Analyse séquencée de la micro-distribution lexicale

Jean-Marie Viprey

Archives, Textes, Sciences des Textes
Université de Franche-Comté – 25030 Besançon Cedex – France
jean-marie.viprey@univ.fcomte.fr

Abstract

Once one has worked out macro- and micro-distribution of the lexical items in the observed corpus, one wishes to study the collocation behavior of specific items, concerning diachrony and/or contrast. The *sequenced* analysis of micro-distributions consists in collecting the occurrence contexts of a chosen item, distributing their elements into the relevant sections of the corpus. The matrix made by this way can pass FCA, which gives a first synthetic view over evolution and/or contrasts inside signification. This method is to be integrated into expert hypertext, in order to increase our knowledge of the corpus, to rationalize critical readings by selecting relevant extracts, and to build ans supervise hypotheses. Applications to the *Monde Diplomatique* (1980-2000) and to the authorship dispute about Corneille and Molière.

Résumé

Une fois établies la macro-distribution et la micro-distribution des items lexicaux dans le corpus à l'étude, on souhaitera observer le comportement co-occurrentiel d'items ou de groupes d'items particuliers, dans une optique diachronique et/ou contrastive. L'analyse *séquencée* des micro-distributions consiste à relever les contextes d'occurrence du pivot et à en distribuer les éléments dans les divisions pertinentes du corpus. La matrice ainsi constituée est soumise à l'AFC, qui donne une première vue synthétique de l'évolution et/ou des contrastes dans la signification. Cette méthode s'intègre à l'hypertexte expert afin de favoriser la connaissance du corpus, de rationaliser les retours au texte par extraits pertinents, et de construire, affiner ou vérifier les hypothèses d'interprétation. Application au *Monde Diplomatique* (1980-2000) et à la controverse d'attribution sur Corneille et Molière.

Mots-clés : analyse multi-dimensionnelle, vocabulaire, diachronie, cooccurrence.

1. Introduction

Viprey (1997, 2000 et 2002a) montre que l'AFC appliquée à des matrices de cooccurrence permet d'asseoir une cartographie lexico-thématique pour l'hypertexte expert dans les bases et corpus littéraires. Viprey (2002b) en est une application aux *Fleurs du mal* de Baudelaire.

Le principe en est de substituer et/ou combiner, aux hypothèses exogènes, des visualisations endogènes sur la structure lexico-thématique afin d'accentuer les vertus exploratoires de l'environnement critique. Ces vues, dont les items sont sensibles à diverses modalités de clic souris, offrent le continuum des relations *isotropiques* (*trepein : pencher vers*), parentés des profils micro-distributionnels.

Par rapport à l'état de l'art du milieu des années 90, nous avons donc cherché à privilégier la micro-distribution, jusqu'alors masquée par la macro-distribution (distribution des items dans les sous-ensembles linéaires du corpus), condition *sine qua non* de l'extraction des structures fines qui SONT le *vocabulaire* comme structure de *textualité*.

On a pu légitimement reprocher à cette méthode de rejeter dans l'ombre les perspectives diachroniques et contrastives. Sans doute ce rejet a-t-il d'abord été voulu, afin de grossir le trait quant au changement de perspective recherché. Mais nous avons très tôt œuvré à la réintégration des critères de temps et de contraste, indispensables à toute entreprise d'analyse des discours. Il n'était pas moins nécessaire que cette réintégration se fasse dans le cadre de la méthode nouvelle et non par retour aux techniques antérieures. Aux JADT 2002, nous avons présenté un premier moyen dédié aux corpus à dimension diachronique : une animation présentant les transitions entre *n* AFC micro-distributionnelles calculées sur les tranches successives chevauchantes du corpus.

Des travaux plus récents, notamment sur des corpus de discours institutionnels fortement diachroniques, ont exigé la mise au point d'un autre outil dans le cadre de ®Astartex. Il s'agit d'une combinaison plus étroite et plus directement lisible de la micro- et de la macro- distributions, qui permet de visualiser les aspects les plus remarquables de l'évolution diachronique de la cooccurrence autour d'un pivot lexical.

Nous en présentons ici le principe et deux applications, l'une dans le domaine du discours politico-médiatique (exploration du corpus du *Monde Diplomatique* 1980-2000), l'autre dans la sphère littéraire (apport d'éléments de jugement dans le cadre du dossier de l'attribution de certaines pièces de Molière à Corneille).

2. Principes du repérage et application au *Monde Diplomatique*

Le corpus d'étude est celui du mensuel *Le Monde Diplomatique*, de 1980 à 2000 soit 21 années, 7000 articles pour 17 millions de mots environ. Il n'est pas étiqueté.

La distribution des items les plus récurrents par année d'édition est éloquente (fig. A1, en annexe[1]). Les positions respectives d'items lexicaux tels que (*travailleurs, gauche, état*) et de groupes d'années très bien délimités (80 à 83 par exemple), nous donnent des indications précieuses, tant sur la signification dont ces items lexicaux se chargent dans la configuration de ce corpus, que sur la signification globale du corpus : à partir de nos connaissances sur la place qu'occupe *Le Monde Diplomatique* dans le discours politique et social, que pouvons-nous supposer de l'évolution générale de ce discours ? Compte tenu de ce que nous savons sur l'évolution du discours politique et social de 1980 à 2000, comment pouvons-nous repenser la position qu'y occupe *Le Monde Diplomatique* ?

La micro-distribution de ces mêmes items (dans les limites de la phrase) est tout aussi intéressante (fig.A2, en annexe). On y voit clairement apparaître les dominantes lexico-thématiques les plus significatives (on n'a retenu ici que les 2 premiers facteurs sur plus de 200). On constate que la ventilation des items sur chacun des deux graphes est très différente ; elle a d'ailleurs une signification évidemment différente.

Nous constatons que *femmes* (qui est aussi un pôle du graphe macro-distributionnel) occupe une position polaire sur le graphe micro-distributionnel d'ensemble. Cette forme a un profil voisin de ceux de *social(e)(s), société, enfants, jeunes, vie, culture, travail*, etc. Elle appartient à cette *isotropie* à l'échelle de l'ensemble du *Monde diplomatique* 1980-2000. Mais comment en savoir plus sur la signification diachronique de cette forme ?

Nous nous proposons de relever les contextes des 3603 occurrences de *femmes*, dans les limites de la phrase toujours (et de 25 mots à gauche et à droite), et de distribuer les coccur-

[1] Les AFC de cette étude sont réalisées avec le programme écrit dans SciLab par Alain Lelu (LASELDI, Université de Franche-Comté).

rences dans une matrice diachronique (une colonne par année). Un seuil de cooccurrence est fixé (au moins égal à 33 % du nombre de partitions – années, dans la limite des 200 premiers cooccurrents).

L'AFC de cette matrice nous éclaire sur la tendance cooccurrentielle de *femmes*, donc sur un aspect important de l'évolution de sa signification, dans *Le Monde Diplomatique* (et dans la mesure du possible, dans le discours politique et social tel que ce mensuel y contribue). On y voit se dessiner deux zones polaires assez bien différenciées, correspondant l'une aux années 80 à 83, l'autre aux années 95 à 2000, les années intermédiaires se distribuant de manière plus confuse..

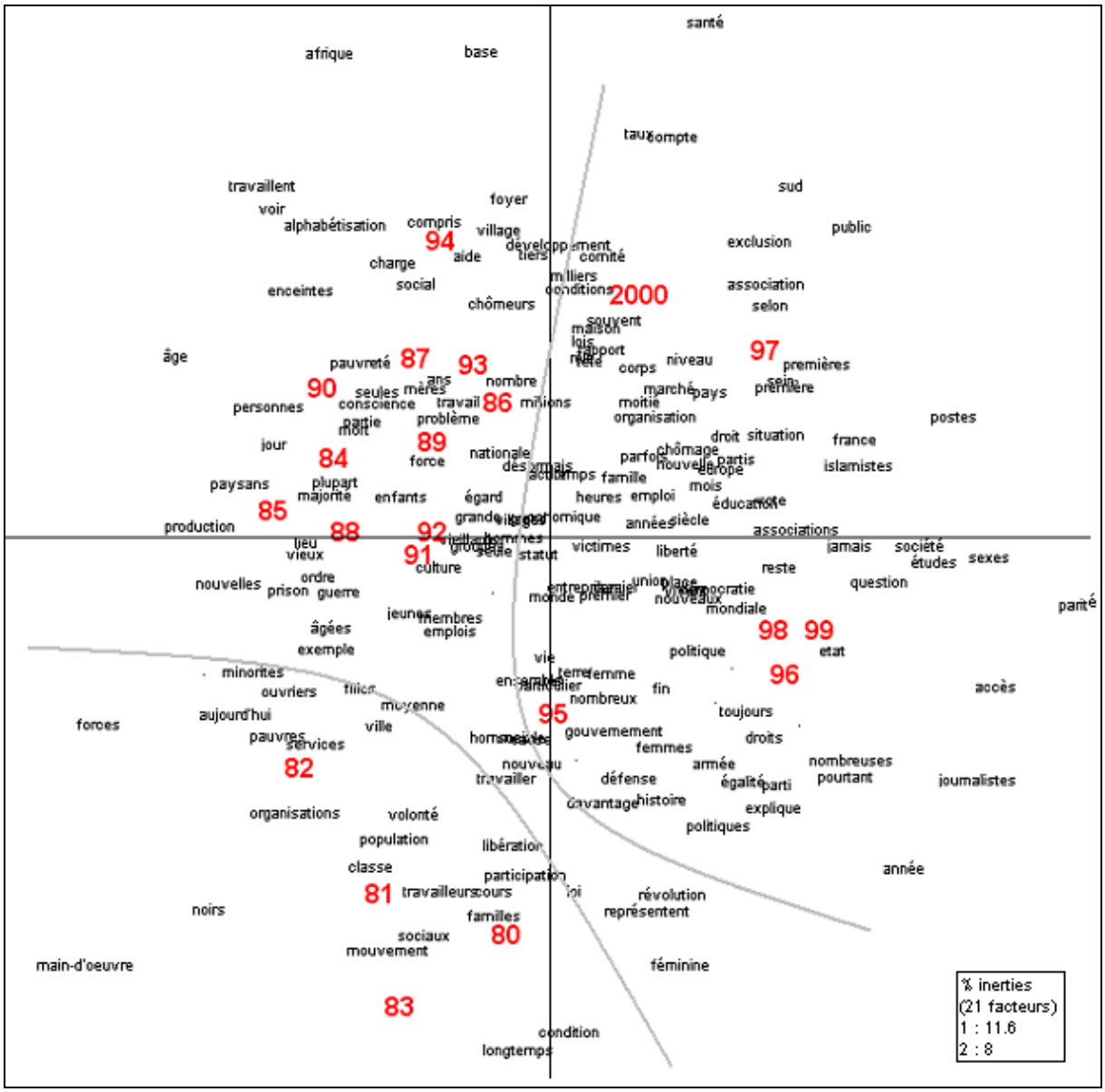

Figure 1. Graphe macro-distributionnel des cooccurrents de femmes*. 186 formes lexicales*

L'intérêt majeur de ces graphes étant de s'intégrer à un hypertexte expert, facilitant le retour immédiat et structuré aux contextes pertinents, nous ne pouvons en donner ici qu'une idée trop générale. Il est néanmoins patent que les cooccurrents caractéristiques des années 80-83 font référence à la sphère des rapports sociaux « classiques » telle qu'elle se définissait encore

alors, tandis que ceux de 95-2000 semblent se rapporter plus confusément aux questions de statut.

Si l'on détermine, par l'écart-réduit à l'équidistribution des cooccurrences, quels sont les plus forts cooccurrents de *femmes* dans les quatre périodes successives de cinq ans :

1980-84	cc	er	1985-90	cc	er	1991-95	cc	er	96-2000	cc	er
mouvement	23	4,03	enfants	106	3,63	enfants	105	2,73	parité	19	3,25
problèmes	12	3,86	hommes	201	3,26	alphabétisation	9	2,41	postes	22	2,69
main-d'oeuvre	9	2,54	aujourd'hui	16	2,45	terre	10	2,28	société	47	2,64
travailleurs	15	2,34	noirs	15	2,38	rues	9	2,11	sexes	20	2,62
ouvriers	12	2,32	jour	15	2,29	ans	41	2,10	accès	22	2,53
longtemps	9	2,25	afrique	12	2,28	travaillent	14	2,02	droits	67	2,49
aujourd'hui	15	2,16	enceintes	12	2,17	familles	11	1,95	études	17	2,22
paysans	9	2,11	âge	14	1,94	développement	14	1,92	question	26	2,16
lieu	11	2,03	mères	10	1,90	taux	10	1,89	selon	24	2,06
population	17	1,99	majorité	17	1,79	main-d'oeuvre	8	1,78	france	33	2,01

Figure 2. Graphe les cooccurrents forts de femmes *sur 4 périodes de 5 ans du* Monde Diplomatique

on pourra notamment favoriser le retour vers l'ensemble des phrases caractéristiques de ces périodes, celles où ces cooccurrents forts sont concentrés.

Ainsi pour la période 1980-84 : (Patrick Tissier, « LA RELANCE DE L'ÉCONOMIE », Octobre 1983)

> Les primes liées au travail aux pièces renforcent les disparités entre travailleurs à la production et travailleurs auxiliaires : par exemple, dans l' industrie lourde, la majorité de ces derniers sont des femmes moins payées que les hommes, surtout engagés dans la production.

Ou pour 1996-2000 : (Olfa Lamloun, « DISCOURS MODERNISATEUR POUR RÉGIME RÉPRESSIF », Juin 1998)

> Pour autant, l'hymne au Sâna'al Tahauil occulte le rôle des tunisiennes et la mobilisation des féministes, qu'il s'agisse du club Tahar Haddad, de la commission d'études de la condition des femmes travailleuses au sein de l'Union Générale des Travailleurs Tunisiens, de la commission de défense des droits des femmes de la Ligue Tunisienne des Droits de l'Homme, ou encore de l'Association des Femmes Démocrates.

On comprendra qu'il s'agit ici d'une pratique essentiellement pré-exploratoire, qui n'a d'autre objet que de soulever des questions en faisant, par principe, monter les faits saillants de la structure elle-même. C'est ainsi que le choix des pivots peut être lui-même plus ou moins automatique. Rappelons qu'en l'occurrence nous avons porté notre attention sur *femmes* en raison de sa position deux fois polaire (en micro- et en macro-distributions).

On peut bien sûr objecter qu'il est dans une certaine mesure normal que le vocabulaire cooccurrent de *femmes*, qui représente en masse, selon les paramètres choisis, 110'000 mots environ soit 0,65 % de l'ensemble du corpus, suive les tendances diachroniques d'ensemble. Cela n'enlève rien à l'utilité de ce graphe, qui nous informe sur la contribution particulière de *femmes* à la macro-distribution diachronique d'ensemble (l'un ne pouvant en aucun cas être indépendant de l'autre).

Mais cela exige de compléter l'information par un autre aperçu d'ensemble, où l'évolution des grandes masses aura été compensée de manière à voir comment, compte tenu même des emplois croissants et décroissants, parmi ce vocabulaire en mouvement, *femmes* modifie la « sélection » de ses cooccurrents. Pour cela, il nous faut considérer, pour chaque cooccurrent dans chaque partie du corpus (année), la proportion de l'ensemble de ses emplois qui appartiennent au contexte de *femmes*. Cette proportion peut être alors appliquée à une matrice « idéale » de 21 parties égales en volume, où les cooccurrents se répartiraient équitablement.

Ainsi, pour *travail* qui a 8806 occurrences, la base de calcul pour une partie sera de 8806/21, soit 419,3. Dans l'année 1987, on constate que 7 des 273 occurrences effectives de *travail* sont en contexte avec *femmes*, soit une proportion de 0,0256. En appliquant 0,0256 à 419,3, on établit la cooccurrence compensée de *femmes* et de *travail* dans l'année 1987, qui est donc de 10,75 (que l'on arrondit à 11).

L'AFC de la matrice des données ainsi compensées offre ce graphe :

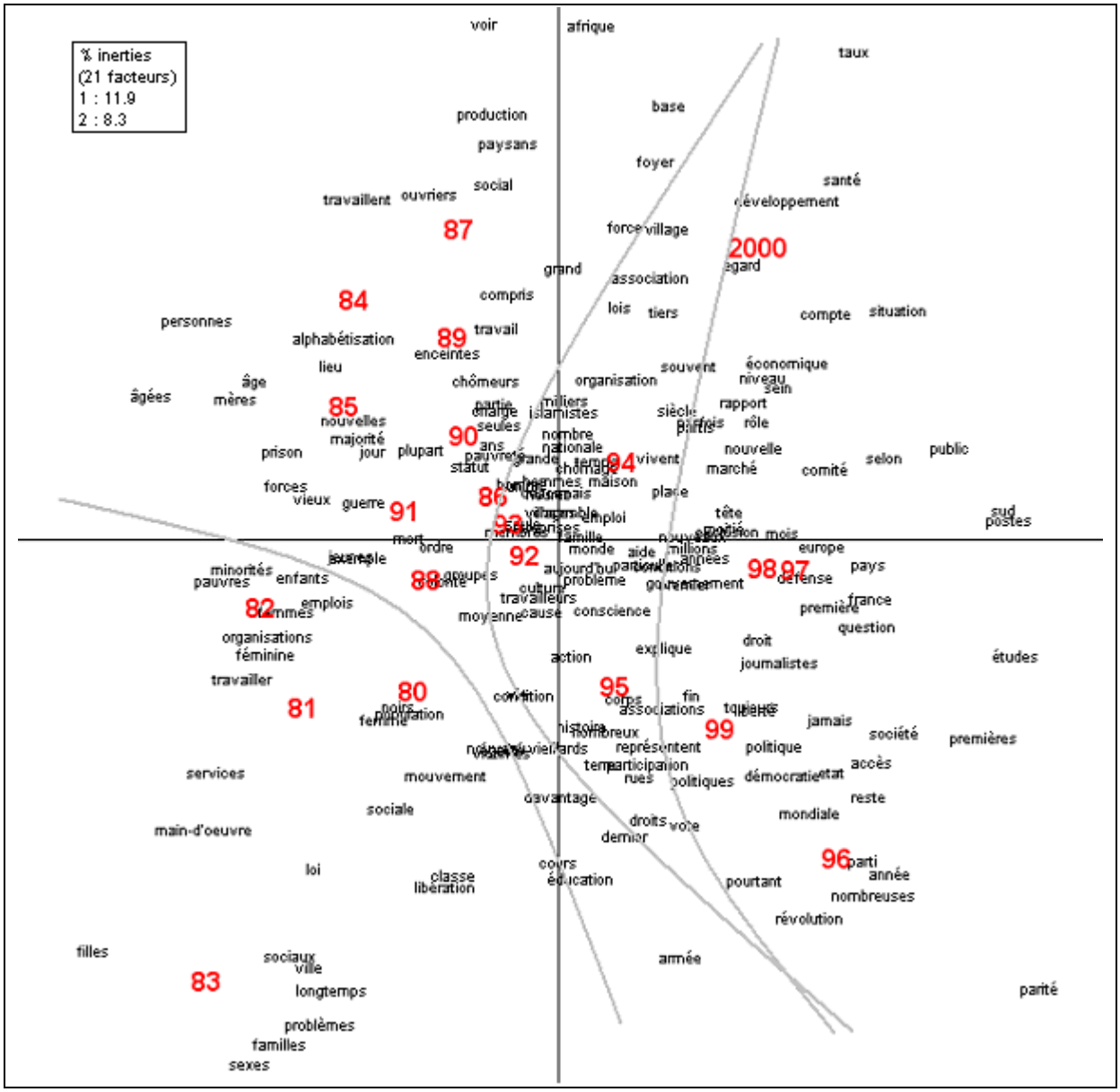

Figure 3.
Graphe macro-distributionnel des cooccurrents [matrice compensée] de femmes. *186 formes lexicales.*

La différence est assez minime. L'orientation diachronique est confirmée, voire renforcée par la restriction du cadre d'étude. L'évolution de ses cooccurrents est relativement autonome par rapport à celle des grandes masses du vocabulaire du *Monde Diplomatique*. On peut dès lors parler de diachronie propre de l'item lexical dans la temporalité du corpus. Il y a bien, dans ce secteur du discours politico-institutionnel, une modification des valeurs de l'item, telles que l'attestent les environnements différents.

Savoir quelle est cette modification relève d'une étude minutieuse des contextes dont il n'est pas question de rendre compte dans les limites ici imparties. Cette étude est facilitée par la collecte ergonomique qui en sera faite par simple click sur un cooccurrent indiqué sur la carte.

Par contraste avec *femmes*, nous présenterons plus sommairement encore le cas de *réalité(s)*, qui a retenu nos premières attentions du fait qu'il joue (avec d'autres), dans le discours politico-institutionnel, un rôle certes obscur, mais néanmoins irremplaçable, de vecteur de l'implicite socio-économique (plus encore que *loi(s)* ou *contrainte(s)*, plus patents).

Comment le discours d'un journal marqué à gauche (et cela de façon croissante entre 1980 et 2000, au point de le faire classer aujourd'hui dans la gauche « radicale »), se tire-t-il de l'emploi délicat (et *a priori* peu auto-surveillé) de ces formes ? On compte 3670 occurrences du pluriel et 751 du singulier.

On note tout d'abord que cet emploi tend à s'élever, comme le suggère l'histogramme des écarts-réduits à l'équi-distribution par année :

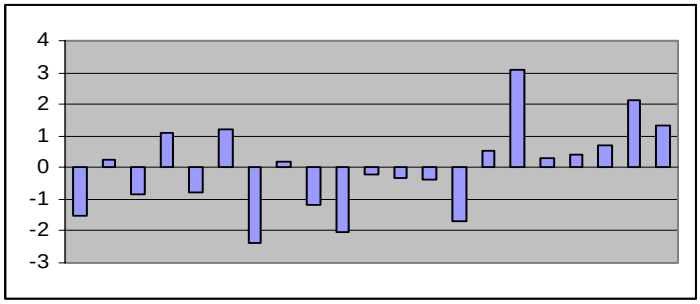

Figure 4. Histogramme des écarts-réduits à l'équi-distribution de réalité(s) dans le corpus LMD

Pour les 7 dernières années (94 à 2000), l'occurrence constatée est 1635, contre 1511 dans l'hypothèse d'équi-répartition : l'écart-réduit est de 3,2. Par ailleurs, la proportion des emplois au singulier s'élève assez sensiblement à partir de 1990 (de 15 à 18 %).

Surtout, si l'on compare sur des cartes « muettes » (où seules les années sont indiquées, et non les formes recensées), les distributions diachroniques de ses cooccurrents, « brute » à gauche,

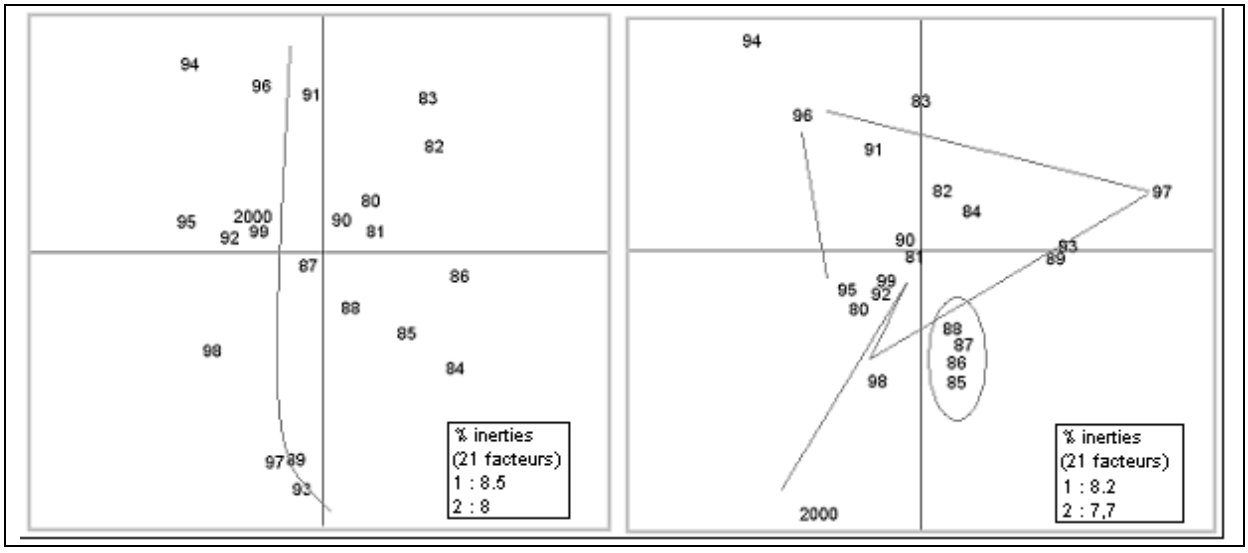

Figure 5. Graphes macro-distributionnels des cooccurrents de réalités(s).
À gauche, sans compensation ; à droite, avec compensatio.

« compensée » à droite, on relève que l'orientation encore sensible en « brut » (c'est-à-dire lorsqu'elle récupère une part notable de l'inertie de la distribution totale), se dissipe presque entièrement, ou du moins se fragmente en « compensé » (on repère encore quelques séquences d'années, mais on voit surtout de grands et brusques écarts). La participation de *réalité(s)* à l'évolution diachronique des masses du vocabulaire n'est donc pas clairement distincte de sa propre évolution distributionnelle. En d'autres termes, ils ne jouent pas un rôle cohésif dans le discours du *Monde Diplomatique* comme diachronie signifiante globale.
Cela n'implique pas que l'étude diachronique de leurs collocations soit sans intérêt, au contraire, puisqu'il y a sans doute moins de maîtrise idéologique sur l'emploi de ces formes : donc, à tout le moins, une attention spécifique à leur porter (surtout, compte tenu de l'augmentation significative de leur occurrence).

3. Application à une controverse d'attribution : Corneille/Molière.

Nous avons montré *supra* un exemple d'application purement pré-exploratoire.

Cette méthode d'analyse séquencée de la micro-distribution peut aussi être convoquée pour dialoguer avec une hypothèse de travail constituée. Dans son argumentation destinée à soutenir que Corneille a réellement écrit 16 des pièces attribuées à Molière, D. Labbé (2003 : 111-116) développe notamment (en substance) que les collocations d'items essentiels (comme *amour* par exemple) sont proches chez Corneille et dans ces 16 pièces.

Il nous semble possible de vérifier cette assertion de manière assez objective et précise, en employant l'analyse séquencée de la micro-distribution, dans une perspective non plus diachronique, mais simplement contrastive. Relevons, au fil des 67 pièces du corpus, les contextes de pivots lexicaux mentionnés par D. Labbé, et distribuons-les dans une matrice de 67 colonnes. Prenons l'exemple des vocables *amour* et *aimer* (cumulés), dans les limites de la phrase (25 mots à droite et à gauche)[2], pour lesquels nous avons, comme exposé *supra*, sélectionné les 201 premiers cooccurrents (en cooccurrence brute).

Avec 67 sections de corpus, il ne peut être question de visualiser en même temps la distribution des items et la distribution dans les pièces. Nous montrerons donc seulement la carte « muette », avec les seules positions des pièces. Dans l'environnement ®Astartex, ce sont des codes couleurs qui permettent de visualiser les différentes classes qui nous intéressent. Sur ce papier, ce sont des conventions graphiques. Les titres (abrégés) des pièces de Molière sont soulignés. Ceux des pièces non attribuées à Corneille sont estompés en gris. Ceux des comédies de Corneille sont en italiques. Nous avons en outre marqué de l'astérisque les pièces de Molière antérieures à 1662.

Cette distribution (fig. 6, p. suiv.) ne prouve certes rien, mais elle apporte des éléments considérables pour indiquer au contraire que les vocables *amour* et *aimer*, comme la plupart d'ailleurs des vocables significatifs du corpus, sont traités de façon fort différente par les deux auteurs, à l'exception des cas de *Dom Garcie de Navarre*, proche du genre de la comédie héroïque et depuis longtemps connue par les spécialistes pour avoir été très fortement influencé par Corneille, et des premières pièces de Molière, qu'elles soient ou non réattribuées à Corneille, et qui se rapprochent en effet dans cette distribution des comédies de ce dernier.

[2] Le corpus est celui de D.Labbé, lemmatisé par ses soins.

Ce n'est pas le critère de la réattribution supposée à Corneille (titres en noir *vs* titres en gris) qui justifie l'éloignement des comédies de Molière par rapport au noyau de celles de Corneille (en haut du graphe), mais la diachronie.

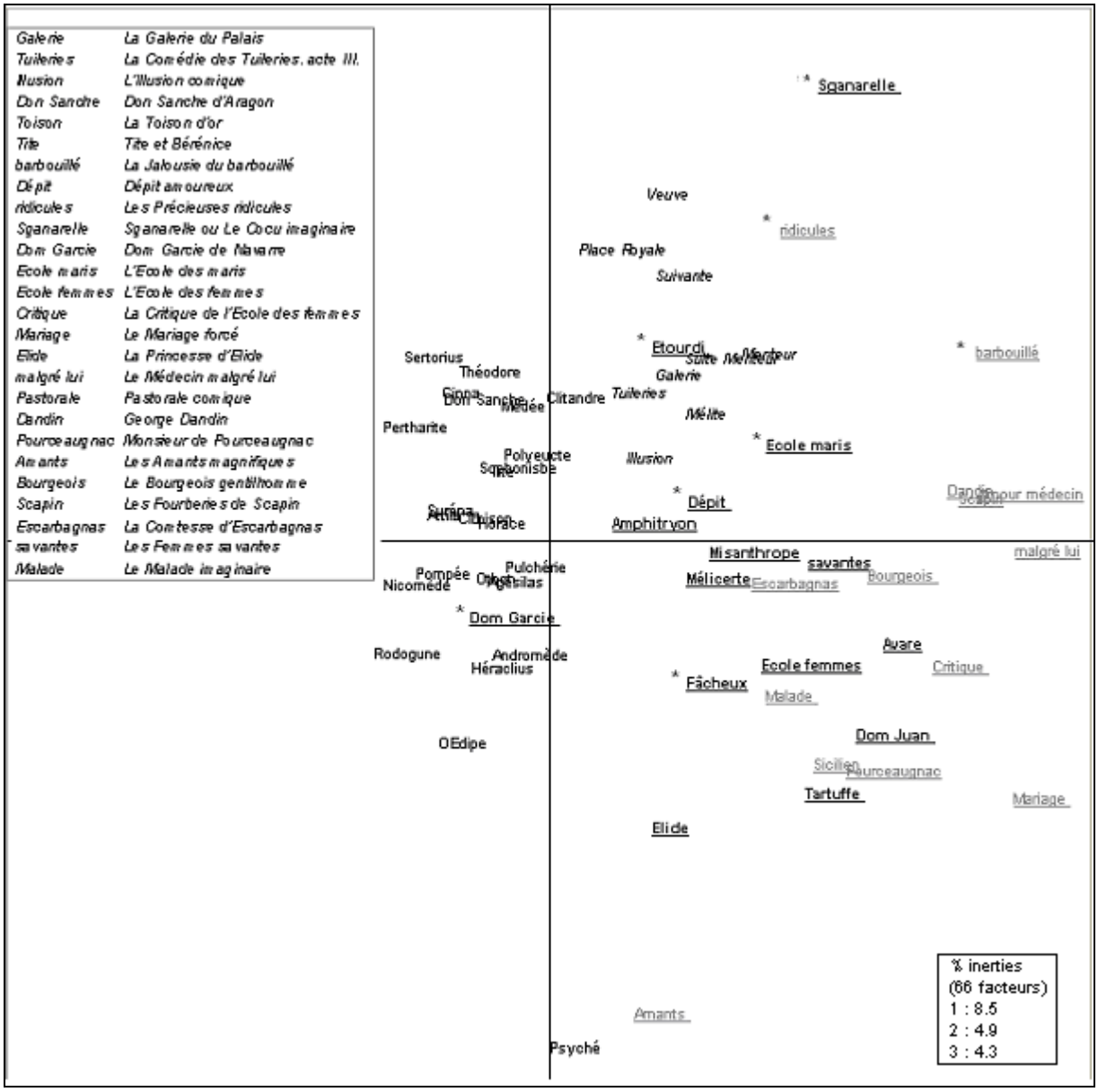

Figure 6. Distribution des co-occurrents de amour *et* aimer *dans le corpus Corneille/Molière*

En revanche, si l'on se libère de l'optique inférentielle pour se replacer dans une perspective exploratoire, on peut maintenant observer, à partir de la cooccurrence séquencée, d'authentiques variations quant au comportement de certains vocables dans cette comparaison. C'est ainsi que, pour le couple *père-fils*, les proximités repérées par D. Labbé semblent plus actives, ainsi que le montrent les positions respectives des pièces de Molière « réattribuées » (+) ou non (-) par rapport au sous-nuage Corneille, ainsi que la position des 2 *Menteurs*. On peut en déduire, avec toute la prudence de mise, que certains vocables sont bel et bien responsables, plus que d'autres, des proximités lexicales fines, et non massives, entre nos deux auteurs (fig. 7, p. suiv.).

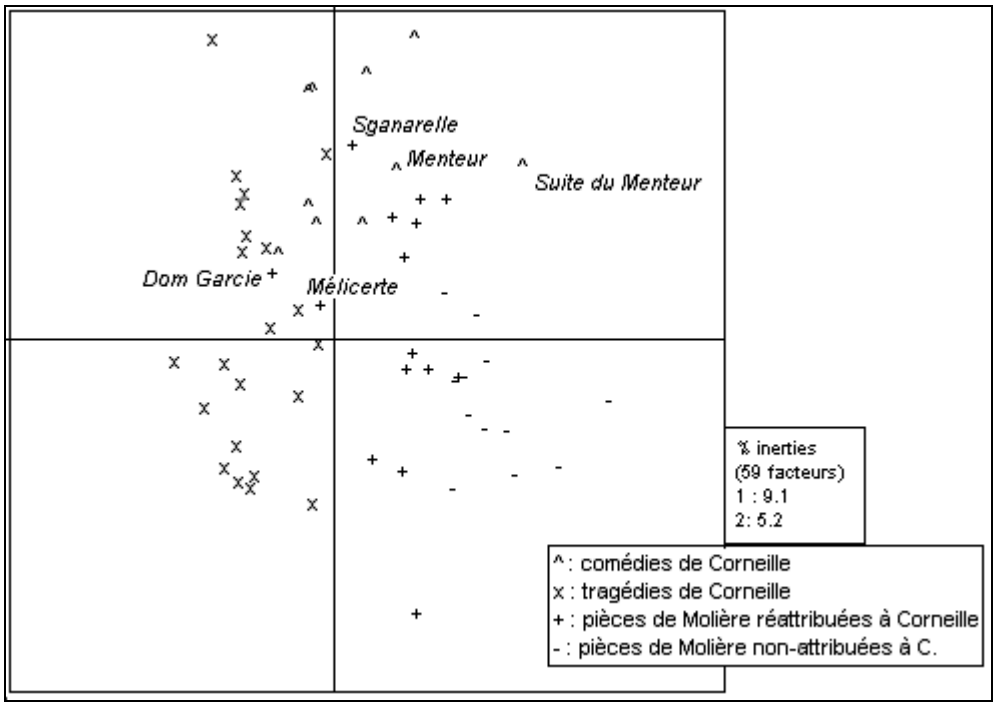

Figure 7. Distribution des co-occurrents de père *et de* fils *dans le corpus Corneille/Molière*

Il reste un test à produire à partir de ces collectes de données. Pour *amour/*aimer, comme pour *père/*fils, nous disposons de matrices de distribution des cooccurrences, dont nous pouvons regrouper les données selon une classification pertinente des pièces du corpus. Nous créerons 4 classes : les comédies de Corneille (CC), les autres pièces de Corneille (CT), les pièces de Molière réattribuées à Corneille par D. Labbé (MA), celles qu'il juge douteuses ou n'attribue pas du tout (MN). Nous créerons également 3 classes mixtes, où nous regrouperons CC et CT en C, MA et MN en M, C et MA en CMA.

Pour chacune des 7 classes, nous attribuerons un rang à chacun des cooccurrents de *amour/aimer*, selon sa cooccurrence brute. Puis, nous calculerons la corrélation de ces rangements (indice de Spearman[3]). Bien sûr, toutes les corrélations seront positives puisque l'ordre décroissant de cooccurrence dépend très largement de l'occurrence totale. Ainsi, *faire* occupe-t-il toujours et partout le 1er rang et *voir* ne descend-il jamais au-delà du 4e. Cependant, les variations de l'indice de Spearman sont tout à fait éclairantes. Une seule corrélation s'élève significativement au-dessus des autres : celle qui résulte de la comparaison entre les deux sous-corpus de Molière, qui montrent ainsi leur très grande homogénéité de ce point de vue. Les autres corrélations sont très voisines les unes des autres, et notamment deux qui sont décisives du point de vue de l'hypothèse explorée : entre les comédies de Molière réattribuées à Corneille et (1) les comédies de Corneille (2) l'ensemble de Corneille.

	MA/C	M/C	MA/MN	M/CC	CC/CT	MA/CC
amour+aimer	0,5037	0,4880	0,7519	0,5199	0,4915	0,5056
père+fils	0,4559	0,4132	0,7631	0,5652	0,4175	0,5766

Figure 8. Corrélation (Spearman) des rangs des plus forts occurrents dans les sous-corpus indiqués

[3] L'indice de Spearman oscille entre +1 – identité de rangement -- et +1 – rangement inverse. 0 indique une corrélation nulle. L'interprétation de l'indice est développée par Muller (1992-A : 153 ssqq). Nous nous contentons ici de comparer les indices pour des couplages donnés de sous-corpus.

Plus précis encore : ce tableau nous montre bien que, contrairement à *amour+aimer*, *père+fils* présentent une meilleure corrélation entre MA et CC, qu'entre MA et C, ce qui en même temps confirme le graphe n°6 (parenté relative de MA et des comédies de Corneille du point de vue de l'emploi de *père+fils*), mais aussi *relativise* cette parenté : elle reste très en-deça de celle constatée, pour les mêmes items, entre les deux sous-corpus de Molière. Avec la corrélation MA/CC, nous excluons le facteur *genre* : il n'y a plus, si nous acceptons cette simplification, que le facteur *auteur*…

4. Conclusions

L'analyse séquencée de la cooccurrence lexicale permet de combiner les analyses micro-distributionnelle et macro-distributionnelle, au profit d'une exploration plus poussée des corpus textuels.

Elle permet d'affiner notamment l'étude diachronique et/ou contrastive de la signification d'un item, ou d'un groupe d'items, telle qu'elle se développe dans les limites d'un corpus constitué, à partir de la norme endogène de celui-ci. Les items étudiés étant eux-mêmes extraits du vocabulaire grâce à des processus endogènes (micro- et macro-distributions globales dans le corpus), on a toutes les chances, tout en conservant un niveau élevé d'objectivation, d'être orienté rapidement et ergonomiquement, dans le cadre d'un hypertexte expert, vers des phénomènes hautement signifiants.

Cette démarche permet aussi de dialoguer utilement avec des hypothèses fortes, dans l'objectif de les vérifier certes, mais surtout de les raffiner pour le meilleur profit de l'analyse de textes.

5. Références bibliographiques

Coll. (2001). *Le Monde diplomatique 1980-2000 : 21 ans d'archives*. CEDROM-SNI, Québec.

Labbé D. (2003). *Corneille dans l'ombre de Molière*. Les Impressions nouvelles.

Muller Ch. (1992a). *Initiation aux méthodes de la statistique linguistique*. Champion.

Muller Ch. (1992b). *Principes et méthodes de statistique lexicale*. Champion.

Schepens P., Lelu A. et Viprey J.-M. (2004, à paraître). *Essais d'analyse de discours sur des corpus d'entretiens cliniques*. 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 *Actes des JADT 2000* : 535-538.

Viprey J.-M. (2002a). « Dynamisation de l'analyse micro-distributionnelle des corpus textuels ». In *Actes des JADT 2002* : 779-790.

Viprey J.-M. (2002b). *Analyses textuelles et hypertextuelles des* Fleurs du mal. Champion.

Annexes :

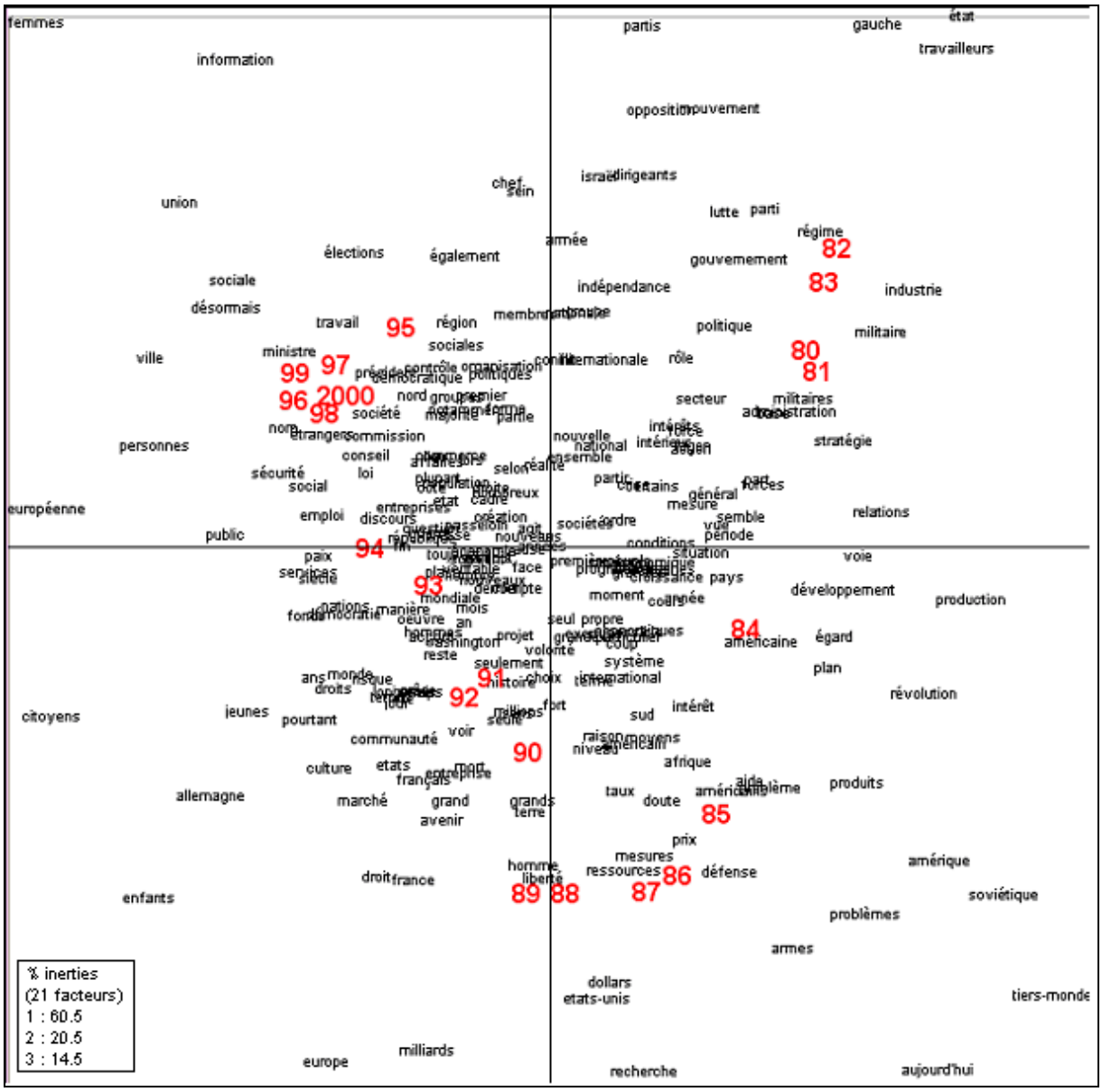

Figure A1. Graphe macro-distributionnel. 240 formes lexicales du corpus Monde Diplomatique

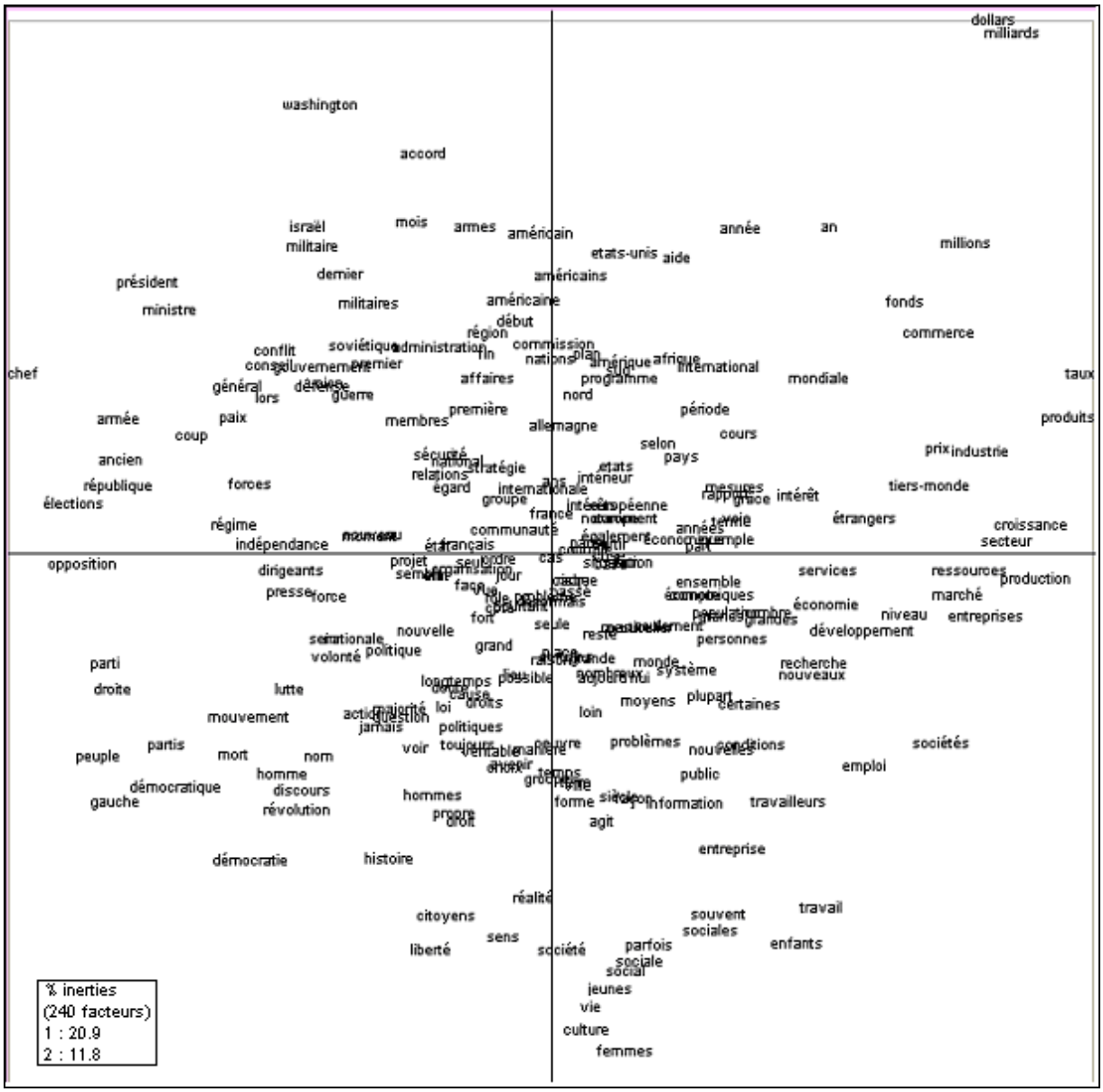

Figure A2. Graphe micro-distributionnel. 240 formes lexicales du corpus Monde Diplomatique