Les séquences (suite)

Etienne Brunet

Laboratoire BCL (UMR 6039), Université de Nice, 98 Bd Herriot, 06 204 Nice

Abstract

The present study targets an unattainable objective, which has been previously treated in a presentation for JADT 2006. The goal is to account for well-known isotopies radiating through literary texts, which we hope to be able to isolate in a study of word sequences. Emphasis is placed on the inventory and the treatment of collocations. Four relatively convergent methods, in addition to the one previously proposed, are presented.

Résumé

La présente recherche poursuit un objectif insaisissable, auquel s'est attaché un précédent exposé aux JADT 2006. Il s'agit de rendre compte des fameuses isotopies qui rayonnent dans les textes littéraires et qu'on espère isoler dans l'étude des séquences. L'accent est mis sur le recensement et le traitement des cooccurrences. Quatre méthodes, s'ajoutant à la précédente et relativement convergentes, sont exposées.

Mots-clés : séquences, cooccurrences, isotopies, proxémie.

1. Introduction

On pardonnera à un ancien latiniste la figure étymologique cachée dans un titre allusif qui fait référence aux JADT 2006, où les organisateurs, méfiants à l'égard des bavards, m'avaient accordé la dernière place. Avant que mon exposé ait touché à sa fin, la cloche avait sonné qui avait précipité les participants vers la gare de Besançon. Je suis donc tenté de reprendre mon sujet et de l'achever (prenez le mot dans le sens qui vous plaît). Quoi de plus naturel en somme que de traiter les séquences en séquences distinctes, comme on faisait jadis pour les romans-feuilletons.

Rappelons que, selon une problématique initiée par Pierre Lafon dans sa thèse, les séquences étaient opposées aux fréquences. En réalité c'est là un large boulevard emprunté depuis longtemps par les chercheurs qui s'intéressent moins à la lexicométrie qu'au traitement automatique du langage. Dès que la perspective s'ouvre vers la documentation, le data mining, les systèmes-experts, la traduction, le résumé, on sort du cadre étroit où s'enferme la lexicométrie, condamnée à comparer les textes à l'intérieur d'un corpus et à manipuler des fréquences. Les textes sont certes des séquences, mais, au moins dans le domaine littéraire, leur empan est trop large pour que la cooccurrence de deux mots ou objets linguistiques ait un sens précis si elle est observée à longue distance (même si parfois le texte littéraire contient des rappels et des échos qui se répercutent de loin en loin). Les séquences impliquent une segmentation courte, qui est celle, non des textes, mais des phrases, des paragraphes, voire des pages ou des fenêtres, glissantes ou successives, de *n* mots. Le voisinage, immédiat ou proche, de deux mots, y prend une signification qui échappe au hasard, qu'il s'agisse d'une contrainte syntaxique, d'une aimantation sémantique, d'une convenance prosodique ou de quelque autre raison attachée à la situation ou à la langue. Rien n'empêche d'ailleurs, une fois que les relevés ont été faits dans les séquences, de les projeter dans l'espace partitionné du

corpus où l'on retrouve la division en textes. Les méthodes traditionnelles de la lexicométrie reprennent alors leurs droits et leur matériau de base, les fréquences, mais elles ne s'appliquent plus à des mots ou à des codes individuels mais à des combinaisons des uns et/ou des autres, à des profils, à des faits observés dans les séquences. Au fond il ne s'agirait là que d'une extension, d'un perfectionnement de la lexicométrie, qu'on a attendu trop longtemps parce que les traitements préalables de désambiguïsation et de lemmatisation n'étaient guère disponibles, sinon de façon manuelle et limitée. Maintenant encore le traitement des séquences reste trop artisanal et un long chemin reste à faire pour accéder à la standardisation, même si on voit la trace à suivre qui est celle du codage XML et des filtres sophistiqués d'interrogation, à base de grammaires spécialisées et d'*expressions régulières*[1].

Sans aller jusqu'à construire une telle grammaire, nous proposons ici une démarche exploratoire où cinq fonctions THEME, CORRÉLATS, ASSOCIATIONS, ALCESTE et TOPOLOGIE relèvent de la même approche, orientée vers l'étude des séquences plutôt que des fréquences. On y considère les mots (ou d'autres objets) dans leur environnement immédiat (paragraphes ou pages) en ignorant la partition en textes.

1 - La fonction TOPOLOGIE représente la distribution, aléatoire ou non, d'un ou de deux objets dans l'espace du corpus, et, le cas échéant, mesure la distance entre les deux distributions. Ce point, déjà développé dans les *Actes* des *JADT 2006,* ne sera pas repris ici. Mais on suppose acquis le calcul exposé, qui relève de la loi hypergéométrique.[2]

2 - La fonction CORRÉLATS regroupe les substantifs ou les mots sémantiques qui sont les plus fréquents dans le corpus et établit la carte synthétique de leurs cooccurrences (par une analyse factorielle de correspondance).

3 - La fonction ALCESTE établit un pont avec le logiciel ALCESTE. Elle lui fournit les données convenablement formatées, en lui transmettant la liste des substantifs les plus fréquents qu'on trouve associés dans un contexte étroit, paragraphe après paragraphe. Après traitement elle en reçoit les résultats sous forme de « classes ».

4 - La fonction THEME recense et assemble tous les passages où un mot (ou autre objet) est rencontré dans le corpus et oppose ces passages au reste du corpus. Il en résulte une liste de spécificités associée à l'objet de la recherche, graphie ou lemme. Ces mots associés au mot-

[1] L'exemple le plus achevé d'un traitement des séquences est celui de *Frantext*. Le logiciel de consultation *Stella* propose non seulement les opérateurs booléens, les jokers et les expressions régulières, mais aussi une grammaire évoluée qui mêle constantes et variables et rend le filtrage aussi fin et aussi souple que l'on veut. Dommage que les fonctions statistiques de *Frantext* ne soient pas à la hauteur des fonctions documentaires et ne s'appliquent pas à des relevés aussi finement établis.

[2] On avait expérimenté deux autres indices, *le Rapport de Vraisemblance* de Dunning et *l'Information mutuelle* de Church, tous les deux issus de la formule de Jaccard et utilisant les mêmes ingrédients – a : nombre de cooccurrences des deux mots dans le champ exploré (ici le paragraphe – b : nombre d'occurrences du premier mot en l'absence du second – c : nombre d'occurrences du second mot en l'absence du premier – d : nombre d'occurrences des autres mots.
Or la convergence n'est pas observée dans toute l'échelle des valeurs On s'en est donc tenu à la méthode hypergéométrique qui n'est pas la plus économique mais qui reste la plus sûre. Il est en effet possible d'obtenir directement la probabilité de la coccurrence observée, grâce à une itération du calcul hypergéométrique pour cumuler les probabilités partielles de 1 à k (k étant la valeur observée). Voir le détail de ce calcul dans une communication de Serge Heiden *Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex*, publiée dans *JADT04, Le poids des mots*, p 578-588 et commodément accessible sur Internet à l'adresse : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_055.pdf. Ajoutons que cet article apporte une contribution fondamentale au traitement des cooccurrences. On y trouvera en particulier sous le nom de lexicogramme une représentation graphique très séduisante des réseaux lexicaux.

pôle peuvent avoir entre eux des liaisons qui sont explorées, phrase après phrase dans le texte. Il en résulte un tableau de cooccurrences, représenté dans un graphe.

5 - La fonction ASSOCIATIONS généralise cette démarche et l'étend au corpus entier. En s'appuyant sur la fréquence, une liste des mots pleins est d'abord constituée et donne lieu à un tableau carré de cooccurrences. Quand le tableau est rempli par un balayage complet du corpus, le détail des associations deux à deux est trié et analysé, et une représentation, sous forme de graphe, est proposée pour rendre compte des liens préférentiels qui tissent un réseau autour de chaque élément du tableau.

2. Les corrélats

Le programme CORRÉLATS commence par établir une liste de mots (au moins les substantifs, adjectifs ou verbes qui sont les plus fréquents) et enregistre toutes leurs rencontres, occasionnelles ou insistantes, dans la même page. Un lien est établi entre deux mots quand ils ont tendance à se donner rendez-vous. La « tendance » tient compte du nombre de cooccurrences (compte tenu de la fréquence respective des deux mots). Le registre est tenu dans un tableau carré où les mêmes éléments sont portés sur les lignes et les colonnes.

L'option en faveur du paragraphe ne permet pas d'échapper totalement aux contraintes syntaxiques mais l'élimination des mots fréquents et des mots-outils concourt à privilégier les relations sémantiques ou thématiques plutôt que les rapports de dépendance grammaticale. On notera que la division en textes est ignorée. La cohabitation à longue distance dans un même texte n'entre pas dans le calcul. Seule compte la proximité immédiate dans la même page, là où l'on a le plus de chances de relever les isotopies.

Le choix des termes est enclenché par le bouton *Prépare* de la page *Corrélats*. Compte tenu de l'étendue du corpus, le programme de sélection s'arrange pour retenir entre 300 et 400 candidats parmi les mots-pleins (substantifs, adjectifs et verbes, ensemble ou séparément)[3]. Ensuite vient une phase, assez longue, d'exploration séquentielle du corpus. Dans chaque paragraphe on teste la présence ou l'absence des éléments de la liste, en notant les cooccurrences. Le tableau final est soumis à un programme d'analyse factorielle de correspondance, plus puissant que celui qu'Hyperbase utilisait jusqu'ici (ANCORR.EXE). Écrit en fortran par Ludovic Lebart dans les années 70, le programme LX2ACL.EXE n'est pas limité comme le précédent à 75 colonnes. On a fixé à 400 le seuil supérieur mais on pourrait doubler ce chiffre, n'était la nécessité de rendre lisible le résultat. Celui qu'on obtient pour notre corpus de démonstration[4], en sollicitant le bouton *Graphique,* est déjà suffisamment encombré (figure ci-dessous). L'interprétation en est pourtant assez claire : de la gauche à la droite on passe du concret à l'abstrait[5]. Quant au deuxième facteur, il paraît séparer ce qui relève de la société et ce qui appartient à l'individu.

[3] La liste une fois établie reste modifiable. On peut y supprimer les indésirables. Même lorsque les calculs ont été exécutés, il est possible de les reprendre, en neutralisant soit un seul élément, soit une ligne entière (ou colonne) du tableau.

[4] Ce corpus présenté aux JADT 2002, puis aux JADT 2006, rassemble 22 romans, de Marivaux à Proust, à raison de deux textes par écrivain. Constitué pour expérimenter la distance hypertextuelle et différencier les auteurs, il n'a pas l'homogénéité qu'on requiert généralement dans un corpus.

[5] Pour la lisibilité de la figure, on a réduit à une centaine le nombre des mots proposés, dans la gamme des fréquences moyennes. Quand l'échantillon, plus large, enveloppe les hautes fréquences, le graphique devient plus probant.

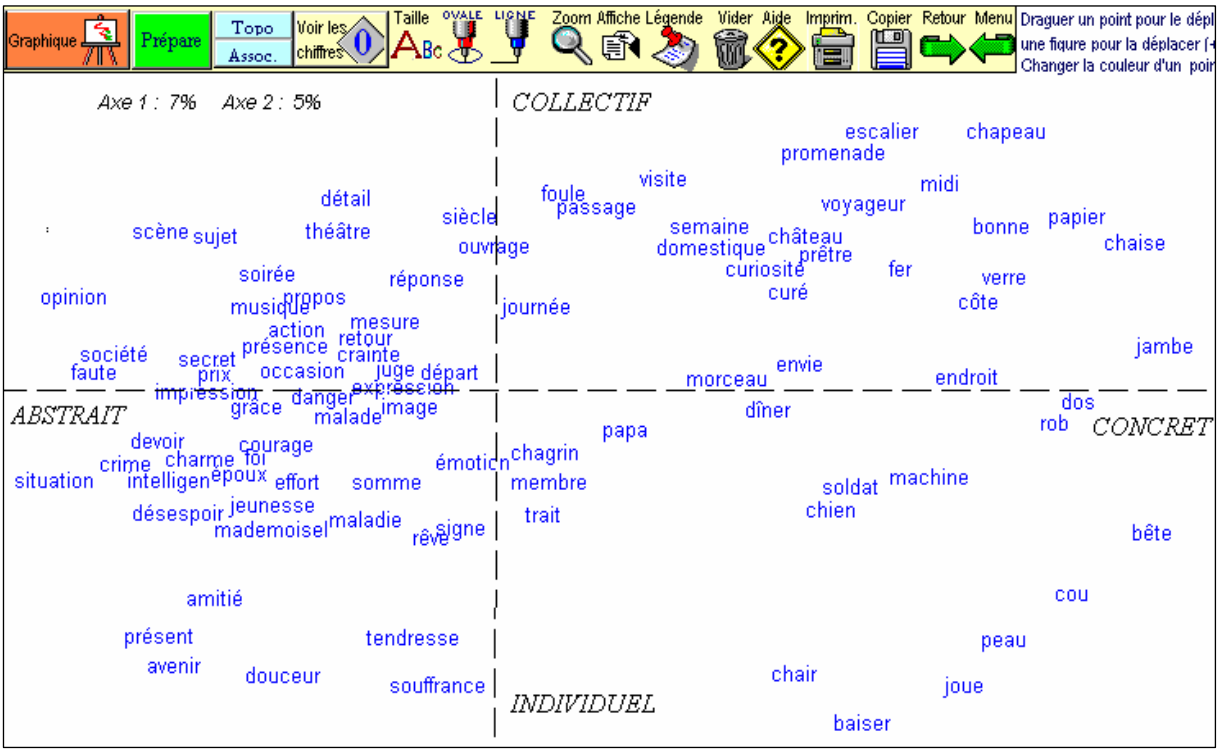

Figure 1. Analyse factorielle des corrélats dans le corpus EXEMPLEM

Chose curieuse, cette structure se retrouve dans des monographies plus cohérentes, comme celles de Flaubert, de Stendhal, de Zola, de Proust, de Gracq. Sans doute doit-on y voir non pas le partage des mêmes thèmes mais un reflet de la composition romanesque, qui dispense dans le même roman des développements narratifs, descriptifs, dialogués ou réflexifs et qui s'impose d'un auteur à l'autre. Cette spécialisation de l'écriture, dictée par la situation où la plume intervient dans le cours de la rédaction, agit comme la loi du genre, mais à l'échelle de la microstructure.

L'exploitation du tableau généralisé des cooccurrences ne se réduit pas à cette synthèse rapide. Il sert de base à des représentations moins synthétiques mais plus fines que nous exposerons plus loin.

3. Un pont vers ALCESTE

La procédure qu'on vient d'exposer donne un avant-goût de ce que réalise *Alceste*. Le point de départ est le même : un réseau de mots associés. Mais dans *Alceste* la notion de cooccurrence est en principe plus étroite, puisqu'elle s'exerce dans des unités plus courtes de 2 ou 3 lignes, et non à l'échelle du paragraphe ou de la page, du moins lorsque les données fournies sont des textes suivis. De plus le calcul ne porte pas sur un échantillon de 400 mots-pleins, mais sur l'ensemble du vocabulaire. Enfin les résultats sont décantés par des filtres discriminants qui séparent les classes et les thèmes, au lieu que notre programme de *Corrélats* présente les alliances et les oppositions en une chaîne continue où les thèmes se succèdent en fondu-enchaîné.

Aussi bien avons-nous jeté un pont vers *Alceste*, sans pouvoir, hélas, fournir ce logiciel qui est un produit du commerce. Ceux qui le possèdent n'auront pas à se soucier de préparer les données. *Hyperbase* s'en charge si l'on actionne le bouton *Préparation des données*. Comme précédemment un seuil minimal et maximal est fixé selon la taille du corpus pour constituer

un échantillon d'un millier de substantifs[6]. Et de la même façon chacune des pages est explorée et réduite à une suite d'une dizaine de mots, qui appartiennent à la liste préétablie et qui sont présents dans la page considérée. *Alceste* considèrera ces extraits comme des *unités de contexte élémentaires*, sur lesquelles s'exerce son algorithme quand l'ordre de *lancer Alceste* est donné. Précisons que le paramétrage est le plus simple et qu'il n'y a pas lieu de cocher la case relative à la lemmatisation, puisque les données sont déjà lemmatisées. Dès lors l'utilisateur quitte momentanément Hyperbase et peut à loisir recueillir et commenter les résultats produits par *Alceste,* comme nous l'avons fait pour les 5 000 pages de l'oeuvre romanesque de Flaubert.

Huit classes ont été distinguées, auxquelles on doit donner un nom qui les résume au mieux, comme on fait pour les facteurs d'une analyse factorielle. Mais la liste des mots qui constitue une classe est suffisamment suggestive pour expliciter la classe (ou le thème), d'autant que le programme délivre une indication précieuse : les textes où le thème est exploité. Qu'il s'agisse des textes ou des mots, un Chi2 mesure l'appartenance plus ou moins étroite à la classe en question. Dans l'exemple de la figure 2, un extrait, même court, de la liste suffit à isoler les questions philosophiques et religieuses qui préoccupent Flaubert dans ses premiers écrits et qui se maintiennent dans les trois versions de la *Tentation de Saint Antoine*.

VARIABLES DE LA CLASSE N°2

Identification	u.c.e total classées	u.c.e. dans la classe	Khi2
*49Antoine	635	308	695.43
*Smarh	230	137	407.65
*56Antoine	318	137	232.57
*74Antoine	286	81	49.96
*Mémoires	135	40	27.37
*Novembre	223	39	2.16

FORMES REPRESENTATIVES DE LA CLASSE N°2

Khi2	u.c.e. dans la classe	Formes réduites	Khi2	u.c.e. dans la classe	Formes réduites
366.52	75	luxure	172.91	36	avarice
341.71	68	péché	147.71	68	colère
255.03	58	logique	137.27	49	seigneur
240.29	52	jésus	131.01	63	chair
205.14	46	éternité	124.55	51	foi
201.15	47	néant	108.00	36	création
198.57	44	enfer			
181.99	42	christ			

Figure 2. Une classe isolée par Alceste dans le corpus de Flaubert

[6] On voit que la limite de 400 éléments a été reculée par rapport au programme *Corrélats*. Cela tient au fait qu'il n'est plus nécessaire de représenter les résultats dans une figure unique, où mille mots ne peuvent pas trouver place sans nuire à la lisibilité.

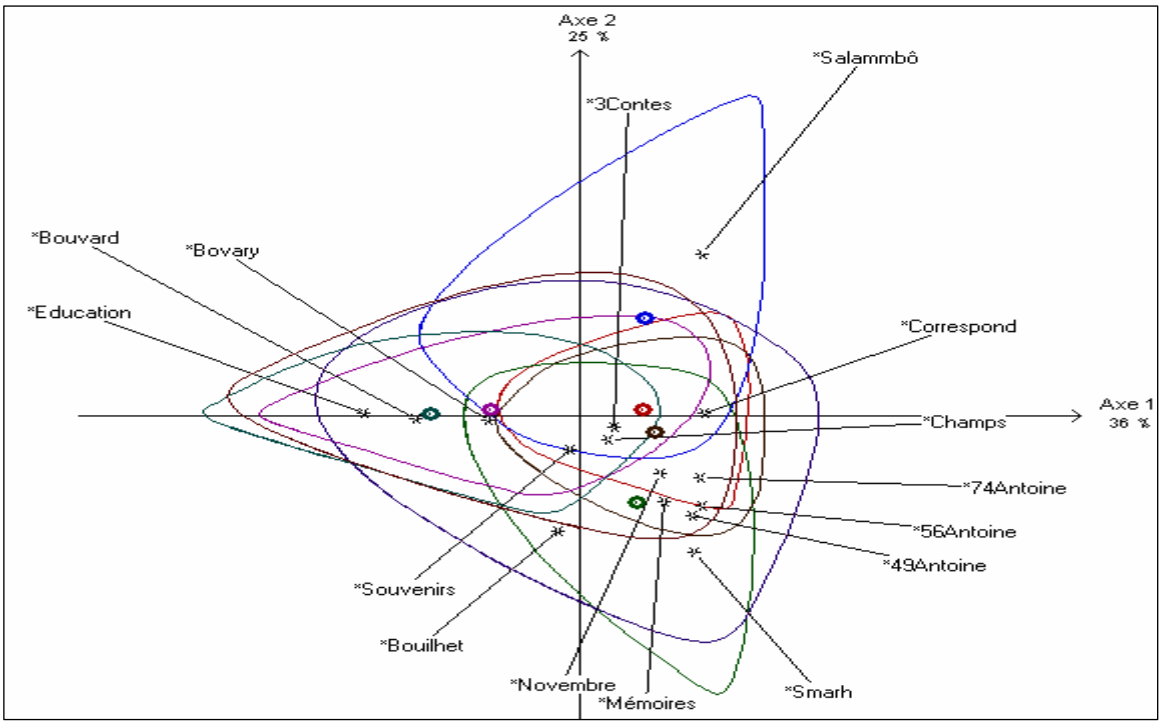

Figure 3. Analyse factorielle faite par Alceste sur les substantifs dans l'œuvre de Flaubert

Les résultats sont distribués par *Alceste* dans une multitude de fichiers où l'utilisateur peut se référer en différé. Il en est un qu'*Hyperbase* rapatrie plus particulièrement : c'est le résumé de l'analyse, qui détaille le contenu des classes et dresse la carte des thèmes en y incorporant les « variables étoilées » c'est-à-dire le nom des textes du corpus. Certes les jalons textuels n'ont eu aucune influence sur les calculs, mais une fois que les classes ont été établies, les textes sont invités à choisir leur camp. C'est ce que montre l'analyse factorielle ci-dessus, où les huit classes occupent un espace particulier du graphique (avec un point de la même couleur au centre de gravité de cet espace).

Chaque mot du corpus peut y prendre place (on s'en est abstenu pour éviter l'encombrement) mais aussi les textes eux-mêmes qui prennent position selon leurs affinités avec les classes établies : dans la moitié supérieure la trilogie des romans modernes, à gauche, s'oppose à *Salammbô,* à droite ; dans la partie inférieure se retrouvent les textes autobiographiques des débuts de l'écrivain, à gauche, et, à droite, les tentations répétées de *Saint Antoine*.

L'analyse brute proposée dans les corrélats apparaît comme une ébauche affinée dans *Alceste*. La première n'est qu'un nuage de mots, sans autre repères que les points cardinaux. Dans le seconde des lignes de démarcation apparaissent, délimitant des constellations lexicales identifiables, tandis que la carte des textes, en surimpression, facilite l'interprétation. On aurait pu en rester là et s'en tenir à une relation client-fournisseur, *Hyperbase* préparant les données pour *Alceste* avant d'en recevoir les leçons.

4. Les associations privilégiées

Si l'exploration entière du corpus est nécessaire pour offrir à *Alceste* les données particulières qu'il réclame, en revanche on n'a pas à renouveler ce long balayage, pour approfondir le

réseau des associations, du moins si le tableau général des cooccurrences a déjà été constitué[7]. La recherche sur les associations s'appuie en effet sur le tableau des cooccurrences, dont la fonction CORRÉLATS a d'abord fourni une vue d'ensemble, sous forme d'analyse factorielle. La carte thématique du corpus y apparaît très claire, mais peut-être trop, car elle souligne assez trivialement les oppositions qui se font jour dans le lexique entre concret et abstrait, collectif et individuel, et qui se réalisent habituellement dans le discours romanesque. Il convient donc de répondre à des questions plus ciblées et de proposer des zooms sur des zones précises du vocabulaire.

Pour une base nouvelle, en supposant qu'on a déjà créé la liste des substantifs (ou verbes ou adjectifs) retenus et qu'on dispose du tableau général des cooccurrences, il faut procéder au calcul et au tri de tous les indices qui évaluent la distance entre les mots pris deux à deux. Ce rôle est joué par le calcul hypergéométrique, comme expliqué plus haut. Le seuil minimal de cet indice est établi par défaut à une valeur convenable vu la taille du corpus. Mais on peut le modifier et renouveler les calculs, ou plus simplement choisir un seuil plus lâche ou plus sévère au moment de la représentation graphique. Quand le calcul a été exécuté pour l'ensemble du tableau, le résultat n'en est pas un tableau de même taille où les éléments nuls seraient majoritaires et encombrants, mais une liste épurée qui détaille les associations privilégiées et abandonne les autres. C'est cette liste ordonnée qui est désormais consultée pour les recherches ultérieures.

test motl mot2		
620.71 point vue	87.34 ciel soleil	71.38 mort vie
550.48 hôtel maître	81.28 expression visage	68.87 jour lendemain
549.35 chef œuvre	80.76 femme mari	67.95 chambre porte
308.61 gens monde	78.61 lèvre sourire	66.96 eau mer
199.20 art oeuvre	77.28 oeil visage	66.78 homme monde
178.59 maison maîtresse	75.72 coup oeil	64.81 ombre soleil
159.07 bord mer	75.69 fenêtre soleil	64.45 amour madame
155.86 femme homme	75.10 mémoire souvenir	64.23 cheveu oeil
124.38 mer soleil	73.87 ciel mer	64.10 madame vie
124.24 oeil regard	73.84 heure matin	63.43 salle table
119.08 lumière soleil	73.84 duc frère	63.43 amour femme
118.47 mère père	73.80 escalier porte	62.88 matin soir
101.67 chambre lit	73.23 duc prince	62.85 frère soeur
97.68 regard sourire	73.03 chambre fenêtre	61.35 amour sentiment
90.58 eau soleil	72.46 mot sens	
89.03 bord eau	72.11 larme oeil	

Tableau 4. Les associations que Proust privilégie parmi les substantifs (extrait très partiel)

Dans l'extrait qui en est livré dans le tableau 4 et qui est relatif à la *Recherche du temps perdu*, on ne s'arrêtera pas aux premières associations qui relèvent de la phraséologie et même du lexique et qu'une lemmatisation étendue aux mots composés aurait dû éliminer. Mais ces scories (*point (de) vue, maître (d') hôtel, chef (d')œuvre, œuvre (d')art*) n'entachent que la tête de liste, comme une écume mal dissoute. Dès que le coefficient échappe aux cooccurrences triviales et fixées dans la langue, des couples solides apparaissent, *femme-homme, père-mère, femme-mari, mort-vie, matin-soir, frère-sœur*, dont le lien tient à la sémantique et à l'attirance magnétique que les mots opposés exercent l'un sur l'autre comme les pôles d'un aimant. Mais le plus souvent les couples se forment par le partage de goûts et

[7] Si ce n'est pas le cas, la fonction ASSOCIATIONS déclenche le départ de l'exploration et lance le programme CORRÉLATS.

de sèmes communs, par quelque raison métonymique, comme la relation de la partie au tout ou de la cause à l'effet, ou la proximité dans l'espace ou le temps.

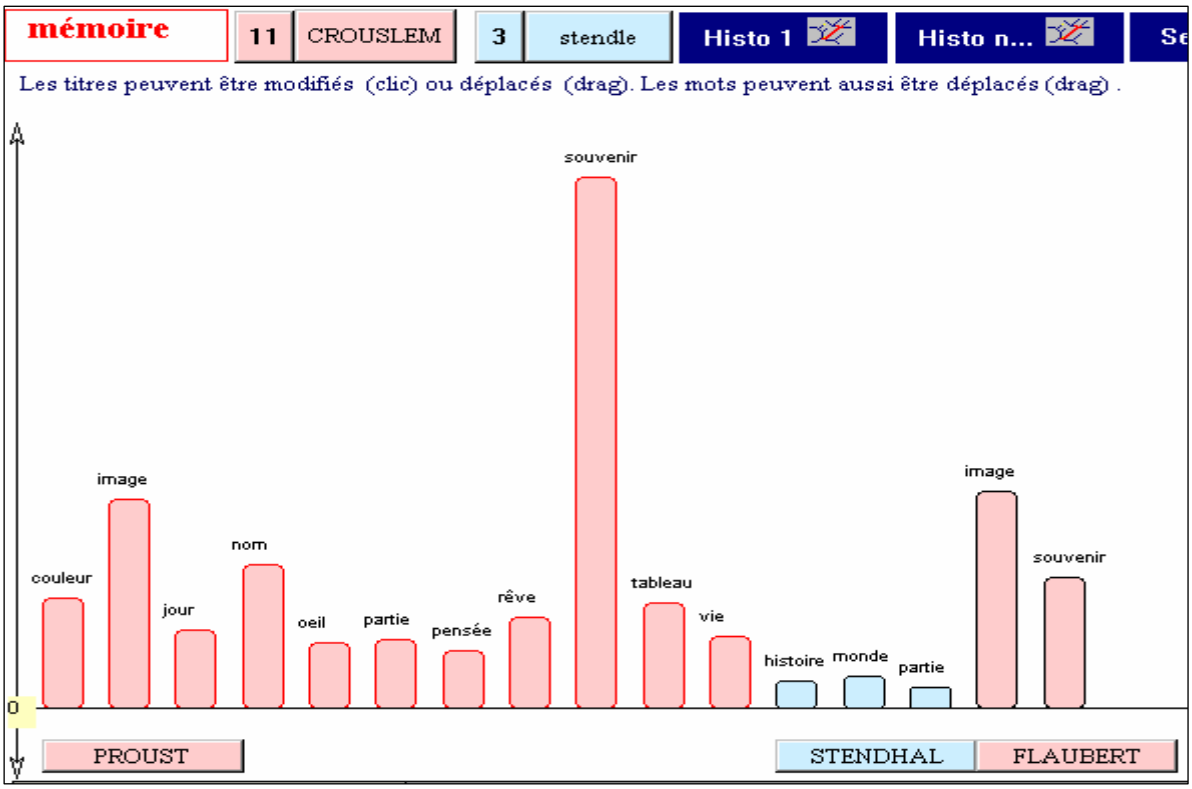

Figure 5. La constellation lexicale autour de la mémoire chez Proust, Stendhal et Flaubert

1 – On peut tout d'abord isoler une ligne du tableau (en cherchant ce qu'il en reste dans la liste, quand tous les éléments nuls et non significatifs ont été expurgés) et la transposer dans un histogramme. Cette représentation simple est disponible, quoique réductrice. Elle a pourtant son intérêt si le mot représenté dans son environnement lexical est recherché de la même façon dans d'autres corpus. Les fréquences brutes du mot en question peuvent être semblables ou différentes dans ces corpus comparés, ce n'est pas là ce qui compte. On ne se soucie que de confronter leur entourage respectif, selon le principe « dis-moi qui tu fréquentes et je te dirai qui tu es ». S'agissant de Proust, la *mémoire,* qui fait l'objet de la figure 4, étend ses synapses dans une large zone du vocabulaire, où l'on retrouve les isotopies attendues (*souvenir, image, nom*) mais aussi les connotations esthétiques (*tableau, œil, couleur*) ou morales (*vie, rêve, pensée*) de la conscience proustienne. La zone de la *mémoire* est bien plus étroite chez Flaubert, et chez Stendhal elle est moins personnelle que sociale et se confond avec l'*histoire*.

2 – Entre la vue lointaine de l'analyse factorielle (figure 1) et le détail myope de l'histogramme (figure 4), il y a place pour un échelon intermédiaire : comme précédemment on commence par s'attacher à un mot parmi les 400 disponibles. Une fois que l'hameçon est accroché, on tire sur le fil et on sort de l'eau non seulement les mots-amis qui sont liés au

mot-pôle, mais aussi ceux qui sont proches de ces proches. L'enquête qui s'étend donc aux amis des amis vise à dessiner un réseau complexe autour du pôle[8].

Nous prendrons le même mot *mémoire* à l'intérieur de la *Recherche du temps perdu*. Les liens représentés dans le graphe 6 sont en rouge s'ils concernent le mot-pôle (ce sont ceux qui ont donné matière à l'histogramme 3), ils sont en bleu s'ils concernent les mots liés au pôle et en noir dans les autres cas. Les mots eux-mêmes sont différenciés par la couleur : le rouge est réservé aux noeuds fréquentés, le noir aux noeuds isolés (moins de 5 liaisons). La force des liaisons influe sur l'épaisseur des traits et la taille des caractères. Si les relations collatérales (en noir) encombrent sans profit le graphique, on peut les faire disparaître (ou les rétablir) en faisant appel au bouton SIMPLIFIER/ENRICHIR. Le calcul du graphe arborescent et de la position des noeuds et des arcs est assuré par le logiciel libre *GRAPHVIZ* (licence GNU) aimablement communiqué par Serge Heiden. Les données sont fournies à ce programme selon les spécifications du langage *DOT* et les résultats bruts, enregistrés dans un fichier au suffixe.DOT, sont repris par *Hyperbase* dans une présentation graphique qui tient compte non seulement des positions mais aussi des pondérations[9].

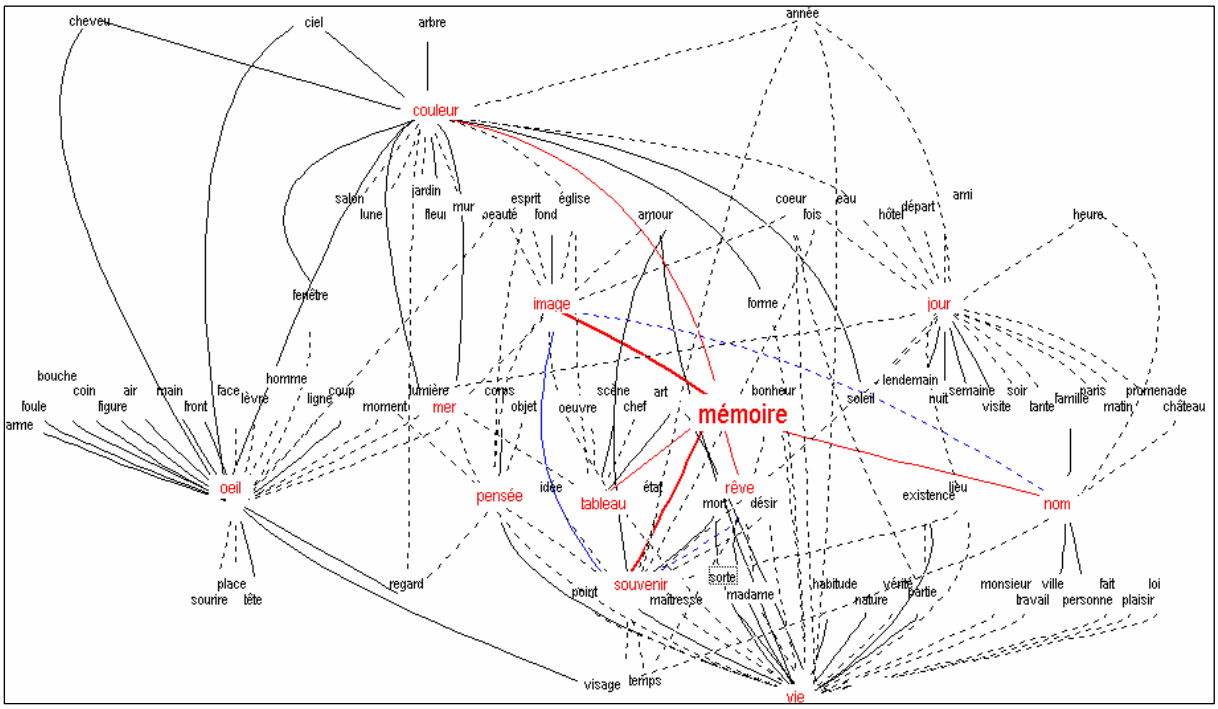

Figure 6. Le graphe de la mémoire dans la Recherche du temps perdu

En réalité le logiciel *GRAPHVIZ* ignore les poids et les pondérations et ne veut connaître pour chaque élément du tableau des cooccurrences qu'une information grossière du type

[8] Des raisons pratiques nous ont dissuadé d'approfondir encore le champ exploré et d'envisager un troisième niveau. À chaque étage le champ s'élargit en effet comme le carré du précédent et on aurait vite atteint les limites d'un tableau pourtant gros de 16000 éléments. En outre la polysémie qu'on peut rencontrer dans le mot-pôle et à chaque étage du réseau produit beaucoup de dispersion et le nuage des points s'effiloche au gré des courants et diversions polysémiques.

[9] Nous n'avons utilisé GRAPGVIZ que pour le calcul de la position des arcs et des points. Quant au dessin des arcs, nous avons aménagé leur courbure pour faciliter l'analyse. Il est possible d'accentuer ou d'atténuer cette courbure avec la souris et de déplacer légèrement un pot lorsqu'il se produit un recouvrement gênant.

présence/absence, comme si l'on circulait dans un réseau binaire avec des portes ouvertes ou fermées mais non entrebâillées ou grandes ouvertes. Comme pour chaque arc nous connaissons la force d'attraction calculée par l'hypergéométrique, nous avons pu réintroduire cette information en épaississant les traits ou en grossissant les caractères. Mais on aurait aimé que le dessin du graphe soit ordonné en tenant compte non seulement de l'existence d'un lien mais aussi de la mesure de son intensité. Ces sortes de graphe sont vite illisibles et on regrette qu'un élément de clarté ait été négligé.

3 – On a constitué un tableau carré où les valeurs de proximité se substituent à la mesure brute des cooccurrences. Les effectifs absolus sont en effet trop dépendants de la fréquence des mots. Dans l'approche qui précède, c'est ce tableau entier qui est exploré, les ramifications pouvant aller loin quand elles se communiquent de voisin à voisin. Mais on peut fixer une barrière à cette propagation, en constituant un sous-tableau, carré comme le grand, et qui porte en marge des lignes et des colonnes la liste des mots directement liés au pôle. Un tel tableau contient les cooccurrences pondérées des uns avec les autres, en excluant précisément le pôle et en neutralisant les liens de chacun avec ce pôle. En somme une séance à huis clos, où les gens en relation avec l'intéressé sont invités à porter leur témoignage en son absence.

Ce sous-tableau est alors soumis aux méthodes habituelles : analyse factorielle de correspondance et analyse arborée. Deux boutons sont disponibles à cet effet et s'appliquent au mot qui a été choisi pour pôle. Nous éclairerons cette procédure avec le mot « nuit », emprunté à la base EXEMPLEM. Le graphe obtenu est encore plus complexe que celui de la mémoire. Des zébrures dans tous les sens y traversent l'espace et font penser à un feu d'artifice nocturne, au point qu'on a renoncé à le montrer. Tout se simplifie pourtant dans l'analyse arborée qui en garde la trace et en souligne la structure : la nuit peut être considérée d'abord sous l'aspect temporel ; elle voisine alors avec les autres unités de temps, celles qui sont à sa mesure, comme le *jour*, le *soir* et le *matin* et celles qui sont à une échelle différente comme l'*heure* et la *semaine*. La nuit s'exprime aussi dans l'espace : elle évoque d'une part le lieu clos de la chambre à coucher (*maison, chambre, lit, fenêtre*) et d'autre part l'atmosphère nocturne à ciel ouvert (*ciel, lune, feu, milieu, ombre, silence*).

Certes ce schéma n'a pas la même précision qu'obtient Bruno Gaume, avec des méthodes semblables, dans la représentation graphique des verbes du *Grand Robert*. Mais nous partons ici non d'un dictionnaire mais de textes littéraires, où les mots voguent en liberté sans s'enfermer dans des définitions circulaires[10].

[10] Gaume B. (2006) *La proxémie : vers un modèle de sémantique lexicale pour un Traitement automatique des langues à ergonomie cognitive*, http://www.limsi.fr/Individu/habert//04-05/inex.html. B Gaume a mis au point un logiciel graphique, qui semble grandement supérieur à Gr*aphviz*. On aimerait que sa diffusion soit rendue possible.

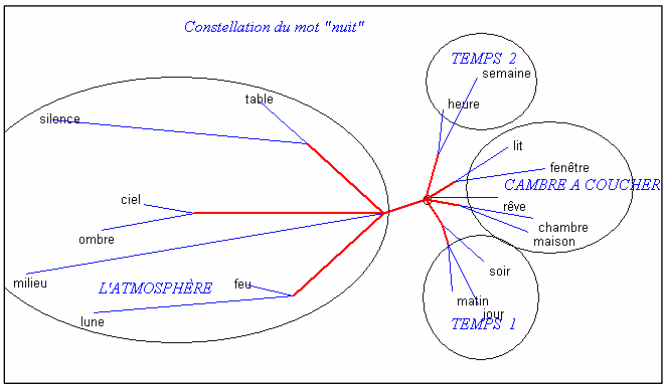

Figure 7. La constellation de la nuit dans la base Exemplem

5. La recherche thématique

Les outils qu'on vient d'employer (hypergéométrique, histogramme, analyses arborée et factorielle) peuvent encore servir une ambition plus pure. Car il reste une part d'arbitraire dans l'approche précédente. Pourquoi ne retenir que 400 substantifs ? Comme s'il suffisait de côtoyer le même nombre de députés à l'Assemblée nationale pour connaître la France. On peut certes choisir un autre échantillon, admettre les verbes et les adjectifs, doser des parités égalitaires, élargir le critère censitaire de la fréquence. Le filtre n'en est pas moins réducteur même s'il s'applique à toute la population des mots du corpus. Y a-t-il moyen d'agrandir assez les mailles du filtre pour qu'aucun mot ne soit rejeté, même les mots-outils, tout en maintenant l'exploration sur le corpus intégral ? En somme on voudrait recenser toutes les combinaisons possibles, et cela dans le texte entier. C'est ce que fait l'indexation, non pour les combinaisons de mot, mais pour les mots individuels. C'est ce que fait le logiciel *Lexico* au moins pour cette espèce particulière de combinaison qui explore les suites de mots adjacents et qu'on connaît sous le nom de segments répétés[11]. C'est enfin ce que tente le logiciel *Alceste* avec un succès certain. Notre ambition est plus modeste : l'exploration reste bien exhaustive, et les combinaisons sans limite, mais il y a une contrainte initiale. Le rayon laser peut balayer tout l'espace, mais il est attaché à une position. Il faut adopter un point de vue, c'est-à-dire partir d'un mot ou de quelque objet linguistique précis. Le plus performant est le lemme.

La première étape consiste à réunir tous les contextes où se rencontre le mot choisi, en veillant à conserver les passages sous la forme d'une suite de lemmes. Ces contextes cousus les uns aux autres forment un sous-ensemble qu'on soumet à l'indexation. Il en résulte une liste de fréquences qui est comparée au dictionnaire des fréquences établi pour le corpus entier. On aboutit alors à une liste de spécificités, qui met en relief les mots associés le plus souvent au mot choisi pour pôle. Cette procédure n'est pas nouvelle et elle rend des services depuis dix ans dans le logiciel *Hyperbase*. Mais on a songé à lui donner des prolongements nouveaux. Il suffit de considérer le sous-ensemble comme un corpus autonome et d'y rechercher la liste des spécificités avec les procédures établies pour les corrélats et les

[11] C'est aussi la même voie que nous suivons dans l'examen des « codes répétés ». Il s'agit de segments constitués non de mots, mais de codes adjacents. Nous avons appelé improprement cet objet structure, ce qui supposerait une analyse plus approfondie. Ces séquences grammaticales se prêtent à toutes les manipulations statistiques, et notamment les plus simples, réduites à deux ou trois éléments (bicodes et tricodes)

associations privilégiées. Naturellement on ne cherche pas les cooccurrences avec le mot-pôle, puisqu'on les connaît déjà et que l'écart réduit les a désignées et mesurées. Mais ce qu'on sait moins ce sont les relations que les spécificités réunies autour du pôle peuvent avoir entre elles. Il peut se faire que le pôle soit polysémique et que la tribu qui l'entoure ne soit pas homogène comme il arrive dans les familles recomposées où il reste des grumeaux dans la pâte familiale. Dans cette analyse spectrale des cooccurrents il est cependant préférable – mais non obligatoire - d'écarter les mots-outils dont la distribution obéit à des contraintes extérieures, sans qu'on saisisse toujours le rapport précis avec le mot-pôle[12], d'autant que leur fréquence intempestive tend à écraser le reste[13]. Mais ce reste est riche de tous les mots sémantiques, sans que le code ou la fréquence puisse justifier leur exclusion.

La figure 8 montre le gain en précision et en extension, bien que pour des raisons de lisibilité on ait réduit à 26 mots cooccurrents de la *nuit* les 69 propositions du calcul[14]. Outre les substantifs fréquents qui tournent autour de la *nuit* (*jour, ciel, heure, ombre, lune*) et qui figuraient déjà parmi les corrélats dans l'analyse précédente, on en distingue d'autres qui ne sont pas dans la liste des notables : *brouillard, astre chemise*. Surtout, avec l'intervention des verbes (*passer, éclairer, rentrer, réveiller, coucher, dormir*) et des adjectifs (*clair, pâle, noir, tombant, mille*), l'interprétation est invitée à ne plus se contenter des rapports sémantiques. Car c'est la phraséologie qui place en tête le verbe *passer* et qui explique la présence de *mille*. On peut aller plus loin encore dans cette direction en acceptant les mots-outils.

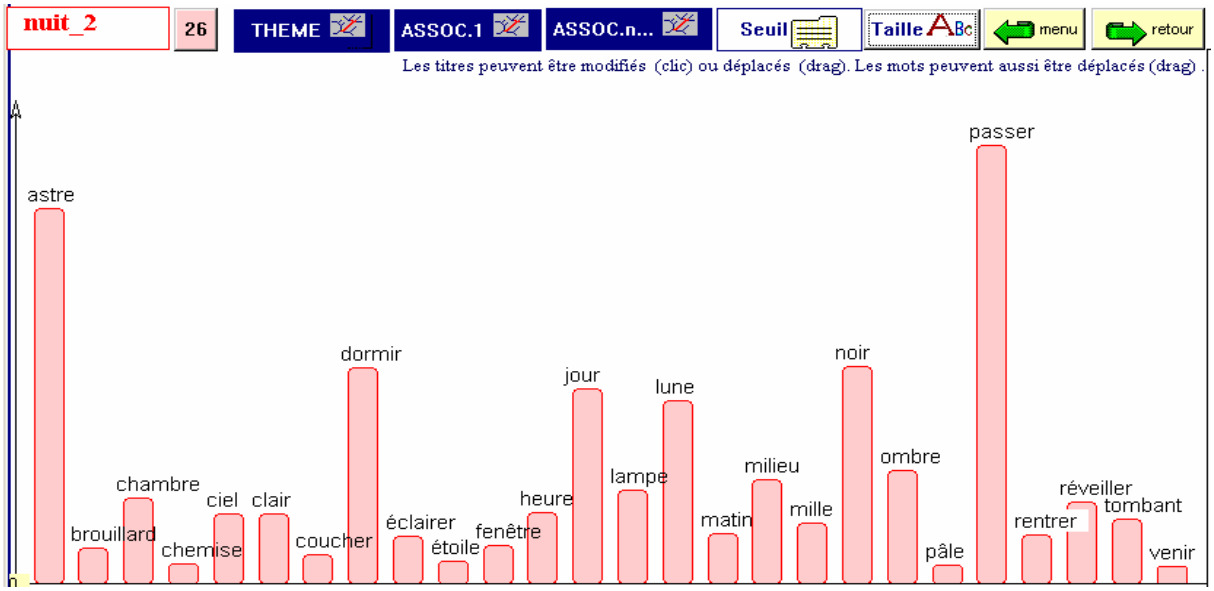

Figure 8. Les cooccurrents de la nuit dans la base Exemplem

[12] Ceux qui se trouvent dans la liste des spécificités ne sont pourtant pas là par hasard. Si certains cas restent obscurs, beaucoup s'expliquent pour des raisons triviales : ainsi *la* et *une* sont des acolytes inévitables d'un pôle féminin ; *à, de* s'introduisent systématiquement dans le cercle des cooccurrents si le pôle entre dans des mots composés.

[13] On s'est toutefois prémuni contre l'effet de taille en pondérant les données. Le tableau des cooccurrences brutes est d'abord converti en un tableau des écarts réduits, en ne conservant que ceux qui sont positifs.

[14] On peut sélectionner les mots qui atteignent une certaine fréquence en laissant les autres dans l'ombre.

Bien entendu le programme Graphe s'applique à de telles données et offre une seconde dimension à la représentation « à plat » du graphique 8. Il rend compte des relations complexes que les cooccurrents de la nuit établissent entre eux[15] (ce sont les traits en bleu dans le graphique 9) ou avec d'autres mots qui n'appartiennent pas au premier cercle (ils n'apparaissent pas dans le graphique 9 qui a été volontairement simplifié, en vertu d'une option disponible).

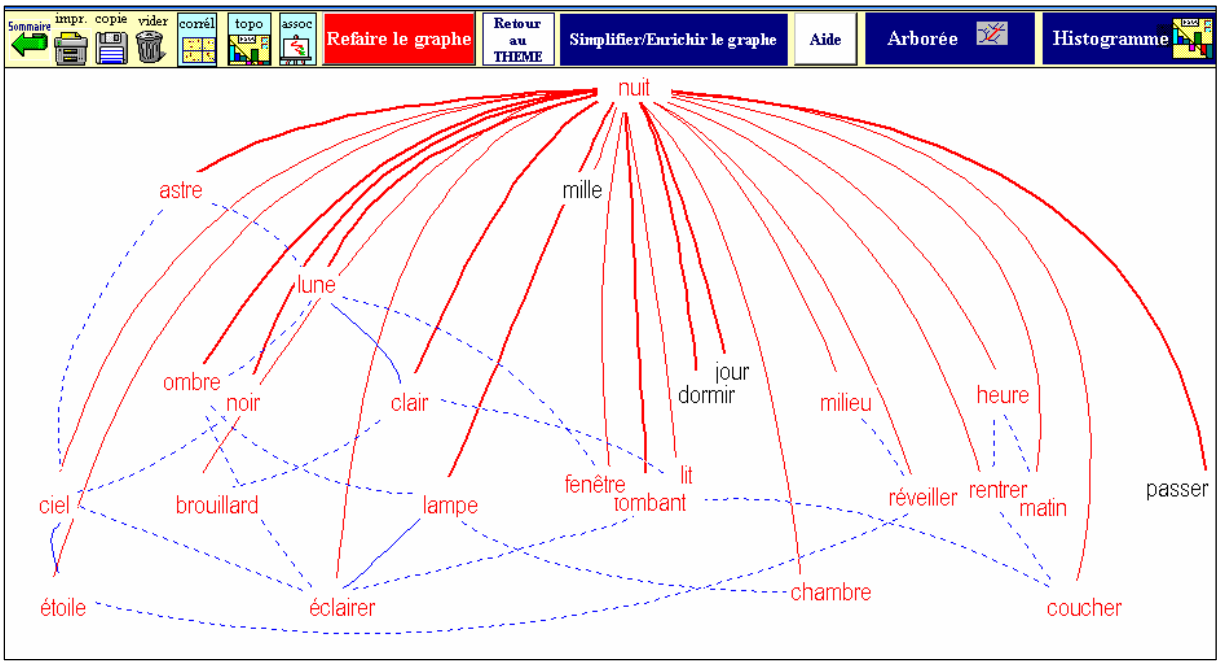

Figure 9. Le graphe de la nuit dans la fonction THEME

Quand un mot a peu de liaisons, cela signifie souvent qu'il est attaché au mot-pôle par un lien exclusif qui le rend indifférent au reste et relève de la phraséologie (*passer la nuit, les mille et une nuits*). Dans le cas contraire, les relations sélectives qu'il peut avoir l'orientent dans un réseau ou dans un autre. On en distingue deux dans le graphique : à gauche c'est le réseau extérieur de la nuit, celui des constellations du ciel nocturne (*ciel, astre, lune, étoile, ombre, clair, noir*). À gauche la nuit est plus humaine, elle est liée au sommeil (*dormir, coucher*), au cadre intime qui entoure le repos (*fenêtre, lampe, chambre, lit*) et aux moments de transition quand on entre dans la nuit ou quand on en sort (*rentrer, soir, heure, matin, réveiller*).

Pour confirmer cette interprétation, il reste une méthode générale que nous avons déjà utilisée dans l'étude des corrélats et qui est affranchie de tout seuil ou paramètre : l'analyse factorielle de correspondance. Livrons à la machine deux éléments : le sous-corpus bâti autour d'un mot et la liste des spécificités qui en est extraite. Le traitement va constituer le tableau des cooccurrences des mots de cette liste et le livrer au calcul factoriel. Le résultat, toujours pour le mot *nuit*, apparaît ci-dessous :

[15] Le calcul est aussi plus complexe. Rien n'est changé pour les couples où entre le mot-pole (ici la *nuit*). Les paramètres *N,f,t* et *k* sont ceux du corpus, puisque le corpus a été intégralement dépouillé pour tous les couples où le mot-pôle est partie prenante , comme expliqué dans la note 2. Mais il n'en va pas de même pour les relations où le mot-pôle n'est pas directement en cause. Celles-ci ne sont explorées et comptabilisées que dans le sous-ensemble constitué par le programme *Contexte* et non dans le corpus entier. Les paramètres *N,f* et *t* doivent donc être adaptés et correspondre au sous-ensemble.

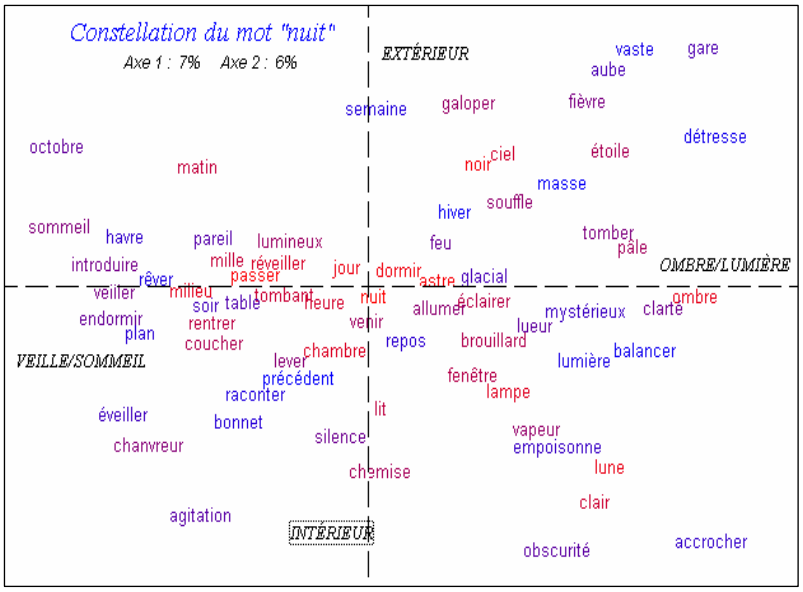

Figure 10. La constellation de la nuit, à travers l'analyse factorielle

Comme on pouvait s'y attendre, la *nuit*, qui enveloppe par définition tous les paragraphes, s'installe au centre de la toile et attire au voisinage immédiat ses acolytes les plus dévoués : *jour, dormir, astre, tombant*. La couleur aide à comprendre les choses puisque les mots les plus proches de la nuit dans la liste des spécificités sont en rouge et les moins proches en bleu, avec tous les dégradés intermédiaires. Le rouge vif est visiblement plus présent dans la zone centrale. Mais dès qu'on s'éloigne de l'origine les couleurs se mélangent et d'autres forces interviennent pour la localisation des points. Le facteur 1 représente l'opposition entre deux faces de la nuit : l'une est faite d'ombre et de lumière, l'autre veille sur le sommeil. Le second facteur s'oriente de l'intérieur à l'extérieur. En bas c'est le cadre, intime et chaud, de la chambre à coucher, en haut l'espace froid des étoiles.

6. Conclusion

En conclusion, on n'est pas certain que les calculs puissent nous faire découvrir le sens des mots. Les collocations ne sont pas nécessairement des connotations et la phraséologie ordinaire joue dans beaucoup de cooccurrences un rôle majeur qu'on peut trouver gênant, sauf si c'est là ce qu'on cherche. Si le sens d'un mot est la somme de ses emplois, il faut pourtant recenser ces emplois, en espérant ramasser dans le filet électronique quelques-unes des isotopies, dont le rayonnement parcourt les textes littéraires sans être facilement observable. Quel dommage que la physique ait tant d'outils pour déceler les multiples rayonnements dans le ciel étoilé au dessus de nos têtes et que la critique en ait si peu pour les rayonnements textuels.

Références

Brunet E. (2006). Navigations dans les rafales. In *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles*. Besançon, Presses Universitaires de Franche-Comté, pp.15-29. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_EB.pdf

Gaume B. (2006). *La proxémie : vers un modèle de sémantique lexicale pour un Traitement automatique des langues à ergonomie cognitive*.

Heiden S. (2004). Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. In *JADT04, Le poids des mots.* Louvain, Presses Universitaires de Louvain, p.577-588. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_055.pdf

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris, Slatkine-Champion.