Structure non-séquentielle des textes

0. Enjeux

Le titre de cette livraison de *Langages* nous invite à réfléchir d'un même pas sur les statuts et modalités de l'*unité-texte* et des *unités de texte*. L'objet du présent article est d'y apporter un point de vue particulier, construit d'une convergence des domaines de l'*analyse textuelle des discours* (Adam & Heidman 2005), de la *linguistique de corpus* (Condamines 2005) et de l'*analyse statistique des données textuelles* (Purnelle et al., 2004). Le point de départ est un double constat. Primo, la situation observée et traitée par Adam & Heidman (2005) – déjà repérée, sous une formulation différente (les *déficits*) par Rastier (2001) - d'une relative ignorance mutuelle des sciences du discours et des sciences du langage. Secundo, le développement inégal du programme proposé par Halliday & Hasan (1976) dans le champ des *linguistiques discursives* (*linguistique textuelle, analyse de discours, sémantique des textes*) (Paveau & Sarfati, 2003), au détriment de la composante lexicale de la *cohésion textuelle* et, au sein même de cette dernière, au détriment de la *collocation*.

Consacrant l'essentiel de nos travaux[1] à montrer la cooccurrence comme facteur primordial de la textualité, comme marque de l'interdiscours et comme vecteur de la lecture hypertextuelle des corpus en sciences humaines, nous nous proposons de l'introduire dans la présente perspective collective. Après avoir évoqué cette partie du programme de Halliday & Hasan, nous indiquerons les enjeux de la cooccurrence lexicale généralisée pour la problématique *Unité(s) du texte*, en illustrant la démarche proposée autour d'un exemple, tiré de Balzac[2].

1. Rapide retour sur *Cohesion in English* (Halliday & Hasan)

> We now come to the most problematical part of lexical cohesion, cohesion that is achieved through the association of lexical items that regularly co-occur. (Halliday & Hasan, 1976 : 284).

---

[1] Après ou avec d'autres, voir notamment Salem (1987), Heiden & Lafon (1998), en cherchant particulièrement (Viprey 1997 et 2000 ; on pourra aussi se reporter aux contributions de l'auteur à Adam & Heidman 2005, Condamines 2005 et Purnelle et al. 2004) à mettre en évidence la *cooccurrence généralisée* comme facteur global de textualité et comme marque du discours difficilement accessible sans une formalisation statistique.

[2] Le choix d'un exemple n'est certes jamais neutre. Il serait trop long de le justifier et surtout d'en examiner les implications. Précisons que l'auteur et son équipe travaillent dans les mêmes perspectives sur la diversité des genres impliquée par la recherche interdisciplinaire en sciences humaines (presse, histoire, clinique, etc).

Dans *Cohesion in English*, les 4 pages de leur chapitre 6 (*Lexical cohesion*) que Halliday & Hasan consacrent à la notion de *collocation* n'ont peut-être pas, trente ans après, retenu toute l'attention ni surtout tous les développements qu'elles méritent. Il est vrai que cette rubrique paraît *a posteriori* très brève, dans le cadre d'ailleurs d'un chapitre lui-même exceptionnellement succinct[3]. Notons la prudence de la formulation liminaire citée en exergue ci-dessus, dont les motifs expliquent sans doute cette brièveté ; il s'agit d'une mention programmatique plus que d'un développement, mais ce caractère programmatique est très fortement emphatisé par les auteurs :

> The analysis and interpretation of lexical patterning of this kind is a major task in the further study of textual cohesion (Halliday & Hasan, 1976 : 287).

Notons surtout que les auteurs passent lentement, mais sûrement, d'une certaine conception de la collocation à une autre, substantiellement différente, qui est la nôtre, mais qui ne se dégage pas distinctement et conclusivement chez eux. La première conception est très contrainte par les *a priori* d'une sémantique lexicale exogène : sont d'abord sélectionnées des collocations de synonymes et antonymes, puis de séries ordonnées[4] – *ordered series* -, puis non-ordonnées[5], après quoi les auteurs font ce prudent pas « en avant » (nous soulignons) :

> The members of any such set stand in some kind of semantic relation to one another, <u>but for textual purposes it does not much matter what this relation is</u> (Halliday & Hasan, 1976 : 285).

Toutefois la phrase suivante réintroduit, de façon plus diffuse, le privilège à une sémantique *en langue*, vs *en discours* (nous soulignons) :

> There is always the possibility of cohesion between any pair of lexical items which are in some way associated with each other <u>in the language</u> (Halliday & Hasan, 1976 : 285).

C'est en fait par une seule formulation sans ambiguïté (autre que le contexte très contraignant où elle s'inscrit) que Halliday & Hasan nous semblent franchir un pas décisif vers le *vocabulaire*, c'est-à-dire vers la composante lexicale du texte comme champ irréductible à une *langue* posée comme *a priori*, donc dotée d'un degré principiel et

---

[3] L'ouvrage compte 8 chapitres, parmi lesquels 5 concernent les types de liens cohésifs (*cohesive ties*, p.4 et *passim*) ; les 4 premiers (2,3,4,5) décrivent la cohésion grammaticale (*grammatical cohesion*, p.274) : *reference, substitution, ellipse, conjunction* et comportent de 50 à 80 pages chacun. Le chapitre 6, le seul à évoquer la cohésion lexicale, ne compte que 19 pages.

[4] Exemples : *Tuesday … Thursday, north … south* etc.

[5] Exemples : *basement … roof, car … brake, chair … table* etc.

opératoire d'autonomie par rapport au *lexique* (fait de langue). Encore est-il restrictivement question dans cette formule de phrases adjacentes :

> In general, any two lexical items having similar patterns of collocation […] will generate a cohesive force if they occur in adjacent sentences (Halliday & Hasan, 1976 : 285).

Notons enfin que la bibliographie de Halliday & Hasan comporte une référence aux *Discourse Analysis Reprints* de Harris (1963), sans que pourtant les principes de ce texte fondateur, qui traite précisément en grande partie de la *collocation*, soient évoqués dans les 4 pages en question. Or la traduction en français, alors relativement récente, dans *Langages* n° 13, et son positionnement très programmatique dans le champ de l'Analyse de Discours en France, ont sans doute contribué à l'occultation ou à une lecture biaisée de cette partie-ci de *Cohesion in English*.

Beaucoup de discussions[6] dans la sphère de l'analyse textuelle des discours ont évoqué, au sujet du texte de Harris, un certain degré de contresens ou plus exactement, à la réflexion, de contre-emploi : J.Sumpf, J.Dubois et les éditeurs de *Langages* n° 13 auraient en quelque sorte forcé ou optimisé le sens du texte de Harris en l'insérant dans l'intertexte de ce numéro, à partir d'une équivoque lexicologique et traductologique posée par le terme *discourse*, le faisant dériver du champ sous-disciplinaire d'une linguistique « nucléaire », « dure », vers celui des sciences du discours[7]. Il est vrai que beaucoup de développements ultérieurs à partir de Harris, orientés vers le TALN, portent à la textualité, et *a fortiori* à la discursivité, proprement dites une attention plus proclamée qu'efficiente.

Il est d'autant plus important de relever le défi que constitue la collocation comme aspect décisif de la texture. D'autant plus essentiel aussi de ne pas marginaliser cet aspect, comme une sorte de zone de non-droit, de flou, en en faisant par exemple l'objet privilégié d'une sous-discipline à part, disons, comme au hasard, la stylistique. Cette contribution vise à prévenir un tel risque et à tenter de remettre la structuration hyperdimensionnelle du vocabulaire au cœur des problèmes de la textualité (et, au-delà, de la discursivité).

L'article de Dominique Legallois dans ce même numéro, envisage la collocation comme une propriété des phrases, dont la plus ou moins grande densité leur confère une capacité cohésive. Nous souhaitons explorer, pour notre part le mode selon lequel la

---

6 Voir notamment à ce sujet Guilhaumou, Maldidier & Robin 1994, pp.77, 178 et *passim*.

7 J'ai discuté très elliptiquement ce thème (Viprey, 1996 : 18:26), et je dirais volontiers aujourd'hui que, comme et avec les propositions de Halliday & Hasan, celles de Harris constituent un enjeu conceptuel jusqu'au sein des compétitions internes aux sciences du langage, enjeu que nos domaines (sciences textuelles et discursives) ne sauraient trop estimer.

collocation modifie et constitue les vocables en tant que tels, non tellement comme individus que comme constituants d'un tout engréné, influant ainsi sur la cohésion par l'intérmédiaire du palier lexical intégralement considéré (et non disséminé phrase par phrase) ; et, au-delà de la cohésion, sur la signification.

2. Séquentiel, non-séquentiel.

Le propos n'est certes pas de nier les propriétés séquentielles du texte, qui sont abondamment évoquées dans ce numéro et, beaucoup plus largement, dans les travaux qui dynamisent le champ des linguistiques discursives. Il s'agit d'éviter que cette classe de propriétés ne « bénéficie » d'un privilège outrancier, voire exclusif, et de montrer qu'un tel « privilège » se révèlerait à terme un handicap pour la séquentialité elle-même.

Nul ne saurait mettre en doute qu'un texte se manifeste dans l'ordre du temps et/ou de l'espace orientés, se caractérise par un début, un milieu, une fin, ordonnés et non interchangeables, et ce à quelque échelle que ce soit, de l'organisation macro-séquentielle à la fine succession des périodes. Mais lorsque Halliday & Hasan écrivent :

> Typically, in any text, every sentence except the first exhibits some form of cohesion with a preceding sentence […] (Halliday & Hasan, 1976 : 293).

derrière un apparent truisme (*except the first*) pourrait se cacher une doxa criticable[8]. Si certes la première phrase d'un texte ne peut faire référence à une précédente[9], en tant que *première phrase* elle anticipe (cataphore implicite) l'ensemble et elle ne cessera jamais de l'avoir fait. On montre par exemple que les débuts de la plupart des textes narratifs ont des propriétés lexicales caractéristiques, en termes statistiques, qui parasitent notamment l'étude de l'*accroissement lexical*. Dans *La Comédie humaine*, de Balzac, plus un texte est court, moins ses 500, 1 000, 2 000 etc premières occurrences représentent une grande variété de vocables. En d'autres termes, les propriétés d'ensemble les plus globales du texte qui débute, contraignent anticipativement ce début[10].

Même si l'on se cantonne, dans un premier temps, à la cohésion grammaticale, on constate dès l'abord que les relations anaphoriques des 4 ordres distingués par Halliday & Hasan installent un réseau immensément profus, dont ils rendent compte notamment

---

8 Que Halliday & Hasan préviennent certes en permanence.

9 A condition bien sûr de laisser de côté l'intertextualité.

10 Pour les 1000 premiers mots, le coefficient de corrélation de Spearman (Muller, 1992A : 153 ssq ;180) entre le classement par longueur et le classement par nombre de formes diverses, est de 0.306 pour 85 degrés de liberté, soit une probabilité de 0.005. Pour les 3 000 premiers mots, le coefficient de Spearman est de 0.283, soit une probabilité inférieure à 0.01.

pp.296 ssqq (*tight and loose texture*) mais en privilégiant toujours les liens de voisinage au détriment de liens potentiellement beaucoup plus distants (bien qu'en tout état de cause la portée moyenne efficace de ces liens soit réduite). Ce réseau prend nécessairement appui sur la structure linéaire (chacun des liens unitaires qui le constituent ne peut être décrit que comme reliant une séquence à une autre) mais il la subvertit et la déforme en chacun de ses points. Cela devient particulièrement sensible au niveau du vocabulaire, car contrairement aux niveaux « grammaticaux », on ne peut même pas envisager de commencer la description du niveau lexical sans le délinéariser.

A la question « qu'est-ce qu'un vocable ? », on répondra que c'est une forme[11] qui fait occurrence (*occurs*). Très longtemps, la statistique lexicale s'est attachée à quantifier cette occurrence (Guiraud, 1954) et à en comparer les proportions avec celles d'un « corpus » de référence (notamment, la base de textes du *Trésor de la langue française*, aujourd'hui *Frantext*). On identifiait ainsi des *mots-thèmes* (les plus fréquents en valeur absolue, à l'exclusion des catégories non-lexicales), et des *mots-clés* (qui, moyennant un certain seuil d'occurrence absolue, se caractérisaient par leur sur-emploi relatif). Ces listages ordonnés ont été les premières vues synthétiques non-linéaires sur les textes.

Les index en fascicules papier[12] (héritiers comme les précédents de la tradition philologique multi-séculaire des spécialistes de la Bible) représentaient sans toujours le formaliser explicitement les textes comme des faisceaux d'éléments discontinus, troués, emboîtés ou mieux *engrénés* entre eux tous par leurs « perforations ». A chaque vocable est affecté une série de numéros de *mot* (au sens lexicométrique du terme, Muller 1992B), série qui à la fois résume sa distribution au sens de répartition, et préfigure sa distribution

---

[11] Par *forme*, on entend substantiellement faire référence à la discussion de Rastier, prolongée dans ce même numéro, et notamment à la distinction *forme/unité*, que nous interprétons (d'une façon sans doute partiellement hétérodoxe aux yeux de Rastier lui-même) comme une distinction dialectique, permettant aux analyses, voire aux pré-traitements, de prendre appui sur le *discret, stable, identique à soi-même* en vue justement d'y articuler la variation et l'altération. Pour mieux cerner notre positionnement quant à la dialectique *répétition/altération*, on se reportera très utilement à Jean Peytard (1993) et à l'ensemble du numéro 8 de la revue *Semen* (*Configurations discursives*), avec les contributions de Jean-Michel Adam, Magid Ali-Bouacha, Dominique Maingueneau, Denise Maldidier, Marie-Françoise Mortureux et Jean Peytard, ainsi qu'au n° 12 de la même revue (2000).
Cette notion de *forme* est à distinguer de celle employée en lexicométrie, qui renvoie à l'identité d'une séquence graphique de la « surface » textuelle.

[12] Et de manière plus implicite encore les index-concordances en volumes, où les contextes étroits listés masquent le caractère « emboîtable » des grilles d'occurrences.

au sens harrissien. Dans le 2ème quintil de « Moesta et errabunda »[13], l'index positionnel du vocable MER peut être schématisé ainsi :

_x___x____/_____x___/_____/_____/_x__x_____

où l'on discerne les espaces, les vides, disponibles pour les emboîtements, par exemple ceux de VASTE :

____z_____/_____/_____/_____/____z_____

Avec une résultante de cet ordre (fig.1) :

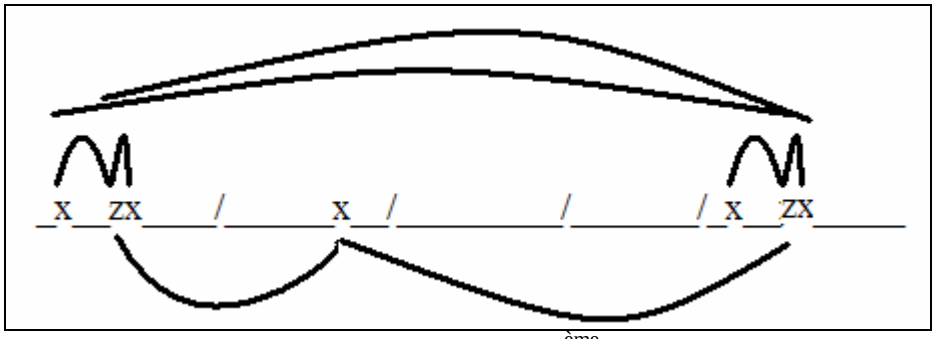


Fig.1. Collocation mer(x)/vaste(z) dans le 2ème quintil de *Moesta et errabunda.*

Faute d'envisager les choses ainsi, on tendra toujours à se représenter la collocation (cooccurrence) comme un fait au mieux pluri-occurrent, mais néanmoins cantonné à la phrase, isolé des autres faits du même ordre, alors qu'en réalité tous ces ensembles positionnels fonctionnent comme des *champs*, au sens le plus topologique du terme, des espaces mutuellement interagissants. Si une occurrence, en tant que telle, n'est rien, ne signifie pas isolément, il en est encore de même pour la chaîne de toutes les occurrences d'un même item. La signification se construit dans l'interaction des contextes, que l'on doit considérer selon des paramètres d'étendue assez variables.

On peut montrer que dans tout ensemble textuel le vocable se caractérise par deux *profils*, figurables par des graphes (courbes ou histogrammes) : un profil macro-distributionnel (mode selon lequel son occurrence se distribue dans les séquences de grandes dimensions), et un profil micro-distributionnel (mode selon lequel il se distribue dans les contextes des autres vocables[14]). On peut ensuite généraliser cette approche du vocable en entreprenant de classifier les profils ainsi obtenus et de faire servir cette

---

13 La mer, la vaste mer, console nos labeurs !
Quel démon a doté la mer, rauque chanteuse
Qu'accompagne l'immense orgue des vents grondeurs,
De cette fonction sublime de berceuse ?
La mer, la vaste mer, console nos labeurs !

14 Relation strictement réversible et symétrique : on aurait pu écrire *selon lequel les autres vocables se distribuent dans son contexte*.

classification à l'exploration textuelle dans le cadre d'une *herméneutique matérielle* (Rastier 2001, Adam & Heidman 2005).

3. L'engrenage des vocables et leur figuration synthétique : l'*isotropie*.

On s'appuiera, à partir de ce point, sur l'exemple *Père Goriot* de Balzac[15]. Dans tout ensemble textuel, les vocables présentant une occurrence conséquente (sans que le seuil puisse en être précisément fixé) peuvent être décrits sous l'aspect de leur profil cooccurentiel, c'est-à-dire de la quantification pondérée de leurs collocations avec les autres vocables. Par comparaison avec l'hypothèse nulle d'équidistribution, on établit quels sont les vocables Vc_1, Vc_2, Vc_3, Vc_n en excédent et en déficit dans l'environnement du vocable « pivot » Vp. Ces excédents et déficits s'expriment en écart-réduits[16].

Le verbe VENDRE, par exemple, a pour cooccurrents excédentaires « forts » (c'est-à-dire présentant un écart-réduit positif significatif au modèle équidistributif) :

PAYER, SOMME, SOIF, DIAMANT, BIJOU, BLÉ, SÛRETÉ, VIAGER (>5),
FRANC, MILLE, CENT, DERNIER, VENDRE[17], MÈRE, PLEURER, RENTE, HABIT, HIER, TANTE, COUVERT, CHANGE, VERMEIL, OPINION (>3),
CONNAÎTRE, FAMILLE, MARI, LETTRE, IDEE, SECRET, ROBE, BOIRE (>2),

et ces données constituent (avec les excédentaires « modérés », les équirépartis, les déficitaires) le profil cooccurrentiel, autrement dit *micro-distributionnel*.

L'analyse factorielle[18] permet de comparer entre eux simultanément les dizaines, centaines, milliers de profils ainsi déterminés, de les classer, d'en quantifier la saillance, et de « cartographier » cette classification. Sa projection « sphérique » en un continuum sans interruption est une des voies possibles, opératoires, pour le passage du *discontinu* logico-grammatical (pour reprendre les termes de Rastier) au *continu* sémantique. Non seulement ce graphe est un continuum sans rupture, mais pour tout item que l'on y repère, on peut

---

15 Edition Furne, aimablement mis à disposition par les Editions Champion. Sauf précision contraire, les relevés de collocations sont effectués dans les seules limites du paragraphe et de 50 mots à droite et à gauche du pivot.

16 Sur l'équidistribution, la variance, l'écart-réduit et maintes notions de lexicométrie, voir Muller (1992A).

17 Cas de multiple occurrence dans le contexte paramétré. L' « auto-collocation » est régulièrement positive, le plus souvent très positive.

18 Choisie parmi un ensemble de méthodes statistiques multidimensionnelles notamment pour les qualités spécifiques des sorties graphiques qu'elle détermine. Sur l'AFC, voir Lebart & Salem (1994) et Cibois (1994).

prolonger l'analyse en limitant l'examen à ses contextes, où l'on fera apparaître des disjonctions elles-mêmes non-dichotomiques[19].

La fig.2 présente une telle cartographie, pour les 230 vocables les plus occurrents du *Père Goriot*. Une fois précisé que carte se lit comme un planisphère[20], la proximité entre les items figure leur parenté de profils collocatifs, leur éloignement, leur dissimilarité (la plus grande dissimilarité étant exprimée par des positions antipodiques et un niveau de gris soutenu). Les items sont d'autant plus estompés (en niveaux de gris) qu'ils sont à une plus faible distance du centre de la sphère, et que leur position est donc moins significative, plus aléatoire.

Les relations lexicales ainsi cartographiées, j'ai proposé[21] de les désigner sous le terme global d' *isotropie*. L'isotropie s'entend (a) comme une valeur graduée : deux vocables sont plus ou moins isotropiques selon qu'ils sont plus ou moins proches sur le graphe ; (b) comme une sous-structure du vocabulaire - une isotropie est un groupe de vocables proches sur le graphe, pouvant être commodément désignée par un vocable pris comme attracteur, en raison notamment de son fort éloignement du centre de la sphère : on peut observer ici, par exemple, l'isotropie de VENDRE[22] (c) comme un continuum complexe et engrené lisible sur le graphe (d) et surtout, comme un vecteur d'exploration par retour au texte.

Le terme d'*isotropie* n'a pas été choisi – ni construit – au hasard de l'étymologie, seulement parce qu'il renvoie au préfixe *iso* et au radical *tre(o)p* (inclinaison/ation). Il est mis en compétition et/ou en coopération avec le terme et la notion d'*isotopie*. Compétition, s'il s'agit de l'emploi totalement perverti qu'en fait la critique académique : un équivalent jargonnant du *champ lexical* ou *notionnel*, déjà si confus, si peu opératoire et si trompeur.

---

19 Nous en donnons un exemple, pour REGARD et REGARDER dans l'ensemble de *La Comédie humaine*, dans notre contribution à Condamines (2005).

20 On prend en compte les 3 premiers facteurs (dont les inerties sont ici 7.47%, 5.18% et 4.14%). Les « longitudes » figurent l'angulation relative sur le plan des 2 premiers facteurs ; les « latitudes » figurent l'angulation avec ce plan, exprimée par le facteur n°3. Les niveaux de gris expriment la distance trigonométrique combinée par rapport à l'origine commune des axes. Sur cette figuration en planisphère, il faut concevoir les bords gauche et droit comme se rejoignant et le choix du « méridien » d'origine comme conventionnel. La figure peut tourner horizontalement de manière à offrir à la vue les relations de proximité entre veuve et rente, par exemple, assez analogues à celles de l'Est Sibérien et de l'Alaska… Les coordonnées sphériques des items sont donc indiquées conventionnellement, pour le repérage visuel, en longitudes et latitudes. Bien sûr, les coordonnées réelles sur les axes bi-orientés de chaque facteur sont accessibles dans un environnement assisté.

21 Viprey (1996 : 154 ssqq).

22 Env. 136°E, 7°N. La distance au centre est 0.32 (sur un maximum de 1 et une moyenne de 0.25).

Coopération, complémentarité, si l'on suit le travail de (re)définition qu'a mené Rastier depuis vingt ans, aboutissant à ce jour à ceci (nous soulignons) :

> *Isotopie sémantique* : effet de la récurrence d'un même sème. Les relations d'identité entre les occurrences du sème isotopant induisent des relations d'équivalence entre les sémèmes qui l'incluent (Rastier, 2001 : 299 - *glossaire*).

Même si l'on entend bien que le sème isotopant (et le sémème isotopé) s'actualisent dans le cadre d'une sélection, contrainte par le texte (ce qui est notamment le thème majeur de *Sens et textualité*), ces catégories n'en restent pas moins exogènes et projetées, à partir de descriptions du lexique, sur le vocabulaire. Comme telles utiles et nécessaires, elles n'en sont pas pour autant suffisantes pour décrire, modéliser et grammatiser la réticulation matérielle d'un texte singulier.

Les catégories de l'*isotopie* ainsi restructurées rendent compte d'invariants qui, des divers « extérieurs » du texte, peuvent contribuer à l'interpréter, au même titre bien qu'à d'autres niveaux, que les catégories lemmatisantes, morpho-flexionnelles, et de façon générale : grammatisantes. Elles modélisent notamment l'interprétant discursif-situationnel, la mise en lecture standard ou moyenne, le lambda synchronique.

Elles ne sauraient jamais se substituer à la formalisation descriptive des singularités sémantiques du texte à l'étude, dont la collocation généralisée nous semble une phase cruciale. Descriptive, cette formalisation n'en est pas moins une opération interprétative première, celle où le texte s'interprète en quelque sorte lui-même, se déploie, se restructure et objective les lignes de force de son vocabulaire.

En d'autres termes encore, des vues synthétiques de faits non-linéaires sont susceptibles d'ajouter à l'appareil textuel d'autres artefacts que ceux que l'imprimerie et l'édition moderne ont institués (mise en page et en volume, préfaces, avertissements, sommaires, tables, index, etc). Ces artefacts si durablement implantés qu'ils semblent faire partie d'une « nature » du texte contemporain, ont tous eu à voir avec la spatialisation extra-linéaire qui est l'une des sources de la textualité. La numérisation des textes eux-mêmes, mais bien plus largement encore, et antérieurement, celle de l'environnement éditorial, participent du même mouvement, qu'elles amplifient.

C'est pourquoi l'on peut admettre que si la préoccupation des unités non-linéaires (tabulaires et réticulaires) n'a pas attendu la fin du XX^{ème} siècle pour donner lieu à de fécondes recherches, la numérisation et l'expansion des ressources numérisées appelle et alimente un recentrement sur ce champ d'études. En effet, l'exploration systématique de ces dimensions ne peut se satisfaire qu'à titre préliminaire ou très ponctuel, des méthodes

intuitives. C'est l'une des vertus de l'informatique (à côté de l'exigence de formalisation explicite), que de permettre la suspension de l'activité interprétative, garant formel de l'exactitude et de l'exhaustivité des relevés programmés. Cette vertu heuristique fondamentale a trop souvent été perdue de vue, au profit d'une illusion holistique (la *boîte noire*) et/ou restrictivement cognitiviste (l' *intelligence artificielle*).

Contrairement à ce qu'une solide *doxa* à propos des statistiques textuelles est susceptible de véhiculer, l'étude de l'isotropie n'a pas pour objet de quantifier la cohésion textuelle ni la collocation, mais essentiellement de les *qualifier*. L'option quantitative (variance globale, « richesse » et « accroissement » lexicaux) entre certes dans la caractérisation du texte par contraste (avec d'autres textes, avec un corpus), mais appliquée à la cohésion, elle risque constamment de sous-tendre une sorte de « question de la preuve », l'idée selon laquelle on pourrait « prouver », par la statistique, la cohésion donc le statut de texte. Nous adoptons le point de vue diamétralement opposé, qui ne contredit pas l'option fonctionaliste de Halliday & Hasan mais en interdit les interprétations réductrices : le texte est un objet empirique du discours et des sciences discursives, qui n'a pas à être « prouvé » ; ainsi que l'écrivait Rastier dans le n°12 des *Cahiers du Crisco* qui a servi de matrice à la présente publication, nous *pren*[ons] *pour objet les textes et discours dans leur production et leur interprétation*, sans nous poser *les problèmes de la référence et de la vérité*.

4. Exemple d'observation *isotropique* : VENDRE dans *Le Père Goriot*.

VENDRE (repéré *supra*) possède un profil moyennement saillant dans le vocabulaire du roman. En témoigne sa mise en forme sur le graphe, d'un niveau de gris intermédiaire. Néanmoins, sa position relative aux autres items est susceptible de frapper l'attention au moment où elle se porte sur ce secteur du graphe. En effet, une lecture guidée par l'attente isotopante comparera interprétativement la position effective de l'item, avec celle qu'il pourrait avoir, plus haut, dans la circonscription des vocables FORTUNE, FRANC, RENTE, PAYER, ARGENT, RICHE ; d'un autre point de vue très légèrement différent, on pourra estimer qu'il est « au bord » du secteur de ces vocables, d'un certain côté, sans doute « attiré » vers d'autres groupements ou formant avec eux une intersection.

En l'occurrence, le groupement *isotropique* lisible au « sud » de VENDRE est lui-même, toujours du point de vue d'une attente « isotopante », manifestement bi-composite. L'une des composantes (*grosso modo*, un ensemble de vocables évoquant les liens de

parenté) amène même deux de ses items au voisinage précis de VENDRE. Les positions de VENDRE, MÈRE et FILLE apparaissent très proches.

Ce n'est pas principalement que les trois termes soient très cooccurrents, en triplet ou par paires. La seule cooccurrence forte, déjà signalée *supra*, est celle de MÈRE et de VENDRE (écart-réduit +4.8). Les deux autres couplages sont… légèrement déficitaires, même MÈRE/FILLE (-0.11) ! Ce qui les apparente, ce sont leurs profils collocatifs proprement dits, autrement dit la constitution lexicale de leurs contextes d'occurrence, tout juste ce que Halliday & Hasan nomment, tout en semblant ne pas en avoir une conception aussi ample, *similar patterns of collocation*.

Il serait trop long d'entrer ici dans le détail d'une telle exploration, qui n'est permise et ne prend sens que dans des environnements assistés tels que les développe l'analyse de textes « par ordinateur ». On peut néanmoins commencer à comprendre, en observant (dans l'un de ces environnements) les (37) contextes[23] de VENDRE, quels facteurs généraux, d'ordre thématique, s'expriment dans sa position sinon inattendue, du moins non-triviale, sur le graphe issu de l'AFC de collocation. En 5 cas seulement, il s'agit du commerce proprement dit. Tous les autres emplois sont en contexte dysphorique, dont la majorité ont pour objet des personnes, des biens personnels et (10 fois) des biens précieux. Liquidation, trahison, prostitution.

Significations qui renvoient à un exogène au moins aussi pertinent que les constructions lexicologiques ou lexico-sémantiques hors corpus. Cet exogène est l'interdiscours construit de la fiction narrative « réaliste » du XIX^{ème} siècle, où dans des configurations certes toujours mouvantes et plurivoques[24] les formes produites par l'isotopie de *l'argent* se combinent le plus souvent aux valeurs dépréciées, funestes ou morbides.

On observera un significatif invariant, en confrontant les fig.2 et 3 aux fig. 5 et 6, qui présentent les mêmes sorties graphiques d'AFC pour l'ensemble du corpus de *La Comédie*

---

[23] Dont nous ne donnons ici (fig.4) que l'état le plus résumé, la concordance. L'environnement d'exploration hypertexte est destiné à permettre l'alternance raisonnée des divers *états de texte* que sont plein-texte, concordances, contextes étendus, graphes, dictionnaires.

[24] L'intérêt des visualisations du vocabulaire en collocation réside aussi dans la possibilité de les comparer d'un texte et d'un ensemble textuel à l'autre, ainsi que de confronter et historiciser les interprétations qu'elles suggèrent dans le retour au texte, comparaisons concevables seulement dans un environnement assisté. La comparaison permet de saisir au mieux combien ces visualisations sont (a) irréductibles l'une à l'autre, malgré un invariant plus ou moins conséquent (b) en cela même, non triviales.

humaine[25]. Comme on a sélectionné ici les items les plus occurrents de l'ensemble de l'œuvre, le graphe global est sensiblement différent. Pourtant, l'entour *isotropique* de VENDRE met en place, avec de sensibles différences lisibles, une configuration comparable (et certes destinée à la comparaison).

Cette expérimentation très embryonnaire éclaire d'un nouveau jour la problématique de l' *unité du texte*, puisque nous mettons en regard des vues synthétiques sur deux vocabulaires, composantes/paliers de deux textes, qui entretiennent entre eux des rapports d'inclusion. Les deux textes ont des *cohésions* sensiblement distinctes, distinction plus saisissable, en envisageant les critères livrés par l'histoire littéraire, que celles se présentant dans les œuvres de Dumas et *a fortiori* de Zola (où la dimension de cycle narratif est affichée).

Par ailleurs, dès lors que l'on vérifie que la configuration *isotropique* de l'ensemble ne contraint que partiellement celle d'une de ses parties, on mesure à quel point le statut des *unités du texte* est endogène : il varie selon la dimension de l'ensemble auquel ces *unités* se rapportent, d'une manière très analogue à la différence d'interprétation que connaît un vers de La Fontaine ou de Baudelaire selon le degré et les angles de connaissance des œuvres extensivement conçues.

5. Au-delà de la séquentialité : une interaction de champs, l' *isotropie*.

Ainsi, la projection (toujours paramétrée et re-paramétrable) des lignes directrices du vocabulaire comme engrenage multidimensionnel, sous la forme d'un continuum graphique sans les ruptures dichotomiques que présentent la plupart des méthodes classificatoires, intuitives ou algébriques, manuelles, assistées ou automatiques, offre-t-elle un accès non-trivial à des niveaux textuels jusqu'ici peu disponibles pour un examen approfondi.

Si l'on veille bien à cantonner ces méthodes dans une fonction exploratoire et en aucun cas inférentielle, à ne les appliquer qu'à l'approfondissement de la praxis des textes, elles pourraient contribuer à une réalisation plus complète du programme de l'analyse textuelle des discours, en s'attachant à son objet le plus complexe et intriqué, le vocabulaire.

---

[25] Une différence importante est que nous n'avons pas à ce jour lemmatisé l'ensemble de *La Comédie humaine* et que son vocabulaire est, par conséquent, observé en formes graphiques. Nous avons cependant regroupé, pour l'occasion, toutes les occurrences fléchies de VENDRE, ainsi que celles de MÈRE et de FILLE.

Pour reprendre quelques questions soulevées au passage, les méthodes statistiques, intégrées à un dispositif d'analyse textuelle prenant en compte l'ensemble des niveaux, visent notamment, dans leur domaine délimité d'application, à renouveler l'articulation du logico-grammatical et du rhétorique-herméneutique (qui ne recouvre que partiellement celle du texte et du discours).

Ainsi, à propos de l'*unité du texte* (1er versant de la problématique soulevée), nous pouvons conclure que par méthode, le type de graphe ici introduit rend compte d'une unité qui n'a pu être *vérifiée* que par ailleurs, dans les processus philologique, d'une part, socio-historique de l'autre. Sa lecture experte dans le dispositif textuel d'ensemble, où il s'intègre, qualifie la cohésion lexicale et, à l'échelle macro-textuelle, confère la *texture* aux niveaux grammaticaux qui, sans celle-ci, ne sauraient accéder à la textualité.

Quant aux questions liées aux *unités du texte*, nous partageons à leur propos l'essentiel de la préoccupation de Rastier envers une conception de celles-ci comme *discrètes, identiques à soi* et *isonomiques*, et du texte comme *résult*[ant] *d'un enchaînement d'unités*, et son souci de *ne rapporte*[r] *pas exclusivement les formes sémantiques à des localisations spatio-temporelles*. Cependant, il nous semble diverger de lui en ceci : ce n'est pas la localisation spatio-temporelle qui est en cause par elle-même, mais le privilège à la localisation linéaire, séquentielle (index positionnel). Les collocations relèvent bel et bien d'une topologie, et comme telles sont en rapport nécessaire avec la spatio-temporalité du texte ; mais pas comme avec une spatio-temporalité « déjà là », imposée par une saisie première qui serait séquentielle, ce qui est un abus de méthode. Au contraire, avec un espace en *champs* qu'elles engendrent complémentairement.

Il est donc d'autres *unités du texte* que les unités séquentielles, qui ne sont pas les collocations elles-mêmes, mais des champs qui interagissent, discontinus sur le plan séquentiel et continus sur le plan textuel, puisque leurs silences mêmes ici et là signifient autant que leurs implantations explicites.

Ces unités se construisent et se décrivent dans un processus endogène, et elles n'ont pour ainsi dire de réalité palpable qu'opératoire. Il ne fait aucun doute qu'elles sont à l'œuvre dans la textualité et les *effets de texte*, mais elles ne sont guère marquées par les technologies du texte, qui s'appliquent pour l'essentiel au découpage linéaire, à l'exception notable des index, qui préfigurent depuis longtemps le marquage d'une diffusion. Au-delà des éditions scientifiques et techniques désormais traditionnelle, les environnements informatisés accentueront cet effet et exigeront une formalisation toujours plus poussée.

Il reste, entre autres, à examiner comment ces analyses retentissent sur la séquentialité elle-même. Nous n'en évoquerons que deux aspects principaux. D'abord, le support linéaire des vocables est une série d'occurrences localisées, résultats d'une certaine segmentation et de l'affectation de certaines propriétés morpho-syntaxiques (parfois phonétiques et/ou métriques) aux segments ainsi définis, le tout inspiré au moins initialement par une certaine grammaire. Tout examen, même rapide, d'une matrice de cooccurrence extraite d'un texte, permet de comprendre que les collocations les plus excédentaires peuvent toujours être interprétées comme des problèmes de segmentation. On tentera vainement de séparer les erreurs flagrantes portant sur des lexies composées[26], encore permises aujourd'hui malgré la sophistication des ressources, de cette zone compressible mais irréductible de flou entre lexique et phraséologie que lexicologie et analyse de discours ont en commun. L'analyse des collocations d'un texte obligera donc à reconsidérer cet aspect élémentaire de la problématique des *unités de texte*. Elle retentira aussi sur l'annotation de ce texte, modifiant par enrichissement (« décoration ») son système endophorique même.

Enfin, et surtout, les lignes saillantes de la collocation lexicale seront des interprétants de la structure séquentielle beaucoup plus féconds que la simple réitération. Formaliser les enchaînements et les combinaisons d'*isotropies* devrait considérablement éclairer l'étude de la progression thématique et les principes mêmes de cette catégorie du texte[27].

Références bibliographiques

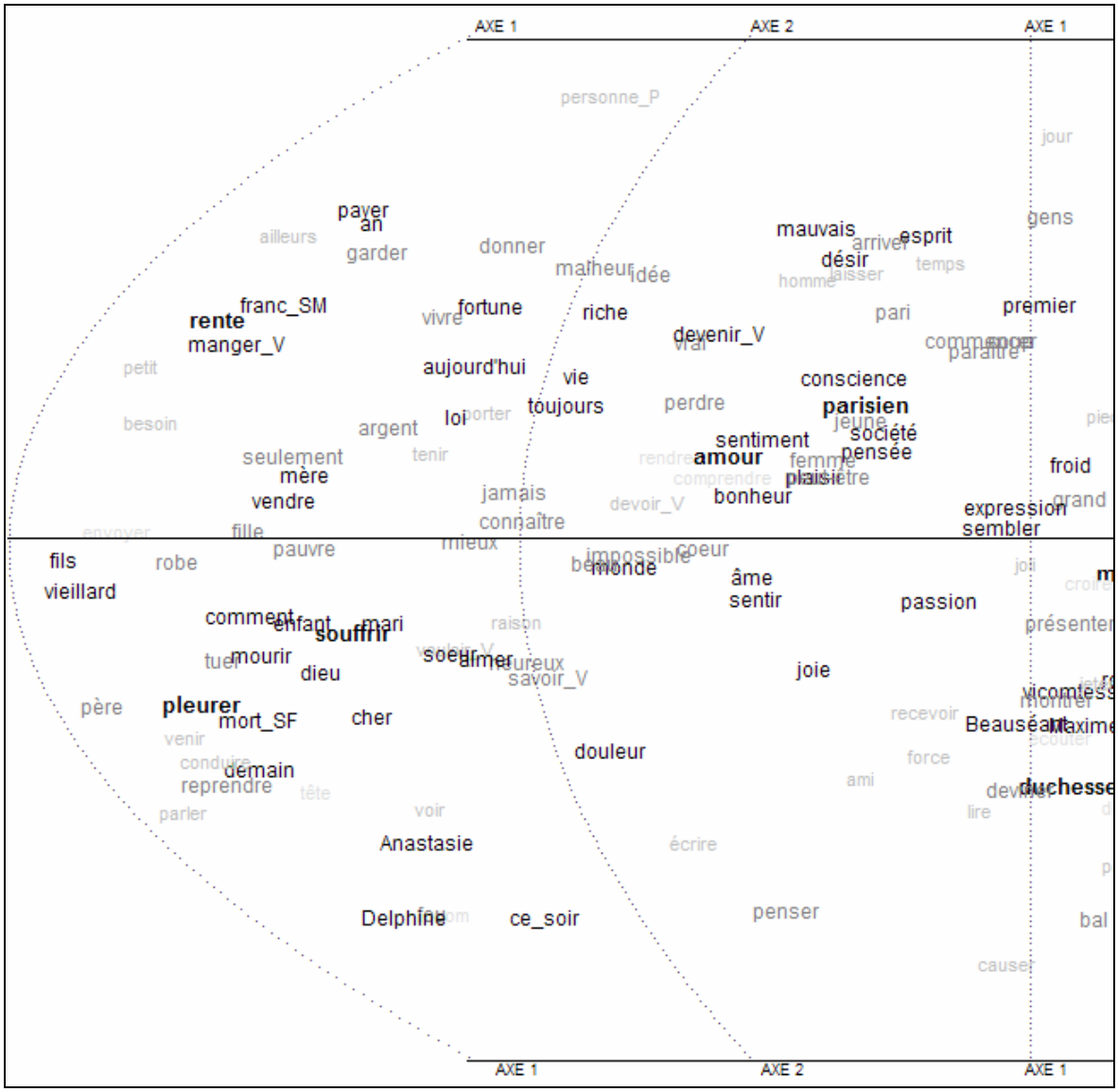
ADAM Jean-Michel 2005 : *Linguistique textuelle* Paris, Colin.

---

26 Une collocation récurrente comme celle de JEUNE et de FEMME, pour peu que l'empan cotextuel ait été paramétré court, sera régulièrement très excédentaire, jusqu'à « écraser » de son poids toute l'information contenue par ailleurs dans la matrice, rejetant tous les autres items en un nuage si concentré qu'on ne peut pas le lire.

27 Une mise au point et une insistance s'imposent à la fin de cette contribution, qui pourrait suggérer que nous exagérons considérablement les vertus scientifiques et heuristiques d'une méthode statistique particulière, alors qu'il en va tout autrement. L'AFC, le type de sorties graphiques et leur insertion dans le dispositif exploratoire sont employés dans le but de soulever un questionnement et d'exposer une hypothèse fondamentale, à propos de la structuration lexicale (ce terme pris dans une large extension, visant aussi bien les morphèmes que de très larges emprises phraséologiques) des textes en vue de leur analyse en discours. Nous avons « rencontré » l'AFC dans le cadre d'une quête de la textualité, et non l'inverse. Il nous a semblé y entrevoir un moyen de dépasser le paradoxe du vocabulaire et de son analyse : fait irréductiblement global et réticulaire, il s'est jusqu'ici laissé saisir par des entrées essentiellement locales. L'appareillage reste, tout aussi paradoxalement, rudimentaire sous un certain angle, « éblouissant » sous un autre. Ainsi restons-nous sans doute quant aux collocations comme facteur de cohésion textuelle et de cohérence discursive, mais à notre rang et dans un contexte scientifique et technique renouvelé, aussi programmatiques que Halliday & Hasan.

FIGURES 2 à 4 :


AXE 1
AXE 2
AXE 1
personne_P
jour
payer
an
ailleurs
garder
donner
gens
mauvais
esprit
arriver
désir
malheur
idée
temps
homme
laisser
franc_SM
rente
manger_V
vivre
fortune
riche
pari
premier
devenir_V
vrai
paraître
petit
aujourd'hui
vie
conscience
besoin
argent
loi
porter
toujours
perdre
parisien
jeune
société
sentiment
pensée
seulement
tenir
rendre
amour
femme
mère
comprendre
plaisir
être
froid
vendre
jamais
devoir_V
bonheur
grand
connaître
expression
sembler
envoyer
fille
pauvre
mieux
impossible
cœur
fils
robe
monde
âme
vieillard
sentir
passion
comment
enfant
souffrir
mari
raison
présenter
tuer
mourir
dieu
soeur
aimer
malheureux
savoir_V
joie
père
pleurer
mort_SF
cher
recevoir
Beauséant
maxime
venir
douleur
conduire
force
demain
ami
reprendre
tête
duchesse
parler
voir
lire
Anastasie
écrire
Delphine
ce_soir
penser
bal
cause
AXE 1
AXE 2
AXE 1

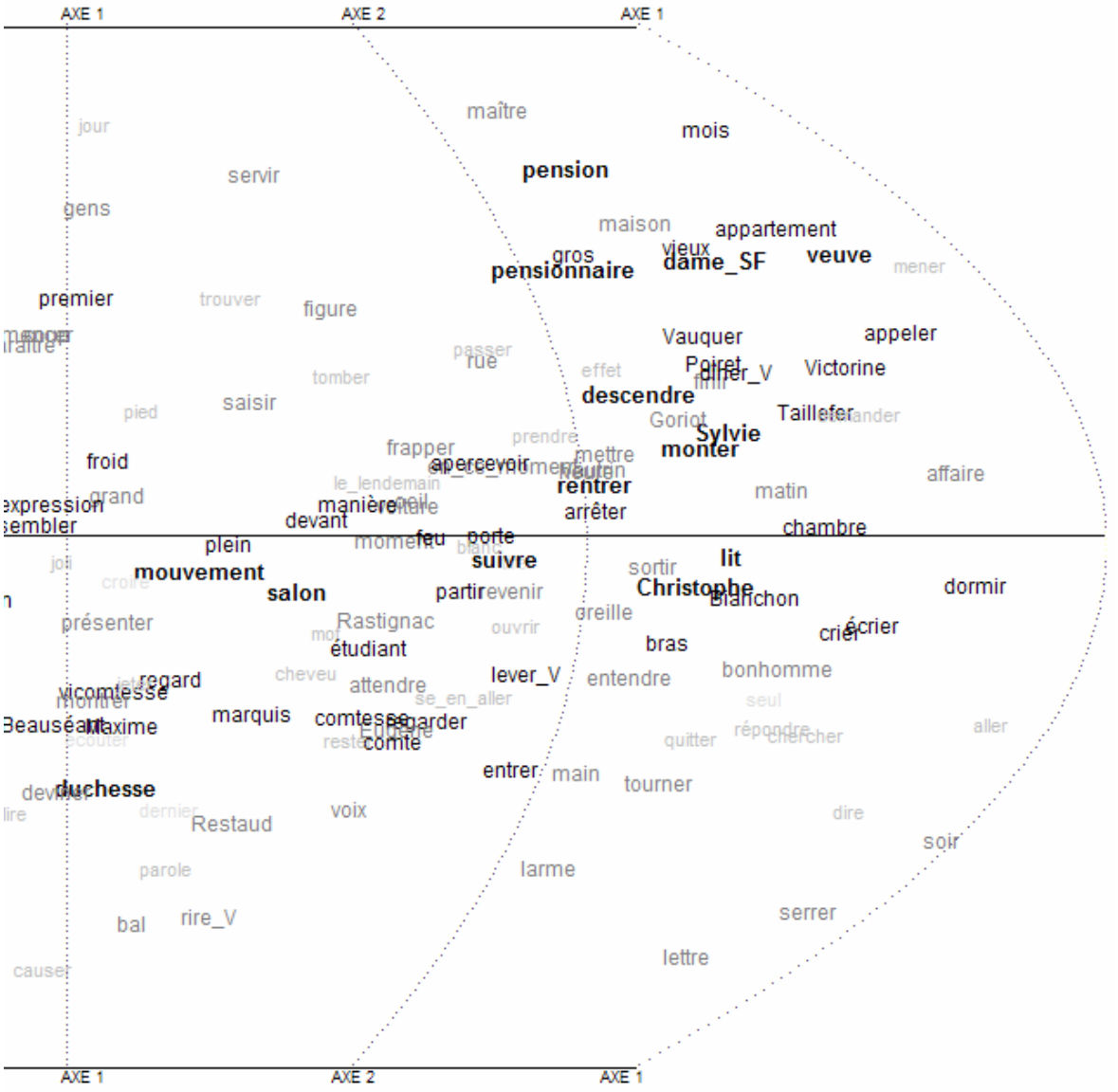


Fig.2. AFC de la cooccurrence de 230 vocables du *Père Goriot*, dans un cotextuel de 50 mots à gauche et à droite sans limitation à la phrase.

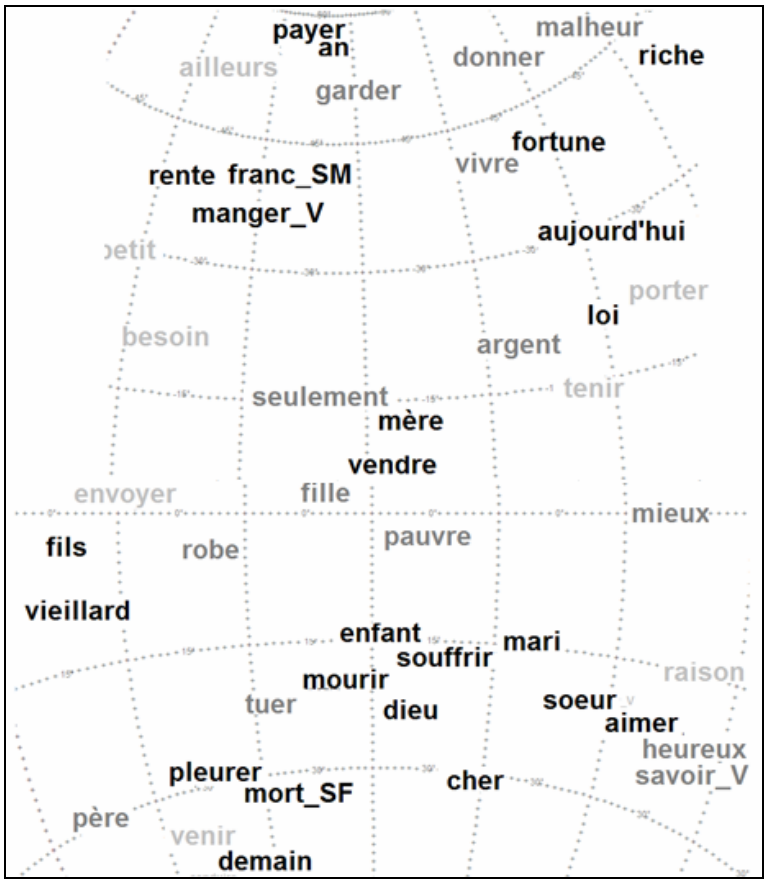


Fig.3. Extrait de la sortie graphique vue en fig.2, selon une projection « sphérique ».

ait au père Goriot tordant son vermeil et le	**vendant** pour aller payer la lettre de change de sa fille . -
disait-il . Ta tante a pleuré sans doute en	**vendant** quelques-unes de ses reliques ! De quel droit maudir
n ; ceux-ci pêchent des consciences, ceux-là	vendent leurs abonnés pieds et poings liés . Celui qui revie
uvent entretenir leur luxe effréné, elles se	vendent . Si elles ne savent pas se vendre, elles éventrerai
ma dernière angoisse . Si quelques femmes se	vendent à leurs maris pour les gouverner, moi au moins je su
à vos paroles . Quand on vous les demandera,	vendez-les . Un homme qui se vante de ne jamais changer d'op
	- Vendez une parure, lui écrivit-il chez le concierge, et que
us irez demeurer d'ici à trois jours . Ne me	vendez pas . Elle veut vous faire une surprise, mais je ne t
mi lesquels il se rencontre des farceurs qui	vendraient leur famille pour monter d'un cran . Si le métier
ine, et souvent croupie ; pour en boire, ils	vendraient leurs femmes, leurs enfants ; ils vendraient leur
vendraient leurs femmes, leurs enfants ; ils	vendraient leur âme au diable . Pour les uns, cette fontaine
ile pendant qu'elle souffrait ; moi, moi qui	vendrais le Père, le Fils et le Saint-Esprit pour leur évite
Vauquer pour renaître en Goriot . Se marier,	vendre sa pension, donner le bras à cette fine fleur de bour
aves, gagner quelques bons petits millions à	vendre mes boeufs, mon tabac, mes bois, en vivant comme un s
Vieux habits, vieux galons, vieux chapeaux à	vendre ! - A la cerise, à la douce ! La palme fut à Bianchon
mme, il n'y a pas d'autres ressources que de	vendre les dentelles de ma tante, dis-lui que je lui en enve
achetais son argent au prix où il veut me le	vendre ! Comment, moi riche de sept cent mille francs, me su
elles se vendent . Si elles ne savent pas se	vendre, elles éventreraient leurs mères pour y chercher de q
ameuse disette, et a commencé sa fortune par	vendre dans ce temps-là des farines dix fois plus qu'elles n
. Ce que j'entends par des sacrifices, c'est	vendre un vieil habit afin d'aller au Cadran-Bleu # manger e
ux les réaliser promptement . Ma bonne mère,	vends quelques-uns de tes anciens bijoux, je te les remplace
enfant, dis-moi si c'est Fil-de-Soie qui m'a	vendu ! Je ne voudrais pas qu'il payât pour un autre, ce ne
il en se frappant le front . Je sais qui m'a	vendu maintenant . Ce ne peut être que ce gredin de Fil-de-S
e de vieux couverts et des galons . Il lui a	vendu pour une bonne somme un ustensile de ménage en vermeil
oulu . L'intendant de ma grand'mère lui en a	vendu pour des sommes immenses . # Ce Goriot partageait sans
ortune, durera plus de six mois . Bon . J'ai	vendu mes treize cent cinquante livres de rente perpétuelle
deux, je me suis rafistolé, requinqué ; j'ai	vendu pour six cents francs de couverts et de boucles, puis
étien . D'ailleurs ce n'est pas toi qui m'as	vendu . Mais qui ? - Ah ! ah ! vous fouillez là-haut, s'écri
re petite, que ne venait-elle ici ! j'aurais	vendu mes rentes, nous aurions pris sur le capital, et avec
s l'avons égorgé : mon pauvre père se serait	vendu s'il pouvait valoir six mille francs . J'aurais été le
uivi . Dans cette extrémité, ma soeur aurait	vendu ses diamants à un juif, ces beaux diamants que vous av
ugit . Eugène ! Eugène, si vous l'aviez déjà	vendue, perdue & Oh ! cela serait bien mal .
	- Je les ai vendues en me réservant ce petit bout de revenu pour mes bes
s siens, les miens, tout, je les ai vendus .	Vendus ! comprenez-vous ? il a été sauvé ! Mais, moi, je sui
trer à tout Paris les diamants qu'on prétend	vendus par elle . Peut-elle dire à ce monstre : - Je dois mi
qui fit un bond . Les diamants n'ont pas été	vendus cent mille francs . Maxime est poursuivi . Nous n'avo
staud, les siens, les miens, tout, je les ai	vendus . Vendus ! comprenez-vous ? il a été sauvé ! Mais, mo

Fig.4. Concordance du verbe vendre

ADAM Jean-Michel, HEIDMAN Ute, éds. 2005 : *Sciences du discours en dialogue : Textualité & comparaison*, Genève, Slatkine.

CIBOIS Philippe 1994 : *L'Analyse factorielle*, Paris, PUF.

CONDAMINES Anne, éd. 2005 : *Sémantique et corpus*, Paris, Hermès(Lavoisier).

GUILHAUMOU Jacques, MALDIDIER Denise, ROBIN Régine 1994 : *Discours et archive*, Liège, Mardaga.

HABERT Benoît, NAZARENKO Adeline, SALEM André 1997 : *Les linguistiques de corpus*, Paris, Colin.

HALLIDAY M.A.K., HASAN Ruqaiya 1976 : *Cohesion in English*, London, Longman.
HARRIS Zelig S. 1963 : *Discourse analysis reprints*, La Haye, Mouton.
HARRIS Zelig S. 1969 : « Analyse du discours » in *Langages* n°13 J.Dubois & J.Sumpf éds., Paris, Larousse.
HEIDEN Serge, LAFON Pierre 1998 : *Cooccurrences : la CFDT de 1973 à 1992*, Fontenay, ENS.
LEBART Ludovic, SALEM André 1994 : *Statistique textuelle*, Paris, Dunod.
MULLER Charles 1992(A) : *Initiation aux méthodes de la statistique linguistique*, Paris, Champion.
MULLER Charles 1992(B) : *Principes et méthodes de statistique lexicale*, Paris, Champion.
PAVEAU Marie-Anne, SARFATI Georges-Elia 2003 : *Les grandes théories de la linguistique*, Paris, Colin.
PEYTARD Jean 1993 : « D'une sémiotique de l'altération », in *Semen n°8 : Configurations discursives*, Peytard & Moirand éds., Paris, Les Belles-Lettres, 145-177.
PURNELLE Gérald, FAIRON Cédrick, DISTER Anne 2004 : *Le poids des mots : Actes des 7èmes Journées internationales d'Analyse des Données Textuelles JADT* , Louvain, UCL.
RASTIER François 2001 : *Arts et sciences du texte*, Paris, Seuil.
RASTIER François 1989 : *Sens et textualité*, Paris, Hachette.
SALEM André 1987 : *Pratique des segments répétés*, Paris, Klincksieck.
VIPREY Jean-Marie 1997 : *Dynamique du vocabulaire des* Fleurs du mal, Paris, Champion.
VIPREY Jean-Marie 2000 : « Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle », in *JADT 2000,* Rajman & Chappelier éds., Lausanne, EPFL, 535-539.