Fréquences et séquences. Mise en œuvre dans Hyperbase.

Étienne Brunet

Laboratoire BCL (UMR 6039), Université de Nice, MSH, 98 Bd Edouard Herriot, BP 3209, 06204 Nice cedex 3

RESUMÉ. On tente ici 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. Plusieurs méthodes, relativement convergentes, sont exposées, dont certaines reprennent les voies initiées par Pierre Lafon, Max Reinert, Serge Heiden et André Salem. On décrit leur implémentation dans le logiciel Hyperbase.

MOTS-CLÉS : *Séquences, fréquences, cooccurrences, corrélats, topologie, proxémie, thématique, constellation lexicale, Alceste*

1. Introduction

L'opposition entre les séquences et les fréquences est clairement établie dans la thèse de Pierre Lafon [Lafon 1984]. Les fréquences imposent un traitement tabulaire du corpus, où les mots sont répartis en lignes et les textes en colonnes et la statistique est omniprésente dans le traitement de ce tableau, soit que l'on compare les mots à travers leur distribution dans les textes, soit que l'on compare les textes à travers les mots qu'on y trouve, soit qu'on poursuive en même temps les deux perspectives. Les séquences abolissent au contraire la partition en textes et traitent le corpus d'un seul tenant, en y cherchant les phénomènes de répétition, de concentration, de nouveauté, de voisinage ou de répulsion. Elles ne donnent pas nécessairement lieu à des comptages. Parfois elles entrent dans une procédure purement documentaire où l'on cherche un exemple d'un profil ou pattern jugé rare.

En réalité tout est séquence dans le discours. Les mots sont des séquences de lettres, comme les textes sont aussi des séquences de mots. Tout est affaire de segmentation. La combinatoire les lettres (ou des phonèmes) n'a guère de sens que si on définit une unité qui limite le champ, par exemple l'espace compris entre les délimiteurs, blanc ou ponctuation. Cet espace limité se confond souvent avec le mot mais ce peut être un découpage arbitraire, comme il arrive dans le cas des n-grammes, ou séquences de n caractères, et chaque fois que la chaîne à déchiffrer est inconnue ou traitée comme telle, ce qui est le cas de l'ADN. La combinatoire dépend de la longueur des segments, mais elle dépend aussi de la richesse de l'alphabet. Dans les langues naturelles, à l'exception des idéogrammes, la longueur et la richesse se comptent en unités ou en dizaines, sans atteindre jamais la centaine. Dans la deuxième articulation du langage la combinatoire est ainsi limitée et n'excède pas la taille d'un dictionnaire.

Mais quand on a affaire à la première articulation formée de «mots » ou unités signifiantes, la combinatoire explose, si l'on ne prend pas soin de limiter soit la taille des segments, soit le nombre d'unités recherchées, aptes à constituer une combinaison.

1 - La première solution, qui est celle du logiciel *Alceste*, réduit à une ou plusieurs lignes le champ d'observation où les combinaisons, c'est à dire les cooccurrences, sont retenues. Si en effet les segments exagérément prolongés ont la taille des textes, quelle peut être la valeur d'une combinaison, quand elle concerne deux mots dont l'un se trouve à la première page du texte et l'autre à la dernière ? Cette cohabitation à longue distance, sans doute fortuite et peu significative, peut contribuer – peut-être illégitimement – à rapprocher les deux textes dans le calcul de la distance intertextuelle. Mais on ne saurait en conclure à quelque accointance entre les deux mots. Pour que la cooccurrence prenne sens, il faut qu'elle prenne place dans une séquence de faible empan, qu'elle se produise parmi des mots qui se suivent, dans un environnement de courte portée : une phrase, un paragraphe, une page, un poème, ou, comme dans *Alceste*, un segment aussi arbitraire que la ligne. Quand les mots se touchent et se frottent, qu'il s'agisse d'un heurt ou d'une familiarité, un courant passe de l'un à l'autre qui se perd lorsque les mots ne se voient plus. Les appareils de mesure enregistrent ces rencontres sans décider si la cause est dans la syntaxe, dans la sémantique, ou dans la situation. Pour un segment donné de n mots, un tableau triangulaire de (n*n)/2 éléments suffit à recevoir toute la variété des cooccurrences. Mais la phrase suivante utilisant des mots différents, c'est un tableau cumulatif qui enregistre les rencontres, au fur et à mesure qu'elles se produisent localement. Ce tableau est gigantesque mais plein de trous, puisque la plupart des rencontres théoriquement possibles ne sont pas observées. Alceste utilise un algorithme original pour traiter les zéros multipliés d'un tel tableau. Une solution plus modeste consiste à négliger les mots qui sont rares et dont les cooccurrences sont de ce fait plus rares encore[1]. C'est ce que nous allons développer dans le programme ***Corrélats*** et dans la préparation des données pour ***Alceste***.

2 - L'autre voie invite à renoncer à la segmentation pour explorer le corpus entier. Mais si le champ s'élargit, l'objet de la recherche se rétrécit : on n'envisage plus le réseau exhaustif des cooccurrences, mais certaines d'entre elles, nommément désignées. Le cas le plus simple est celui de la cooccurrence du même au même, c'est-à-dire de la répétition d'un même terme dans le texte étudié. Le hasard s'étonnerait que les occurrences d'un même mot soient distillées de façon trop régulière ou dispersées de manière trop inégale. Le calcul permet de dresser deux listes parmi les mots, en distinguant les distributions «normales », sans relief particulier, et les distributions en rafales. Ce point de départ ne sera pas traité ici, l'ayant été aux JADT de 2006 [Brunet 2006].

On peut envisager la répétition non plus seulement pour un mot isolé, mais pour une séquence de mots. C'est ce que fait le programme Lexico d'André Salem, auquel nous renvoyons le lecteur. La fonction connue sous le nom de « segments répétés » peut être étendue à d'autres objets que les mots et nous le montrerons avec les « **codes répétés** », bicodes, tricodes ou structures syntaxiques récurrentes.

Il faut faire un pas de plus pour traiter les profils ou les patterns. Cette fois l'objet de la recherche n'est plus homogène mais un mélange de variables et de constantes, de codes, de mots et de jokers. L'interrogation est nécessairement plus complexe et nécessite l'emploi «d'expressions régulières » et de grammaires spécialisées. *Frantext* et le *TLFI* fournissent un modèle exemplaire pour des recherches de ce type. Mais plus les contraintes sont nombreuses et précises, moins les résultats sont abondants, ce qui augmente leur valeur documentaire mais affaiblit leur exploitation statistique. Le logiciel Stella ne donne d'ailleurs que de médiocres

[1] Un hapax n'aura guère que quelques *1* perdus dans une forêt de *0*. En l'absence de segmentation courte, au contraire, la ligne (et la colonne) qui lui est allouée sera uniformément remplie de 1 et le 0 ne s'observera nulle part, ni dans cette ligne, ni dans les autres.

ressources, pour l'analyse quantitative de ces profils. Mais de telles ressources sont disponibles dans le logiciel *Weblex* de Serge Heiden[2].

3 - Segments répétés ou patterns complexes, ces groupes de mots ont un ordre séquentiel qui donne à la phraséologie et à la syntaxe un rôle prépondérant. Ce n'est pas toujours ce que l'on cherche. Quand on vise à discerner les rapports sémantiques ou les connotations stylistiques qui lient les mots, il peut être utile de casser l'ordre des mots et de rechercher les cooccurrences sans s'occuper de leur place dans le segment considéré. Il importe d'abord d'établir le calcul pour deux mots en cooccurrence, puis de généraliser la procédure en associant, tour à tour et deux à deux, sinon tous les mots du texte, du moins tous ceux que leur fréquence ou leur statut met en avant. On obtient alors une liste triée **d'associations privilégiées**, dont l'exploitation peut faire appel à diverses techniques : graphe du réseau des associations, analyse arborée et analyse factorielle.

4 - On se propose en dernier lieu d'approcher la notion d'isotopie ou du moins de cerner les éléments concrets dans lesquels s'enferme – et se découvre peut-être – la relation isotopique. En reprenant une fonction dite « thématique », qui existe depuis toujours dans notre logiciel Hyperbase, nous lui avons donné un prolongement. Pour un mot donné (ou quelque objet du texte) tous les contextes où il apparaît constituent une sorte de texte puissamment orienté, dont on tire les spécificités en le comparant à l'ensemble du corpus. Là s'arrêtait jusqu'ici l'aide de la machine qui laissait au chercheur le soin d'interpréter la liste obtenue. Dans ce texte unifié et polarisé autour d'un mot, les techniques appliquées précédemment au corpus peuvent s'appliquer avec un rendement supérieur, qu'il s'agisse de recherche de corrélats, d'associations privilégiées ou de graphe du réseau associatif.

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, non seulement dans la même page mais aussi dand le même paragraphe. 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 le même espace restreint, 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

[2] On observera que ces objets linguistiques complexes se prêtent comme les mots simples aux opérations statistiques traditionnelles. Une fois que les relevés ont été établis, dans l'ensemble du corpus et dans chacun des textes qui le composent, on peut comparer l'effectif obtenu à sa valeur théorique et appliquer les lois connues et les méthodes multidimensionnelles. La statistique n'a pas changé de nature, mais son objet n'étant plus proprement lexical, certains proposent d'appeler logométrie ou textométrie ce que la lexicométrie désignait jusqu'ici. Le traitement du langage ne se contente plus en effet d'isoler des graphies. La lemmatisation s'impose de plus en plus même si les programmes sont encore déficients qui réalisent cette opération. L'ambiguïté est en partie dissoute, les liens syntaxiques soulignés et on dispose parfois aussi d'un codage sémantique, quoiqu'encore insuffisant dans *Cordial.*

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. Quant au deuxième facteur, il paraît séparer ce qui relève de la société et ce qui appartient à l'individu.

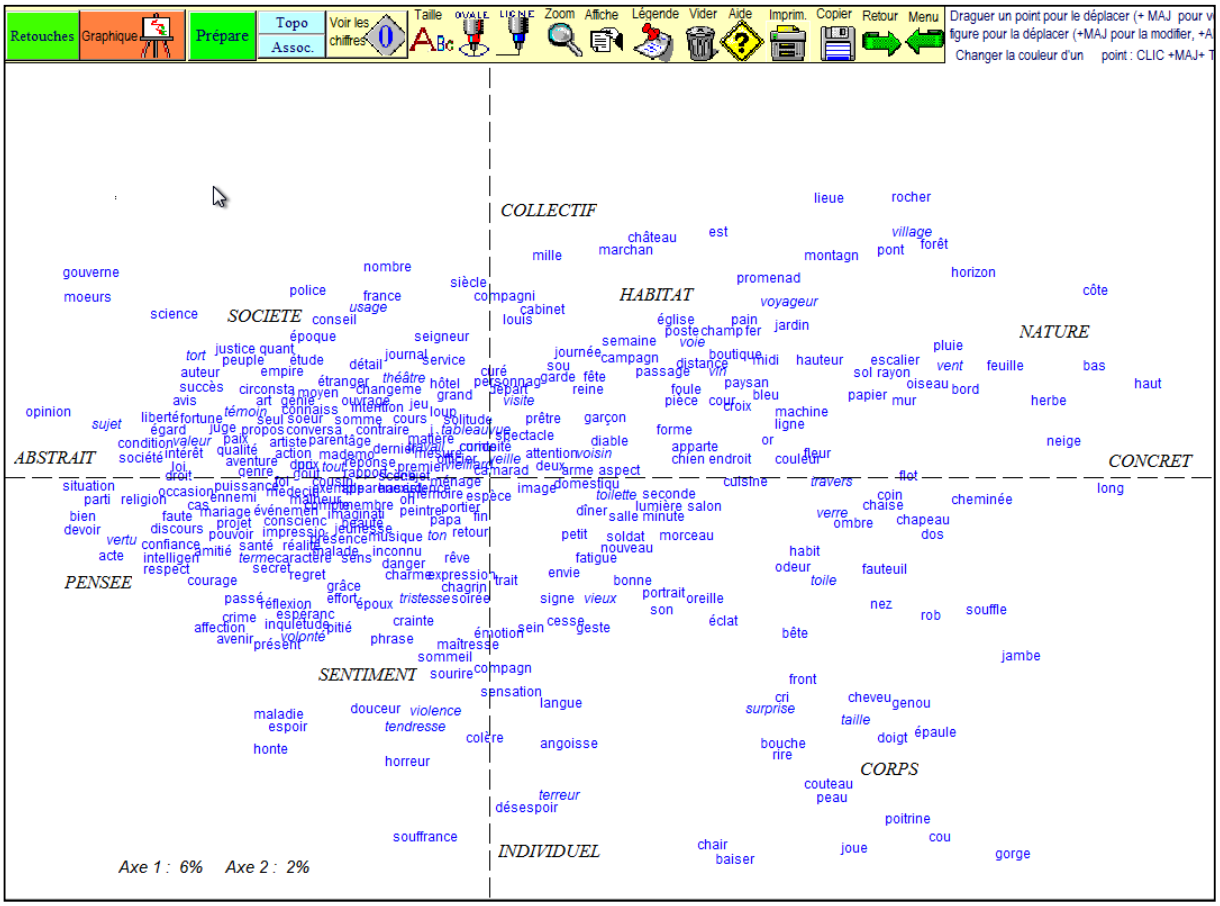

Figure 1. Analyse factorielle des corrélats dans le corpus EXEMPLEM

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

[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] Sous le nom d'*Exemple* (version standard) ou d'*Exemplem* (version lemmatisée), ce corpus rassemble 22 œuvres romanesques, de Marivaux à Proust, à raison de deux textes par auteur.

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 d'environ deux 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[5]. Et de la même façon chacune des pages est explorée et réduite à une suite variable d'une dizaine de mots en moyenne, si ces mots appartiennent à la liste préétablie et sont présents dans la page considérée. *Alceste* considérera 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 5000 pages de l'œuvre 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 Chi^2 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*.

[5] 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.

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*

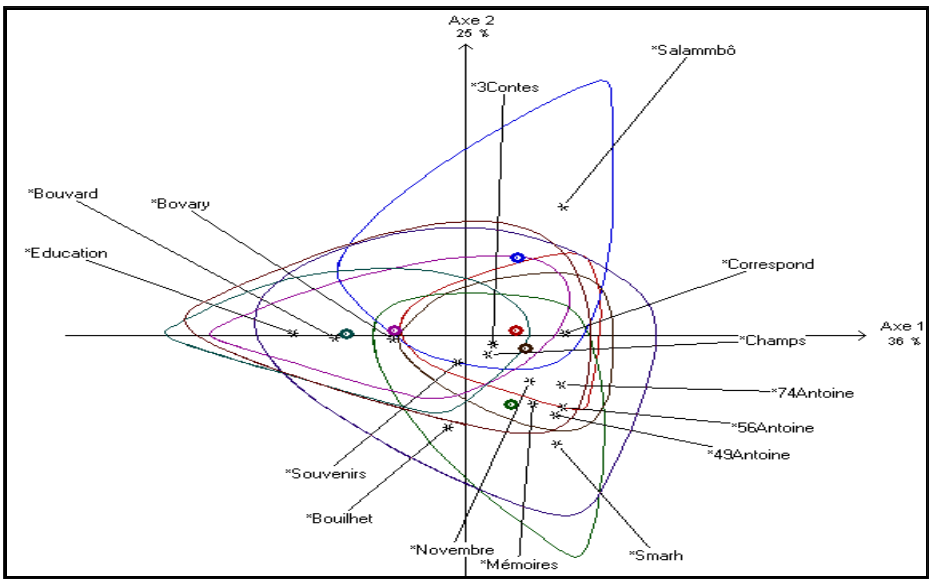

Figure 3. Analyse factorielle faite par Alceste *sur les substantifs dans l'œuvre de Flaubert*

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 simple ébauche, nettement affinée dans Alceste. La première n'est qu'un nuage de mots, sans autres repères que les points cardinaux. Dans la seconde des lignes de démarcation apparaissent, délimitant des constellations lexicales identifiables, tandis que la carte des textes, en surimpression, facilite l'interprétation.

4. Les codes répétés

Ici nous prenons pour modèle non plus *Alceste* mais *Lexico,* logiciel développé par André Salem et son équipe. Il y a dans *Lexico* une fonction puissante qui permet non seulement de relever les segments répétés mais de les intégrer dans les traitements statistiques : calcul des spécificités et analyse factorielle. Certes on trouve dans Hyperbase des fonctions documentaires capables de trouver des combinaisons de mots, expressions ou cooccurrences selon que les mots sont ou non adjacents. Mais encore faut-il proposer la combinaison. Il n'y a pas d'exploration systématique et exhaustive pour toutes les combinaisons, quelle qu'en soit la longueur et quels qu'en soient les éléments. Le calcul est en effet très long et peu réalisable quand le texte du corpus n'est pas tout entier en mémoire centrale.

Mais si la combinatoire porte sur une liste limitée et préétablie d'éléments elle peut être considérablement réduite et ne pas alourdir exagérément le traitement. C'est le cas des parties du discours qui dans la plupart des langues occidentales n'ont guère qu'une dizaine d'espèces. Or les lemmatiseurs, d'une manière ou d'une autre, fournissent cette information, généralement en tête du code de description grammaticale, au moins dans les deux logiciels dont Hyperbase fait usage : *Cordial* et *TreeTagger*. Le lemme qu'isole le lemmatiseur a besoin de cette information complémentaire, pour empêcher l'ambiguïté et ne pas confondre par exemple un *le* article et un *le* pronom. C'est pourquoi nous avons ajouté aux lemmes un indice, noté de 0 à 9, pour préciser la partie du discours dont relève le mot en question. Ainsi le lemme *le* est transcrit *le_7* quand il s'agit de l'article et *le_5* quand il s'agit du pronom. Un codage parallèle, cette fois transcrit en lettres, enregistre le flux des séquences codiques en s'arrêtant aux ponctuations. Ainsi la séquence *Mon tailleur est riche* est notée *dnva (d = déterminant, n = substantif, v= verbe, a = adjectif, r=adverbe, s=préposition, p=pronom, l = relatif, m=numéral, c= coordination, b = subordination, i = interjection).* Ces assemblages de codes sont traités comme les assemblages de lettres que sont les mots. Ils donnent lieu comme les mots individuels et les codes individuels à des calculs de distance, qui reflètent la structure syntaxique des textes, indépendamment du sens des mots. Leur autonomie est assurée même à l'égard des codes individuels, puisque l'ordre et le rythme des codes successifs sont pris en compte. Ainsi le même texte est présenté sous quatre formes rigoureusement alignées et pareillement indexées : les graphies, les lemmes, les codes grammaticaux, et les séquences ou structures syntaxiques. Ces éléments disjoints du même objet linguistique se différencient comme la forme, la couleur, le poids et la matière d'un objet naturel. Ils n'en donnent pas moins la même image, comme si quelque lien unissait ces propriétés en principe indépendantes. Dans un corpus romanesque qui réunit onze écrivains avec deux textes de chacun, on voit que le calcul des distances[6] fondé sur les séquences syntaxiques rapproche les textes qui sont issus de la même plume, même si ces textes sont éloignés chronologiquement dans la carrière de l'écrivain. Le même schéma est observé si le calcul porte sur les graphies, les lemmes ou les codes.

[6] La représentation est celle de l'analyse arborée, méthode mise au point par Luong et Barthélémy.

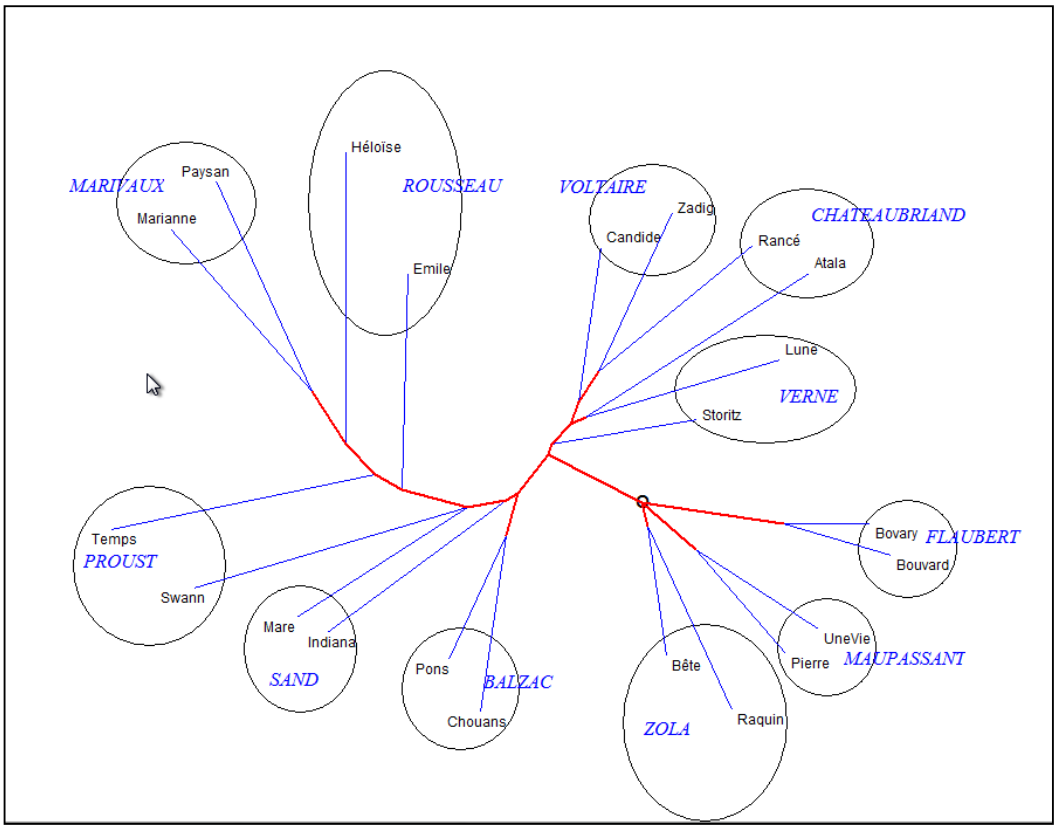

Figure 4. La distance intertextuelle établie sur les séquences syntaxiques.

Les séquences syntaxiques sont souvent longues entre deux ponctuations et l'éventail des combinaisons si vaste que la plupart des structures complètes relevées sont des hapax. Mais rien n'interdit de considérer des sous-chaînes à l'intérieur des séquences, de même qu'on peut isoler des syllabes ou des combinaisons de lettres à l'intérieur d'un mot. Prenons pour exemple une combinaison simple, à deux éléments *n* (substantif) et *a* (adjectif).

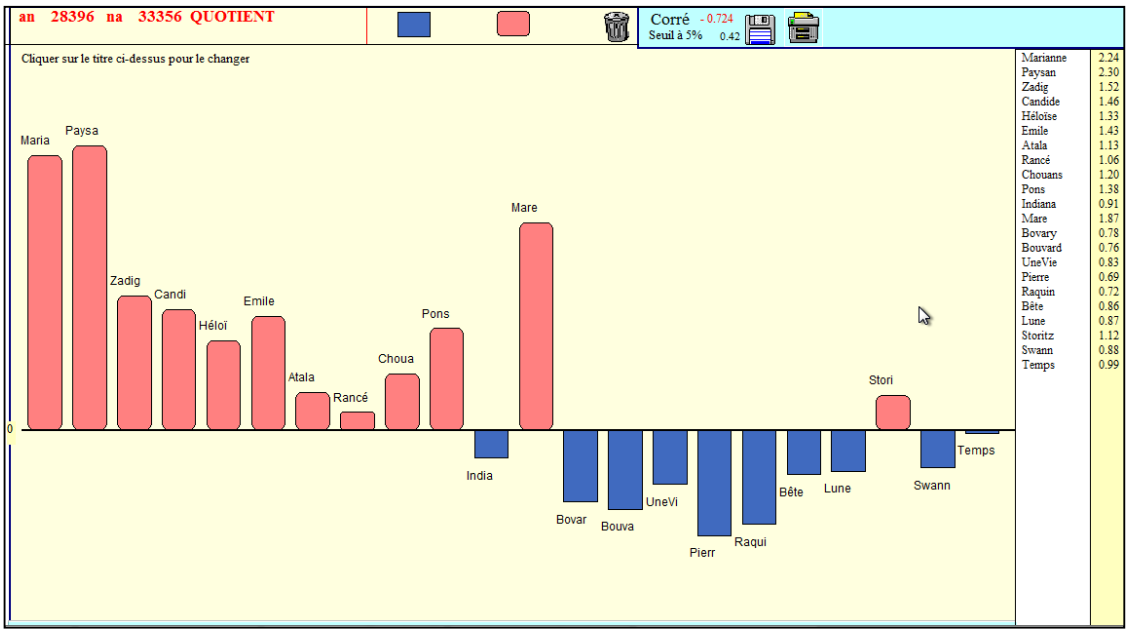

Figure 5. L'adjectif antéposé et postposé de Marivaux à Proust[7].

[7] Le graphique est fondé sur le rapport : antéposition / postposition.

En réalité comme l'ordre importe on a affaire à deux séquences *an* et *na* qui représentent respectivement l'adjectif antéposé et postposé. La seconde avec 33 356 occurrences dans le même corpus romanesque de 2 millions de mots l'emporte sur la première (28 396 occ.). Et surtout le rapport *an/na* n'est pas constant comme le montre la figure 5 qui voit l'antéposition céder le pas à la postposition au cours du temps. Si l'on se fonde sur 12 parties du discours, l'ensemble des bicodes représente 144 variétés qui remplissent les lignes d'un tableau où les textes sont en colonne. Cela conduit à une analyse factorielle qui confirme en la précisant celle qu'on obtient avec les parties du discours prises isolément.

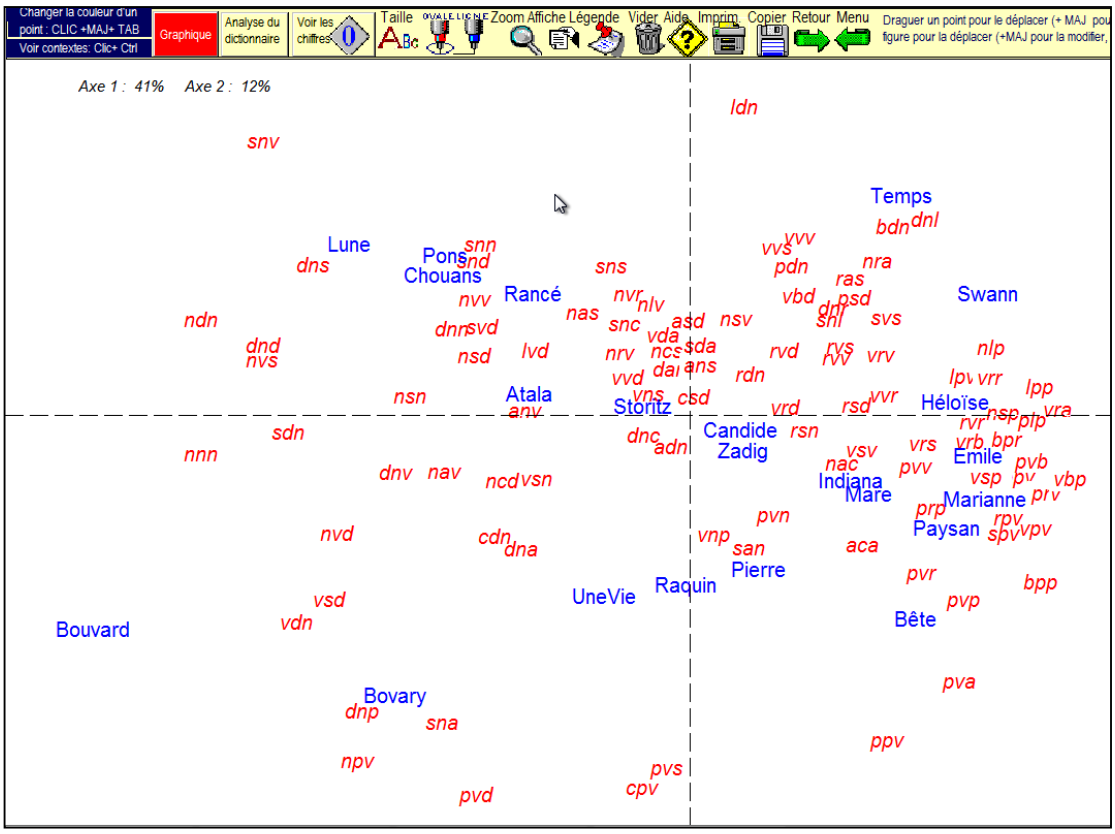

Figure 6. Analyse factorielle des tricodes (segments répétés de trois codes)
(d = déterminant, n = substantif, v= verbe, a = adjectif, r=adverbe, s=préposition, p=pronom, l = relatif, m=numéral, c= coordination, b = subordination, i = interjection)

La précision est encore plus forte lorsqu'on étend l'enquête aux tricodes ou segments répétés de trois codes adjacents. Parmi les 12^3 =1728 variétés possibles, certaines ne sont pas observées parce que la langue n'admet pas certaines séquences agrammaticales, comme la succession de trois relatifs ou de trois coordonnants. Ne figurent donc dans le tableau que les combinaisons suffisamment représentées. L'analyse factorielle de la figure 6 montre une disposition des textes assez semblable à celle de la figure 4. Les deux textes du même auteur ne sont jamais loin l'un de l'autre. La disposition des tricodes et celle des textes s'expliquent mutuellement : d'un côté se regroupent les combinaisons que domine le verbe (c'est là que s'installent les auteurs du XVIIIe siècle, mais aussi Sand et Proust), de l'autre celles où règne le substantif (Flaubert s'établit là, avec la plupart des écrivains du XIXe). Les acolytes suivent le chef du clan : prépositions, déterminants autour du substantif, adverbes, pronoms et subordonnants autour du verbe.

4. Les associations privilégiées

La recherche sur les associations s'appuie sur le tableau des cooccurrences, dont la fonction *Corrélats* a fourni une vue d'ensemble, sous forme d'analyse factorielle. La carte théma-

tique 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 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[8]. 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 la plupart des recherches ultérieures. Dans l'extrait qui en est livré ci-dessous (tableau 7) et qui est relatif à l'œuvre de Stendhal, on ne s'arrêtera pas aux associations qui relèvent de la phraséologie et même du lexique et qu'une lemmatisation étendue aux mots composés aurait dû éliminer. Mais ces scories (*coin (de) rue, chef (d')œuvre, gens (d')esprit, garde (du) corps et corps (de) garde*) 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, femme-mari, fils-père, fille-mère, matin-soir,* dont le lien tient à la sémantique et à l'attirance magnétique que les mots opposés ou symétriques exercent l'un sur l'autre comme les pôles d'un aimant. Mais le plus souvent les couples se forment, sur le modèle de *amour-bonheur*, par le partage de goûts et 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[9].

LISTE HIÉRARCHIQUE	10.11 matin soir	10.11 femme homme	9.18 amour coeur
Test mot1 mot2	10.11 mur pied	10.11 femme mari	9.16 chambre porte
	10.11 âge an	9.96 amour bonheur	9.13 amant amour
10.11 an fille	10.11 chambre femme	9.94 père tante	9.11 fois jour
10.11 fils père	10.11 chef oeuvre	9.94 amour femme	8.94 fenêtre rue
10.11 fois semaine	10.11 coeur homme	9.89 jour lendemain	8.90 air oeil
10.11 fois vie	10.11 coin rue	9.69 pied terre	8.89 porte soldat
10.11 heure jour	10.11 comte prince	9.63 cheval route	8.86 chambre heure
10.11 heure lendemain	10.11 amant femme	9.59 esprit femme	8.83 salon soir
10.11 heure matin	10.11 corps garde	9.56 âme bonheur	8.82 heure porte
10.11 heure soir	10.11 cour prince	9.52 fille mère	8.81 jour matin
10.11 homme monde	10.11 cour femme	9.51 maison rue	8.71 an homme
10.11 jardin porte	10.11 esprit gens	9.48 jardin mur	8.66 âme amour
10.11 larme oeil	10.11 esprit homme	9.32 chambre fenêtre	8.61 lèvre main
10.11 lendemain matin	10.11 expression oeil	9.26 bord mer	8.50 homme vie
10.11 madame salon	10.11 amour passion	9.24 mariage marquis	

Tableau 7. Les associations que Stendhal privilégie parmi les substantifs (extrait très partiel)

[8] On a 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 si l'on note bien une convergence étroite pour donner une valeur théorique à la cooccurrence de deux mots, la mesure est incertaine dès qu'il faut apprécier les écarts. On s'en est donc tenu à la méthode hypergéométrique qui n'est pas la plus économique mais qui reste la plus sûre. Elle a été proposée pour la première fois par [Lafon 1984 : 163] et reprise par [Heiden 2004 : 578].

[9] On ne s'étonnera pas de voir ex æquo les premiers couples de la liste. Cela est dû à un artifice d'écrétage, lorsque les valeurs extrêmes sont atteintes et qu'il n'est plus utile de prolonger les calculs. Le tableau 7 est un extrait du fichier qui recense la liste des cooccurrences et qui porte le même nom que la base, avec l'indice 4 remplaçant la dernière lettre et le suffixe.txt.

4.1. On peut tout d'abord isoler une ligne du tableau , en éliminant tous les éléments nuls (où la cooccurrence n'a pas été rencontrée) 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 ». La *vie*, qui fait l'objet de la figure 8, apparaît plus souriante et plus active chez Stendhal, plus sombre et plus méditative chez Proust. Proust associe la *mort* à la *vie*, alors que c'est le *cœur* qui vient en tête chez Flaubert et curieusement le mot *fois* chez Stendhal, sans doute parce que dans son univers d'apprentissage et de conquête tout arrive pour la première fois. On retrouve les isotopies attendues chez les trois écrivains (*bonheur, amour, cœur, homme, mort, jour, monde*), mais avec des dosages différents. Certaines associations ne sont pas partagées : le *sentiment* est stendhalien, l'*art* proustien et *dieu* flaubertien.

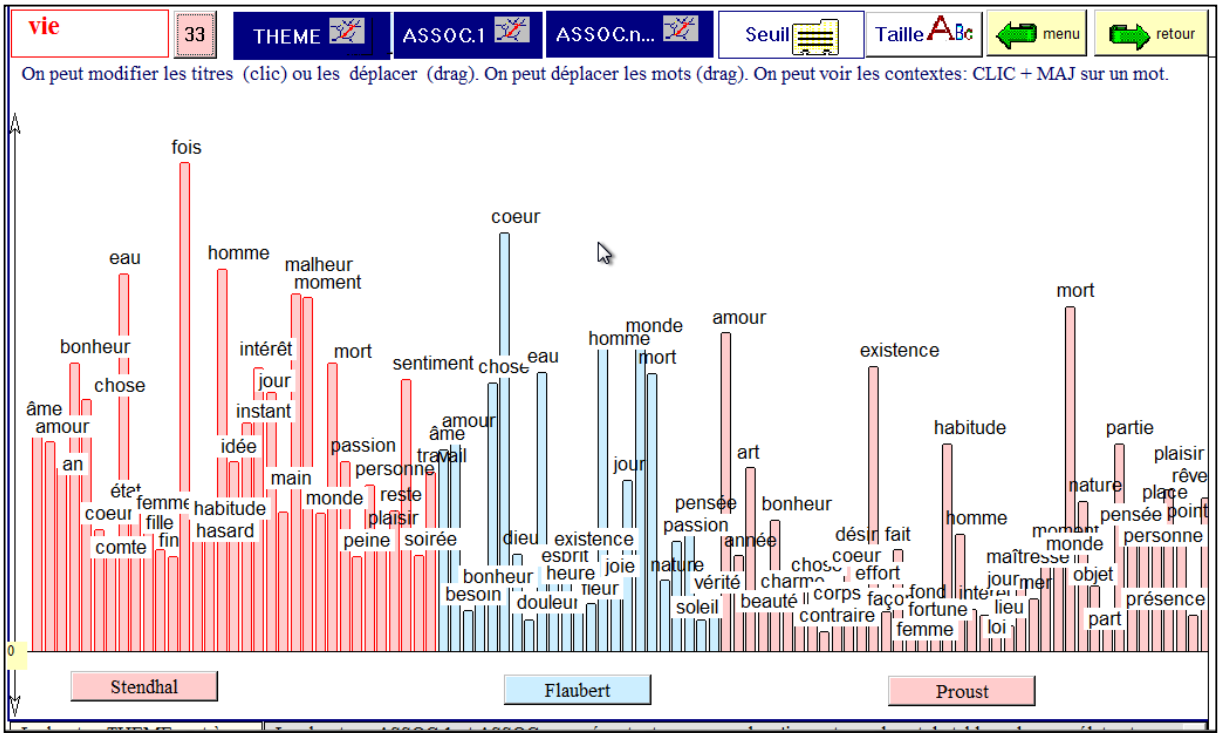

Figure 8. La constellation lexicale autour de la vie *chez Stendhal, Flaubert et Proust*

La comparaison peut s'étendre à bien d'autres mots, et par exemple être centrée sur la *mémoire*, laquelle prolonge ses synapses dans une large zone de la conscience proustienne, avec l'obsession de la vie passée (*souvenir, image, nom*) et de riches connotations esthétiques (*tableau, œil, couleur*) ou morales (*vie, rêve, pensée*). 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*[10].

[10] Pour que le rapprochement de corpus différents ait un sens plus riche, la même liste de 400 substantifs peut être explorée à chaque fois, avec les mêmes méthodes. On propose le choix entre deux listes, selon la nature des textes traités : l'une est littéraire et fondée sur une vingtaine de romans, de Marivaux à Proust ; l'autre est socio-polique et résulte de l'exploitation de deux années du journal *Le Monde*. Mais si l'on s'enferme dans un seul corpus, la liste peut être constituée des 400 substantifs (ou autres catégories) les plus fréquents dans ce corpus, dont un grand nombre se retrouveront dans d'autres corpus, ce qui permet encore la comparaison.

4.2. Entre la vue lointaine de l'analyse factorielle (figure 1) et le détail myope de l'histogramme (figure 8), il y a place pour un échelon intermédiaire : comme précédemment on commence par s'attacher à un mot parmi les 300 ou 400 disponibles. Une fois que l'hameçon est accroché, on tire sur la ficelle 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[11].

Nous prendrons le même mot *mort* à l'intérieur du corpus Stendhal. Les liens représentés dans le graphe sont en rouge s'ils concernent le mot-pôle, 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 nœuds fréquentés, le noir aux nœuds isolés (moins de 5 liaisons). La force des liaisons influe sur l'épaisseur des traits et la taille des caractères. Le calcul du graphe arborescent et de la position des nœuds 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[12].

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 présence/absence, comme si l'on circulait dans un réseau binaire avec des portes nécessairement ouvertes ou fermées. 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 dans la conception du programme.

Cependant un œil attentif jeté sur la figure 9 montre que la mort n'est pas un thème stendhalien, comme il l'est dans la déploration majestueuse de Chateaubriand. Les morts n'ont pourtant pas manqué sur les champs de bataille où Stendhal a suivi Napoléon. Mais chez Stendhal la mort n'est guère qu'un événement, à peine un accident. Elle est liée au temps *(vie, an, mois, jour, fois, moment, instant*) et elle frappe tour à tour *hommes* et *femmes, comtes, ducs* et *princes*, et tous les membres de la famille *(père, mère, fils, fille, frère, famille, maison*). C'est une nécessité de la narration, mais non un objet de description et encore moins de réflexion et de commisération. C'est à peine si le mot *peine* est évoqué dans un réseau lexical où figure le mot *bonheur*, mais non la souffrance, le chagrin et le deuil.

Les graphes veulent représenter sur un plan ce qui ne se conçoit que dans l'espace et le réseau des fils entrecroisés peut devenir inextricable. Aussi avons-nous prévu un moyen de simplifier la représentation graphique, en rendant momentanément invisibles les arcs ou liens secondaires qui apparaissent en noir sur le graphique et les nœuds ou mots non directement liés au mot-pôle.

---

[11] 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.

[12] 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 point lorsqu'un recouvrement gênant diminue la lisibilité.

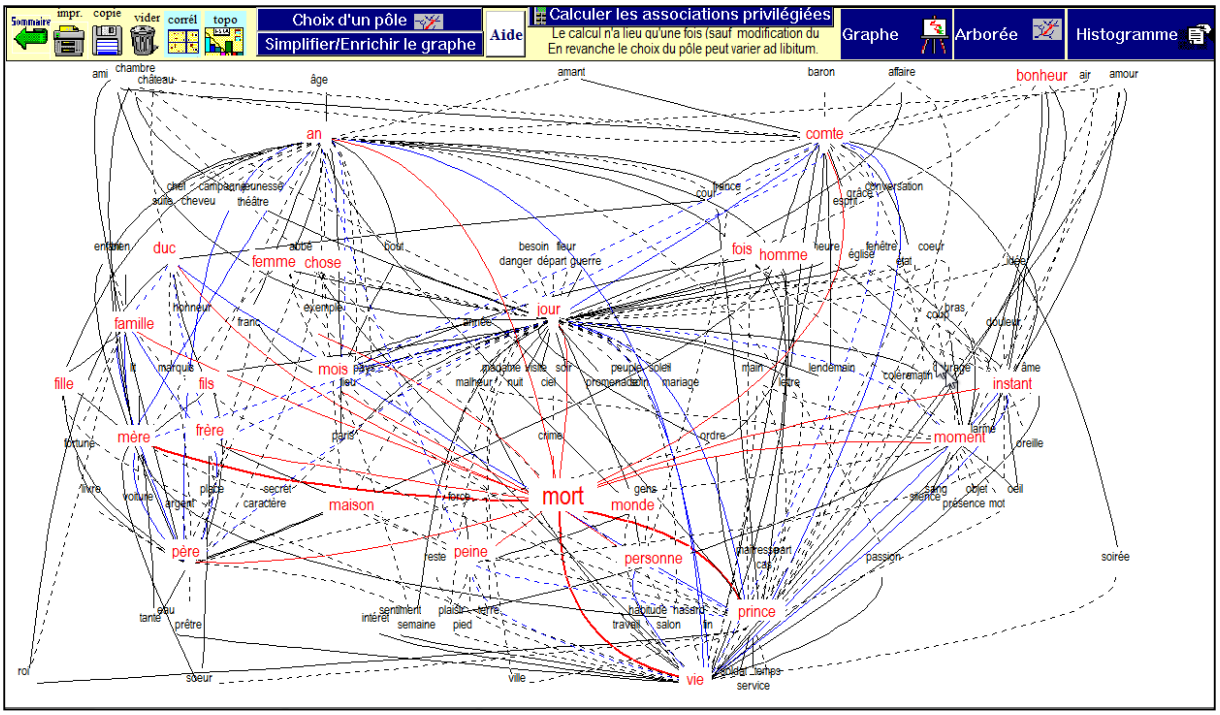

Figure 9. Le graphe de la mort *dans le corpus Stendhal*

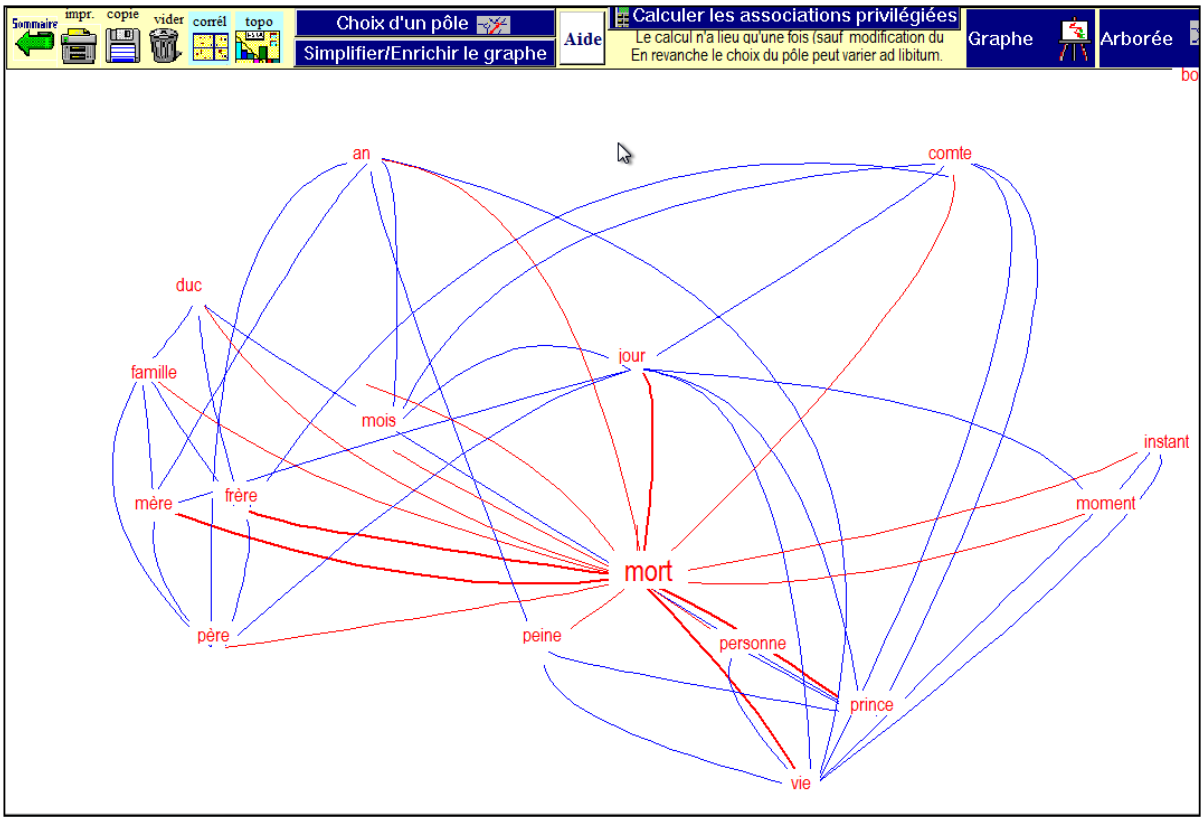

Figure 10. La mort *chez Stendhal. Vue simplifiée.*

Le gain en clarté accompagne la perte en information. Le bouton SIMPLIFIER/ENRICHIR permet de changer d'objectif en noyant ou grossissant alternativement les détails. La figure 10 en donne l'illustration à partir du même exemple.

4.3. Pour extraire la quintessence des cooccurrences croisées qui constituent une constellation lexicale, on a songé à constituer un tableau carré où les valeurs de proximité se substituent comme précédemment à la mesure brute des cooccurrences. Dans l'approche qui précède, c'est le tableau entier qui est exploré, les ramifications pouvant aller loin quand elles se communiquent de voisin à voisin. On fixe ici une barrière à cette propagation, en constituant un sous-tableau, carré comme le grand, et réduit à la liste des mots directement liés au mot choisi pour pôle. Un tel tableau contiendrait 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 à l'analyse arborée. Un bouton est disponible à cet effet et s'applique au mot qui a été choisi pour pôle, par exemple au même mot mort, emprunté au même corpus Stendhal. Le résultat, visible dans la figure 11.

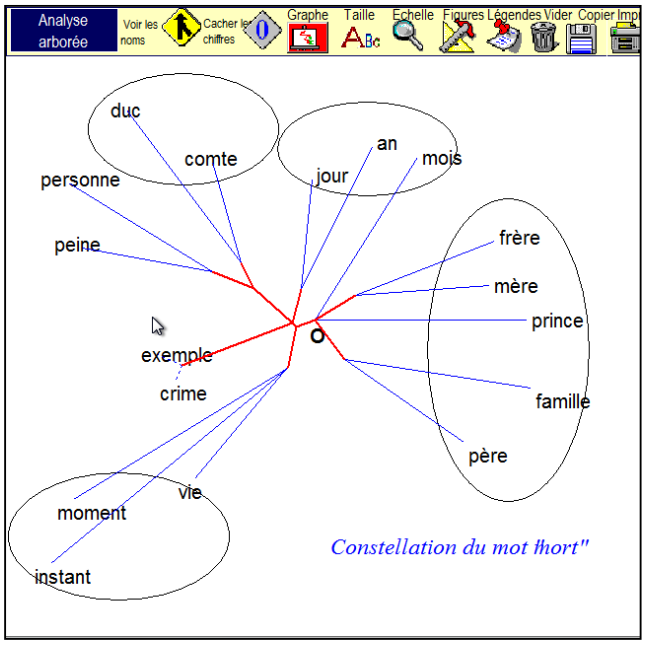

Figure 11. La constellation de la mort *dans la base* Stendhal

Il souligne la relation froide et presque indifférente que la mort entretient chez Stendhal avec le temps et les personnes. 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[13].

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 300 ou 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

[13] Voir [Gaume 2006]. B Gaume a mis au point un logiciel graphique, qui semble grandement supérieur à G*raphviz*. On aimerait que sa diffusion soit rendue possible.

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. 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 étant souvent le lemme.

5.1. 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 (le même mot *mort* dans la figure 12). La liste est triée en ordre alphabétique (champ de droite) et hiérarchique (champ de gauche).

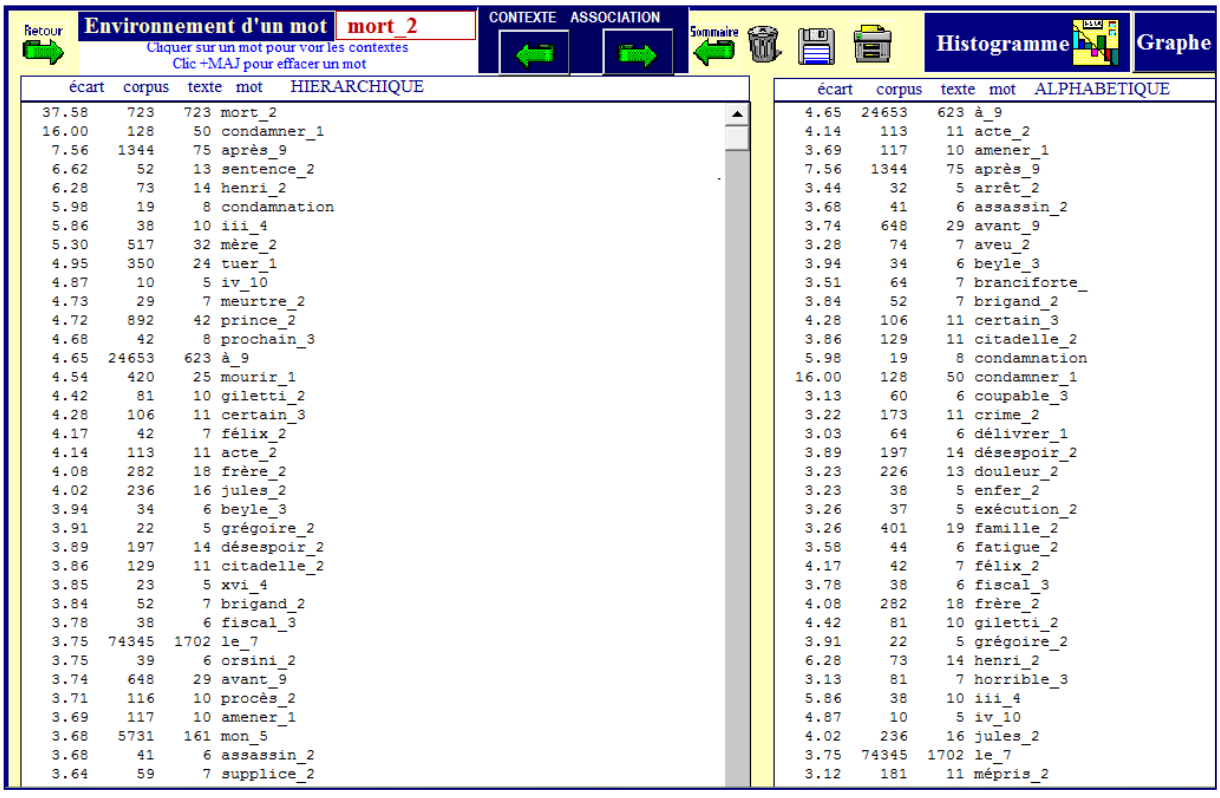

Figure 12. La fonction THEME. Les spécificités qui entourent le mot mort *chez Stendhal*

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 appliquer les procédures établies pour les corrélats et les associations privilégiées. Naturellement on ne cherche pas les cooccurrences avec le mot-pôle, puisqu'on les connaît déjà et que le calcul hypergéométrique 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[14], d'autant que leur fréquence intempestive tend à écraser le reste. Mais ce reste est riche des mots sémantiques, tous acceptés sans exception, sans que le code ou la fréquence puisse justifier leur exclusion.

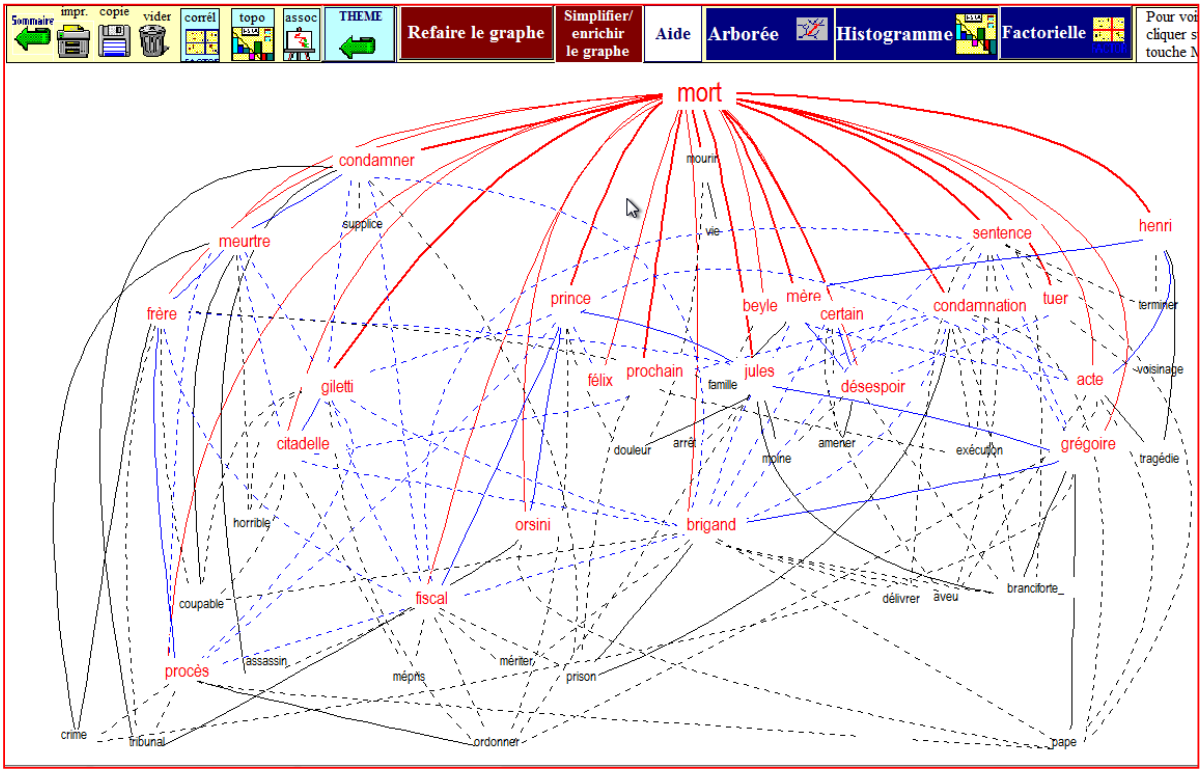

Figure 13. Les cooccurrents de la mort *dans la base Stendhal. Relevé exhaustif.*

5.2. Le programme *Graphe* s'applique à de telles données et offre une seconde dimension à la représentation « à plat » de la liste 12. Il rend compte des relations complexes que les cooccurrents de la *mort* établissent avec le pôle (ce sont les arcs en rouge) ou entre eux (ce sont les traits en bleu) ou avec d'autres mots qui n'appartiennent pas au premier cercle (traits en noir). La figure 13 et le tableau 14 montrent le gain qu'on obtient en précision et en extension, par rapport aux graphes 9 et 10. Apparaissent ici, et certains au premier rang, les *verbes* (*condamner*, *tuer*, *mourir*) et les adjectifs (*coupable*, *prochain, fiscal*) qui avaient été précédemment écartés. Apparaissent aussi les substantifs qui n'avaient pas atteint le seuil exigé pour la fréquence : *condamnation*, *sentence*, *meurtre*, *brigand*, *procès*, *assassin*, *supplice*, *exécution*, *aveu*. Le visage de la mort se tient à l'écart, comme précédemment, des plaintes et du deuil. Mais son aspect violent et juridique se précise. Chez Stendhal et

[14] 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.

particulièrement dans ses essais et dans les *Chroniques italiennes*, la mort est un meurtre ou une exécution. Il faut attendre Chateaubriand pour être invité à pleurer à la mort d'Atala[15].

LISTE HIÉRARCHIQUE			
test *mot1 mot2*	5.11 mort jules	4.26 mort famille	3.76 mort horrible
	5.08 mort certain	4.24 mort supplice	3.74 mort exécution
	4.95 mort frère	4.20 mort assassin	3.67 mort mériter
17.67 mort condamner	4.66 mort meurtre	4.17 mort félix	3.65 mort grégoire
8.50 henri tragédie	4.54 mort acte	4.11 mort fatigue	3.60 mort branciforte_
7.40 mort sentence	4.54 jules branciforte_	4.09 mort crime	3.59 mort tribunal
7.13 mort henri	4.53 mort désespoir	4.07 mort procès	3.56 mort moine
6.78 mort mère	4.47 mort amener	4.00 mort mépris	3.52 mort douleur
6.51 mort condamnation	4.45 mort beyle	3.99 mort terminer	3.36 mort ordonner
6.36 mort tuer	4.37 meurtre coupable	3.91 mort voisinage	3.34 henri acte
6.05 mort prince	4.30 mort fiscal	3.91 mort arrêt	3.21 frère procès
5.75 mort mourir	4.30 mort prison	3.90 mort aveu	3.20 grégoire pape
5.64 mort vie	4.29 mort citadelle	3.89 mort brigand	3.15 mort coupable
5.28 mort prochain	4.27 mort orsini	3.85 prince orsini	
5.16 mort giletti	4.27 mort tragédie	3.83 mort pape	

Tableau 14. Tableau trié de la probabilité des cooccurrences observées chez Stendhal pour le mot mort

[15] Il s'agit parfois de pudeur contenue, comme le prouve ce passage de *Vie de H. Brulard* où le narrateur évoque la mort de sa mère: « Elle périt à la fleur de la jeunesse et de la beauté en 1790, elle pouvait avoir vingt-huit ou trente ans. Là commence ma vie morale. Ma tante Séraphie osa me reprocher de ne pas pleurer assez. Qu'on juge de ma douleur et de ce que je sentis ! Mais il me semblait que je la reverrais le lendemain, je ne comprenais pas la mort. Ainsi il y a quarante-cinq ans que j'ai perdu ce que j'aimais le plus au monde. »

Reste à dire quelques mots sur le calcul de la probabilité des cooccurrences, qui est ici assez complexe. Rappelons que les paramètres dans le calcul direct des cooccurrences sont les suivants dans la terminologie de Pierre Lafon : s= nombre de phrases, f= fréquence du mot-pôle dans le texte, g= fréquence du mot cooccurrent dans le texte et k=cooccurrence observée.

Première difficulté liée au choix du pôle : pour un couple dont on mesure la force d'attraction, les premiers paramètres s, f et g varient suivant que le relevé des contextes est réalisé à partir de l'un ou de l'autre terme, alors que le nombre de cooccurrences k reste le même. Ainsi dans le corpus *Exemplem* la relation *vertu-bonheur* est estimée à 4.62 si l'on part de la *vertu* et à 4.68 si on suit le *bonheur*. Ces mêmes paramètres changent encore si l'on envisage le corpus entier, selon la procédure indiquée plus haut (dans les associations privilégiées) : le nombre de cooccurrences reste le même (14) mais sa probabilité est estimée différemment : 5.35 pour le même couple *vertu-bonheur*.

Seconde difficulté dans le cas des contextes de la fonction *Thème* : le champ d'observation diffère si le mot-pôle entre ou non dans la liaison considérée. Si oui les paramètres sont ceux du corpus, puisque le corpus a été intégralement dépouillé pour tous les couples où le mot-pôle est partie prenante. 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 doivent donc être adaptés et correspondre à ce sous-ensemble.

Dernière difficulté qu'il convient d'élucider : certains esprits pourraient s'étonner que le calcul ne donne pas la même valeur lorsque les spécificités sont évaluées à partir du programme THEME dans la page « contexte », ou à partir du programme GRAPHE de la page « environnement ». Prenons l'exemple du mot *sentence* dont la spécificité est estimée à 6.62 dans le tableau 12, quand on compare les phrases où apparaît le mot *mort* à l'ensemble du corpus. On apprécie certes la force de la liaison *sentence-mort* mais le calcul reste indirect et la fréquence du pôle *mort* ne figure pas parmi les ingrédients : T = taille du corpus, t=taille du sous-corpus, f=fréquence du mot *sentence* dans le corpus, k = fréquence du mot *sentence* dans le sous corpus. Deux seulement de ces paramètres se retrouvent dans le calcul direct de la cooccurrence évoqué plus haut. On y ajoute la fréquence du mot-pôle et la taille du corpus est estimée en nombre de contextes (ou phrases) et non en nombre de mots. Il peut donc y avoir divergence théoriquement entre les deux calculs selon que les phrases sont plus ou moins longues en moyenne. En réalité, qu'on pose d'une façon ou d'une autre les paramètres du calcul hypergéométrique, les résultats sont parallèles et dans le calcul direct la cooccurrence *mort-sentence* est estimée à 7.40, comme indiqué dans le tableau 14. Il sa sans dire que les deux formules de la loi hypergéométrique n'ont pas la même forme : dans le premier cas on a :

$$\text{Prob}(x=k) = \mathbf{(}f\,!\,(T-f)\,!\,t\,!\,(T-t)\,\mathbf{!)} \,/\, \mathbf{(}k\,!\,(f-k)\,!\,(t-k)\,!\,(T-f-t+k)\,!\,T\,!\mathbf{)}$$

et dans l'autre

$$\text{Prob}(x=k) = \mathbf{(}f\,!\,(s+g)\,!\,g\,!\,(f+s)\,!\,\mathbf{)} \,/\, \mathbf{(}\,k\,!\,(f-k)\,!\,(g-k)\,!\,(s+k)\,!\,(f+g+s)\,!\mathbf{)}$$

5.3. Quand un mot a peu de liaisons, cela signifie souvent qu'il est attaché au mot-pôle par un lien quasi exclusif qui le rend indifférent au reste et relève de la phraséologie. Dans le cas contraire, les relations sélectives qu'il peut avoir l'orientent dans un réseau ou dans un autre. En suivant les flèches de proche en proche, on peut circonscrire ce réseau particulier, comme celui qui se situe à l'extrême droite de la figure 13 et qui met en relation *Henri, tragédie, acte*, etc... Un clic sur un mot de ce réseau (en sollicitant aussi la touche MAJ) faisant apparaître les contextes où la cooccurrence est observée (figure 15), on a le moyen d'expliquer les relations relevées : dans ce cas précis il s'agit des réflexions de Stendhal développées dans *Racine et Shakespeare* sur les crimes et fins tragiques qui se multiplient dans le théâtre shakespearien.

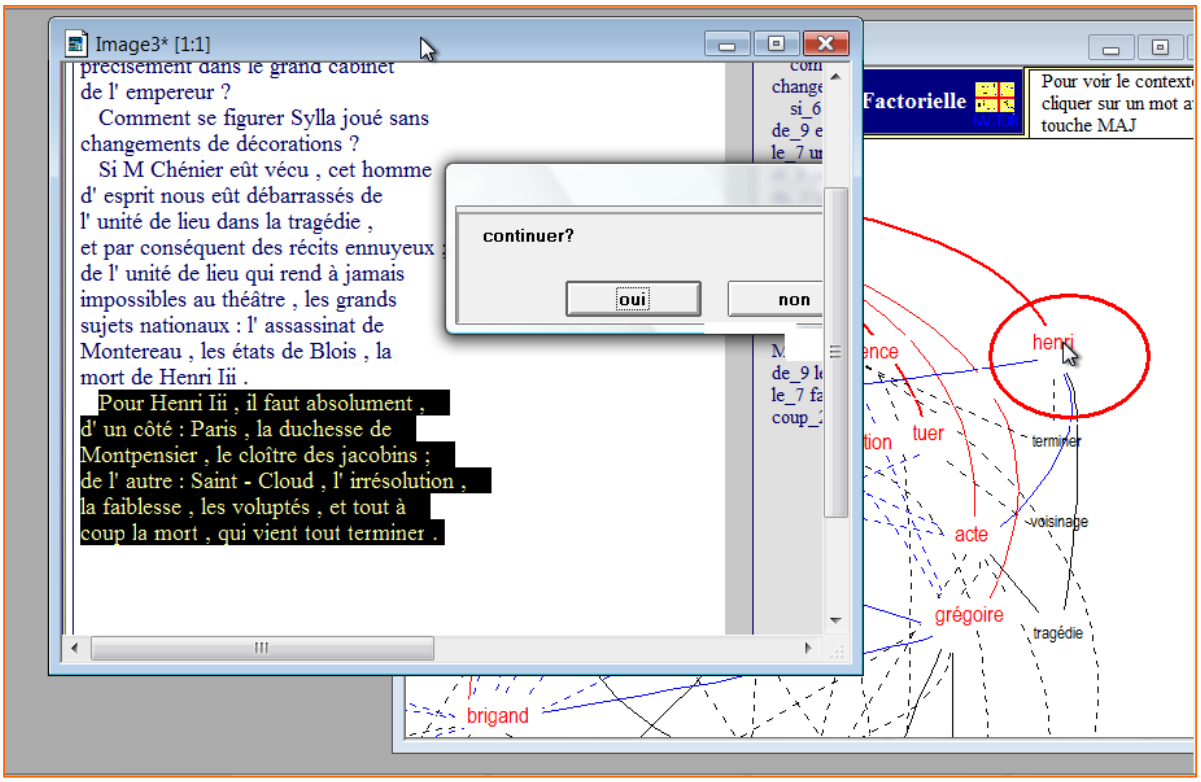

Figure 15. Le retour au texte assuré par un clic sur un mot

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 correspondances. Livrons à la machine deux éléments : le sous-corpus bâti autour d'un mot (c'est le fichier DATA1.txt, quel que soit le nom de la base) et la liste des spécificités qui en est extraite (comme celle du tableau 12) et qui donne lieu à l'histogramme de la figure 16 (pour le mot *argent* dans le corpus *Eugène Sue*). Le traitement va remplir le tableau des cooccurrences des mots de cette liste et le livrer sans autre procès au calcul factoriel. Le résultat, pour le même mot *argent* dans le même corpus, est dans la figure 17.

Visiblement l'argent n'est pas rare dans les feuilletons d'Eugène Sue, surtout dans les *Mystères de Paris*. Il est vrai que c'est avec l'amour l'ingrédient principal de l'intrigue romanesque, de Balzac à Zola. Mais les 458 occurrences de l'argent ne brillent pas du même éclat dans tous les contextes. Tantôt il s'agit de la monnaie et cette acception ordinaire s'établit dans la partie basse et gauche du graphique 17. Tantôt il s'agit du métal et l'argenterie miroite au haut du graphique. Tantôt il s'agit de la couleur, qui illumine le flanc droit. On notera que c'est souvent le vêtement qu'on qualifie ainsi et que la transition entre le

métal et la couleur passe par la broderie où le fil du métal donne à l'étoffe certaines de ses propriétés.

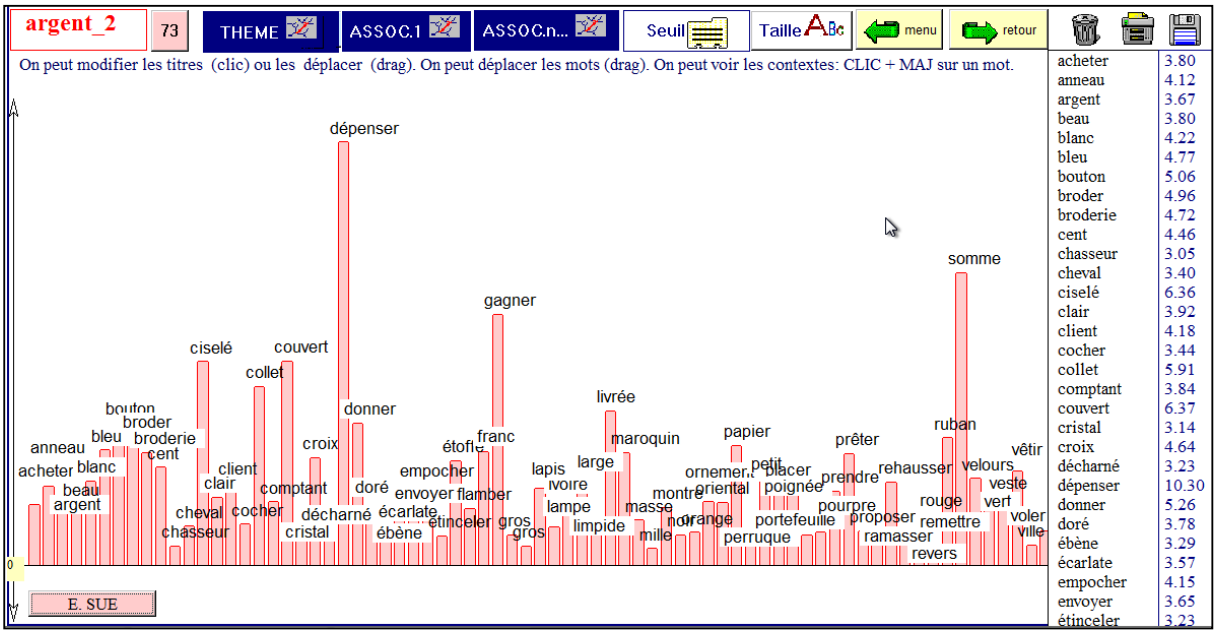

*Figure 16. Histogramme des corrélats de l'*argent *dans le corpus* Eugène Sue

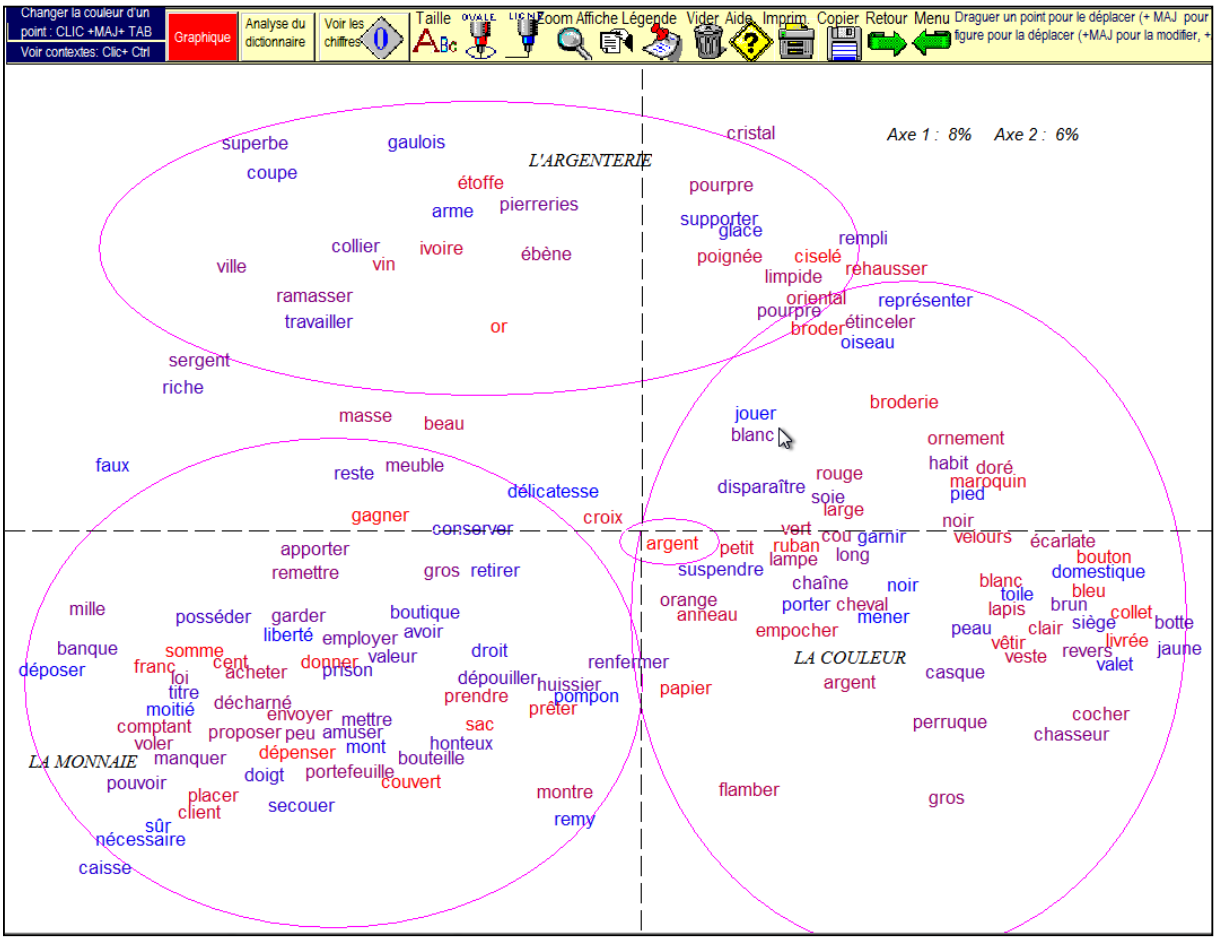

*Figure 17. Analyse factorielle des mots attirés par l'*argent *dans le corpus* Eugène Sue

En conclusion, on n'est pas certain que toutes ces opérations perpétrées sur les séquences et les cooccurrences puissent faire découvrir le sens des mots, sauf dans certains cas où la polysémie peut être facilement décantée, comme dans le cas de l'*argent*. Les champs sémantiques qu'on parcourt en arpenteur sont d'autant mieux délimités qu'ils appartiennent à des séries relativement fermées comme les liens de la parenté ou les parties du corps humain. Mais le bornage des champs est le plus souvent approximatif et les contestations sans solution. En outre 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. Mais si le sens d'un mot est, comme on l'a dit, la somme de ses emplois, on ne peut éviter de 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 qu'on en ait si peu pour les rayonnements textuels.

Références

[BRUNET 2006] BRUNET Étienne, « Navigations dans les rafales », in *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles*, Besançon : Presses Universitaires de Franche-Comté, 2006, pp. 15-29.

[GAUME 2006] GAUME Bruno, *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.

[HEIDEN 2004] HEIDEN Serge « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex », in *JADT04, Le poids des mots*, Louvain : Presses Universitaires de Louvain, 2004, p. 577-588.

[LAFON 1984] LAFON Pierre, *Dépouillements et statistiques en lexicométrie*, Genève-Paris : Slatkine-Champion, 1984