquetage linguistique désautomatisé (alternance de phases automatiques et conviviales), probablement coûteux (il ne s'agit donc nullement des seuls corpus « littéraires », mais au moins de tous les corpus patrimoniaux, au sens large, concernant les sciences humaines).

néanmoins, dans les temps à venir, contribuer aux fondements sémantiques de la théorie des *mots composés* et des *phraséologies*. En outre, la *cooccurrence* dans son acception *généralisée* (voir *supra* microdistribution) nous indique d'avoir à chercher non pas du côté des « mots », des unités discrètes, mais avant tout dans la multidimensionnalité du discours et du texte.

7.7. Bibliographie

Adam, J.-M. & Heidmann, U. (à paraître). Sciences du discours en dialogue : textualité & comparaison. Colloque interdisciplinaire de Lausanne Mai 2004. Genève : Slatkine.

Adam, J.-M. (1999). Linguistique textuelle : des genres de discours aux textes. Paris : Nathan.

Bakhtine. M. (1977). Marxisme et philosophie du langage. Paris : Minuit.

Balpe, J.P. *et al.* (1996). Techniques avancées pour l'hypertexte. Paris : Hermès.

Biber, D. (2004). Conversational texts types : a multidimensional analysis. Le Poids des mots. Actes des 7es journées internationales d'analyse statistique des données textuelles. Louvain : UCL.

Biber, D. *et al.* (1999). The Longman grammar of spoken and written English. New York : Longman.

Bloomfield, L. (1930). Language. New York : Holt. Trad.1970. Genève : Payot.

Bolasco, S. et al.. (1995). JADT 1995 : III Giornate internazionali di analisi statistica dei dati testuali. Rome : CISU.

Cibois, P. (1994). L'Analyse factorielle. Paris : PUF.

Dister, A. (2000). Actes des Troisièmes Journées INTEX in RISSH 2000. Liège : CIPL.

Ducrot, 0. et Schaeffer, J.M. (1995). Nouveau dictionnaire encyclopédique des sciences du langage. Paris : Seuil.

Greimas, A.J. (1970). Sémantique structurale. Paris : PUF.

Guiraud, P. (1954). Les Caractères statistiques du vocabulaire. Paris : PUF.

Guiraud, P. (1960). Problèmes et méthodes de la statistique linguistique. Paris : PUF.

Habert, B., Nazarenko, A. et Salem, A. (1997). Les linguistiques de corpus. Paris : Colin.

Harris, Z.S. (1969). Analyse du discours. Langages, 13,11-65.

Lebart, L. et Salem, A. (1994). Statistique textuelle. Paris : Dunod.

Lelu, A. & Aubin, S. (2001). Vers un environnement complet de synthèse statistique de contenus textuels : Neuronav version 2. – Séminaire ADEST : http : //www.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm

Maingueneau, D. (1991). L'Analyse du discours : introduction aux lectures de l'archive. Paris : Hachette.

Mellet, S. (1998). JADT 1998 : 4es journées internationales d'analyse statistique des données textuelles

Morin, A. *et al.* (2002). JADT 2002 : 6es journées internationales d'analyse statistique des données textuelles. Rennes : IRISA/INRIA.

Muller, C. (1993). Langue française : débats et bilans. Paris : Champion.

Muller, C. (1997a). Initiation aux méthodes de la statistique linguistique. Paris : Champion.

Muller, C. (1997b). Principes et méthodes de statistique lexicale. Paris : Champion.

Pécheux, M. (1990). L'Inquiétude du discours. Paris : Editions des Cendres.

Purnelle G. et al., ed. (2004). Le Poids des mots : Actes des 7es Journées internationales d'analyse statistique des données textuelles. Louvain : PUL.

Rajman, M. & Chappelier, J.-C. (2000). JADT 2000 : 5es journées internationales d'analyse statistique des données textuelles. Lausanne : EPFL.

Rastier, F. (1989). Arts et sciences du texte. Paris : Hachette.

Rastier, F. (2001). Sens et textualité. Paris : PUF.

Riffaterre, M. (1971). Essais de stylistique structurale. Paris : Flammarion.

Silberztein, M. *et al.* (2004). INTEX pour la Linguistique et le Traitement Automatique des Langues : 4es et 5es journées INTEX. Besançon : PUFC.

Viprey, J.M. (1997). Dynamique du vocabulaire des Fleurs du mal. Paris : Champion.

Viprey, J.M. (2000). Pour un traitement textuel de l'allitération. Semen, 12, 245-272.

Viprey, J.M. (2002). Analyses textuelles et hypertextuelles des Fleurs du mal. Paris : Champion.