Nouveau traitement des cooccurrences

Étienne Brunet

Nouveau traitement des cooccurrences dans HYPERBASE[1]

Etienne Brunet
(Bases, Corpus et Langage , CNRS, Nice))

1 – On vise ici à représenter les relations qui lient entre eux les mots d'un texte, ces relations étant considérées dans l'espace étroit du paragraphe et non dans l'espace du texte. Cette fonction, intitulée THEME, est depuis longtemps intégrée à HYPERBASE, dans une approche limitée à un mot (graphie ou lemme). Le mot proposé par l'utilisateur est repéré dans le corpus et tous les paragraphes où on le trouve forment un texte composite qui est comparé à l'ensemble. Un calcul de spécificités met alors en valeur les mots qui s'associent le plus souvent au mot choisi et qui circonscrivent ainsi un « thème ». Ainsi les 5000 passages où l'on croise une FILLE dans le corpus Balzac (figure 1) montrent assez que son sort dépend de la famille et de la réponse qui sera donnée à la question essentielle : à qui la marier ?

2 – On peut varier les questions et multiplier les thèmes, mais cette méthode, même cumulative, reste partielle et trop dépendante des choix proposés. On aimerait que le calcul seul délimite les thèmes exploités dans un corpus, en interrogeant tous les mots sans en privilégier aucun : c'est l'objet de la fonction CORRELATS. En principe on devrait croiser tous les vocables d'un texte et constituer un immense tableau qui aurait autant de lignes et de colonnes que de vocables recensés, soit une matrice de 100 millions de cases pour un petit corpus de 10 000 mots. On préfère limiter l'exploration à un échantillon raisonné, en choisissant les mots les plus fréquents, ceux qui ont précisément le plus de chances de s'accoupler. On commence par constituer un tableau carré où les mots retenus (environ 300 parmi les mots-pleins les plus fréquents) sont portés

1. NDÉ : Article paru dans *Corpus*, n°11, 2012, p. 219-246 (2012b). Il s'agit ici d'une version basée sur le fichier original de l'auteur et révisée par l'auteur, présentant de nombreuses variantes par rapport à la version publiée, ne serait-ce qu'au niveau de la structuration du texte. Le texte ici présenté a été édité par Bénédicte Pincemin, dans *Tous comptes faits, Questions linguistiques*, Champion, 2016, p. 287-311.

à la fois dans les lignes et les colonnes ; à l'intersection on a l'effectif des rencontres où le mot de la ligne *i* est en cooccurrence (dans le même paragraphe) avec le mot de la colonne *j*. Pour être plus précis, il faudrait parler de co-présence : quelle que soit la fréquence f_a du mot *a* et la fréquence f_b du mot *b* dans le paragraphe, la rencontre des deux mots est comptée pour 1. On pourrait aussi légitimement aligner la cooccurrence sur f_a ou sur f_b, ou sur le produit $f_a \times f_b$.

Pour établir la liste des mots à retenir, solliciter le bouton PREPARE. On recommande de se limiter aux substantifs, mais les verbes et les adjectifs sont aussi éligibles. Par défaut la sélection se fait d'abord sur la catégorie, puis sur la fréquence et enfin sur la taille du tableau. En aucun cas le nombre de mots retenus ne doit dépasser 400 (car, au-delà, trop de mots encombrant le graphique, les résultats perdent en lisibilité). Si ce nombre est dépassé lors d'un premier balayage, une seconde exploration, voire une troisième, est déclenchée automatiquement avec un seuil plus sévère. Prendre patience. Le programme est long, chaque paragraphe devant être examiné en séquence. Quand le tableau des cooccurrences est constitué, l'exploitation statistique fait appel à l'analyse factorielle de correspondance. Ce programme (LX3AFC.EXE) a été écrit par Ludovic Lebart dans les années 70. Les résultats de l'analyse sont rapatriés et présentés par le bouton GRAPHIQUE, qu'on peut solliciter plusieurs fois, selon qu'on veut croiser les facteurs 1, 2 ou 3.

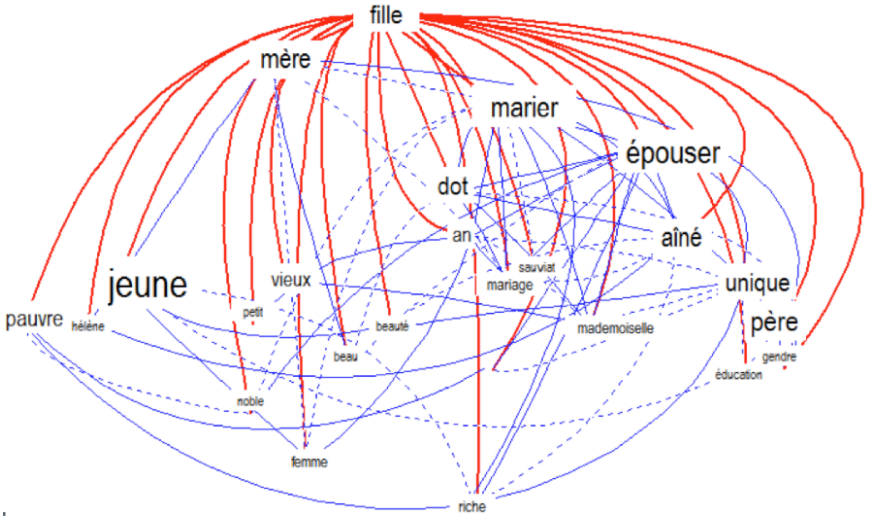

Figure 1. Le dessin d'un thème (fonction ASSOCIATIONS) : autour du mot FILLE chez Balzac

3 – Il arrive parfois que des mots soient si souvent associés par la phraséologie (par exemple ÉTATS et UNIS) que cette liaison triviale déséquilibre le tableau et influence certains facteurs de façon excessive. Il arrive aussi que certains mots accaparent le pouvoir discriminant et qu'on souhaite réanalyser les données en dehors d'eux. Plus souvent encore il reste des scories dans le traitement grammatical et certaines propositions de la lemmatisation peuvent être refusées. Comme le traitement des cooccurrences est long, il vaut mieux nettoyer la liste avant de la livrer à l'exploration. La liste accepte l'effacement des éléments indésirables (un clic suffit). Si un premier traitement laisse des regrets, la liste peut être reprise et corrigée. Mais un nouveau traitement est alors nécessaire, entraînant de longs délais.

4 – Les données enregistrées par le bouton PREPARE se trouvent dans le fichier qui porte le nom de la base avec le suffixe .SOC.

Tableau 2. Les fichiers NERVAL.SOC (effectifs absolus des cooccurrences) et CORAN.DON (conversion en racine carrée)

```
    abord_2              0 0 2 2 1 2 1 1 1 2 0 1 4 1 0 2 1 2 1 1 1
3 0 2 1 3 0 0 1 1 1 2 1 0 1 1 2 1 2 1 1 1 3 0 0 1 1 1 0 2 2 2 0 2 1
1 1 0 0 4 2 8 2 1 2 0 0 1 3 0 1 2 2 3 3 2 0 5 0 2 0 1 5 0 0 1 0 0 3
1 10 0 1 0 1 1 4 3 0 2 2 2 1 1 0 0 5 1 1 0 2 0 2 0 4 1 1 0 0 0 0 4
2 15 3 0 1 7 2 0 0 0 0 2 1 1 0 0 10 0 2 0 1 1 etc.
```

```
           abord_2                 0.0  0.0  1.4  1.4  1.0  1.4  1.0
1.0  1.0  1.4  0.0  1.0  2.0  1.0  0.0  1.4  1.0  1.4  1.0  1.0
1.0  1.7  0.0  1.4  1.0  1.7  0.0  0.0  1.0  1.0  1.0  1.4  1.0
0.0  1.0  1.0  1.4  1.0  1.4  1.0  1.0  1.0  1.7  0.0  0.0  1.0
1.0  1.0  0.0  1.4  etc.
```

Il s'agit des effectifs absolus. Par défaut l'analyse s'appuie plutôt sur la racine carrée de ces données brutes, afin d'amortir l'effet de taille. Mais on peut renoncer à cette pondération et appliquer l'analyse factorielle aux effectifs absolus. Solliciter à cet effet le bouton VARIANTES. Dans tous les cas le fichier de données, pondérées ou non, est transféré, sous le nom de CORAN.DON, au programme d'analyse factorielle LX3AFC.EXE, qui restitue systématiquement les résultats dans le fichier CORAN.SOR. Ces deux fichiers tampons servant à de multiples analyses, veiller à actualiser leur contenu.

5 – Ce même bouton VARIANTES propose une troisième option, qui fonde l'analyse sur des écarts réduits. Chaque cellule du tableau est soumise à un calcul probabiliste mettant en œuvre les lois normale et hypergéométrique (explication dans l'AIDE sur les cooccurrences en bas

à droite de la page TOPOLOGIE, figure 4)[2]. Pour construire ce tableau des écarts, on est invité à solliciter le bouton CALCULER LES ASSOCIATIONS PRIVILEGIEES de la page ASSOCIATIONS. Le traitement génère alors le fichier qui porte le nom de la base (diminué d'une lettre) et la finale `8.TXT` (ici `NERVA8.TXT`) (tableau 3).

Tableau 3. Le fichier `NERVA8.txt`. Conversion en écarts réduits

abord_2	0.0	0.0	1.6	1.3	0.0	2.3	0.0	0.0	0.0	0.6
0.0	0.6	3.1	0.2	0.0	1.9	0.8	1.6	0.8	0.3	0.6
2.2	0.0	0.9	0.2	2.7	0.0	0.0	0.0	0.3	0.0	2.2
0.8	0.0	0.7	0.4	2.0	0.0	2.3	0.7	0.3	0.6	2.1
0.0	0.0	0.2	0.4	0.6	0.0	0.0	0.5	0.0	0.0	1.2
0.6	0.6	0.0	0.0	0.0	2.1	1.7	4.8	1.3	etc.	

Calcul de la cooccurrence théorique

Le calcul de la cooccurrence théorique s'appuie sur le produit de deux probabilités: celle qui est attachée au premier mot et celle qui est propre au second. Chacune de ces probabilités relève du calcul hypergéométrique, les paramètres étant fixés comme suit:

T = Nombre total de mots dans le corpus
t = nombre moyen de mots dans un paragraphe (ou dans une page)
f = fréquence du mot considéré dans le corpus
k = 0 absence de ce mot dans le paragraphe (ou dans la page)

On obtient p1 pour l'absence du mot 1 et p2 pour l'absence du mot 2.

Le complément à l'unité de cette probabilité sert à mesurer les chances de rencontrer au moins une fois le mot dans l'espace considéré,

q1(présence mot1) = 1 – p1 et q2(présence mot2) = 1 – p2

et le produit des deux résultats mesure les chances d'observer la cooccurrence des deux mots à la fois dans le même paragraphe (ou la même page).

p (cooccurrence)= q1 * q2

En multipliant cette probabilité élémentaire par le nombre de paragraphes (ou de pages), on obtient l'effectif théorique des cooccurrences des deux mots dans le corpus. Reste à comparer l'effectif réel à l'effectif théorique, ce dont rend compte le calcul classique de l'écart réduit.

Remarque. La cooccurrence théorique de trois termes pourrait être calculée aussi facilement: avec p3 et q3 pour le mot 3 et p = q1 * q2 * q3

Figure 4. Calcul hypergéométrique de la cooccurrence théorique

6 – On peut observer l'effet obtenu par l'option choisie dans VARIANTES en comparant les trois analyses factorielles qui peuvent être successivement réalisées à partir des mêmes données de départ. À première vue les différences paraissent faibles. Qu'il s'agisse d'effectifs

2. NDÉ : Ce calcul de cooccurrence avait été initialement présenté dans une communication aux JADT 2006 (2006c) qu'on retrouvera dans le volume II des *Écrits choisis*, chapitre 13, « Navigation dans les rafales », p. 259-277 (voir p. 270). Il est maintenant expliqué de la façon la plus complète dans le *Manuel de référence* d'HYPERBASE.

bruts, ou convertis en racine carrée ou soumis au calcul probabiliste, les mêmes constellations lexicales s'ordonnent de la gauche à la droite, en allant du monde physique au monde moral (c'est le premier facteur). Dans les trois cas le facteur 2 oppose l'individu (en haut) à la société (en bas). Les noms qu'on peut donner à de telles constellations sont les mêmes et à la même place : ciel, eau, nature, habitat, corps, société, sentiment, droit, pensée, art (figure 5).

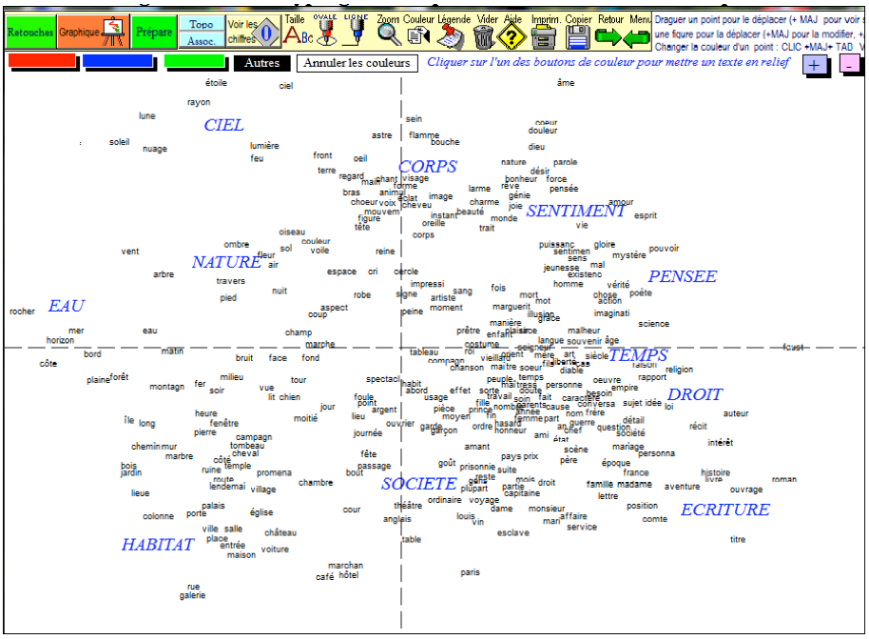

Figure 5. Analyse factorielle du corpus NERVAL sur écarts réduits

Pourtant un examen attentif révèle des déplacements qui pour être de faible portée n'en sont pas moins systématiques. Ils tiennent à la manière de traiter les hautes et moins hautes fréquences. Quoique les mots réunis (des substantifs) figurent parmi les fortes fréquences, puisque ce sont les 300 premières de la catégorie, il peut y avoir de fortes inégalités en passant de la première à la dernière place. Le rapport est au moins de 1 à 10. On fera donc deux lots particuliers dans la population, en isolant les 50 mots les plus forts et parallèlement les 50 plus faibles. Pour opérer cette décantation solliciter le bouton gris marqué d'un plus (les mots fréquents apparaissent en bleu) et le bouton rose marqué du signe moins (les moins hautes fréquences en rouge).

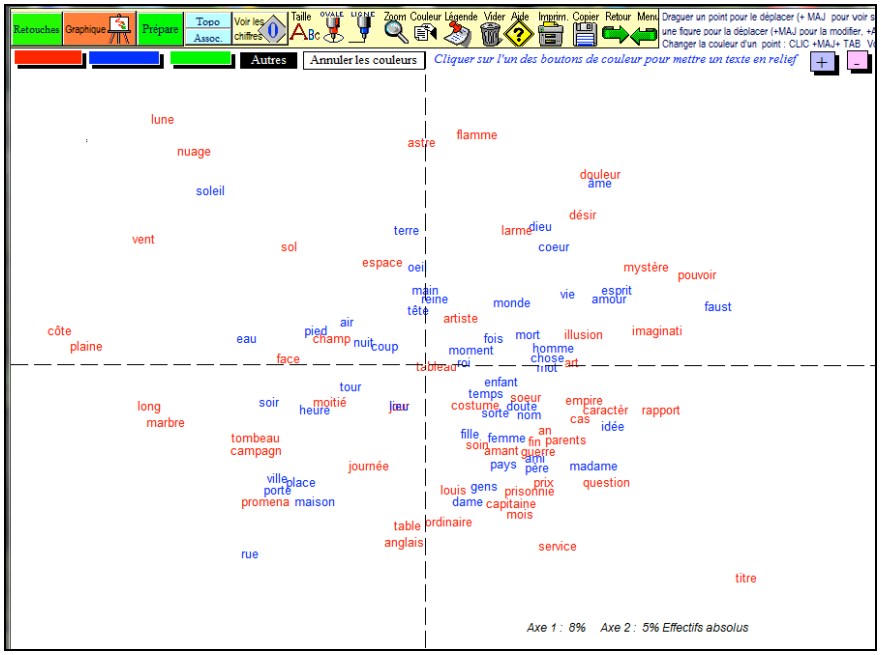

Figure 6. Cooccurrences brutes
Les mots les plus fréquents se rapprochent du centre

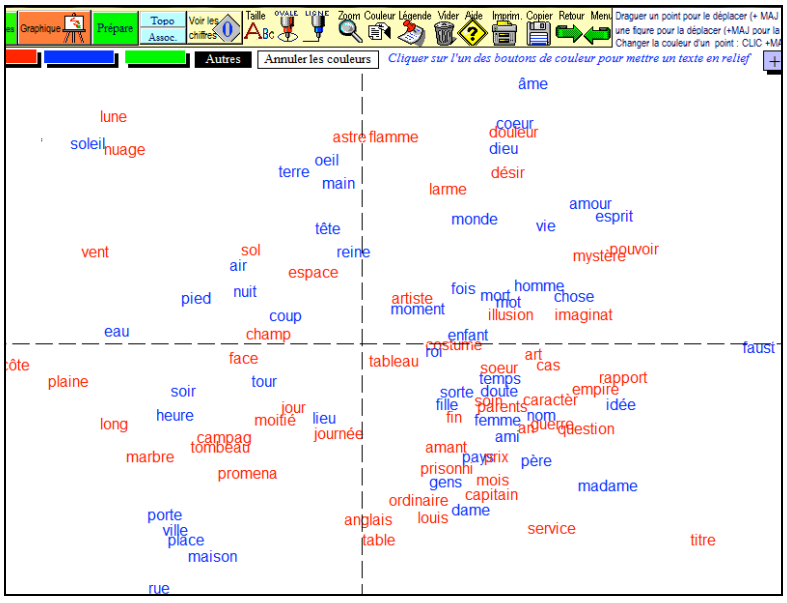

Figure 7. Écarts réduits. Mixité et équilibre des deux classes

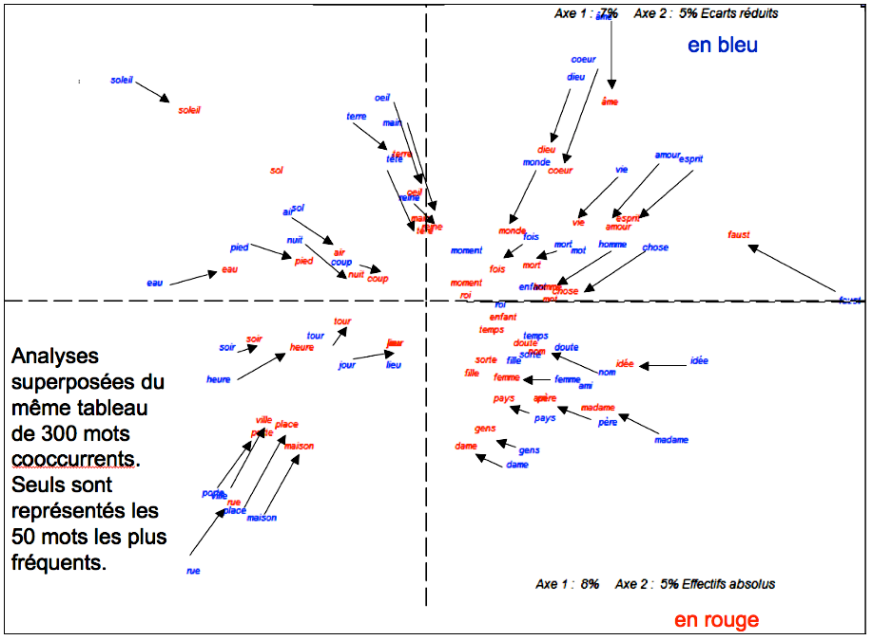

Figure 8. Les éléments lourds dans les analyses brute et réduite

Pour plus de clarté on mettra dans l'ombre les autres mots (bouton AUTRES). Dès lors on voit mieux le mouvement qui oriente vers le centre ou la périphérie les deux lots opposés. Quand les données sont brutes, les hautes fréquences se rapprochent du centre-ville où elles font la loi, en repoussant dans la banlieue les fréquences plus basses. C'est l'effet de taille classique. Sans doute la distance du Chi2, qui règne dans l'analyse de correspondance, tend à corriger ces déviances, mais elles restent visibles dans l'analyse ci-dessous (figure 6), qui rend compte des chiffres absolus. Les mots fréquents revêtus de bleu se concentrent près de l'origine des axes, tandis que les mots moins fréquents font un cercle rouge à l'entour.

Quand au contraire les inégalités trop fortes sont rabotées par la loi fiscale et hypergéométrique, la mixité et l'équilibre des classes tendent à se rétablir. C'est ce que montre la figure 7, fondée sur des calculs probabilistes (écarts réduits). Quant à l'analyse fondée sur les racines carrées des cooccurrences, elle propose un compromis qui atténue les inégalités, sans aligner les faits sur un lit de Procuste.

Pour mettre en évidence le déplacement des populations, attachons-nous au lot des nantis et superposons les positions qu'ils occupent dans

l'analyse brute (en rouge) et l'analyse réduite (en bleu) (figure 8). Ces positions sont toujours voisines et un même mot s'interdit de chevaucher un axe, vertical ou horizontal. Mais le décalage est constant pour chacun des mots : quand il apparaît en rouge, c'est-à-dire dans l'analyse brute, son poids le rapproche du centre par une sorte de gravité. Quand la péréquation est introduite, le même mot, représenté en bleu, est affranchi de la pesanteur et s'écarte de l'origine des axes. Si l'on représente parallèlement le lot des fréquences plus basses, le même mouvement est observé, mais inversé.

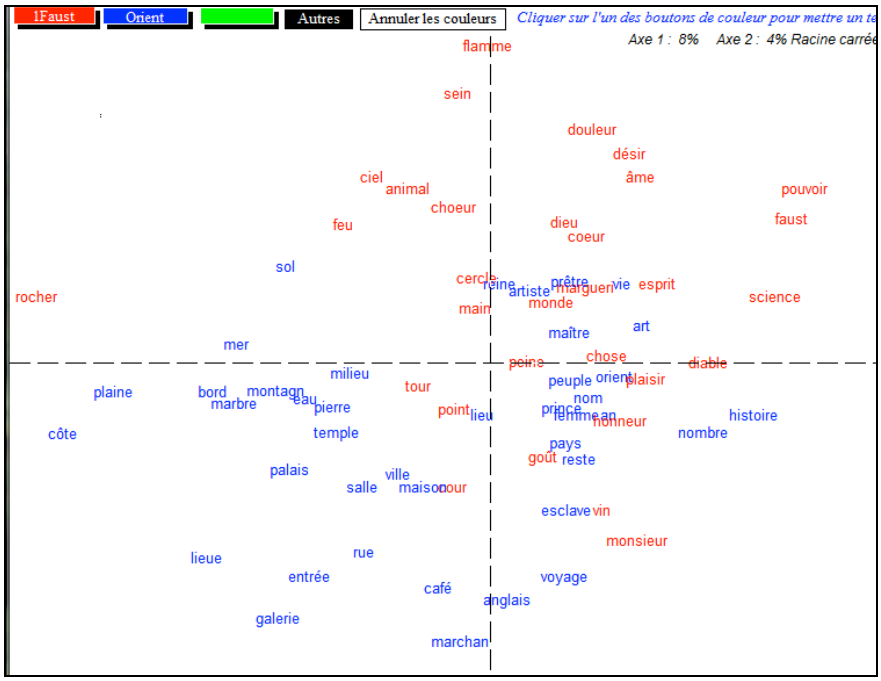

Figure 9. Textes et cooccurrences :
les choix opposés de *Faust* et du *Voyage en Orient*

7 – Le réseau des cooccurrences ignore les barrières des textes et ne considère que le champ étroit du paragraphe. On peut cependant projeter l'image des textes sur le graphique en recourant aux boutons bleu, rouge et vert. On met ainsi en couleur les mots du graphique qui sont spécifiques d'un texte (ou de plusieurs). Faculté est donnée d'isoler ces spécificités en mettant dans l'ombre les autres mots, ou de les rétablir en effaçant ou non les couleurs.

La figure 9 reprend l'analyse de la figure 5 en évacuant tous les mots qui n'appartiennent pas aux spécificités des deux textes choisis. Le premier, relatif à *Faust*, s'installe en rouge dans le quadrant supérieur droit et se cantonne dans le vocabulaire des sentiments et des valeurs où se plaît le débat dramatique. Le second, relatif à *Voyage en Orient*, occupe l'espace opposé et multiplie les curiosités locales du tourisme culturel.

8 – Mais il est un moyen plus ambitieux de rétablir la séparation et la comparaison des textes, dans l'analyse même des cooccurrences. Le bouton FACTOR, à l'extrême droite, tente de répondre à un vœu exprimé par Mayaffre et Rastier : établir la typologie des textes non plus sur les occurrences des mots individuels, mais sur les cooccurrences qui lient les mots entre eux et sont porteurs du sens (le sens des mots n'étant pas dans le dictionnaire, dans la référence, mais dans la préférence, dans le contexte des mots associés).

La difficulté vient du changement d'échelle. Un relevé de cooccurrences a trois dimensions : $N_{textes} \times (N_{mots} \times (N_{mots}\text{-}1) / 2)$, là où un tableau d'occurrences n'en a que deux : $N_{textes} \times N_{mots}$. Notre approche est de se contenter des 300 substantifs les plus fréquents et d'ignorer les cooccurrences non observées. On tente d'imiter la démarche du logiciel *Alceste*, mais en empruntant un autre chemin. Au lieu de multiplier les cases vides d'un tableau encombrant, à trois dimensions, on établit une liste des cooccurrences réellement rencontrées. Un programme d'indexation et de tri s'exerce sur les relevés et en constitue un dictionnaire alphabétique, sous le nom de la base associé au suffixe `.OCC` (par exemple `NERVAL.OCC`).

Tableau 10. Extrait du dictionnaire des cooccurrences NERVAL.OCC

abord_2_affaire_2	2	2056 2086 ,	4 2
abord_2_âge_2	2	2891 3872 ,	5 2
abord_2_air_2	1	4107 ,	6 1
abord_2_amant_2	2	2211 2785 ,	4 1 5 1
abord_2_âme_2	1	1935 ,	2 1
abord_2_ami_2	1	3724 ,	5 1
abord_2_amour_2	1	4559 ,	8 1
abord_2_an_2	2	2032 2446 ,	4 2
abord_2_animal_2	1	2080 ,	4 1
abord_2_année_2	4	1646 1649 2370 2608 ,	1 2 4 2 etc.

Un extrait de ce dictionnaire est présenté dans le tableau 10. Les cooccurrences apparaissent par couple en ordre alphabétique, avec leur fréquence dans le corpus, leurs adresses, et leur répartition dans les

textes. Par exemple le couple `abord_2_année_2` est rencontré 4 fois dans le corpus, dont 2 fois dans le texte 1 et 2 fois dans le texte 4).

Le programme d'analyse factorielle s'empare de ce dictionnaire, sans s'effrayer des milliers de lignes qu'il peut contenir (on peut écarter cependant les hapax et, dans les très gros corpus, les cooccurrences rares, limitées à quelques unités). Le fichier des paramètres de l'analyse réalisée sur Nerval, fait état de 15 190 lignes différentes (figure 11), dont les premières sont montrées ci-dessus (tableau 10). Les résultats sont consignés dans la figure 12. Noter que seuls les textes prennent position sur le graphique, les couples de mots en cooccurrence étant trop nombreux pour figurer sur l'écran (il y a 30 000 couples différents répertoriés dans le corpus GAULLE). On pourrait toutefois les consulter en changeant les paramètres du programme `ANCORR.EXE` (fichier `AFC.PAR`) et en lisant les résultats dans le fichier `ANALYSE.AFC`.

```
Voici les paramètres (qu'on peut changer en intervenant dans le fichier C:\HYPERBAS\ afc.par)

$RUN ANCORR
$L080
$F11=TABLEAU.afc
$PRT=ANALYSE.afc
$PAR=.
TITRE  ANALYSE FACTORIELLE ( écart réduit ) ;
PARAM NI = 15190 NJ = 11 NF = 4  ;
OPTIONS IMPFI=0 IMPFJ=1 NGR=2 ;
GRAPHE X=1 Y=2 GI = 0 GJ=3;
GRAPHE X=3 Y=4 GI = 0 GJ=3;
FLISTE 1Fau 2Fau Bohê Illu Orie Nuit Prom Fill Chim Auré
Pand  ;
(12X,A4,120F5.0) ;
$END
```

Figure 11. Les paramètres de l'analyse factorielle des cooccurrences

Ce n'est pas le lieu ici pour développer des commentaires littéraires sur la typologie des textes nervaliens dont certains appartiennent à l'écriture de la description ou de l'évocation, comme le *Voyage en Orient* isolé à l'extrême gauche, d'autres à l'ordre dramatique (les deux *Faust*), ou poétique (*Chimères*) ou romanesque (*Filles de feu*, *Illuminés*).

Nous importe davantage la confrontation avec les résultats habituels qui portent sur les simples occurrences. Or, mis à part une inversion, sans réelle signification, qui change l'orientation verticale, la représentation que livrent les cooccurrences n'est que l'image en miroir de celle que produisent les occurrences dans la figure 13, où les 2000 mots les plus fréquents sont pris en compte.

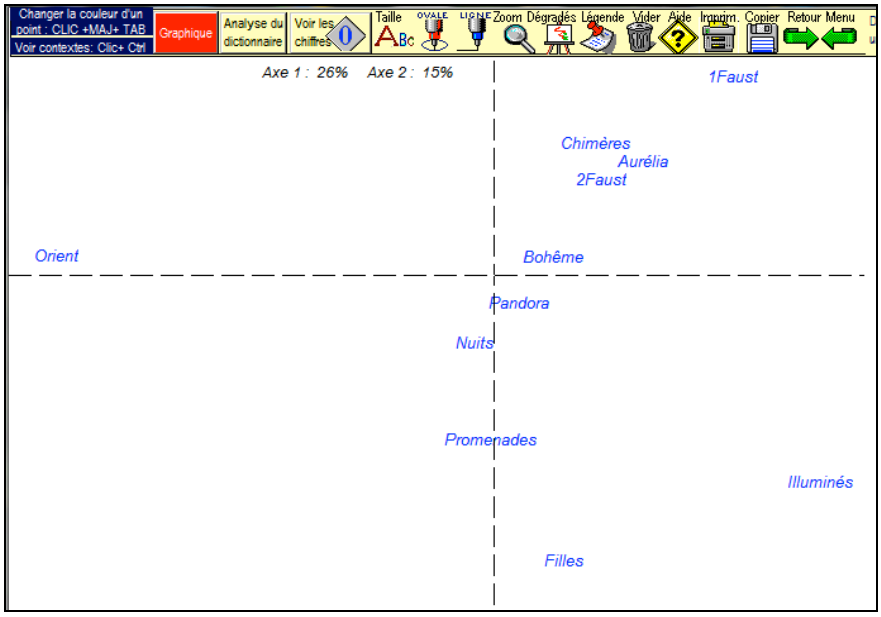

Figure 12. Analyse factorielle des cooccurrences dans les textes

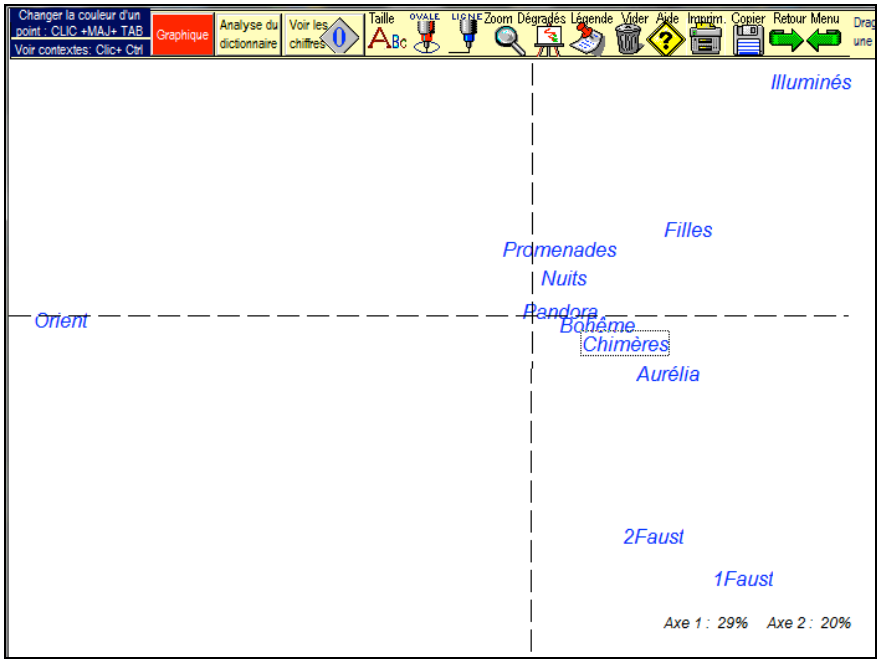

Figure 13. Analyse factorielle des occurrences dans les textes

Précisons que cette dernière analyse est obtenue par l'option ANALYSE DU DICTIONNAIRE de la page FACTORIELLE : dans les deux cas, on traite des profils, les fréquences brutes étant converties en écarts réduits.

C'est le cas aussi de l'analyse des 300 mots du tableau des cooccurrences, traités comme de simples occurrences. Une option du programme FACTOR y conduit. Ce tableau de 300 mots-lignes et de 11 textes-colonnes produit un graphique en tous points superposable au précédent, comme si ces 300 mots constituaient un échantillon admissible de la population.

Ce constat de la convergence est rassurant. On serait inquiet si des forces inconnues venaient à bouleverser l'ordre établi. Mais un brin de frustration accompagne et affadit la satisfaction. On pouvait rêver d'une lexicométrie quantique qui contesterait les observations et les lois de la lexicométrie traditionnelle. De même dans le passé l'accès plus récent au lemme aurait pu affaiblir, voire contredire, le traitement des graphies. Il n'en fut rien. Ainsi, en s'élevant des graphies aux lemmes, et des occurrences aux cooccurrences, la construction lexicométrique, d'étage en étage, reproduit les mêmes structures et s'appuie sur les mêmes fondations. Cela est même vrai quand on suit le chemin inverse et qu'on s'enfonce dans l'infra-lexical : on a démontré que la typologie des textes se maintenait inchangée quand elle portait sur les n-grammes ou séquences de 4 lettres, et qu'elle survivait ainsi au dépeçage des mots et à la ruine de la syntaxe et de la sémantique. Mieux ou pire encore, elle résiste même à la mort de l'alphabet : car l'ossature des textes reste parfaitement visible quand le texte, réduit en cendres, n'a plus qu'un code à trois éléments : voyelle, consonne et blanc.

9 – Dans un dernier effort, grimpons d'un étage encore. Cette fois le panorama s'étend non plus sur la chaîne orientée des textes, mais sur des massifs indépendants, encore plus larges : les corpus. Il va sans dire que le survol simultané de ces corpus implique qu'ils aient été soumis aux mêmes traitements et que chaque base dispose d'un fichier où sont consignés les tableaux de cooccurrences. Un tel fichier porte le nom de la base et une finale en « .soc », à l'image de celui qui accompagne la base NERVAL et dont un extrait a été présenté précédemment dans le tableau 2.

Comme la liste des 300 mots cooccurrents n'est pas identique dans toutes les bases, chacune faisant son choix indépendamment des autres, le traitement est plus épineux que dans le cas simple d'une base unique. Il

convient donc de passer en revue les différents corpus qu'on peut comparer à partir des relations cooccurrentielles.

La chiquenaude initiale part du bouton FACTOR, qui propose de choisir le niveau de l'enquête, entre textes et corpus. Si l'option CORPUS est choisie, s'ensuit une invitation à choisir les différents corpus, rassemblés dans le répertoire `C:\HYPERBAS\`. On utilisera la souris associée à la touche `CTRL` pour épingler dans le même mouvement les fichiers à suffixe `.SOC` qu'on veut retenir parmi ceux qui sont disponibles. Lorsque le choix est fait, il durera ce que dure la séance en cours, avec cependant la possibilité de le modifier. Dans la suite du traitement, l'ordre alphabétique de ces fichiers sera respecté. S'il ne convient pas, ajouter préalablement un numéro d'ordre devant les noms des fichiers à traiter pour imposer le classement désiré.

On procède alors à l'exploration des différentes bases ou plus précisément du fichier où est enregistré le détail des cooccurrences qu'on y trouve. De tableaux remplis de chiffres on tire un long ruban de cooccurrences, répétées autant de fois que nécessaire, avec pour seule structure un jalon indiquant le passage d'un corpus à l'autre. En somme on a constitué un texte réduit aux seules cooccurrences, lesquelles seront considérées comme des mots ordinaires une fois qu'on a collé l'un à l'autre les deux membres de la cooccurrence. Dès lors les programmes habituels d'indexation peuvent être mis en œuvre, pour aboutir à un dictionnaire récapitulatif des cooccurrences, pourvu des sous-fréquences par corpus (tableau 14).

Tableau 14. Extraction et regroupement des cooccurrences à partir de plusieurs corpus

```
&&&1,1,01&&&
abbé_2_abeille_2  abbé_2_abord_2  abbé_2_action_2  abbé_2_action_2
abbé_2_affaire_2 abbé_2_affaire_2 abbé_2_affaire_2 abbé_2_affaire_2
abbé_2_âge_2  abbé_2_âge_2  abbé_2_âge_2  abbé_2_air_2  abbé_2_air_2
abbé_2_air_2  abbé_2_air_2  abbé_2_air_2  abbé_2_âme_2  abbé_2_âme_2
abbé_2_âme_2  abbé_2_âme_2  abbé_2_âme_2  abbé_2_ami_2  abbé_2_ami_2
abbé_2_ami_2  abbé_2_amour_2  abbé_2_amour_2  abbé_2_an_2  abbé_2_an_2
abbé_2_an_2 abbé_2_an_2   etc.
```

Poursuivant son chemin, le programme explore chaque ligne de ce dictionnaire des fréquences et soumet l'ensemble à l'analyse factorielle. Les données étant des effectifs bruts, démunis de probabilités, il n'est guère envisageable d'en tirer des écarts. Afin de limiter l'effet de taille qu'on peut craindre avec des corpus et des mots de poids inégal, les données brutes sont converties en racine carrée. Le résultat apparaît dans la figure 15.

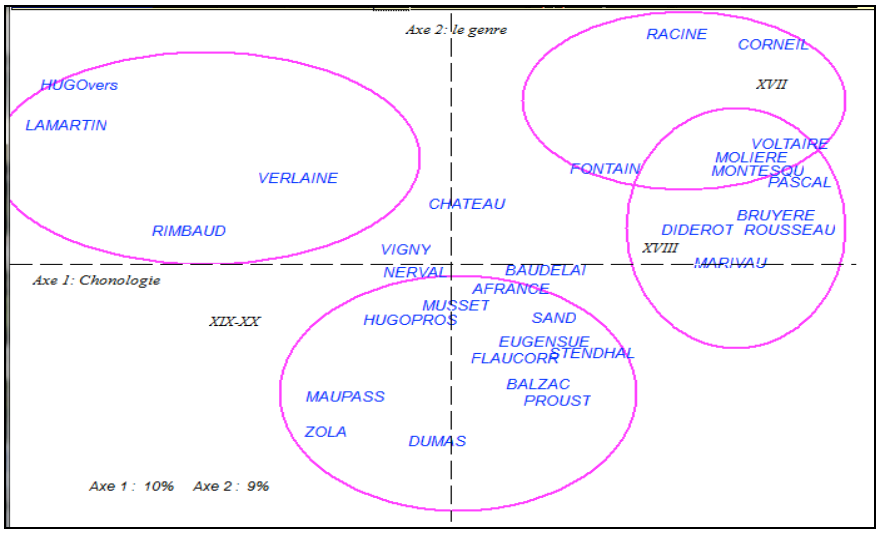

Figure 15. Analyse factorielle de plusieurs corpus selon la distribution des cooccurrences

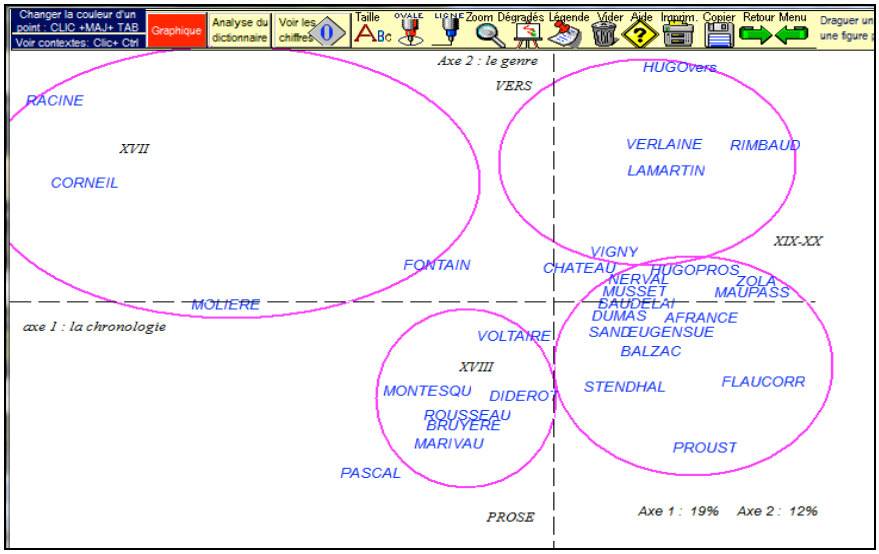

Figure 16. Analyse factorielle de différents corpus selon les occurrences simples, après dégroupement des cooccurrences

On a juxtaposé les résultats obtenus avec les mêmes données, préalablement dégroupées. Les mots qui entrent dans la composition de la cooccurrence sont maintenant traités individuellement, comme de simples

occurrences. Dans l'urne s'agitent des mots indépendants qui se considèrent comme étrangers les uns aux autres, une révolution morale ayant aboli toutes les unions ou accordailles. Ce traitement est assuré par une option du bouton FACTOR (figure 16).

On n'attachera aucune signification à l'inversion qui présente en miroir le premier facteur, qui dans les deux cas est le reflet de la chronologie : d'un côté le XVII^{e} siècle et le XVIII^{e}, de l'autre le XIX^{e} et le XX^{e}. Pareillement dans les deux figures le second facteur rend compte du genre, les vers à la surface, la prose en profondeur. Mais un brouillage s'observe parallèlement dans les deux analyses : l'interférence du genre dans la chronologie. La Bruyère et Pascal se détachent des versificateurs du siècle classique pour rejoindre les prosateurs moralistes du XVIII^{e}. Molière, dont l'œuvre est pour la moitié en prose, subit cette attirance, tandis que Voltaire, à cause de son théâtre versifié, s'oriente dans le sens inverse. La production multi-genre n'est pas l'apanage de Voltaire. Elle est observée chez beaucoup d'écrivains comme Vigny, Musset, Nerval, Chateaubriand ou Baudelaire, qui divisés et indécis se rapprochent par là même du marais central. Les positions seraient plus nettes si la séparation introduite chez Hugo entre les œuvres en vers et les œuvres en prose avaient été généralisée. Mais le genre implique bien d'autres catégories que le vers et la prose et la diversification des données pouvait s'étendre à l'infini.

Mais comme précédemment le but n'est pas d'obtenir une carte fidèle de la littérature française mais de vérifier si l'approche des cooccurrences plutôt que des occurrences apporte une précision supplémentaire, voire une vision renouvelée. La réponse semble devoir être provisoirement évasive, les deux analyses étant superposables. On se gardera d'en conclure que l'examen des cooccurrences est inutile : une confirmation n'est jamais une épreuve superfétatoire. On s'en convaincra avec l'examen du théâtre classique (figure 17). Pratiqué sur les cooccurrences, il reproduit exactement l'image obtenue sur les occurrences : une carte où les trois écrivains délimitent nettement leur territoire, sauf dans les rares occasions où les lois du genre s'y opposent (les *Plaideurs* de Racine au milieu des pièces de Molière, la pièce sérieuse de Molière, *Garcie de Navarre*, parmi les tragédies de Corneille).

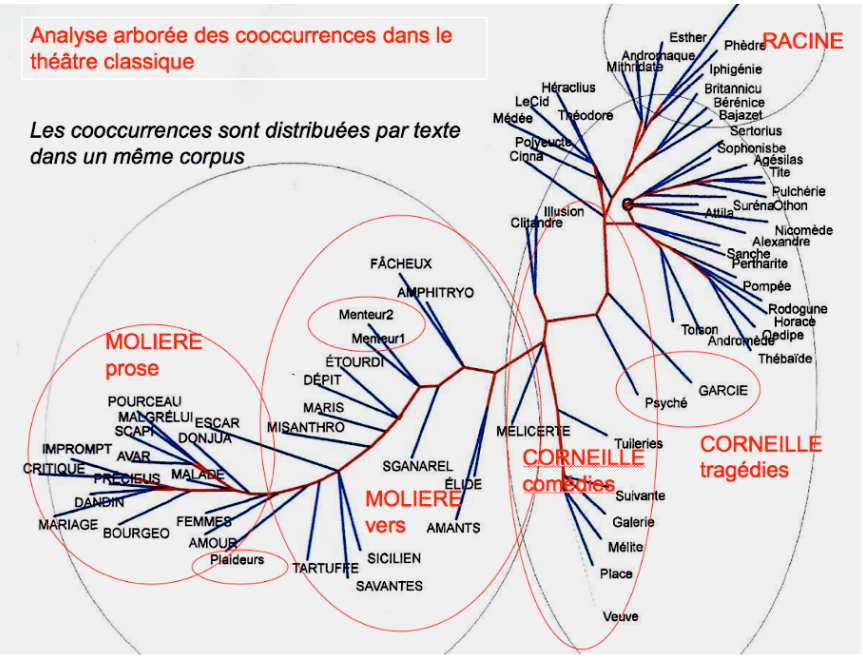

Figure 17. La distance intertextuelle, calculée sur les cooccurrences

10 – Mais si l'image globale n'est pas bouleversée, le détail de la distribution apporte souvent des indications lumineuses sur telle ou telle thématique. Pour illustrer cette approche dans HYPERBASE, on doit activer le bouton ASSOCIATIONS qui voisine avec le bouton CORRELATS dans le menu principal. En réalité ces deux dénominations recouvrent la même chose : les cooccurrences. Mais avec les « corrélats » on envisage l'aspect global du phénomène dans le cadre d'une sélection représentative, et avec les « associations », même « généralisées », on privilégie leur étude détaillée, dans le même cadre. Quand la page ASSOCIATIONS invite à « choisir un pôle », proposons par exemple le mot BONHEUR.

Parmi les 300 mots de la sélection, le BONHEUR a la bonne fortune de trouver des alliés (et des antonymes). Or un lien peut être établi avec les autres bases disponibles, pour observer si les relations cooccurrentielles y sont semblables ou différentes. Le bouton HISTOGRAMME, non content de reproduire sur un graphique les relations préférentielles du mot-pôle dans le corpus exploré, invite à y ajouter d'autres corpus, au moins ceux qui ont des données comparables à propos du mot considéré. Ainsi voit-on dans la figure 18 que chez Pascal le BONHEUR est dans l'au-delà et dans

un rapport à Dieu tandis que Proust y voit un REVE impossible associé au CHAGRIN, à la SOUFFRANCE, au TEMPS destructeur, à la TRISTESSE et bien sûr au DESIR et à l'AMOUR. Si l'on juxtaposait les données de Stendhal, grand assoiffé de bonheur, l'AMOUR serait aussi au premier plan, mêlé à l'ENNUI, à l'ORGUEIL et à la VANITE.

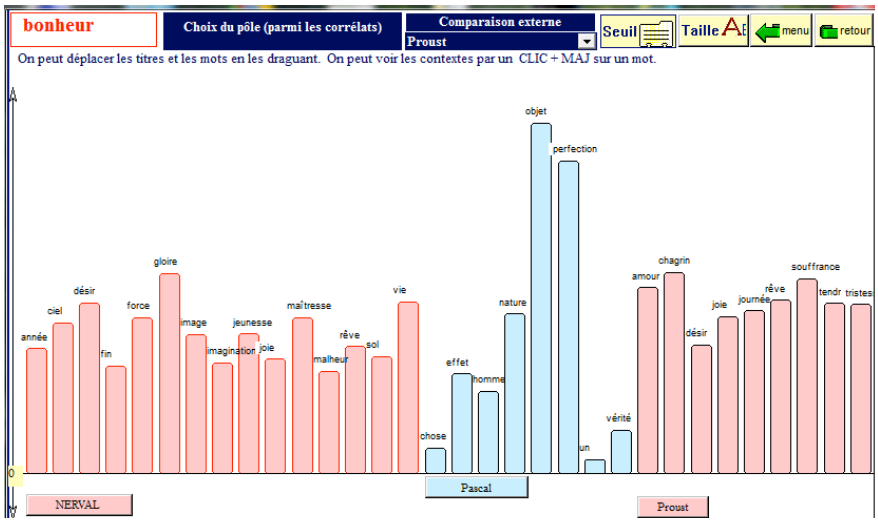

Figure 18. L'environnement du bonheur chez quelques écrivains

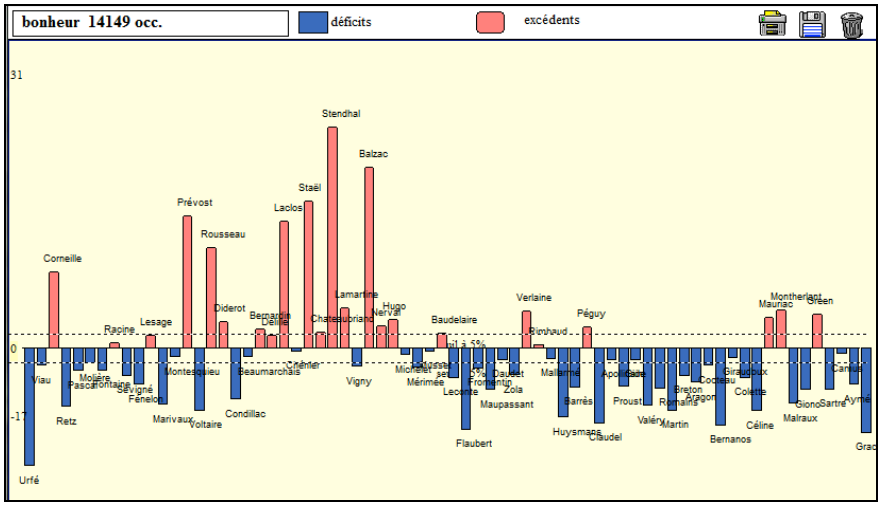

Figure 19. Le bonheur sans assaisonnement. Répartition chez 75 écrivains

Cette excursion des cooccurrences hors du corpus semble très riche d'information et renouvelle l'approche comparative. Si l'on se contentait

des occurrences simples relevées sans tenir compte du voisinage, la courbe du BONHEUR serait beaucoup plus pauvre et sèche, sans les connotations ou harmoniques que permet l'environnement cooccurrentiel. Le graphique 19 a beau répertorier 14 000 emplois du mot chez 75 écrivains, on voit bien que Stendhal est le champion de la catégorie, mais la notion de bonheur n'y reçoit aucun éclairage. Même sans polysémie trop violente, le mot reste enveloppé dans un brouillard que la cooccurrence aiderait à percer.

11 – La tentative précédente est limitée à un zoom spectaculaire mais instantané, qui n'éclaire qu'un mot à la fois. C'est aussi le reproche qu'on peut faire à l'entreprise de *Google Books*, qui a canalisé les textes de l'édition mondiale dans un réservoir monstrueux, dont 44 milliards de mots pour le domaine français !. Mais à la sortie de cet immense bassin, il n'y a qu'un robinet étroit qui distribue les mots un par un, au compte-gouttes. C'est pourquoi en puisant dans le même bassin (généreusement téléchargeable), on a cru devoir constituer une base plus ouverte (appelée GOOFRE) où puissent s'employer les outils de la statistique et notamment les analyses multidimensionnelles.

Ne pourrait-on pas tenter une expérience semblable avec les cooccurrences ? Damon Mayaffre dans un colloque tenu à Rome en octobre 2011 regrettait qu'il n'existe pas pour les cooccurrences une méthodologie lexicométrique analogue à celle qui a cours pour les occurrences. *A priori* on pourrait estimer que c'est la même, mais à une autre échelle, la seconde étant de l'ordre du carré de la première (plus précisément à la dimension n correspond la dimension $n\times(n\text{-}1)/2$). Deux obstacles liés se dressent sur la route : d'une part la taille démesurée des tableaux et la longueur des calculs, et d'autre part la faiblesse des effectifs, la plupart étant nuls. Pour échapper à ce dilemme, on propose de diminuer le nombre de mots étudiés, en réponse à la première aporie, et d'augmenter la taille des données, pour répondre à la seconde. La solution a été entrebâillée dans les pages qui précèdent : ne s'intéresser qu'à une fraction restrictive mais représentative de la population, les 300 substantifs les mieux représentés dans chaque corpus. Et parallèlement étendre l'enquête à tous les corpus qu'on voudra comparer.

Pour n'être pas lié à une fonction particulière d'un corpus particulier, on a créé une version nouvelle d'HYPERBASE, dénommée HYPOCCUR. C'est un modèle riche de programmes et vide de données, comme HYPERTAG ou HYPERCOR. Quand on le lance, il génère une copie que l'on est invité à remplir. Il ne demande qu'une seule espèce de données :

non pas des textes, mais des tableaux de cooccurrences, expressément consignés par HYPERBASE dans des fichiers à suffixe .SOC. La procédure initiale est donc la même que celle qu'on a décrite précédemment (tableau 14). Mais le traitement diverge ensuite. Au lieu de créer seulement un dictionnaire des fréquences, on constitue une véritable base, semblable à celles qu'on peut réaliser avec HYPERBASE. La suite des mots en cooccurrence est enregistrée à la queue-leu-leu, comme s'il s'agissait d'un véritable texte en continu. Quand on passe d'un corpus à l'autre, un jalon s'interpose, qui joue le même rôle que la barrière qui sépare deux textes dans un corpus normal. Dès lors tout s'accomplit jusqu'au terme sans changement notable dans le traitement, les couples cooccurrents devenant des mots et les corpus des textes.

Une fois la base constituée, les fonctions documentaires et statistiques d'HYPERBASE sont accessibles, même si certaines perdent leur justification, comme la CONCORDANCE ou la TOPOLOGIE, vu le statut particulier des données. On peut certes renouveler le calcul de l'analyse factorielle, présentée dans le graphique 15, mais beaucoup d'autres, sur choix partiel ou différent, peuvent présenter de l'intérêt. La page LISTE offre toutes les options possibles tant pour la sélection des données que pour le choix des outils. Parmi ceux-ci les figures 20 et 21 offrent deux variétés de représentation arborée, l'une fondée sur la « connexion » de Muller, l'autre sur la formule d'Évrard. Dans les deux cas, il s'agit d'apprécier la distance intertextuelle à partir des relations cooccurrentielles. Dans les deux graphiques, des régions périphériques au contour net, circonscrites autour de corpus homogènes (XVII^{e}, XVIII^{e}, Roman, Vers) enveloppent une zone centrale, plus floue, où le mélanges des genres crée l'indécision.

À côté des vues synthétiques, factorielles ou arborées, des faits saillants peuvent sortir d'un simple histogramme, comme celui du couple MERE-ENFANT, que les siècles classiques semblent ignorer, jusqu'à ce que Rousseau s'indigne et s'enflamme dans l'*Émile* à propos de l'allaitement. Tous ceux qui l'ont précédé dans la littérature s'inscrivent dans la zone négative. Les valeurs de la maternité et de l'enfance se déploient avec et après Rousseau dans la génération qui, de Diderot à Balzac, a eu à souffrir des nourrices et des internats. On observera que sous ses airs de garçonne Georges Sand trône au sommet de la courbe 22, comme symbole non seulement de la féminité mais de la maternité.

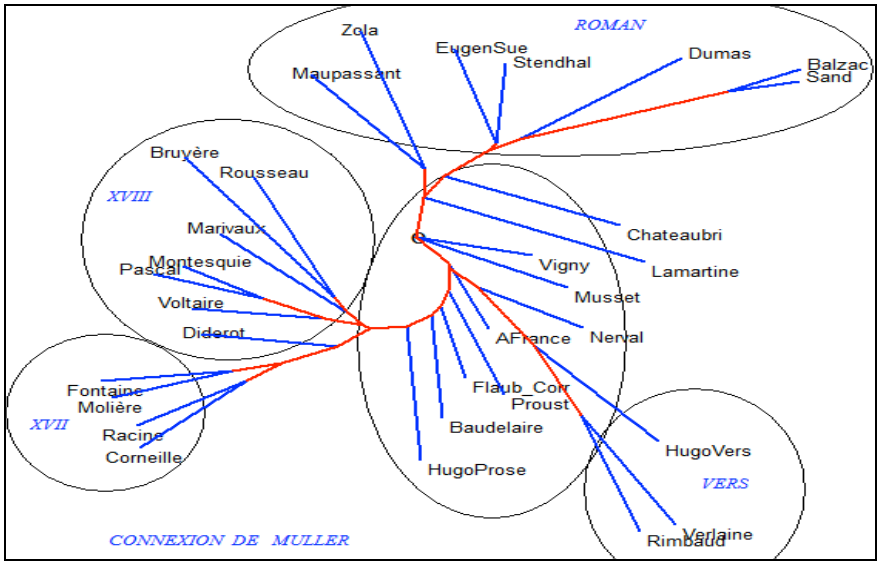

Figure 20. Analyse arborée selon la « connexion » de Muller

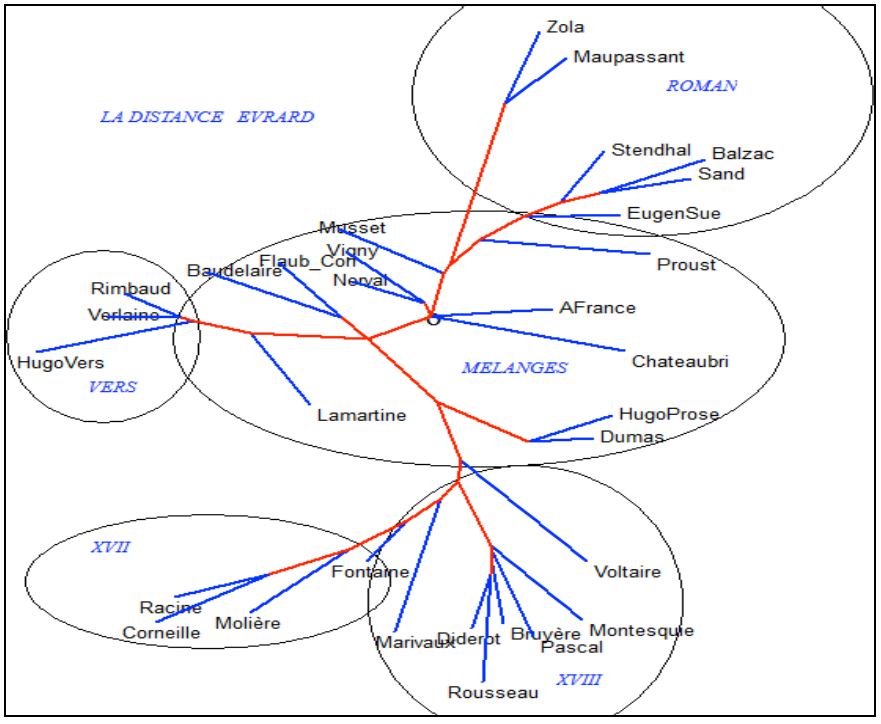

Figure 21. Analyse arborée selon la formule d'Étienne Évrard

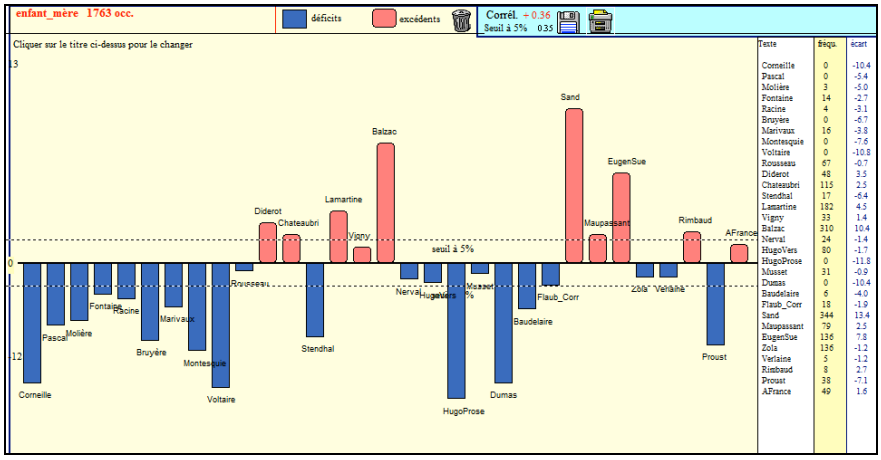

Figure 22. Graphique du couple mère-enfant

Tableau 23. L'évolution des couples lexicaux
En progression (à gauche) et en régression (à droite)

Progression	Fréquence	Forme	Régression	Fréquence	Forme
+ 0.672	137	air_travers	- 0.627	81	gloire_lieu
+ 0.666	301	air_soir	- 0.626	107	coeur_discours
+ 0.609	167	année_heure	- 0.616	166	oeil_vertu
+ 0.602	286	air_fleur	- 0.609	176	jour_vertu
+ 0.602	235	air_matin	- 0.605	162	âme_objet
+ 0.595	143	fleur_heure	- 0.603	76	état_honneur
+ 0.591	319	chambre_nuit	- 0.602	170	lieu_roi
+ 0.583	308	chambre_soir	- 0.598	75	faveur_grâce
+ 0.582	143	soir_table	- 0.586	300	esprit_lieu
+ 0.582	107	côté_soir	- 0.585	396	coeur_vertu
+ 0.579	102	face_monde	- 0.585	101	faveur_homme
+ 0.575	101	doute_nuit	- 0.584	139	gloire_vertu
+ 0.573	237	chambre_milieu	- 0.583	86	chose_pouvoir
+ 0.569	158	heure_joie	- 0.579	191	honneur_vertu
+ 0.566	142	saint_soir	- 0.575	105	coeur_faveur
+ 0.565	155	année_fille	- 0.573	125	coeur_pouvoir
+ 0.564	554	chambre_lit	- 0.572	118	honneur_lieu
+ 0.563	230	fenêtre_nuit	- 0.572	114	âme_ordre
+ 0.563	92	eau_face	- 0.570	138	âme_soin
+ 0.562	117	chambre_lumière	- 0.568	91	cour_lieu
+ 0.562	93	bonheur_rêve	- 0.562	98	honneur_intérêt
+ 0.560	307	mère_soir	- 0.562	83	coeur_ennemi
+ 0.560	91	mère_table	- 0.558	79	âme_bien
+ 0.557	111	chambre_fleur	- 0.554	107	coeur_cour
+ 0.556	113	fille_souvenir	- 0.553	171	force_vertu
+ 0.554	88	regard_table	- 0.553	131	amant_moment
+ 0.554	84	année_doute	- 0.550	166	coeur_soupir
+ 0.553	352	air_pied	- 0.550	95	bien_honneur
+ 0.553	155	matin_tête	- 0.548	95	jour_pouvoir
+ 0.552	123	pied_table	- 0.547	252	esprit_peine
+ 0.552	114	air_argent	- 0.546	145	bien_coeur
+ 0.551	237	heure_travail	- 0.545	240	coeur_objet
+ 0.551	82	doute_soir	- 0.545	85	coeur_mérite

L'évolution des mœurs qu'on touche du doigt à propos d'un couple comme MERE-ENFANT ou TRAVAIL-FAMILLE, peut être appréhendée globalement à partir du coefficient de corrélation, calculé pour tous les couples. Ceux qui gagnent ou perdent en crédit apparaissent dans le tableau 23.

Ce faible extrait d'une liste beaucoup plus longue montre assez la tendance littéraire qui s'éloigne des abstractions un peu laiteuses dont s'abreuvaient les siècles classiques : la VERTU, l'HONNEUR, la GLOIRE, l'AME, la GRACE, le CŒUR, et même l'ESPRIT, qui se mêlent à loisir dans les combinaisons cooccurrentielles de l'époque, voient leur cours dévalué dans la colonne des régressions. Inversement la colonne de gauche souligne l'invasion du concret, du cadre de vie (CHAMBRE, FENETRE, LIT, TABLE, FLEUR, EAU, LUMIERE), du corps humain (PIED, TETE, REGARD), des relations de famille (MERE, FILLE) et surtout l'obsession du temps : sur les 32 couples les plus prisés de notre époque, le SOIR compte 6 mentions, l'HEURE 4, la NUIT 3, le MATIN 3 et l'ANNEE 2.

On n'ignore pas que dans ces changements, dont témoignent les cooccurrences, le genre, lui-même soumis aux variations du temps et de la mode, a une large influence qui s'impose à l'écrivain. C'est pourtant l'écrivain qui choisit son genre ou qui l'invente, suivant ses goûts et ses dons. Et les spécificités de chacun apparaissent plus clairement dans les couples que dans les mots isolés. Ainsi dans le tableau 24, qui met en parallèle Corneille et Rousseau, on trouverait sans difficulté des mots appartenant aux deux listes, comme CŒUR par exemple. Mais chez Corneille le CŒUR a ses raisons et ses cooccurrents qui ne sont pas les mêmes chez Rousseau. On chercherait en vain un couple qui soit identique chez les deux écrivains. Les éléments peuvent être communs aux deux, non les combinaisons. Bien entendu si l'on avait rapproché Corneille de Racine, les rencontres seraient plus nombreuses, notamment celles que la rime engendre : AME-FLAMME, CHOIX-ROI, CONQUETE-TETE, GLOIRE-VICTOIRE (ce sont les quatre premiers couples chez Corneille), GLOIRE-VICTOIRE, PRINCE-PROVINCE, CRIME-VICTIME, GLOIRE-MEMOIRE, CŒUR-VAINQUEUR, ALARME-LARME (ces couples sont en tête des spécificités de Racine).

Parmi les outils disponibles l'analyse factorielle reste la pièce maîtresse. Elle peut ici s'appliquer à la phraséologie qui tourne autour d'un mot. On s'est longtemps battu sur le crédit incertain qu'on pouvait donner au lemme, en invoquant les nuances de sens qui s'attachent au singulier et au pluriel (PEUPLE *versus* PEUPLES, LIBERTE *versus* LIBERTES). Lemme ou simple graphie, un mot isolé n'a guère de substance. Il vit de son environnement. Or la cooccurrence, en lui restituant une part de son entourage, lui redonne sens et signification. Ainsi le débat sur LIBERTE et LIBERTES s'efface quand mille libertés différentes jaillissent du texte en demandant classement et typologie.

Tableau 24. Les spécificités cooccurrentielles de Corneille et de Rousseau

Retour Sommaire **Rousseau (positif)** Trier

Cliquer sur un mot pour connaître sa spécificité dans les

écart	corpus	texte	mot
25.3	704	200	état_homme
22.3	231	109	auteur_livre
20.7	230	101	état_nature
19.6	50	50	autrui_homme
18.0	92	60	corps_membre
17.8	64	51	éducation_homme
17.1	317	95	devoir_homme
16.8	81	53	citoyen_droit
16.7	618	127	besoin_homme
16.6	129	63	droit_nature
16.0	72	48	corps_volonté
15.9	118	58	citoyen_loi
15.7	106	55	forme_gouverneme
15.3	357	90	espèce_homme
15.0	65	43	autorité_raison
15.0	31	31	attachement_coeu
14.8	240	73	coeur_objet
14.8	228	71	auteur_homme
14.6	83	46	autorité_loi
14.6	42	35	autorité_droit
14.4	2176	230	coeur_homme
14.3	108	50	auteur_lettre
14.1	72	42	femme_sexe
14.1	37	32	constitution_hom
14.0	58	38	autorité_homme
13.8	27	27	droit_souverain
13.7	93	45	citoyen_homme
13.6	426	88	coeur_plaisir
13.6	129	51	chef_peuple
13.5	26	26	autrui_mal
13.5	112	48	état_un
13.4	804	121	coeur_sentiment
13.3	46	33	citoyen_magistra
13.2	44	32	éducation_enfant
12.9	72	38	conseil_droit
12.9	513	92	coeur_raison
12.8	55	34	expérience_homme

Retour Sommaire **Corneille (positif)** Trie

Cliquer sur un mot pour connaître sa spécificité dans

écart	corpus	texte	mot
27.8	490	171	âme_flamme
23.7	82	72	choix_roi
21.2	201	90	amour_haine
20.0	60	53	conquête_tête
19.9	238	90	gloire_victoire
19.7	2222	251	amour_coeur
19.7	154	75	envie_vie
19.4	93	61	coeur_rigueur
19.2	1302	185	amour_jour
18.6	213	80	coeur_flamme
18.5	137	67	coeur_vainqueur
18.3	306	91	coeur_roi
17.6	73	50	crime_victime
17.4	41	39	haine_reine
17.2	126	60	choix_coeur
17.1	178	68	coeur_seigneur
17.1	129	60	coeur_empire
16.9	73	48	amour_choix
16.9	220	73	amour_roi
16.8	97	53	amour_ardeur
16.3	82	48	haine_peine
16.3	118	55	coeur_haine
16.2	70	45	choix_loi
16.1	38	35	attentat_etat
16.0	106	52	effort_mort
15.9	103	51	amour_faveur
15.8	236	70	coeur_gloire
15.6	139	56	coeur_voeu
15.4	97	48	amour_voeu
15.4	188	62	douleur_malheur
15.3	57	39	époux_main
15.0	104	48	amour_espoir
14.9	40	33	couronne_personn
14.9	40	33	coeur_tyran
14.9	174	58	âme_roi
14.8	77	42	amour_hymen
14.8	155	55	amour_crime

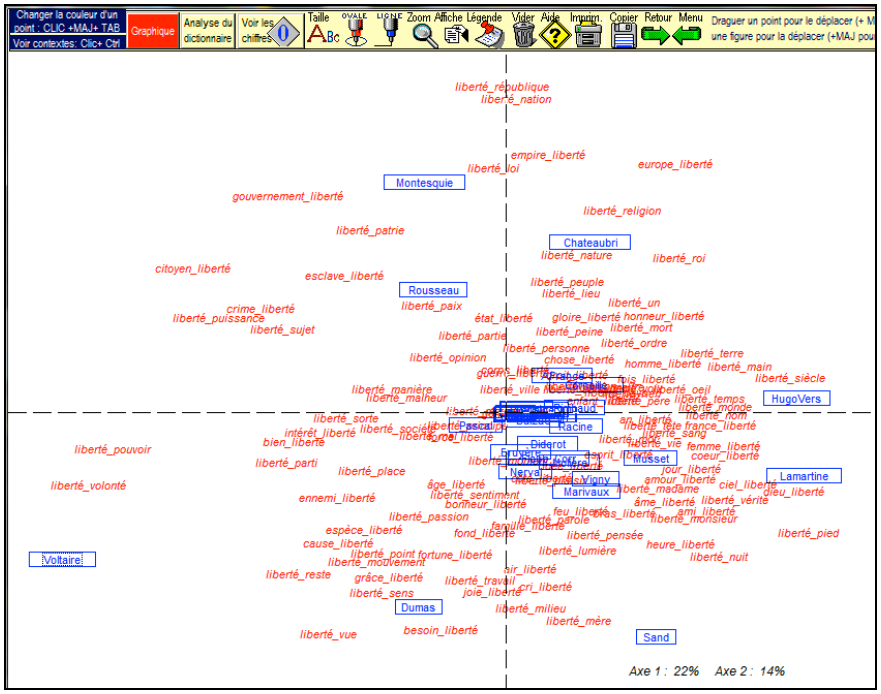

Figure 25. Analyse factorielle des couples où règne la LIBERTE

La figure 25 éclaire le visage multiple où s'incarne le mot au fil des siècles. Les écrivains antérieurs aux Lumières ne lui donnent pas grand crédit et ils s'agglutinent au centre du graphique sans prendre parti. C'est Montesquieu au haut du graphique qui donne au mot ses lettres de noblesse, en l'associant au vocabulaire politique (REPUBLIQUE, NATION, EMPIRE, LOI, GOUVERNEMENT). À sa gauche Rousseau oriente le terme dans un sens plus social et plus affectif (PATRIE, PAIX, ESCLAVE, CRIME, OPINION), tandis qu'à droite Chateaubriand rapproche la liberté des questions religieuses et monarchiques (RELIGION, ROI, PEUPLE, EUROPE). Dans la moitié basse de la figure, l'opposition est encore celle de la chronologie : Voltaire à gauche et Lamartine à droite, avec l'appui de quelques écrivains engagés comme Hugo et Sand. Chez eux la liberté est moins une notion théorique de l'économie politique qu'un étendard sous lequel il convient de se battre, à l'image du tableau célèbre *La liberté guidant le peuple*. Le combat de Voltaire (qui n'a pas oublié son séjour à la Bastille) est celui de la VOLONTE, contre l'ENNEMI, contre le POUVOIR, contre le PARTI. Celui de Lamartine, de Sand et de Hugo, est tout aussi sincère, mais il s'entoure de notions ou sentiments plus vagues où la liberté est un ingrédient qu'on

utilise dans toutes les sauces et qui peut s'associer à l'AMOUR, à la FEMME, à DIEU, à la VERITE, etc. Le cas de Dumas en position centrale est un peu particulier : la liberté y est attachée trop rudement à la PRISON. C'est un élément d'une intrigue qui multiplie les emprisonnements, les évasions et les délivrances.

En conclusion il reste à prolonger l'expérimentation des méthodes fondées sur la cooccurrence. Les segments répétés du logiciel *Lexico* ont ouvert une brèche, et le succès d'*Alceste* a montré qu'il y avait une attente de la communauté scientifique, que voudrait satisfaire, dans une modeste mesure et parmi beaucoup d'autres[3], notre contribution.

[3] Signalons deux logiciels récents, spécialisés dans le traitement des cooccurrences. L'un, *Gephi*, s'appuie sur le tableau des cooccurrences et crée un graphe où les nœuds (c'est-à-dire les mots) et aussi les liens s'articulent en constellations colorées et contrastées qui tendent à reproduire la carte thématique du texte. Le second, *Iramuteq*, a été créé par Pierre Ratinaud et propose des analyses dont l'une reprend la démarche d'*Alceste*. Comme ce dernier logiciel, *Iramuteq* peut constituer le tableau des cooccurrences à partir du texte même. Comme nous l'avions fait pour *Alceste*, HYPERBASE a été aménagé pour fournir à *Gephi* et à *Iramuteq* des données appropriées dans le format requis. Pour *Gephi* on fournit un tableau d'associations au format GML (bouton GRAPHE GEPHI dans la page ASSOCIATIONS).. Pour *Iramuteq* (comme pour *Alceste*) on fournit un fichier constitué de séquences de mots associés, lemmatisés et débarrassés des mots-outils, avec les jalons textuels propres au format *Alceste*. Signalons que *Gephi* et *Iramuteq* (et la bibliothèque statistique *R* sur laquelle s'appuient certains programmes) sont des logiciels libres qu'on peut télécharger.

`https://gephi.org/users/download/`

`http://www.iramuteq.org/telechargement/`