# Dynamisation de l'analyse micro-distributionnelle des corpus textuels. Exemple de *Voyage au bout de la nuit*

Jean-Marie Viprey[1]

[1] GRELIS-LASELDI – Univ. De Franche-Comté - F-25030 Besançon Cedex – France

## Abstract

Micro-distributional analysis which is permitted by FAC applied to large cooccurrences arrays, may apply to corpuses in which diachrony is relevant; in that case it is able to incorporate diachrony even better that macro-distributional methods. ®ASTARTEX, developped in the Grelis at Besançon, includes a function of animation for the CFA outputs; it computes transitions between maps representing successive overlaping sections. Highlighting the most migratory items on the one hand, the large areas of vocabulary at every phase, on the other hand allows these animations to permit the non-specialist user to see the headlines of the dynamic process of text, giving its full efficiency to FAC imagery in the domain of texts. We show the dialectic between an invariant framework (items being subject to little migrations) and widely migrating items, which obviously bear the big thematic variations which animate the text. We underline as well the eventual breaking lines within a global linear structure. Celine's novel *Voyage au bout de la nuit* shows a strong variation in its first half, and next each following section models itself increasingly on the overall structure.

## Résumé

L'analyse micro-distributionnelle permise par l'AFC de grandes matrices de cooccurrence, appliquée à des corpus dotés de diachronie, peut récupérer celle-ci autant et mieux que l'analyse macro-distributionnelle. ®ASTARTEX, du Grelis de Besançon, comporte un module d'animation des sorties d'AFC par calcul des transitions entre les graphes extraits de tranches successives chevauchantes. Eclairées par divers moyens graphiques de repérage concernant les items les plus migrateurs d'une part, et les grandes zones du vocabulaire aux diverses phases, ces animations permettent à un utilisateur non spécialiste de visualiser les traits saillants de la dynamique processuelle du texte, donnant toute sa puissance à l'imagerie AFC dans le domaine des textes. On met ainsi en évidence la dialectique d'une ossature invariante (items subissant de très faibles migrations) et d'items très migrateurs, manifestement porteurs des grandes variations thématiques qui animent le texte. On repère aussi les éventuelles lignes de rupture et une structure linéaire globale. *Voyage au bout de la nuit*, par exemple, présente une forte variation dans sa première moitié, après quoi chacune des tranches ultérieures se calque de plus en plus sur la structure d'ensemble

**Mots-clés :** analyse multi-dimensionnelle, vocabulaire, diachronie, animation graphique, ergonomie

## 1. Introduction

L'AFC appliquée à des matrices (« carrées ») de cooccurrence, où les individus (lignes) et les variables (colonnes) sont les vocables les plus fréquents, nous permet depuis quelques années d'envisager une cartographie lexico-thématique au service de l'orientation hypertexte dans les bases et corpus littéraires. Nous en rappellerons le principe : aux hypothèses thématiques exogènes, projectives, l'on substitue ou l'on combine des propositions endogènes sous la forme des *isotropies*, proximités complexes et continues d'items lexicaux (lemmatisés, étiquetés ou à l'état brut).

Cette approche distributionnelle est beaucoup plus fine que celle permise par l'analyse classique des matrices (« rectangulaires ») de répartition dans les unités successives assez étendues de texte (c'est le sens que nous donnons à l'opposition micro- / macro-

distributionnel). Lorsque ces unités de contexte se raccourcissent à la dimension de la phrase ou du syntagme, comme c'est le cas d'Alceste, nos méthodes ont beaucoup d'affinités. Mais nous privilégions l'AFC parce qu'elle est la seule à nous offrir une sortie graphique sous la forme d'un nuage dans le plan, ou dans l'espace de dimension 3, nuage qui représente remarquablement le continuum des relations *isotropiques*, parentés de profils cooccurrentiels. Nous perdons certes le bénéfice d'une ferme classification hiérarchisée, mais il nous semble clair que dans les grandes matrices (dont la dimension –J- est de l'ordre de 200, voire 500), les dichotomies s'opèrent souvent sur des impondérables. Aussi avons-nous préféré offrir le continuum orthogonal à l'interrogation des chercheurs.

Nous avons donc axé nos efforts sur l'ergonomie de l'offre des résultats. Dans l'environnement ®ASTARTEX, développé au Grelis de l'Université de Franche-Comté, testé notamment sur des éléments de la base ®Basile (Cned-Champion) et sur divers corpus discursifs, nous avons d'abord cherché à disposer les items de la sortie AFC selon les 2 ou 3 premiers axes (d'après ®Anaconda) de manière lisible et directement utilisable, sans recours aux listings classiques. Pour cela, nous avons recalculé les coordonnées des points à partir de l'origine de manière à les disperser au mieux sans modifier le moins du monde leurs positions relatives et en conservant la proportionalité de toutes les directions à partir de l'origine.

Puis nous avons cherché à faire des items et des groupes d'items du graphe des zones sensibles dans le sens hypertexte, c'est-à-dire des outils graphiques de relance de la recherche et d'aide à la saisie des requêtes pertinentes. Enfin, nous avons cherché diverses procédures d'éclairage de ces « cartes » par les opérations immédiatement antérieures (par exemple : éclairage du graphe par les cooccurrents spécifiques d'un pivot individuel ou collectif, ce qui permet de distinguer, dans une zone du graphe, les cooccurrents proprement dits – écart-réduit > 2 ou > 4 – des « synonymes » distributionnels en un sens plus large).

Nous voulons présenter ici un mode d'éclairage nouveau, sur lequel nous nous sommes mis à travailler, à partir de corpus particuliers où la diachronie méritait d'être réintroduite. Ces corpus nous ont été soumis aux fins de diverses opérations d'expertise, dans le cadre d'un projet dans l'Action Concertée Incitative « Cognitique », projet dont le maître d'œuvre est le Dr Noël-Jorand, et ils concernent tout d'abord la psycho-pathologie clinique sous sa dimension verbale-discursive. Mais les procédés mis au point nous ont immédiatement semblé applicables à tout travail expert sur les textes littéraires longs et les corpus d'auteurs, sans parler pour l'instant de bases plus étendues et moins homogènes.

En effet, l'AFC de matrices strictement lexicales (items lexicaux en lignes et en colonnes) évacue nécessairement toute perspective diachronique, même celle du processus linéaire de déploiement de la surface textuelle (une cooccurrence est comptée 1 qu'elle soit au début, au milieu ou à la fin du texte, et la matrice retient *n* cooccurrences, que celles-ci soient groupées plus massivement en quelque secteur, ou non).

On peut bien sûr effectuer des AFC à partir des segments du texte considéré, chapitres ou groupes de chapitres, mais se posent aussitôt deux obstacles : (1) l'inégale longueur, voire l'inexistence de partitions pertinentes *a priori*, comme les parties et chapitres, qui doivent être parfois créées arbitrairement, ou regroupées tout aussi arbitrairement. Et surtout (2) il est très difficile de proposer une mise en regard de ces analyses successives aux fins de comparaison, qui ne dépende pas d'un degré d'expertise statistique très élevé, donc inaccessible à un utilisateur lambda du logiciel d'analyse textuelle.

Nous proposons une solution dont les principes sont fort classiques : le chevauchement. Ce qui nous intéresse ici encore, *mutatis mutandis*, c'est-à-dire en ajoutant la « dimension »

diachronique, est une combinaison de continuité et de rupture, qui puisse être suivie graphiquement et interprétée commodément en terme non pas conclusifs, mais heuristiques (succession et amendement des hypothèses).

On « découpe » donc le corpus en *n* tranches (dans le 1er exemple présenté, 9 tranches), que l'on regroupe en unités de 3 tranches avec balayage (a :1.2.3, b :2.3.4, c :3.4.5, α :n-2.n-1.n). Nous désignerons ces unités de 3 tranches par le terme de *mégatranche*, noté $MT_n$. Ce sont ces mégatranches chevauchantes dont le « vocabulaire » (entre guillemets parce que seul un texte complet et cohérent déploie et structure les relations internes que l'on est pleinement en droit de désigner sous ce nom) est analysé et représenté en graphes.

On dispose ainsi de *n*-3 graphes ou plus exactement de *n*-3 listes de coordonnées, dont il est facile de calculer les transitions graphiques point par point, de a à b, de b à c, etc., de manière à créer une animation. On peut ainsi voir se succéder plastiquement les distributions, ce qui n'est évidemment permis que par le type de continuum spatial de l'AFC. On peut donc suivre la migration des items et des groupes d'items d'une manière qui par ailleurs n'est absolument pas permise par la simple juxtaposition (ou pire : par la superposition) des graphes.Ce dispositif n'aurait guère de sens si l'on ne décidait d'y associer divers mode d'éclairage (repérage par le style et par la couleur notamment).

## 2. Repérage par le style : suivi des items

Pour chaque transition, les items subissant une migration importante sont repérés par une mise en relief spécifique (grossissement et mise en gras). Le programme calcule la moyenne des valeurs absolues des vecteurs de migration pour une phase donnée et une valeur-seuil indexée sur cette moyenne (par exemple : 2 fois).

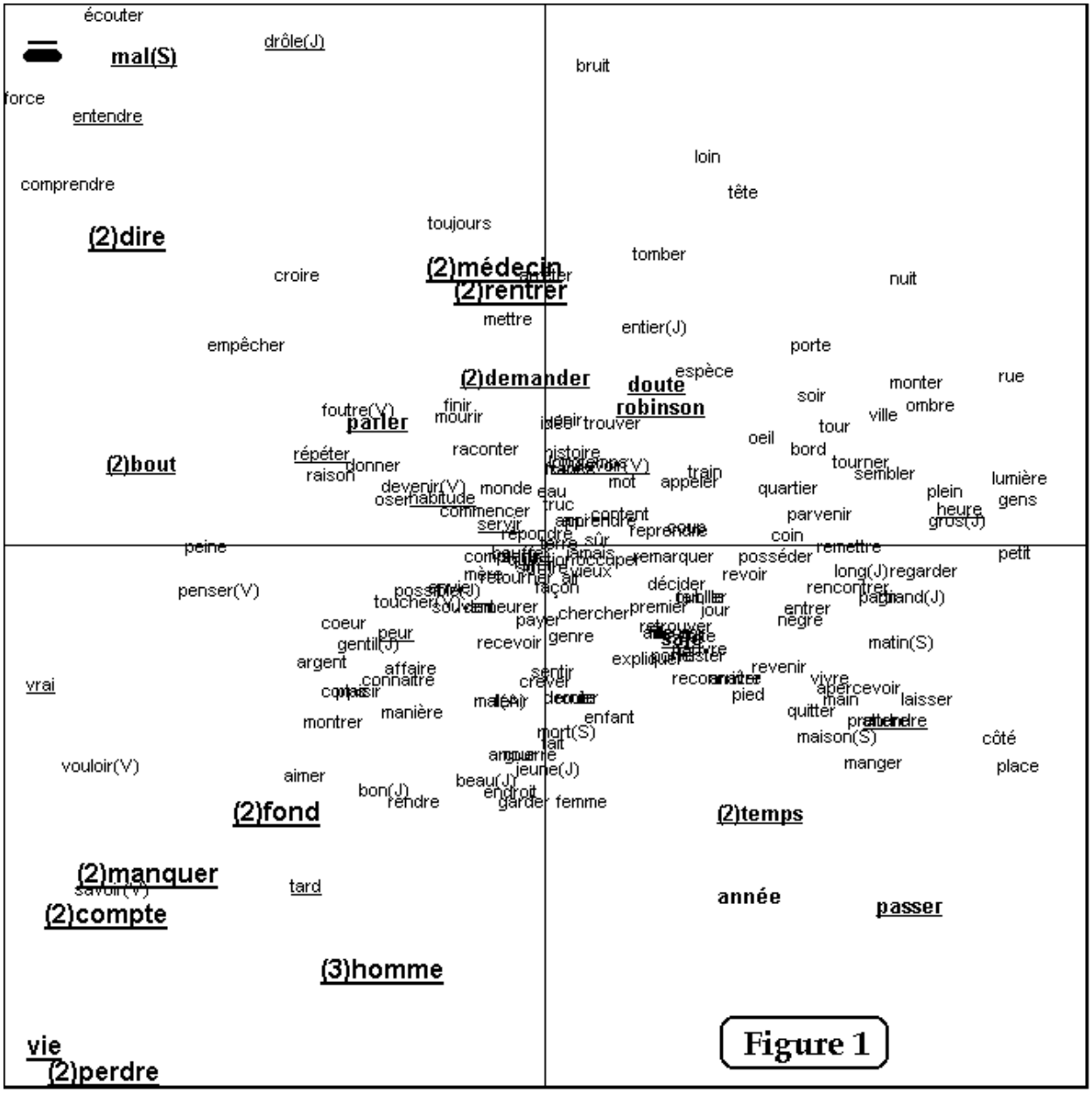
Par ailleurs, on calcule également la migration cumulée (somme des valeurs absolues des vecteurs de chaque phase), valeur que l'on préfère à celle de la migration apparente de la première à la dernière phase, ce qui reviendrait à privilégier sans motif deux phases particulières. On rapporte la migration cumulée d'un item à la moyenne générale selon le même principe que *supra* (valeur-seuil indexée sur la moyenne) et on met en relief les items sélectionnés (soulignement).

Voici les 29 items dépassant le seuil de 1.4 fois la migration totale moyenne dans *Voyage au bout de la nuit*, dont la valeur est 66.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| vie | 159 | | attendre | 114 | | sale | 97 |
| médecin | 158 | | passer | 113 | | répéter | 96 |
| bout | 153 | | mal(S) | 108 | | tard | 94 |
| perdre | 153 | | doute | 104 | | vrai | 94 |
| dire | 147 | | temps | 104 | | entendre | 93 |
| manquer | 147 | | parler | 102 | | heure | 93 |
| rentrer | 142 | | demander | 101 | | servir | 93 |
| homme | 141 | | robinson | 101 | | habitude | 91 |
| compte | 130 | | devoir(V) | 100 | | peur | 91 |
| fond | 126 | | drôle(J) | 99 | | | |

Il est utile de s'assurer que la migration totale d'un item dans cette optique ne présente pas de corrélation significative avec la variance de sa distribution dans les 9 tranches successives découpées (à condition cependant d'avoir éliminé 11 items ayant un bon effectif total mais présentant plus de 2 absences complètes dans une tranche : ce sont des désignateurs de personnages, typiquement épisodiques : *Baryton, Bébert, Docteur, Ferdinand*, etc.)

On peut observer finement l'évolution des relations de cooccurrence, voir des items lexicaux extrêmement stables dans la structure générale, dont ils constituent l'ossature invariante, et d'autres au contraire modifiés dans ce que leur champ d'emploi apporte à leur signification textuelle. Une migration spectaculaire se produit à la phase 3, c'est-à-dire au moment où l'on quitte la 3ème tranche et où l'on prend en charge la 6ème. Cette phase présente un relief interne très fort, puisque 9 items y dépassent la migration moyenne coefficientée par 3. Ces items sont repérés par un grossissement plus important ; la fig.1 montre l'analyse des tranches 3-4-5, les items en gras sont ceux qui s'apprêtent à migrer.




La fig.2 (p.suiv.) montre l'analyse des tranches 4-5-6 et on a matérialisé les migrations significatives par un trait.

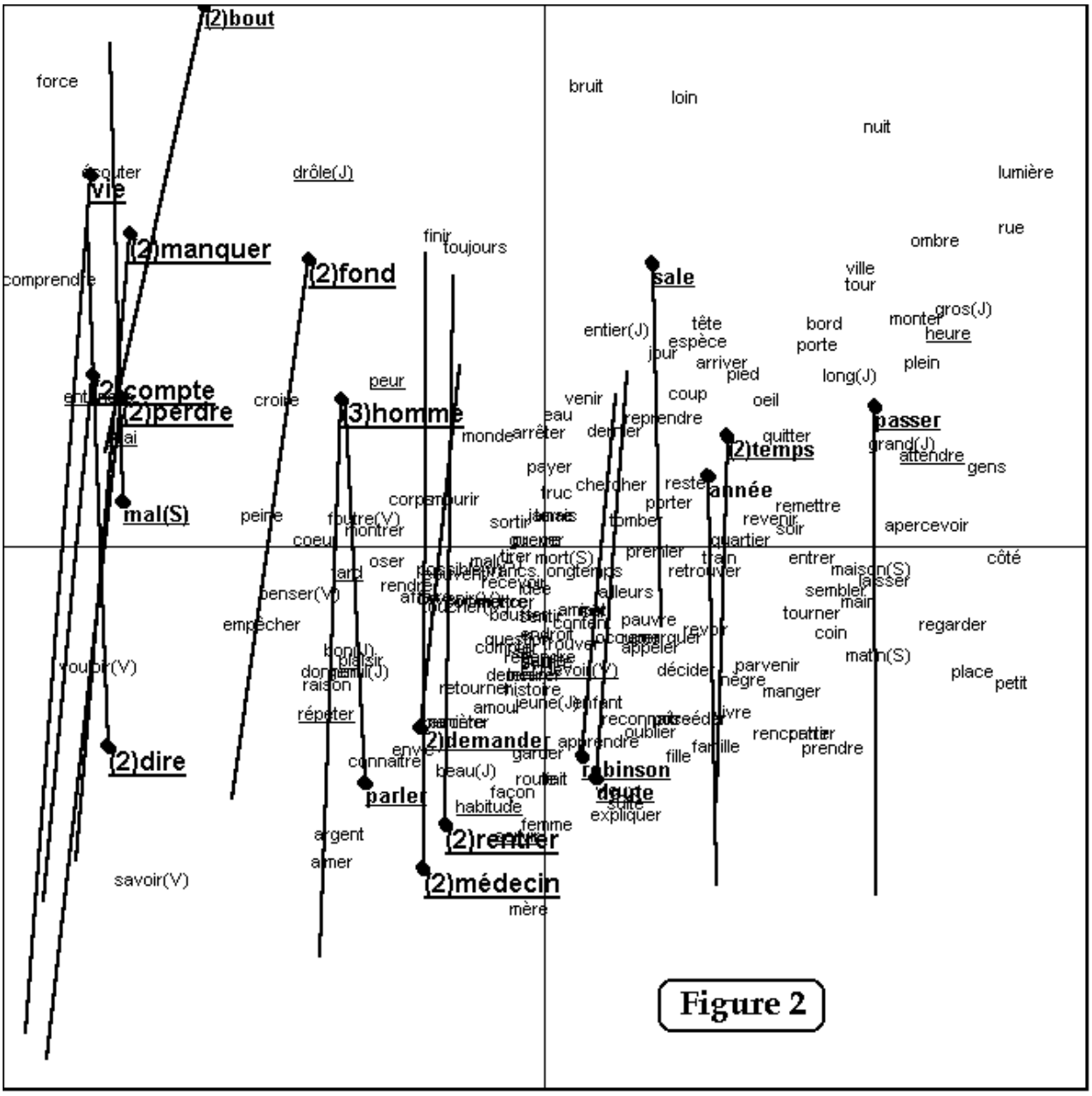



**Figure 2**

Un groupe est particulièrement bien lié : 6 vocables (MANQUER, PERDRE, FOND, COMPTE, VIE HOMME) « montent » sur l'axe 2 sans changer de position par rapport à l'axe 1 (remarquons que toutes ces fortes migrations se produisent strictement par inversion sur l'axe 2, sans affecter l'axe 1). VRAI et PEUR les accompagnent dans un mouvement un peu moins prononcé. Un invariant *isotropique* les « accueille », constitué des vocables COMMPRENDRE, CROIRE, ENTENDRE, PEINE, FOUTRE, eux-mêmes fortement polarisés par un pôle nord-ouest (FORCE, ECOUTER, DROLE, TOUJOURS, FINIR).

Comment interpréter cette migration ? ASTARTEX propose de nombreux outils de retour au texte, qui permettraient de rendre compte précisément de la modification des relations de profils qui se produit dans le 2ème tiers du roman. Mais dans l'espace réduit de cette communication, nous irons droit à l'essentiel, qui relève de la structure globale.

## 3. Repérage par les couleurs : suivi des groupements

On choisira une matrice de référence dont l'analyse servira à la coloration. Ce peut être n'importe laquelle des *n*-3 analyses effectuées sur les mégatranches (sachant que la première et la dernière notamment peuvent présenter ici un intérêt particulier), ou encore l'analyse effectuée sur l'ensemble du corpus. On colore les items d'après le graphe choisi pour référence, par exemple selon 4 couleurs, chacune attribuée aux items d'un secteur de 90° à cheval sur chacun des 4 demi-axes (dans une projection à deux axes), et suffisamment éloignés de l'origine commune pour être significatifs

Reprenons l'analyse de *Voyage au bout de la nuit*. Si nous prenons comme référence le début du roman ($MT_1$ recouvrant le 1/3 du texte), la coloration de départ est évidemment homogène secteur par secteur . La fig.3, comme les suivantes, est une « carte muette » puisque pour la présentation papier nous devons remplacer la coloration des items par un code monocolore (carré = vert, croix = magenta, point = rouge, cercle = bleu). On observera donc sur ces figures les seuls aspects globaux de migration.

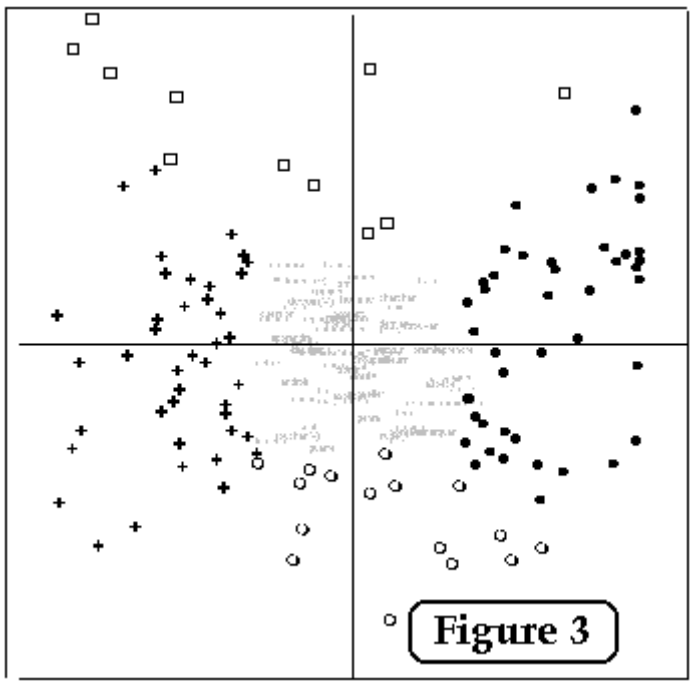


Figure 3

L'attribution de couleur aux items reste la même tout au long de l'animation. Le logiciel calcule la transition entre les coordonnées en (1.2.3) et les coordonnées en (2.3.4), en décomposant le vecteur par un nombre entier (ici, 10) et la met en œuvre grâce à un timer.

Nous nous servirons pour décrire les migrations de l'analogie aux points cardinaux. Sur le graphe de la fig.4 (p.suiv.), la disposition des zones repérées par référence à $MT_1$, dans le graphe de $MT_7$ (tranches 7.8.9, fin du texte) montre une interpénétration des zones SUD et EST, et à un moindre degré de NORD dans OUEST. L'axe 1 reste remarquablement stable en termes de grandes zones (il n'y a aucun croisement), mais l'axe 2 est moins imperméable (migrations NORD-SUD). Intuitivement, lorsque l'on suit l'animation, on repère que l'état final est déjà presque installé au milieu du processus.

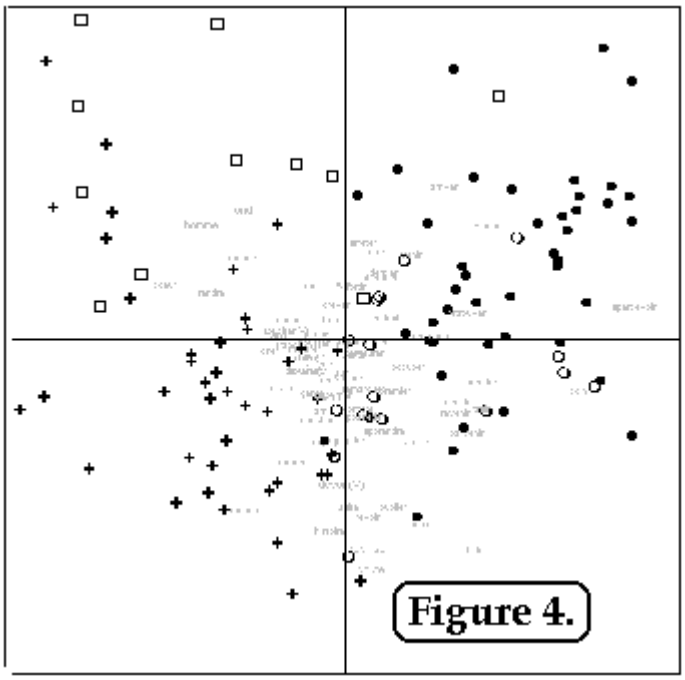


Figure 4.

On le vérifiera au mieux en prenant pour référence de la coloration $MT_4$. Le graphe de $MT_1$ coloré par $MT_4$ présente une grande confusion (fig.5),

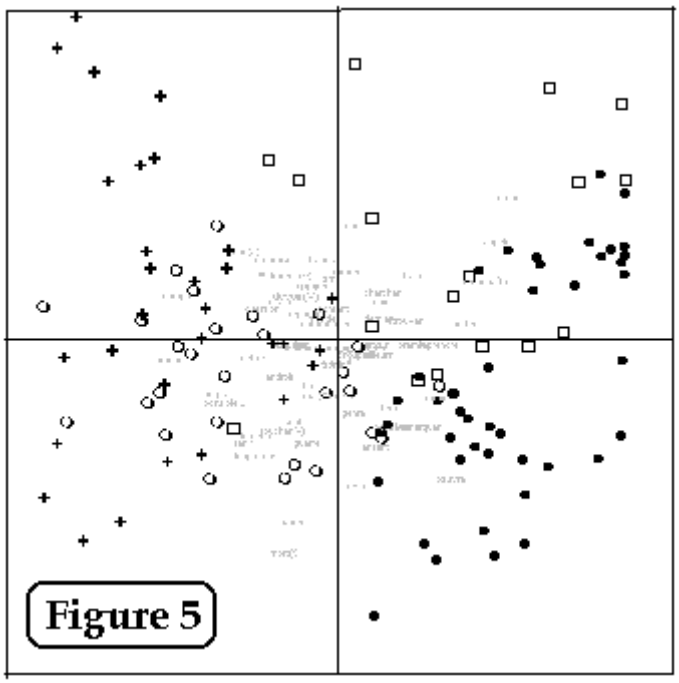
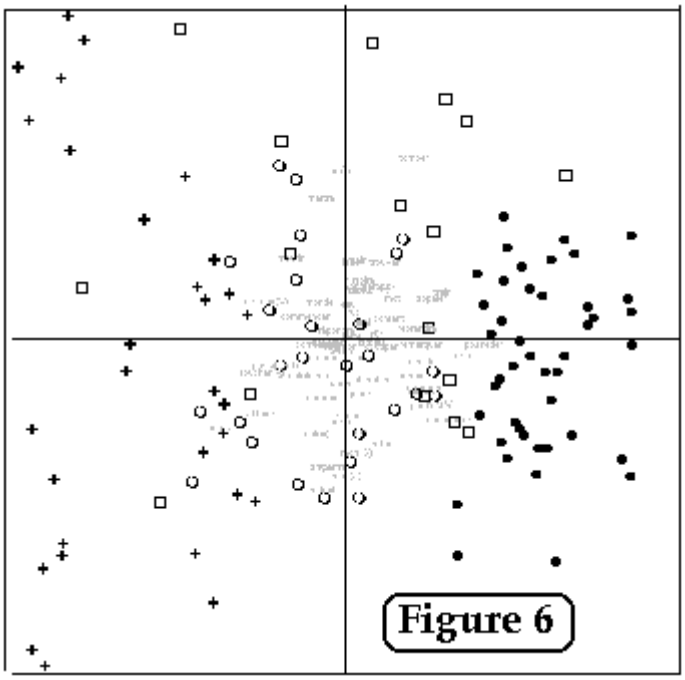
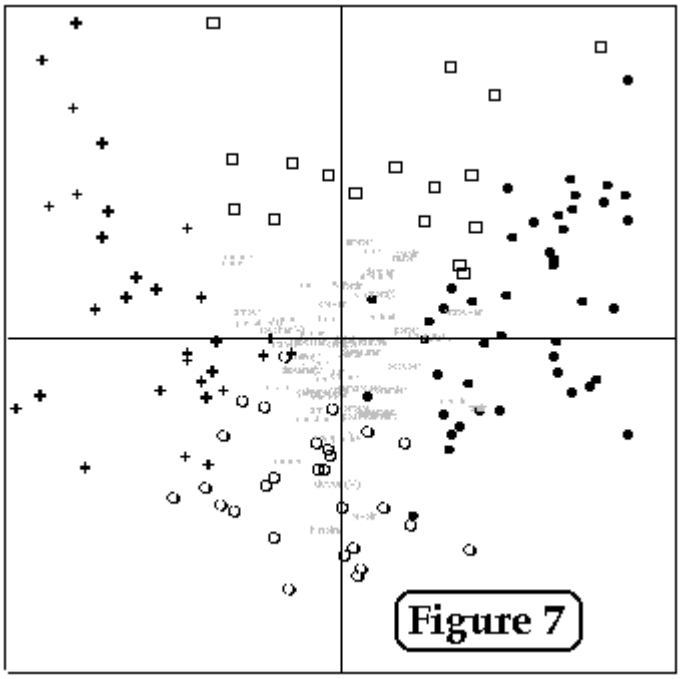


Figure 5

qui est loin d'être résolue en $MT_3$ (fig.6), c'est-à-dire au début de la phase qui conduit immédiatement à la coloration de référence,




où le groupe en bleu –cercles- (SUD en $MT_4$) se confond avec le centre du graphe et où le vert –carrés- (NORD en $MT_4$) est encore très dispersé, avec 3 de ses constituants les plus significatifs (FOND, PEUR et BOUT) engagés loin vers le SUD-EST. A l'inverse, tous les graphes de $MT_4$ à $MT_7$ présentent une remarquable similitude de répartition globale. A titre d'illustration, voyons le graphe $MT_7$ (toujours coloré par $MT_4$) (fig.7). Il y a donc une




remarquable stabilité des relations isotropiques sur l'ensemble de la seconde moitié du roman.

Pour évaluer et commencer à interpréter cette stabilité, employons désormais comme référence l'analyse totale du vocabulaire du roman entier, qui se présente ainsi (fig.8, p.suiv. : il n'y a pas de coloration puisque ce graphe n'apparaît jamais au cours de l'animation).

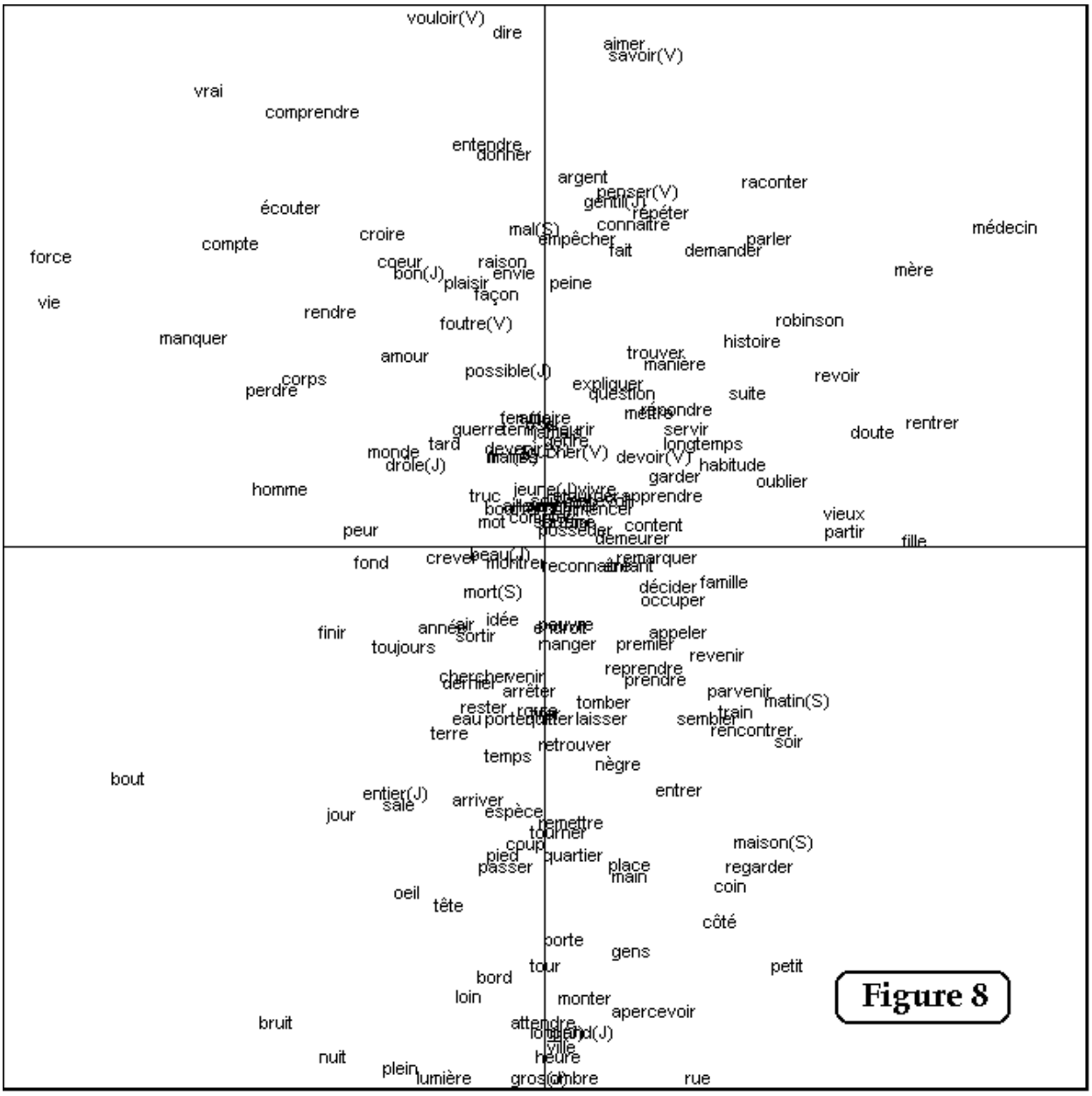
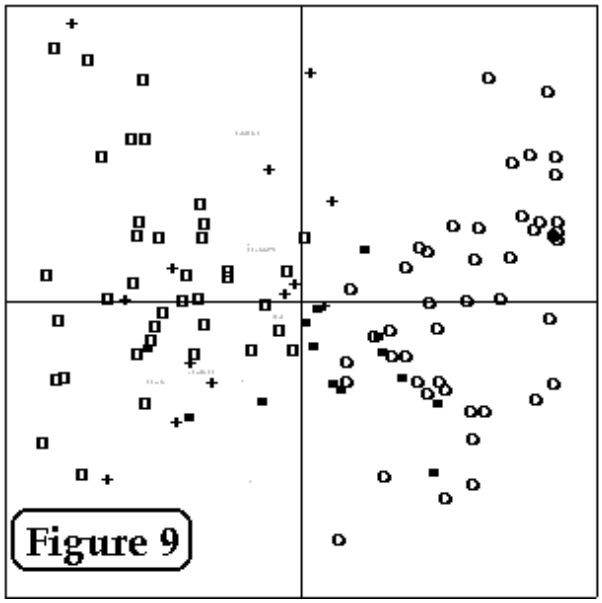


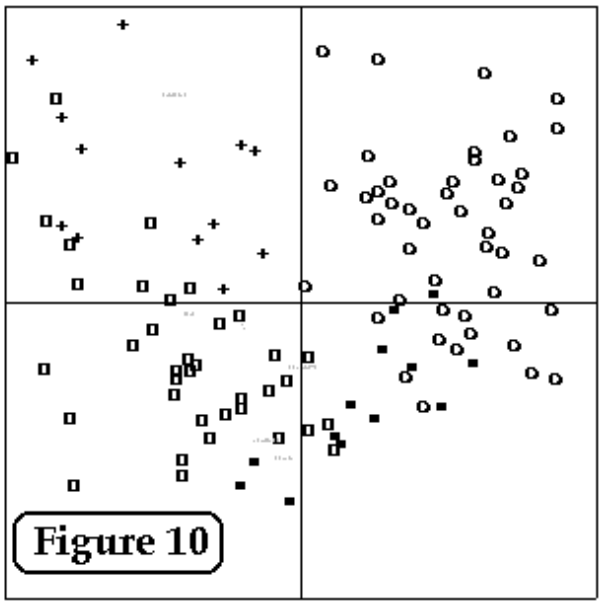

**Figure 8**

Voici (p.suiv.) les états successifs de l'animation (début, milieu et fin), colorée selon les secteurs de ce graphe général (fig. 9 à 11). On y constate aisément un processus de mise en place, presque achevé en $MT_4$(fig.10) et quasiment parfait en $MT_7$(fig.11). Bien sûr, les zones sont nettement transposées, puisque ce qui est au SUD dans le graphe général est au NORD-EST en $MT_7$, NORD passe à l'OUEST-SUD-OUEST, OUEST au NORD-OUEST, EST au SUD-SUD-EST

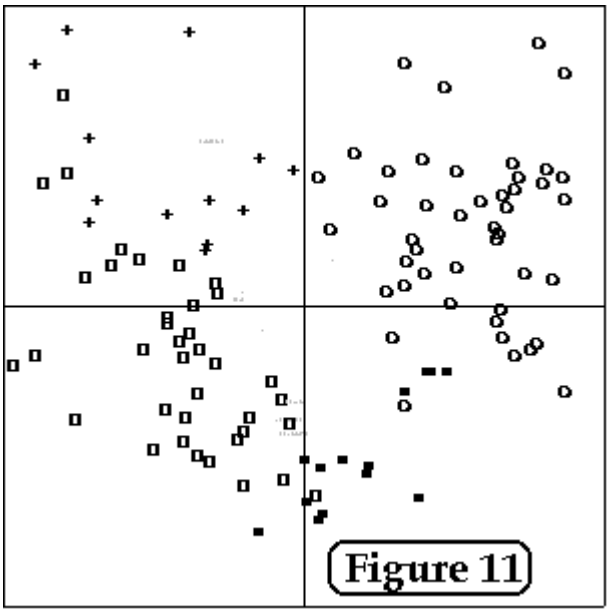
(Au passage, nous y voyons une très vive confirmation de deux intuitions fortes qui étaient les nôtres, notamment dans notre thèse en 1996, mais qui rest(ai)ent à vérifier : -1- la stabilité et la fiabilité de l'AFC appliquée à des matrices carrées, malgré la faible inertie relative des deux premiers axes, qui inquiétait à juste titre les premiers récepteurs : ni des modifications progressives, ni des comparaisons du tout à la partie, ni même des modifications brutales comme par exemple le passage de $MT_1$ à $MT_7$ ne rendent la structure méconnaissable ; on

reconnaît au contraire très bien l'invariant et les variations. –2- la figuration des axes proprement dits, donc leur pertinence géométrique, n'a pas une importance décisive ; puisque




dans cet exemple il est très clair qu'avec de très grandes matrices les mêmes groupements peuvent être repérés sur deux graphes avec une rotation de 45°).




Insistons sur le fait que cela n'était absolument pas sûr. Tous les schémas étaient possibles et le sont effectivement suivant la structure du texte étudié. On peut esquisser une typologie à partir de la notion de conformité, entendue ici comme conformité à la structure du vocabulaire à l'échelle du corpus entier.

| Conformité-type | Commentaire |
|---|---|
| Constante | Grande stabilité profonde de la structure lexicale à l'œuvre |
| Nulle | Régime fort de variation |
| Au début | Dissipation progressive |
| A la fin | Construction, recherche d'un état (téléologie ?) |

Dans le cas de *Voyage au bout de la nuit* (qui renvoie à la dernière ligne du tableau), on invoquera certes le caractère plus nettement picaresque du premier tiers, avec ses vastes déplacements (le Front, l'Afrique, l'Amérique). Mais il est risqué de parler de stabilisation référentielle, car si les déplacements ultérieurs (Rancy, Paris, Toulouse, Vigny) sont moindres en amplitude, ils sont néanmoins narrés comme des ruptures irréversibles et l'ensemble comme une course en avant. De plus, il faut rappeler qu'à tout moment l'axe 1 est préservé, que ce soit dans la coloration d'après une mégatranche particulière, ou d'après l'analyse d'ensemble, ce qui signifie un degré élevé de stabilité ou plutôt une forte extension du noyau de stabilité (la notion de degré de stabilité semblant devoir renvoyer plutôt à la première ligne du tableau de typologie).

## 4. Perspectives globales sur la dynamique

Le profil général de *Voyage au bout de la nuit* peut être confirmé et affiné de deux manières.

Tout d'abord, nous pouvons confronter aux résultats obtenus *supra*, l'AFC de la distribution des vocables dans les 9 tranches. Ce graphe (fig.12) est présenté très schématiquement, avec


**Figure 12**

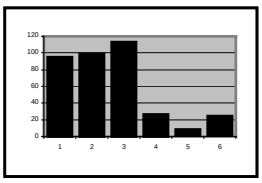
les seuls points-colonnes lisibles, correspondant aux tranches. Il nous donne un résultat au premier abord très paradoxal : on y voit les 3 premières tranches parfaitement regroupées, alors qu'elles sont le lieu d'une très forte variation des relations cooccurrentielles. Cela montre en fait à quel point la distribution linéaire massive des items (par grandes unités) doit être distinguée de leur distribution cooccurrentielle à courte portée, véritable noyau de la structure du vocabulaire. Il faut néanmoins rester prudent : on ne fait pas encore, avec la procédure ici introduite, la part entre l'effet du retrait des tranches antérieures et l'effet de l'apport des tranches ultérieures (glissement). On note que les deux dernières tranches sont, elles aussi, très bien regroupées. Il y a donc ici une certaine symétrie (légèrement inégale), qui exclut en grande partie l'hypothèse selon laquelle la forte variation observée dans l'animation pour la première moitié serait due à l'instabilité distributionnelle par grandes unités du centre du roman. Celle-ci devrait en effet avoir le même effet au début et à la fin du roman, d'après ce que l'on vient d'observer sur le graphe macro-distributionnel.

Ensuite, nous pouvons évaluer la migration globale opérée par l'ensemble des items pour chaque phase successive (il y en a 7, s'il y a 9 tranches). Plutôt que de comparer des moyennes de valeurs absolues de vecteurs (celles des phases et la moyenne générale), nous utiliserons la moyenne générale pour dénombrer les items qui la dépassent dans chaque phase. Nous avons ainsi un effectif de forts migrants pour chaque phase, un effectif moyen pour l'ensemble, à partir desquels il redevient commode de mesurer des écarts-réduits, donc de laisser à la théorie probabiliste le soin délicat de déterminer un seuil de pertinence.

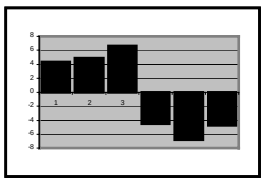
En voici le résultat. La moyenne de référence est 60,83 items dépassant le seuil, par phase.

| | absolu | écart-réduit |
|---|---|---|
| MT1-2 | 95 | 4,35 |
| MT2-3 | 99 | 4,87 |
| MT3-4 | 113 | 6,66 |
| MT4-5 | 26 | -4,48 |
| MT5-6 | 8 | -6,79 |
| MT6-7 | 24 | -4,74 |

absolu



écart-réduit



## 5. Conclusions

Il est difficile de montrer ici en détail les apports à l'interprétation des résultats d'AFC pour les non-spécialistes. Nous insisterons seulement pour finir sur deux points.

L'ergonomisation et/ou la « convivialisation » des outils de réception et de présentation des résultats de la statistique textuelle est une priorité vitale pour la diffusion de ces méthodes dans les milieux professionnels et scientifiques concernés, notamment littéraires. Dans cette optique, l'urgence ne doit pas nous conduire à négliger la nécessaire articulation étroite entre les aspects globaux et les aspects de détail, tout en sachant que les aspects globaux peuvent être présentés avec un luxe de détails, mais que la plus grande globalité, représentée notamment par les deux derniers histogrammes *supra*, exclut les détails, et *vice versa*. Par ailleurs, ®ASTARTEX comporte de nombreux autres modules articulés à celui-ci, qui permettent précisément d'aller explorer le plus fin détail.

La procédure ici décrite s'applique également, *mutatis mutandis*, à des corpus non-littéraires et notamment, dans notre pratique, à des corpus d'entretiens cliniques. Sur un corpus de discours « délirants » étagé dans le temps, nous offrons alors quant à l'évolution clinique des pistes interprétatives supplémentaires, que ce soit par rapport à l'analyse micro-distributionnelle « statique » que nous avons développée précédemment, ou par rapport à l'analyse macro-distributionnelle, qui n'indique que dans les cas d'école une évolution claire. C'est l'occasion de dire aussi que, si le corpus comporte des partitions « naturelles » suffisamment amples et régulières, leur logique peut avantageusement supplanter celle des tranches arbitraires. Dans le cas de *Voyage au bout de la nuit*, les grandes partitions par lieux ne sont pas marquées dans le texte ; elles ont été repérées, sur un seul critère (le lieu) par la critique. C'est pourquoi nous ne les avons pas retenues en première intention.

## 5. Références bibliographiques

Céline L.-F. (1931-1952). *Voyage au bout de la nuit*. Gallimard

Hubert de Phalèse (1993) *Guide de* Voyage au bout de la nuit. Nizet

Massonie J.-Ph. (1986). *Pratique de l'analyse des correspondances*. Les Belles-Lettres

Viprey J.-M. (1997). *Dynamique du vocabulaire des* Fleurs du mal. Champion

Viprey J.-M. (2000). Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle. In *JADT 2000*. EPFL, pages 535-538

Viprey J.-M. (1998). Une norme endogène pour le calcul stylistique du vocabulaire. In *JADT 1998*. UNSA-CNRS, pages 639-648