Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection « géodésique »

Jean-Marie Viprey

ATST, EA 3187

Université de Franche-Comté, Besançon

jean-marie.viprey@univ-fcomte.fr

Abstract

In order to make it more helpful to computer-assisted textual investigation into discourse, so well as to expert browsing into corpuses, we propose to renew the graphical output of FAC. Usual projections onto a bi-factorial plane are uneasy to read, to handle and often deceptive for a humanist aiming at synthetic views and dynamic cartography. We propose the geodesic projection, which enables to build a planisphere and zone views "as seen from the center", in which angulations and distances from the axis origins are respected and clearly pointed out. As regards comparisons, that projection is a complement to matricial distance calculation, more detailed than them and turned towards text.

Résumé

Afin de la mettre réellement au service de l'exploration textuelle assistée des discours, de la navigation experte dans les corpus, nous proposons de renouveler la sortie graphique de l'AFC. Les projections habituelles sur le plan de 2 facteurs sont peu lisibles, mal maniables et trompeuses pour un spécialiste de sciences humaines soucieux de prendre des vues synthétiques et d'acquérir une cartographie dynamique. Nous proposons la projection géodésique, qui permet de construire un planisphère et des zonages « vus du centre », où les angulations et les distances aux origines des axes sont respectées et clairement indiquées. En vue de comparaisons, cette projection est un moyen complémentaire aux calculs de distance matricielle, plus détaillé qu'eux et orientés vers le retour au texte.

Mots-clés : AFC, projection géodésique, hypertexte, isotropie, vocabulaire.

1. Introduction

L'analyse textuelle des discours (ATD, Adam, 2005), le développement de la *philologie numérique* (Rastier, 2002), la réintégration d'une perspective herméneutique dans l'analyse de discours (Adam & Heidman, 2005), exigent une intégration de plus en plus poussée des formalismes informatiques et statistiques aux cadres conceptuels et aux environnements techniques de l'exploration des corpus. C'est le sens des développements cadrés par le pôle *Archive, Bases, Corpus* de la Maison des Sciences de l'Homme de Franche-Comté.

Nous avons montré (Viprey, 1997, 2005) l'intérêt de l'application de l'Analyse Factorielle des Correspondances (AFC) à des relevés de cooccurrence afin d'obtenir des vues synthétiques sur la structure du vocabulaire de grands ensembles textuels. Puis, nous avons suggéré (Viprey, 2000) que les sorties graphiques de ces analyses constituent un élément important de la cartographie de cet ensemble dès lors qu'on se soucie de l'instituer en hypertexte d'exploration, et mis en œuvre une première expérimentation (Viprey, 2002). Il s'agit, dans un environnement d'analyse textuelle fondé sur le primat de la lecture non-

linéaire des corpus, de disposer en sus des listes, concordances et champs d'affichage plein-texte, de graphes à nuages de points où ceux-ci sont cliquables en vue de diverses fonctions de poursuite, optimisant ainsi le caractère dynamique de l'appareillage.

En substance, l'AFC présente l'intérêt majeur de hiérarchiser l'information sur les distributions, sans la dichotomiser, spatialisant les relations (ici micro-distributionnelles) (cooccurrence, collocation) sur le continuum d'un axe, d'un plan, d'un espace à N dimensions.

Dans cette optique, nous nous sommes heurtés à plusieurs obstacles, dont le principal reste la disposition des sorties graphiques telle que nous l'héritons de la tradition benzécriste. C'est de ce dispositif de sortie graphique que nous voulons discuter ici, sur la base de plusieurs expériences, menées sur d'assez grands corpus.

2. Projections sur le plan de 2 facteurs

Les sorties graphiques classiques, programmées dans les environnements comportant un module AFC, sont des projections strictes des nuages de points (colonnes et/ou lignes) sur le plan des deux premiers facteurs ou de deux des trois premiers facteurs.

Le corpus soumis à l'exploration est *Le Monde diplomatique* 1980-2000[1], en texte nu (nous avons seulement appliqué un module automatisé de segmentation et de ponctuation, à l'exclusion de toute autre reconnaissance lexicale). Nous relevons les cooccurrences de 242 formes graphiques retenues parmi les plus fréquentes ; ont été exclues les formes, ambiguës ou non, correspondant à des catégories autres que noms propres, substantifs, adjectifs, verbes et adverbes lexicaux. L'empan cotextuel défini est de 15 mots à gauche et à droite dans les bornes de ponctuations fortes.

La figure 1 présente sur le plan des 2 premiers facteurs le résultat de l'analyse de la matrice brute (242x242). Ce graphe est très représentatif des sorties offertes lorsqu'en toute première intention l'on travaille sur un texte non préparé : seuls quelques items sont différenciés d'un nuage très compact, parce qu'une ou deux cooccurrences très fortes « écrasent » l'analyse. Ce sont notamment des syntagmes lexicalisés, ici : *premier ministre* et *droits de l'homme*. En l'état, ce graphe n'est certes pas utilisable, ni en visualisation directe, ni encore moins dans la perspective d'en rendre cliquables les items constitutifs.

L'expérience la plus élémentaire montre qu'il est vain d'espérer se débarrasser de cette sorte de problème en éliminant les items perturbateurs, que ce soit en les pré-étiquetant (*premier_ministre*) ou en supprimant l'un ou l'autre, ou les deux, de la matrice. D'autres collocations lexicalisées, puis semi-lexicalisées, puis des phraséologies typiques, prendront le relais pour « écraser » le graphe.

Nous avons supposé qu'une autre solution se révèlerait plus élégante et efficace : l'écrêtage. Il s'agit de fixer un seuil d'écart-réduit à l'équidistribution et de calculer, pour les cooccurrences dépassant ce seuil, une cooccurrence « réduite » à ce seuil. On amortit ainsi les saillances perturbatrices. La figure 2 présente sur le plan des 2 premiers facteurs le résultat de l'analyse de la même matrice, ainsi réduite (seuil + d'écart-réduit : 5).

Le graphe est clarifié. On notera que les pourcentages d'inerties des deux premiers facteurs sont considérablement augmentés (en cumul, de 12.5 % à 25.9 %). Nous y avons repéré les items mis en cause *supra*, qui ont ainsi été « libérés » d'une attraction exclusive au profit de

[1] 17 330 000 occurrences (ponctuations comprises).

relations plus diverses et complexes. Un examen plus détaillé permet de constater que les points saillants de la fig.1 conservent une partie de leur saillance suffisante pour être identifiables, et restent organisés sur le plan selon une configuration peu altérée.

Avec un outil de zoom approprié, il est déjà plus pertinent qu'en 1 d'employer un tel graphe comme « carte » dynamique du corpus, notamment par les fonctions hypertexte.

Cependant, trois reproches, assez complémentaires, peuvent être formulées. Seule une discussion approfondie dans la communauté benzécriste permettra de dire quel est le plus fondamental.

Fig.1 : plan des 2 premiers facteurs.

Le premier visera la manipulation des données qu'est l'écrêtage. On a choisi ici le seuil d'écart-réduit 2 pour son statut presque « symbolique » en lexicométrie, largement lié aux travaux et enseignements de Charles Muller (un écart-réduit de valeur absolue 2 indique une déviation à la norme qui a une probabilité de 5% d'être aléatoire). Mais cette valeur n'est pas

« sacrée », et toute autre option produira des déformations plus ou moins marquées du graphe (en pratique, plus le seuil sera élevé, plus le graphe sera « écrasé » par les perturbations identifiées *supra*[2]). Il faut souligner que dès un seuillage à 20 le graphe est très lisible et que

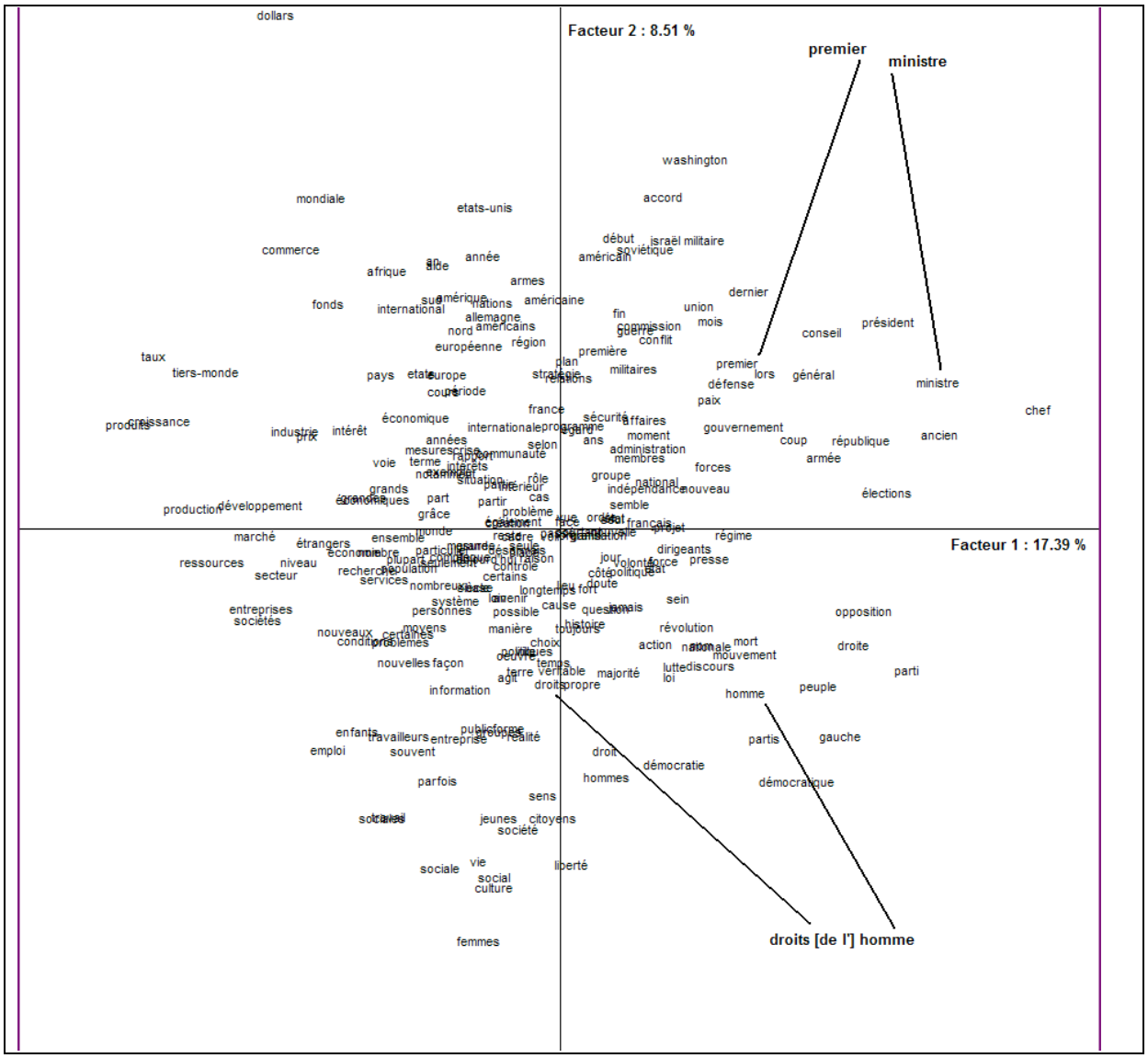


Fig.2 : plan des 2 premiers facteurs, avec une matrice écrêtée à l'écart-réduit 5/-5.

[2] On consultera avec intérêt le tableau de croissance des inerties au fur et à mesure de la réduction du seuil en écart-réduit :

seuil	inertie 1	inertie 2	inertie 3	cumul
aucun	7,05	5,02	4,67	**16,74**
20	12,73	6,7	6,46	**25,89**
15	13,56	7,09	6,7	**27,35**
10	14,53	7,58	6,97	**29,08**
5	16,01	8,08	7,14	**31,23**
4	16,41	8,21	7,14	**31,76**
3	16,86	8,35	7,09	**32,3**
2	17,39	8,51	6,94	**32,84**

les modifications graduelles sont peu significatives, surtout dans la perspective d'un emploi descriptif et exploratoire. On pourrait sans doute convenir d'un seuillage à 5 (probabilité inférieure au millionnième d'un écart aléatoire).

Les deux autres reproches concernent la qualité de l'information représentée.

Tout d'abord, on regrettera forcément de ne disposer que d'une vue sur le plan de 2 facteurs, et on cherchera à obtenir un complément d'information sur le 3ème facteur. La réponse la plus connue à cette demande est celle des visualisations du type *Macspin*[3]. Le principe de cette visualisation était de considérer les trois premiers axes comme les arêtes d'un cube dont on simule la rotation autour de l'origine commune. Sa logique est à la base des effets « 3D » produits par une rotation simulée, permettant de prendre des angles de vues divers sur un solide virtuel ; à ce titre, ses dérivés sont employés dans les logiciels de présentation pour de très divers secteurs d'activité : architecture, design, anatomie, etc. Pour l'AFC, et surtout pour sa prise en main par des chercheurs non spécialistes en statistique multidimensionnelle, les risques de mésinterprétation sont constants et l'utilisation rationnelle impossible. En effet, le point de vue simulé est à l'extérieur du solide et l'infinie variété de combinaisons réglant la rotation sur les 3 axes permet d'amener en superposition deux à deux successivement et indifféremment tous les points et groupes de points. Si l'on considère l'espace déterminé par les 3 premiers axes (linéarisant les 3 premiers facteurs) comme l'intérieur d'une sphère (une *boule*), on comprend que si cette sphère est transparente, vue de son extérieur, et si les seuls objets visibles dans son espace sont les points du nuage, c'est exactement comme si on observait un globe terrestre en verre et si l'on prenait les superpositions (alignements) ainsi engendrées pour des proximités[4].

D'autres solutions ont été explorées, mais très peu publiées, envisageant de figurer, sur un graphe du plan des 2 premiers axes, la position du point sur le 3ème par un niveau de gris. Le principal inconvénient de ces solutions est de donner à des paramètres de même nature (distance orientée sur une dimension de l'espace vectoriel) des expressions hétérogènes. De plus, le niveau de gris permet d'exprimer très bien la distance absolue à l'origine, mais très mal l'orientation, positive ou négative. Enfin, le niveau de gris comme les colorations est fort utile pour les divers éclairages que l'environnement d'exploration peut suggérer.

Troisième reproche, une partie considérable du graphe sur le plan de 2 facteurs est « indûment » occupée par une information que chacun sait non pertinente : plus on se rapproche du centre du graphe, moins les positions individuelles et les proximités sont significatives. On croit donc ne rien pouvoir connaître, grâce à une vue synthétique issue du calcul de l'AFC, du plus grand nombre des items constitutifs.

3. Projections « sphériques » ou géodésiques

Pour toutes ces raisons, nous nous sommes tournés vers la nécessité de représenter les positions sur les 3 axes, sans reproduire les défauts heuristiques du type *Macspin*.

[3] Le logiciel *Macspin*, très en vogue dans les années 90, n'est plus aujourd'hui ni développé, ni commercialisé, ni commenté. Le dernier site connu, celui des concepteurs initiaux (*famille Donoho*) : http://www.ddg.com ne documente plus *Macspin*.

[4] L'auteur de ces lignes croit savoir de quoi il parle. Lors de la rédaction de ma thèse sur le vocabulaire des *Fleurs du mal*, je disposais de *Macspin* et j'ai longuement hésité avant de renoncer (entièrement je crois) au confort apparent et aux interprétations implicites incontrôlées que m'offrait ce logiciel pour mettre en scène les sorties d'AFC. J'ai en revanche employé à plusieurs reprises ces vues dans mes premières présentations. C'est pourquoi je me sens particulièrement responsable pour proposer une alternative réfléchie.

Nous considérons bien l'espace déterminé par les 3 premiers axes comme l'intérieur d'une sphère (étant donné les contraintes de l'algorithme, cette sphère peut être considérée comme de rayon 1, puisqu'il s'agit de la coordonnée-limite sur chacun des demi-axes). Les positions sur les 3 axes déterminent un parallélépipède inclus dans l'un des huit secteurs de la *boule*.

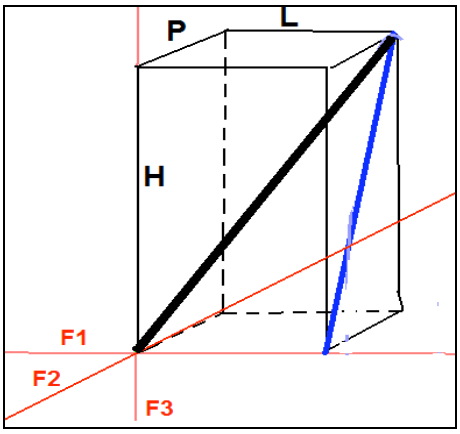
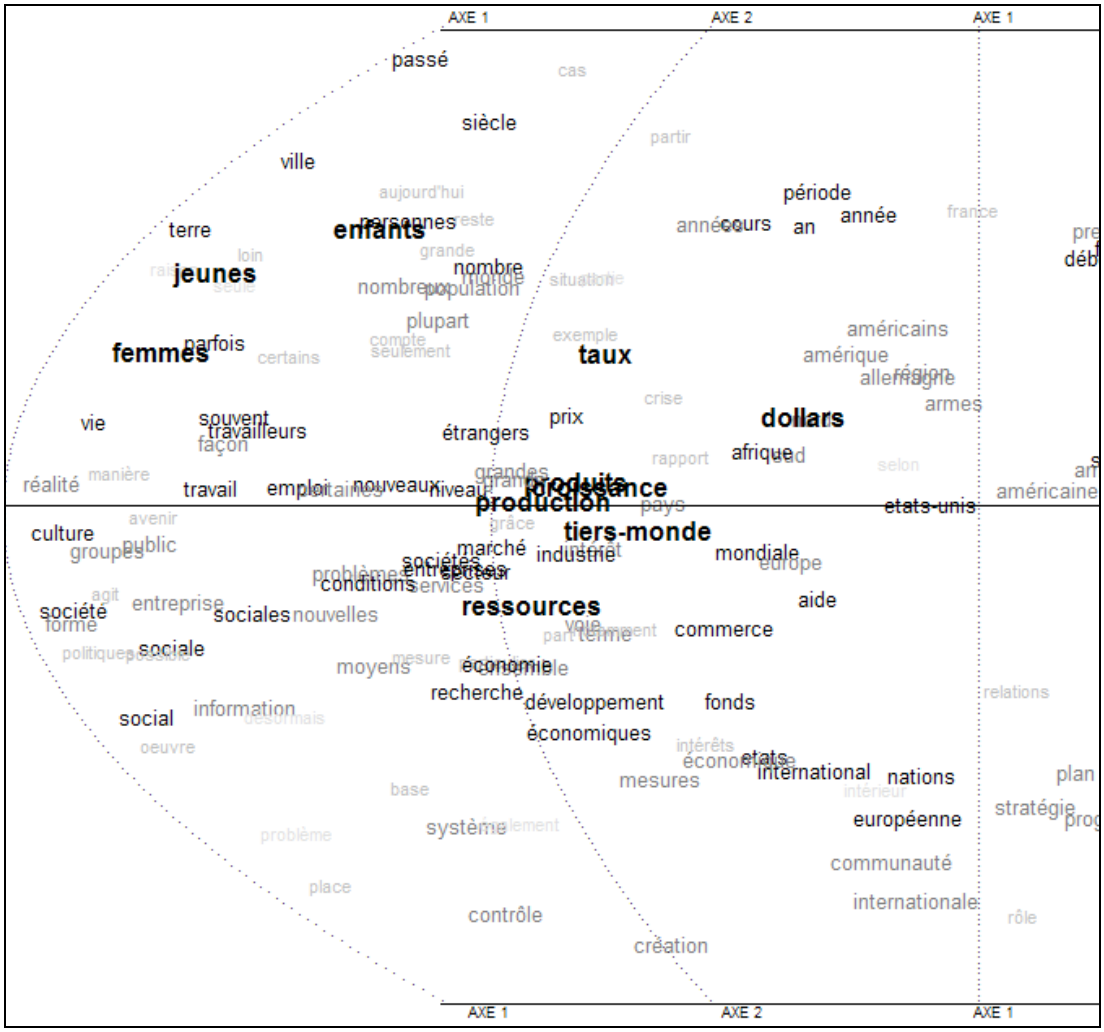


Fig.3 : schéma du parallélépipède :
en gras, la grande diagonale, mesurée à l'unité du rayon de la sphère



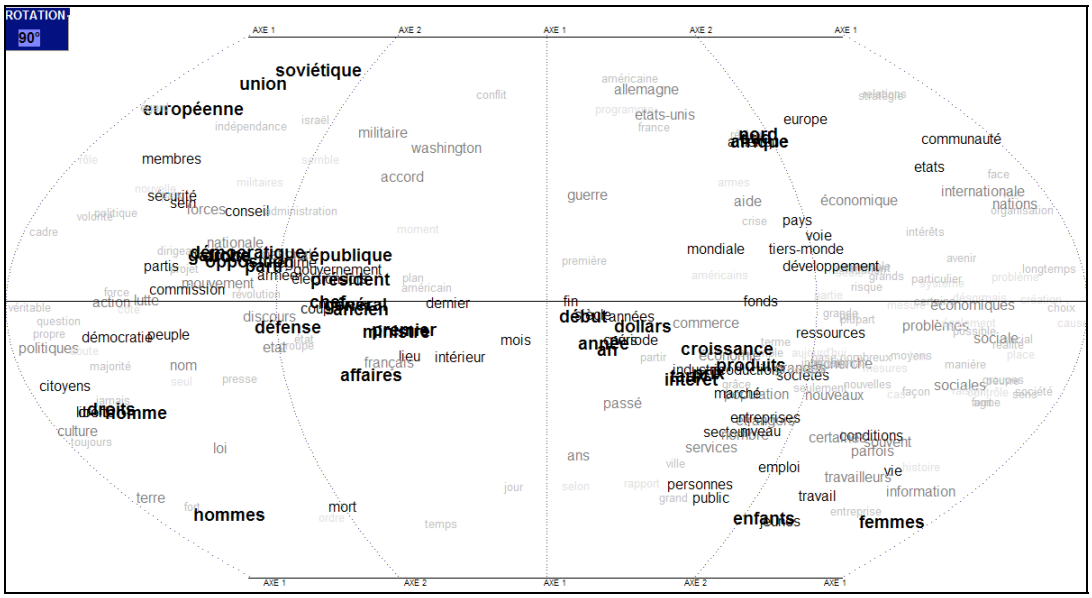


Fig.4 (ci-dessus) : planisphère, données brutes

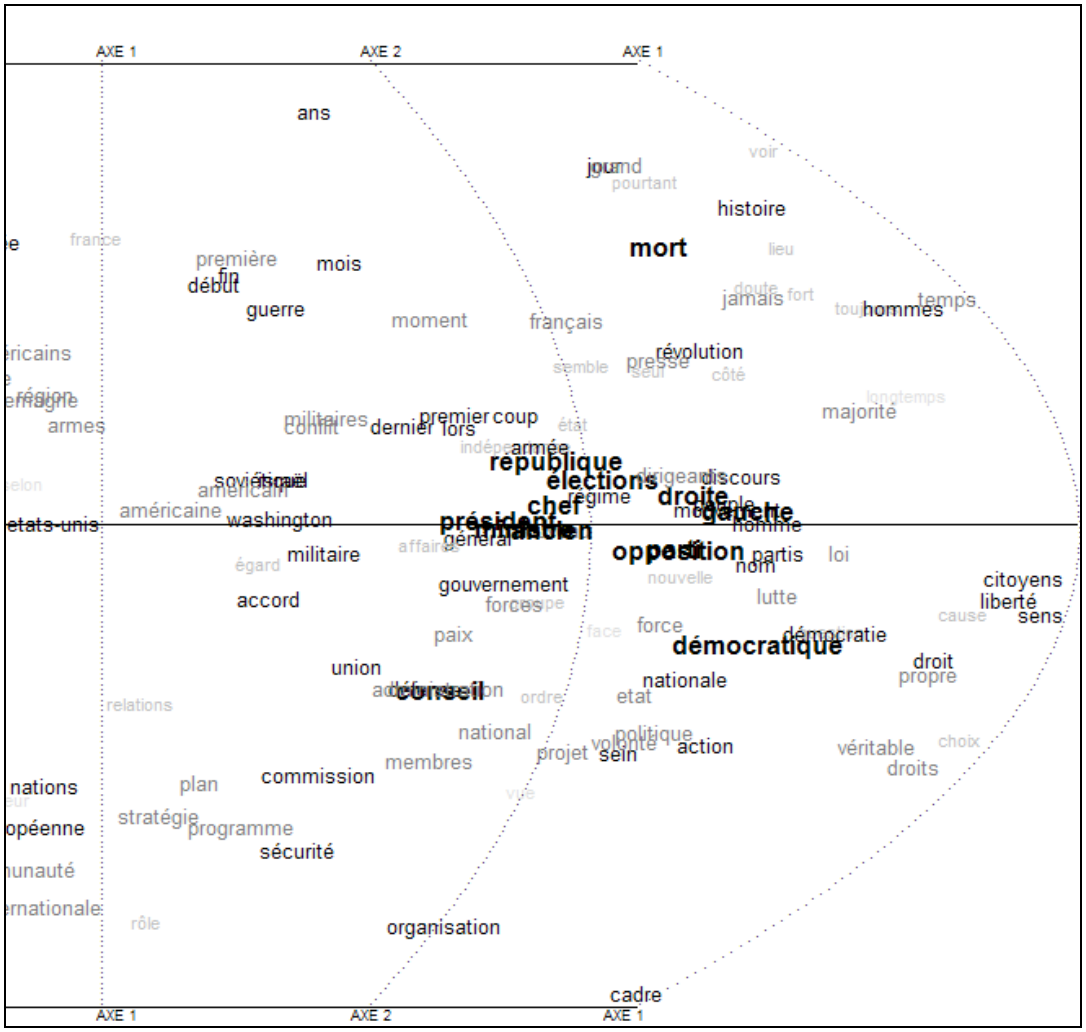


Fig.5 (ci-dessous, sur 2 pages) : planisphère, données écrêtées à l'écart-réduit 5/-5

Les positions sur les deux premiers axes (largeur L et profondeur P du parallélépipède) déterminent une *angulation équatoriale* (une *longitude*) et une *distance à l'origine du plan équatorial*, qui sont exactement exprimées par la vue présentée en fig.2. La position (hauteur H) sur le troisième axe détermine une *angulation polaire* (une *latitude*) et une *distance à l'origine de la boule*, qui est la grande diagonale du parallélépipède, mesurée par l'unité *rayon*.

Ainsi se trouvent déterminées les coordonnées géodésiques des points du nuage, d'après les 3 premiers facteurs. Ces coordonnées sont deux angles, l'un sur l'équateur (de 0 à 360°), l'autre sur le méridien ainsi déterminé (de -90 à +90°).

Considérons alors le point de projection, depuis le centre, de chacun des points sur la sphère de rayon 1. Il reflète sans erreur l'angulation contenue dans les coordonnées calculées par l'algorithme de Benzécri. Au moyen d'une projection de Robinson, nous pouvons en déduire un *planisphère* dont on peut espérer qu'il exprime au mieux, visuellement, les distributions qui ont été compilées.

Nous utiliserons dès lors les niveaux de gris pour exprimer la distance à l'origine dans les 3 dimensions (grande diagonale du parallélépipède), de manière à estomper progressivement les points à mesure qu'ils se rapprochent du centre dans une position de moins en moins significative. La figure 4 présente un tel *planisphère*, calculé de manière entièrement explicite à partir de la même matrice de cooccurrence que la fig.1 (sans aucun écrêtage).

On constate que, même sans le moindre écrêtage, sur des données tout à fait brutes, la projection est déjà utile et permet de repérer des groupements distributionnels (*isotropies*, Viprey, 1997) très bien différenciés. Un zoom permet de remédier aux effets de superposition. Un outil permet d'opérer les rotations équatoriales nécessaires pour visualiser les relations entre les bords latéraux, résultats d'un centrage arbitraire.

Si certains de ces groupements sont structurés autour de collocations lexicales (ici *Union Soviétique* et *Union Européenne*, *Afrique du Nord, droits de l'homme*), d'autres donnent à voir des configurations thématiques beaucoup plus pertinentes.

Avec un écrêtage des données au seuil de 5, le planisphère se présente comme sur la fig.5 (pages précédentes).

4. Zooms « régionaux » : isotropie

Au-delà du zoom local destiné à visualiser des zones enchevêtrées sur le fond du planisphère (loupe), nous proposons d'observer des « régions » entières de la sphère selon un mode géodésique, c'est-à-dire « comme si » l'on regardait vraiment depuis le centre. Il suffit de cliquer un item quelconque pour obtenir une vue centrée sur cet item et présentant l'ensemble de sa périphérie. Par exemple, la « région » centrée sur *élections* (89°E, 8°N), en figure 6.

Ces zonages sont qualitativement différents de ceux que permet la visualisation sur le plan de 2 facteurs. En effet, c'est l'angulation relative de deux items dans l'espace des 3 premiers facteurs (31.2 % de l'inertie totale d'une matrice à 241 facteurs) qui détermine la distance entre les deux points sur le graphe de zone, dans toutes les directions. Alors que sur le graphe plan, les choses sont très déformées selon que l'on s'éloigne, dans une direction sur le « zoom », du centre (origine) ou que l'on s'en approche (pertinence décroissante), voire qu'on le dépasse.

Ils donnent, selon nous, un accès très amélioré à l'observation de ce que nous avons appelé (Viprey, 1997) les *isotropies*, c'est-à-dire des classes non-dichotomiques (dans le continuum

permis par la seule AFC) de parentés de profils micro-distributionnels, configurations les plus fines de la structure globale du vocabulaire, où s'expriment les spécificités thématiques, lexicales, stylistiques, pragmatiques que le texte permet d'assigner.

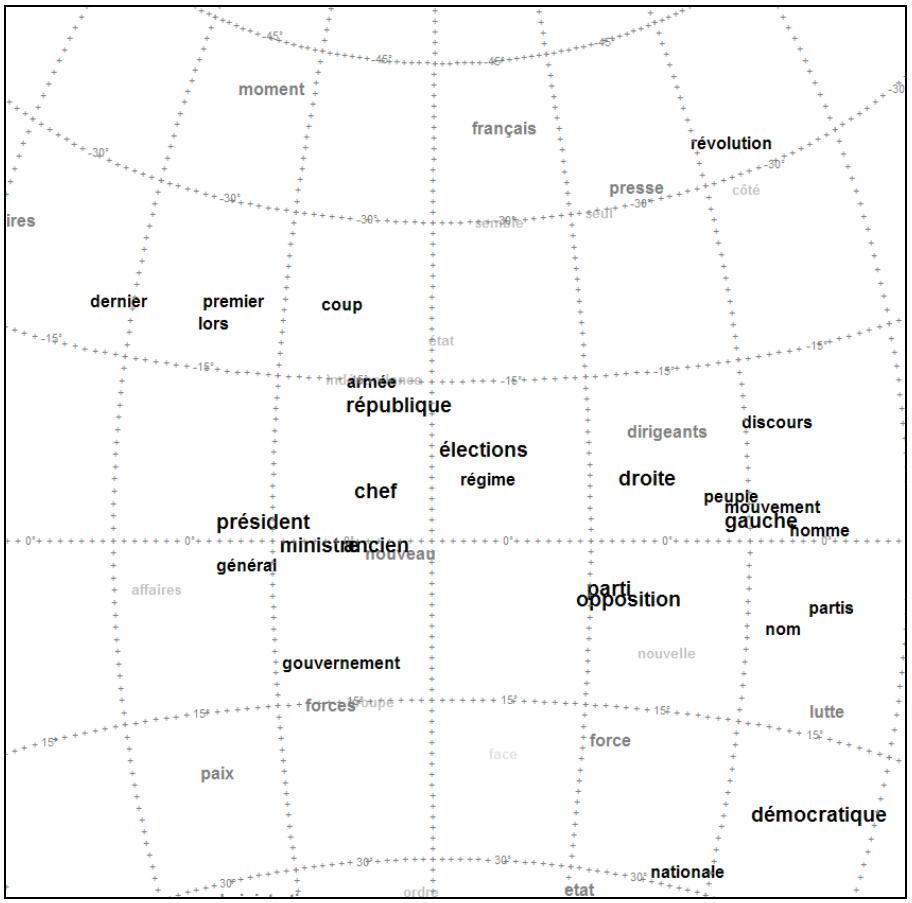


Fig.6 : zonage centré sur élections

Ces zonages sont conçus pour être cliquables selon les mêmes modalités que le planisphère et que les zooms, de manière à accéder au plein-texte, aux concordances et aux listes construites dans le cours de l'exploration.

5. Éclairements : comparaison de sous-corpus et/ou de corpus

On sait, à partir de la géodésique, calculer la distance d'arc entre deux points à partir de leurs coordonnées angulaires. Il est dès lors possible d'envisager de calculer des distances angulaires dans l'espace résumé des 3 premiers facteurs. Cela ne présente pas un grand intérêt, nous semble-t-il, lorsqu'on reste dans les limites d'une même analyse. Les calculs de distance, entre les micro-distributions (paramétrées) de 2 items du corpus, par le chi-2 sont beaucoup plus précis puisqu'ils ne subiront pas la réduction géométrique indiquée *supra*.

Il n'en va pas de même pour ce qui est de la distance entre deux micro-distributions lexicales, dès lors qu'on ne se soucie pas de calculer un indice massif de distance, mais d'offrir une visualisation précise du contraste.

Dans un genre comparable, il semble possible de comparer la constitution lexicale fine de deux ensembles textuels A et B (entendue comme paramétrée : une liste d'items communs, un empan cotextuel et un écrêtage par l'écart-réduit identiques), en procédant comme suit.

Après avoir trouvé, pour la projection « sphérique » de B, la rotation engendrant la moindre distance angulaire cumulée[5] avec A, colorer les items de B selon l'angle qu'ils présentent entre leurs positions en B et A[6]. Si les deux ensembles sont très différemment structurés, la coloration sera nécessairement très confuse. Si en revanche un fort invariant existe, on pourra visualiser celui-ci, ainsi que les altérations de A à B, de façon intéressante.

Nous avons isolé des séries de 2 années consécutives, et observé particulièrement les séries 1985-86, 1986-87 et 1995-96, de manière à comparer deux séries très proches (et même chevauchantes) et deux séries écartées, ainsi qu'à comparer chacune de ces séries à l'ensemble.

Il devient ainsi possible de comparer visuellement ces distributions fines, item par item et surtout par groupes, d'avoir par exemple sur les évolutions de structures collocatives une vue synthétique jusqu'ici très difficile à obtenir. *Le Monde diplomatique* 1980-2000 présente, par exemple, au moins trois noyaux isotropiques d'une grande stabilité, c'est-à-dire constitués d'items toujours colorés en bleu, à travers les expérimentations successives (ne variant pas, en position dans la micro-distribution, de plus de 20° quel que soit le couple de sous-corpus comparés, pas plus qu'entre un sous-corpus et le corpus entier).

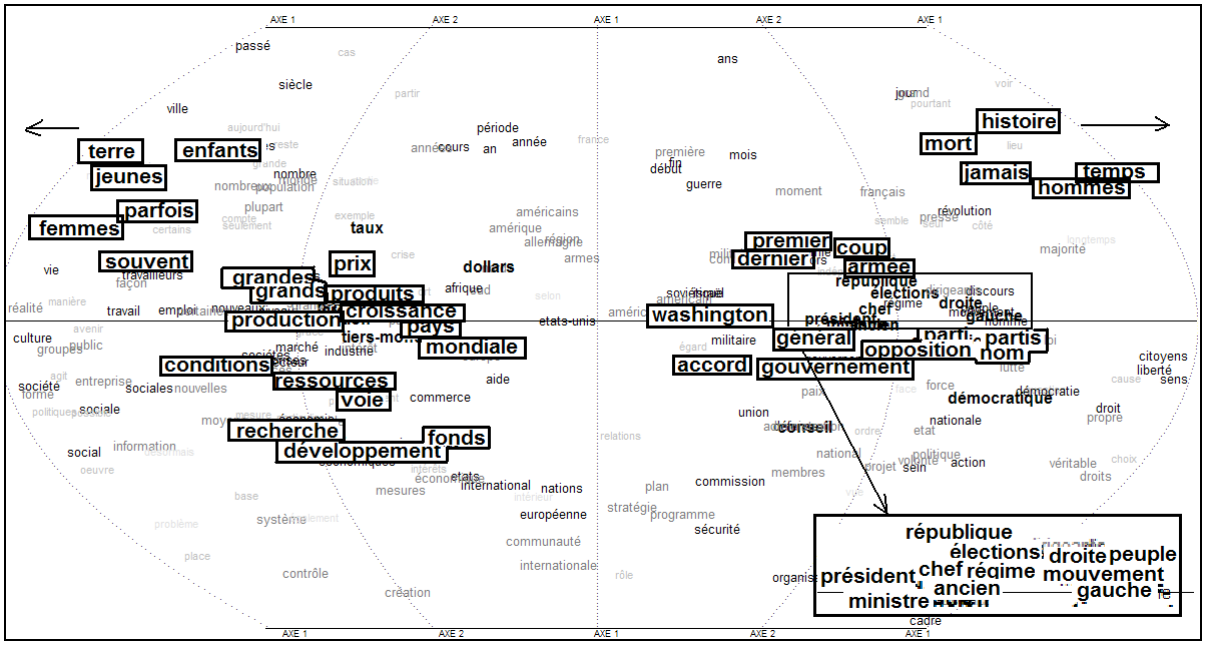


Fig.7 : planisphère, les items variant très peu d'un sous-corpus à l'autre, et des sous-corpus au corpus .

[5] La distance angulaire pour chaque item est coefficientée, afin de donner plus d'importance aux migrations des items éloignés du centre, par le produit des deux distances au centre (en A et en B).

[6] Convention :

α < 10°	bleu foncé
10° < α < 20°	bleu clair
20° < α < 40°	noir
40° < α < 80°	orange
α > 80°	rouge

6. Changements globaux de signification sur fond d'invariance ?

Une autre classe d'observations est permise par la confrontation des projections géodésiques. C'est le repérage d'items lexicaux qui migrent assez distinctement dans la micro-distribution et en laissent une trace significative. C'est le cas, pour *Le Monde diplomatique* entre 1985-86 et 1995-96, de la forme graphique *marché.*

La méthode est la suivante : on calcule les 2 angles que forme *marché* avec chacun des autres items sur la sphère A (1985-86) et sur la sphère B (1995-96). La comparaison de ces 2 angles permet de distinguer les items qui se rapprochent de *marché* dans cet intervalle de 10 ans (qui acquièrent donc un profil collocatif plus semblable) de ceux qui s'en éloignent et de ceux qui ne subissent pas d'évolution notable sur ce point.

Pour la plupart des items, l'observation est pertinente essentiellement dans les limites de leur voisinage (par exemple, sur un *zonage* comme celui de la fig.5 ; au-delà, elle devient le plus souvent confuse et peu interprétable ; ce ne sera pas le cas de l'exemple que nous avons choisi de retenir ici). En deux vues (colorées à l'identique, l'une zoomant dans le graphe 85-86, l'autre dans le graphe 95-96), on obtient une vue synthétique de la tendance du champ cooccurrentiel. Il en va ainsi de *marché* :

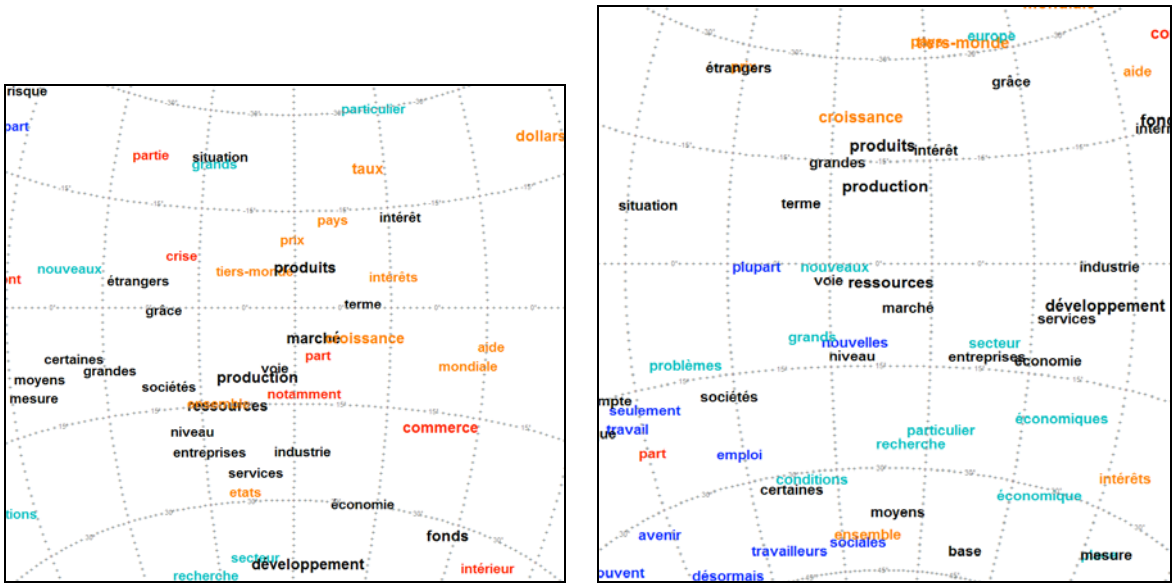


Fig.8 et 9 : zonages colorés 'migration des cooccurrents' centré sur marché, 85-86 (8) et 95-96 (9)

À gauche, le graphe de 85-86, avec une dominance de notations oranges et rouges, indiquant les items qui (sur la décennie) vont s'éloigner de *marché*, et quelques couleurs bleues indiquant des items qui vont se rapprocher (c'est-à-dire, venir très près) ; en noir, les items qui migrent très peu. À droite, celui de 95-96, où l'on voit « arriver » le nouveau matériel cooccurrentiel (en bleu) sur un invariant (en noir) par définition identique ; et l'on croit « voir » aussi que ce qui « arrive » vient du bas, et ce qui s'éloigne se dirige vers le haut à droite…

Impression, confirmée par la carte d'ensemble de 95-96 (fig.10, page suivante).

En réalité, notre pivot a « glissé » sur le fond de la carte, massivement structurée par son mouvement, indiquant nettement par là qu'il s'agit très certainement de l'un des *lieux* discursifs du changement, dans ce journal, opéré dans la première décennie néo-libérale.

7. Conclusion

Ces propositions sont soumises à la discussion des experts en statistiques. Elles relèvent d'une préoccupation majeure et permanente : favoriser la diffusion des méthodes d'analyse assistée sans instaurer d'effets de « boîte noire », valoriser *ad optimum* les vertus de visualisation de l'AFC, encore trop peu diffusées du fait notamment de leur faible ergonomie, et pour cela même, non seulement assurer les fonctions de *retour au texte*, mais les intégrer pleinement à un dispositif d'exploration hypertexte.

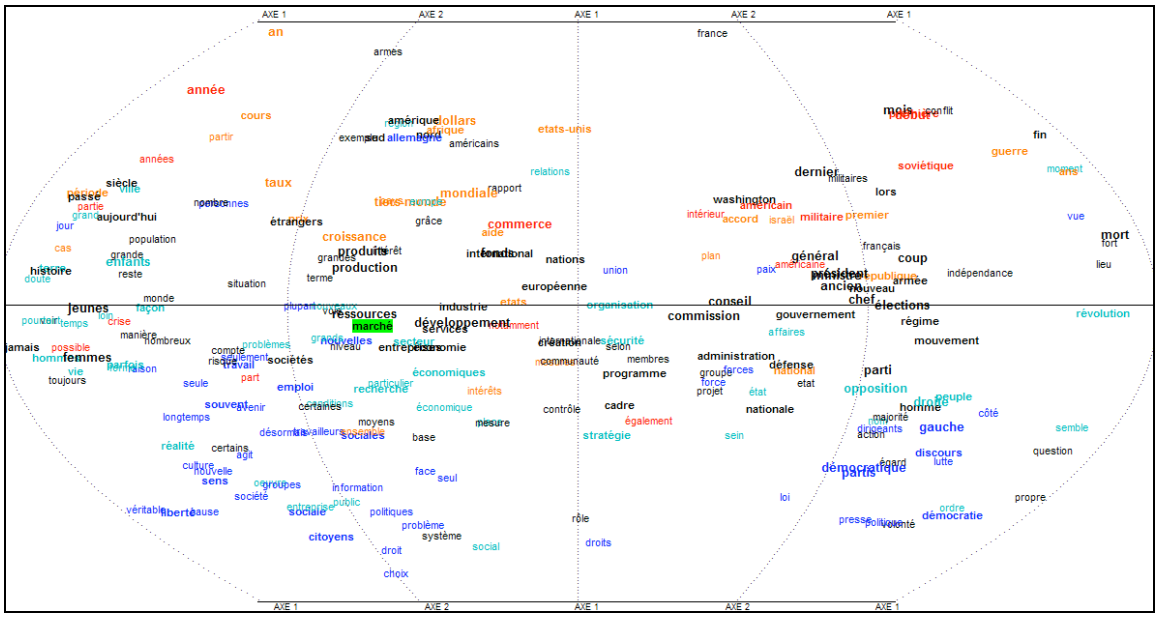


Fig.10 : planisphère coloré 'migration des cooccurrents' 95-96

L'AFC est à ce jour la méthode la plus stimulante pour faire monter d'un corpus et d'un sous-corpus leur structure micro-distributionnelle (c'est-à-dire un composant nodal de leur cohésion, de leur spécificité, de leur composition contrastive et en fin de compte, de leur signification) et, d'un même mouvement, une part notable de leur propre appareil de navigation ; à condition d'en assurer l'ergonomisation graphique.

Références

Adam J-M., Heidman U. (2005). *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité*. Genève, Slatkine.

Lebart L., Salem A. (1994). *Statistique textuelle*. Paris, Dunod.

Muller C. (1992a). *Initiation aux méthodes de la statistique linguistique*. Paris, Champion.

Muller C. (1992b). *Principes et méthodes de statistique lexicale*. Paris, Champion.

Viprey J-M. (2005). Méthodes pour la lecture des corpus. In *Sémantique et corpus*, dir. Anne Condamines, Hermès

Viprey J-M. (2005). Philologie numérique et herméneutique intégrative. In *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité*, dir. Jean-Michel Adam & Ute Heidman, Genève, Slatkine : 51-68.

Viprey J-M. (2000). Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle. In *JADT 2000*, Rajman & Chappelier éds., Lausanne, EPFL : 535-539.

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