Eléments de statistique et de visualisations pour l'analyse textuelle

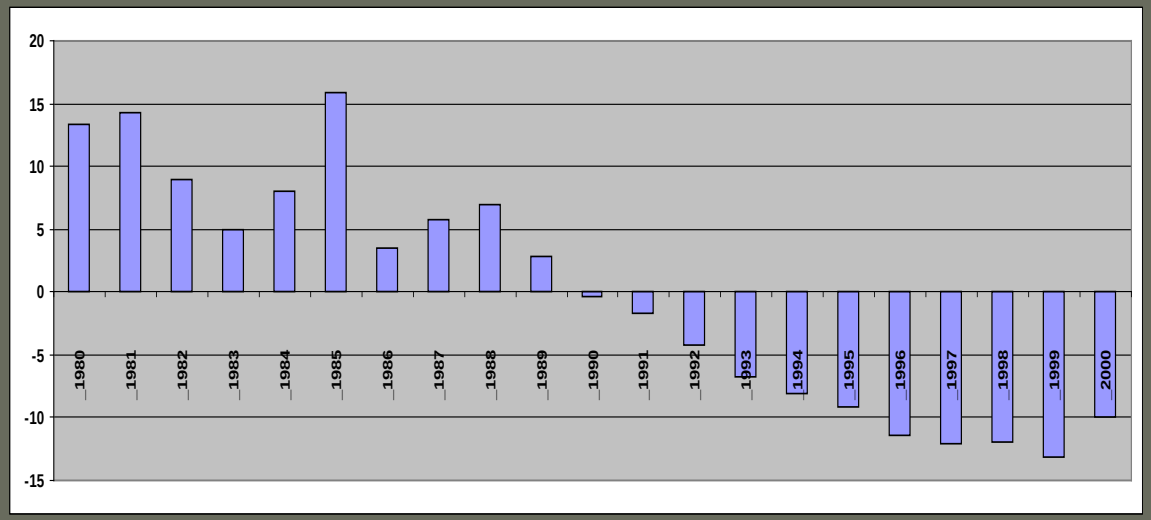
Jean-Marie Viprey

Maison des Sciences de l'Homme de Franche-Comté

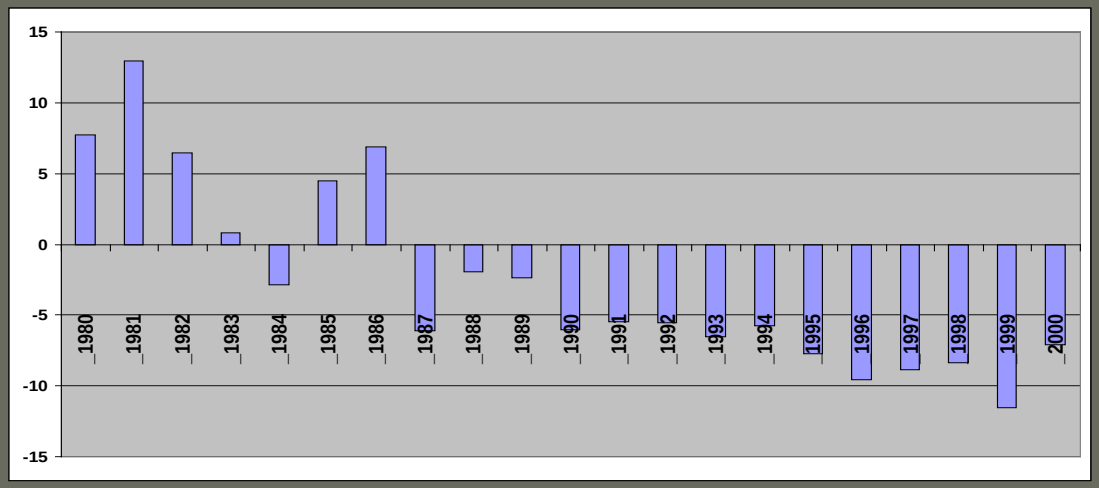
jmviprey@univ-fcomte.fr

Vues d'ensemble sur les déficits et excédents dans Le Monde diplomatique (1980-2000)

tiers-monde

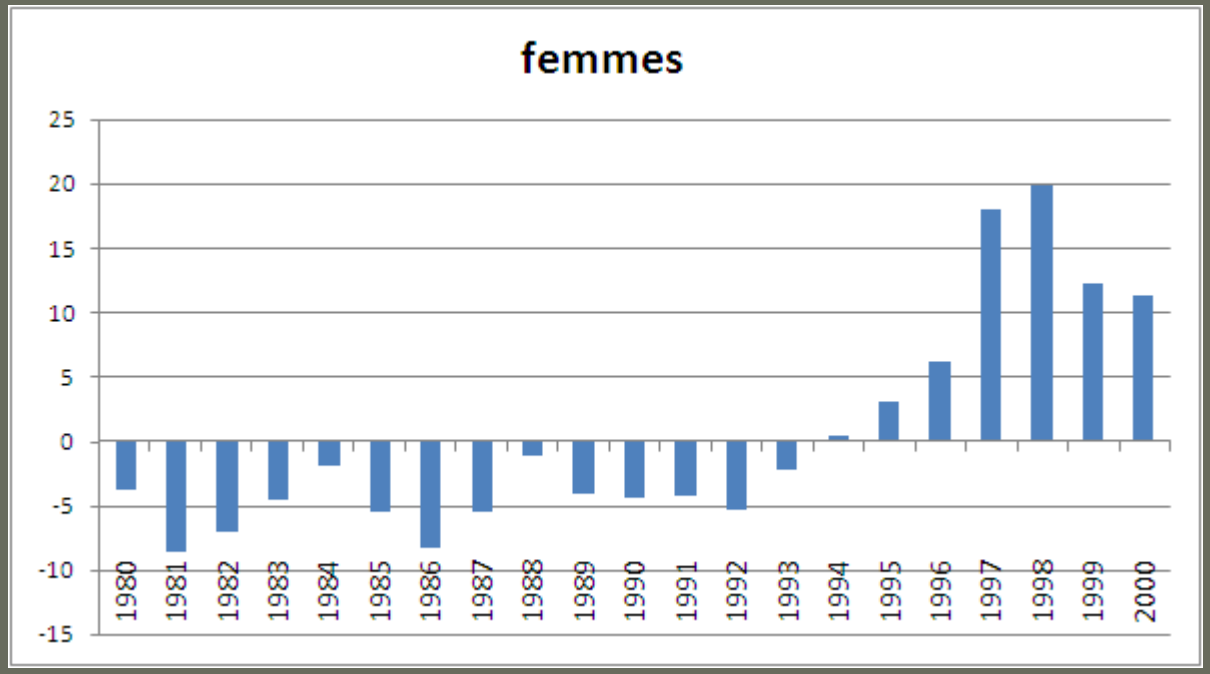


production



Profil (distributionnel).

On nommera ces histogrammes et ce qu'ils figurent des *profils distributionnels* (ici : *diachroniques*).



L’analyse factorielle des correspondances.

Soit le tableau de distribution suivant (avec ses « marges ») :

	80-83	84-87	88-91	92-95	96-2000	TOTAL
développement	3254	2785	2457	2116	1986	12598
gouvernement	4041	3038	2445	2779	3136	15439
guerre	2968	2993	3372	3241	3723	16297
marché	1770	1804	2338	2305	2410	10627
société	2025	1698	2150	2065	2899	10837
tiers-monde	1599	1385	1006	488	203	4681
travail	2056	1836	1768	2077	2911	10648
production	2493	1813	1481	1255	1157	8199
TOTAL	20206	17352	17017	16326	18425	89326

Le Monde diplomatique 1980-2000

L’analyse factorielle des correspondances.

On peut lui confronter un tableau « d’indépendance », où les valeurs sont calculées dans l’hypothèse d’équirépartition

	80-83	84-87	88-91	92-95	96-2000	TOTAL
développement	3254	2785	2457	2116	1986	12598
gouvernement	4041	3038	2445	2779	3136	15439
guerre	2968	2993	3372	3241	3723	16297
marché	1770	1804	2338	2305	2410	10627
société	2025	1698	2150	2065	2899	10837
tiers-monde	1599	1385	1006	488	203	4681
travail	2056	1836	1768	2077	2911	10648
production	2493	1813	1481	1255	1157	8199
TOTAL	20206	17352	17017	16326	18425	89326

contingence

« correspondance »

	80-83	84-87	88-91	92-95	96-2000	TOTAL
développement	2849,73	2447,22	2399,97	2302,52	2598,55	12598
gouvernement	3492,38	2999,1	2941,2	2821,77	3184,56	15439
guerre	3686,47	3165,77	3104,65	2978,58	3361,53	16297
marché	2403,88	2064,35	2024,49	1942,28	2192	10627
société	2451,39	2105,14	2064,5	1980,66	2235,31	10837
tiers-monde	1058,87	909,306	891,751	855,54	965,536	4681
travail	2408,63	2068,42	2028,49	1946,12	2196,33	10648
production	1854,66	1592,69	1561,95	1498,52	1691,18	8199
TOTAL	20206	17352	17017	16326	18425	89326

indépendance

L’analyse factorielle des correspondances.

On peut lui confronter un tableau « d’indépendance », où les valeurs sont calculées dans l’hypothèse d’équirépartition

	80-83	84-87	88-91	92-95	96-2000	TOTAL
développement	3254	2785	2457	2116	1986	12598
gouvernement	4041	3038	2445	2779	3136	15439
guerre	2968	2993	3372	3241	3723	16297
marché	1770	1804	2338	2305	2410	10627
société	2025	1698	2150	2065	2899	10837
tiers-monde	1599	1385	1006	488	203	4681
travail	2056	1836	1768	2077	2911	10648
production	2493	1813	1481	1255	1157	8199
TOTAL	20206	17352	17017	16326	18425	89326

contingence

indépendance

	80-83	84-87	88-91	92-95	96-2000	TOTAL
développement	2849,73	2447,22	2399,97	2302,52	2598,55	12598
gouvernement	3492,38	2999,10	2941,20	2821,77	3184,56	15439
guerre	3686,47	3165,77	3104,65	2978,58	3361,53	16297
marché	2403,88	2064,35	2024,49	1942,28	2192,00	10627
société	2451,39	2105,14	2064,50	1980,66	2235,31	10837
tiers-monde	1058,87	909,31	891,75	855,54	965,54	4681
travail	2408,63	2068,42	2028,49	1946,12	2196,33	10648
production	1854,66	1592,69	1561,95	1498,52	1691,18	8199
TOTAL	20206	17352	17017	16326	18425	89326

On en déduit un tableau des écarts :

	80-83	84-87	88-91	92-95	96-2000
développement	404,27	337,78	57,03	-186,52	-612,55
gouvernement	548,62	38,90	-496,20	-42,77	-48,56
guerre	-718,47	-172,77	267,35	262,42	361,47
marché	-633,88	-260,35	313,51	362,72	218,00
société	-426,39	-407,14	85,50	84,34	663,69
tiers-monde	540,13	475,69	114,25	-367,54	-762,54
travail	-352,63	-232,42	-260,49	130,88	714,67
production	638,34	220,31	-80,95	-243,52	-534,18

L'analyse factorielle des correspondances.

On va travailler sur le tableau des écarts, dit « T_1 »

	80-83	84-87	88-91	92-95	96-2000
développement	404,27	337,78	57,03	-186,52	-612,55
gouvernement	548,62	38,90	-496,20	-42,77	-48,56
guerre	-718,47	-172,77	267,35	262,42	361,47
marché	-633,88	-260,35	313,51	362,72	218,00
société	-426,39	-407,14	85,50	84,34	663,69
tiers-monde	540,13	475,69	114,25	-367,54	-762,54
travail	-352,63	-232,42	-260,49	130,88	714,67
production	638,34	220,31	-80,95	-243,52	-534,18

On cherche quelles marges pourraient, par multiplication, donner un tableau aussi approchant que possible

13,32 * 21.37 = 297.57

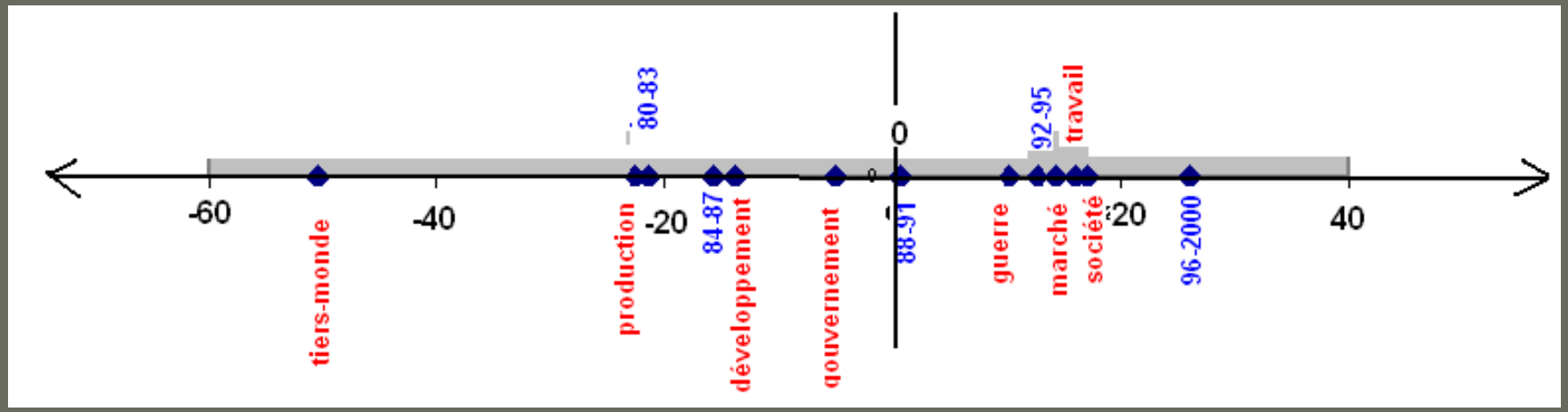
297,56693	217,98981	-10,57413	-177,32007	-364,67200	-13,91846
109,11482	79,93468	-3,87743	-65,02149	-133,72157	-5,10376
-221,14491	-162,00502	7,85845	131,78020	271,01586	10,34388
-307,77042	-225,46461	10,93672	183,40032	377,17650	14,39572
-363,99370	-266,65233	12,93463	216,90376	446,07884	17,02552
1079,48704	790,80417	-38,35990	-643,26608	-1322,92489	-50,49216
-341,71262	-250,32979	12,14286	203,62647	418,77310	15,98334
484,56141	354,97710	-17,21904	-288,75003	-593,83607	-22,66498
-21,37930	-15,66192	0,75972	12,73992	26,20060	

L’analyse factorielle des correspondances.

Une sorte de tableau d’indépendance « dérivé » …

	80-83	84-87	88-91	92-95	96-2000	VECTEUR C
développement	297,56693	217,98981	-10,57413	-177,32007	-364,67200	-13,91846
gouvernement	109,11482	79,93468	-3,87743	-65,02149	-133,72157	-5,10376
guerre	-221,14491	-162,00502	7,85845	131,78020	271,01586	10,34388
marché	-307,77042	-225,46461	10,93672	183,40032	377,17650	14,39572
société	-363,99370	-266,65233	12,93463	216,90376	446,07884	17,02552
tiers-monde	1079,48704	790,80417	-38,35990	-643,26608	-1322,92489	-50,49216
travail	-341,71262	-250,32979	12,14286	203,62647	418,77310	15,98334
production	484,56141	354,97710	-17,21904	-288,75003	-593,83607	-22,66498
VECTEUR L	-21,37930	-15,66192	0,75972	12,73992	26,20060	

On considère chaque valeur du vecteur C comme coordonnée du point ligne correspondant sur un axe bi-orienté, et chaque valeur du vecteur L comme coordonnée du point colonne correspondant :



L’analyse factorielle des correspondances.

L’AFC est algorithme itératif.

La passe n°2 consiste à confronter T1 au tableau « calculé »

	80-83	84-87	88-91	92-95	96-2000
développement	404,27	337,78	57,03	-186,52	-612,55
gouvernement	548,62	38,90	-496,20	-42,77	-48,56
guerre	-718,47	-172,77	267,35	262,42	361,47
marché	-633,88	-260,35	313,51	362,72	218,00
société	-426,39	-407,14	85,50	84,34	663,69
tiers-monde	540,13	475,69	114,25	-367,54	-762,54
travail	-352,63	-232,42	-260,49	130,88	714,67
production	638,34	220,31	-80,95	-243,52	-534,18

T1

Calculé d’après T1

	80-83	84-87	88-91	92-95	96-2000
développement	297,56693	217,98981	-10,57413	-177,32007	-364,67200
gouvernement	109,11482	79,93468	-3,87743	-65,02149	-133,72157
guerre	-221,14491	-162,00502	7,85845	131,78020	271,01586
marché	-307,77042	-225,46461	10,93672	183,40032	377,17650
société	-363,99370	-266,65233	12,93463	216,90376	446,07884
tiers-monde	1079,48704	790,80417	-38,35990	-643,26608	-1322,92489
travail	-341,71262	-250,32979	12,14286	203,62647	418,77310
production	484,56141	354,97710	-17,21904	-288,75003	-593,83607

On en déduit un tableau des écarts …. : ….

	80-83	84-87	88-91	92-95	96-2000
développement	106,70076	119,78882	67,59914	-9,20045	-247,87881
gouvernement	439,50409	-41,03380	-492,32069	22,25500	85,16641
guerre	-497,32019	-10,76468	259,49062	130,63773	90,45194
marché	-326,11157	-34,88064	302,57255	179,31633	-159,17519
société	-62,39136	-140,48642	72,56871	-132,56854	217,60641
tiers-monde	-539,35328	-315,11067	152,60860	275,72563	560,38938
travail	-10,91967	17,90519	-272,63419	-72,74796	295,89661
production	153,78268	-134,67182	-63,72690	45,22918	59,65350

L’analyse factorielle des correspondances.

… sur lequel on va appliquer la même recherche des marges permettant la meilleure approximation

	80-83	84-87	88-91	92-95	96-2000
développement	106,70076	119,78882	67,59914	-9,20045	-247,87881
gouvernement	439,50409	-41,03380	-492,32069	22,25500	85,16641
guerre	-497,32019	-10,76468	259,49062	130,63773	90,45194
marché	-326,11157	-34,88064	302,57255	179,31633	-159,17519
société	-62,39136	-140,48642	72,56871	-132,56854	217,60641
tiers-monde	-539,35328	-315,11067	152,60860	275,72563	560,38938
travail	-10,91967	17,90519	-272,63419	-72,74796	295,89661
production	153,78268	-134,67182	-63,72690	45,22918	59,65350

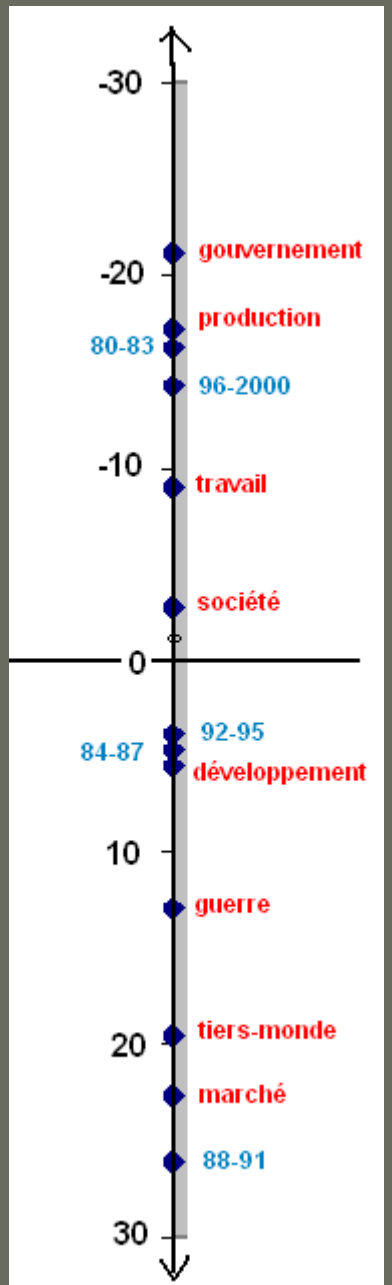
	80-83	84-87	88-91	92-95	96-2000	VECTEUR C
développement	-89,09875	26,01999	143,50418	21,15767	-78,05996	5,5216
gouvernement	341,03313	-99,59375	-549,27460	-80,98279	298,78124	-21,1344
guerre	-208,92120	61,01239	336,49255	49,61108	-183,03716	12,9472
marché	-363,69187	106,21090	585,76920	86,36341	-318,63270	22,5386
société	43,39723	-12,67353	-69,89643	-10,30524	38,02059	-2,6894
tiers-monde	-313,38180	91,51858	504,73882	74,41662	-274,55573	19,4208
travail	276,51335	-80,75169	-445,35779	-65,66172	242,25506	-17,136
production	144,78546	-42,28248	-233,19428	-34,38121	126,84744	-8,9726
VECTEUR L	-16,13640	4,71240	25,98960	3,83180	-14,13720	

L’analyse factorielle des correspondances.

Une sorte de 2ème tableau d’indépendance « dérivé » …

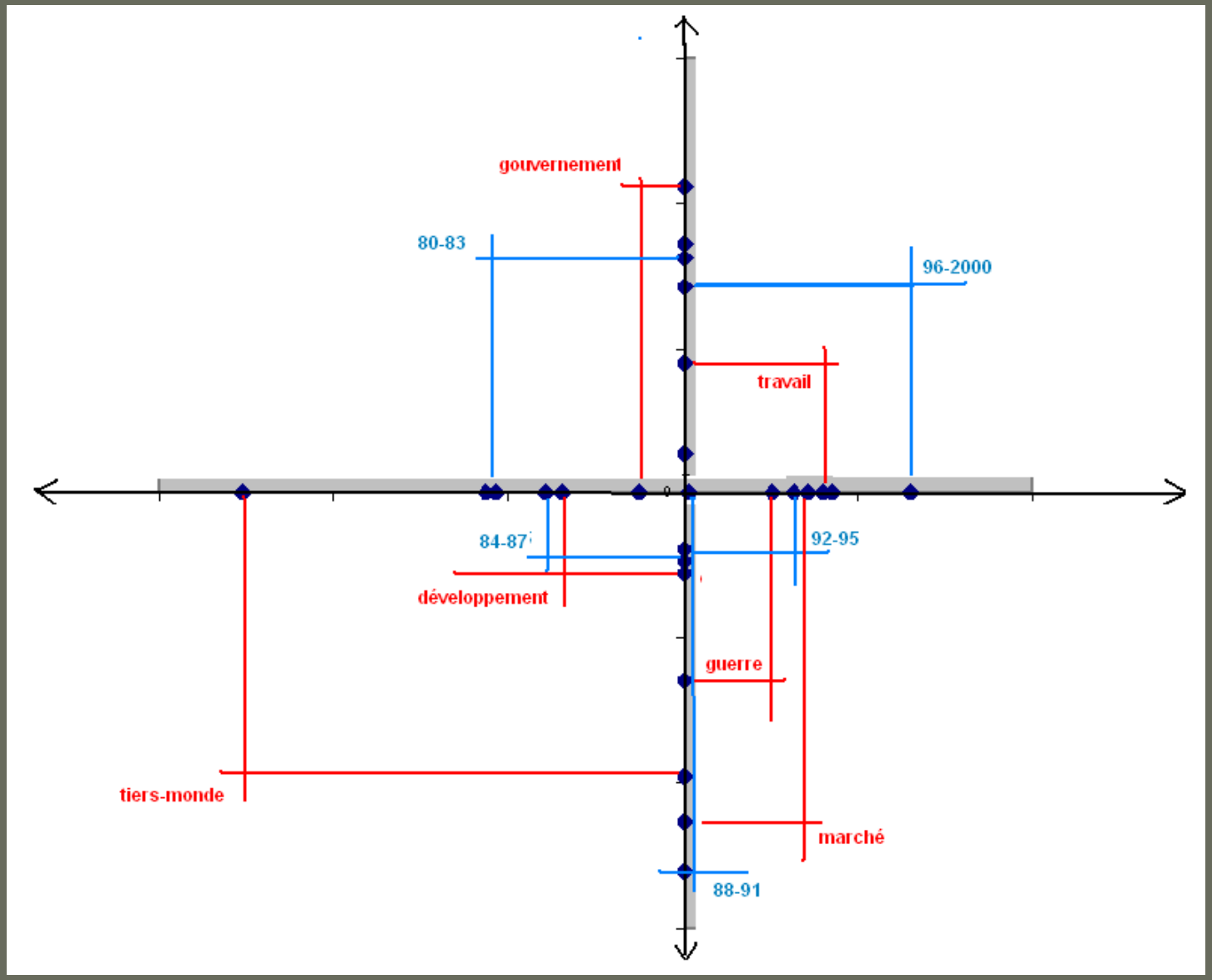
	80-83	84-87	88-91	92-95	96-2000	VECTEUR C
développement	-89,09875	26,01999	143,50418	21,15767	-78,05996	5,5216
gouvernement	341,03313	-99,59375	-549,27460	-80,98279	298,78124	-21,1344
guerre	-208,92120	61,01239	336,49255	49,61108	-183,03716	12,9472
marché	-363,69187	106,21090	585,76920	86,36341	-318,63270	22,5386
société	43,39723	-12,67353	-69,89643	-10,30524	38,02059	-2,6894
tiers-monde	-313,38180	91,51858	504,73882	74,41662	-274,55573	19,4208
travail	276,51335	-80,75169	-445,35779	-65,66172	242,25506	-17,136
production	144,78546	-42,28248	-233,19428	-34,38121	126,84744	-8,9726
VECTEUR L	-16,13640	4,71240	25,98960	3,83180	-14,13720	

On effectue la même projection qu’en 1, sur un 2ème axe, orthogonal au 1er :



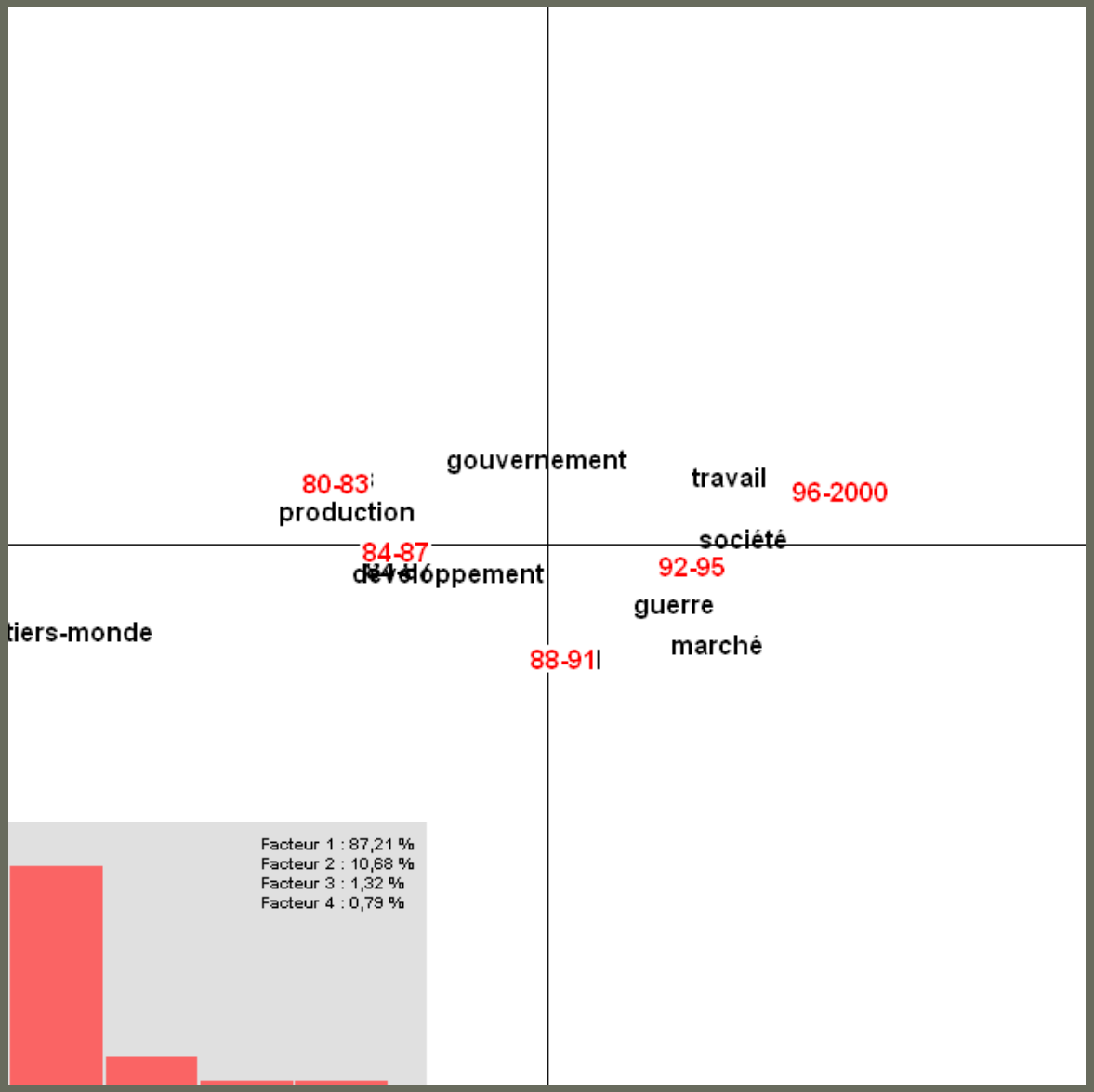
L'analyse factorielle des correspondances.

Un système de coordonnées sur un plan de 2 « facteurs »



L’analyse factorielle des correspondances.

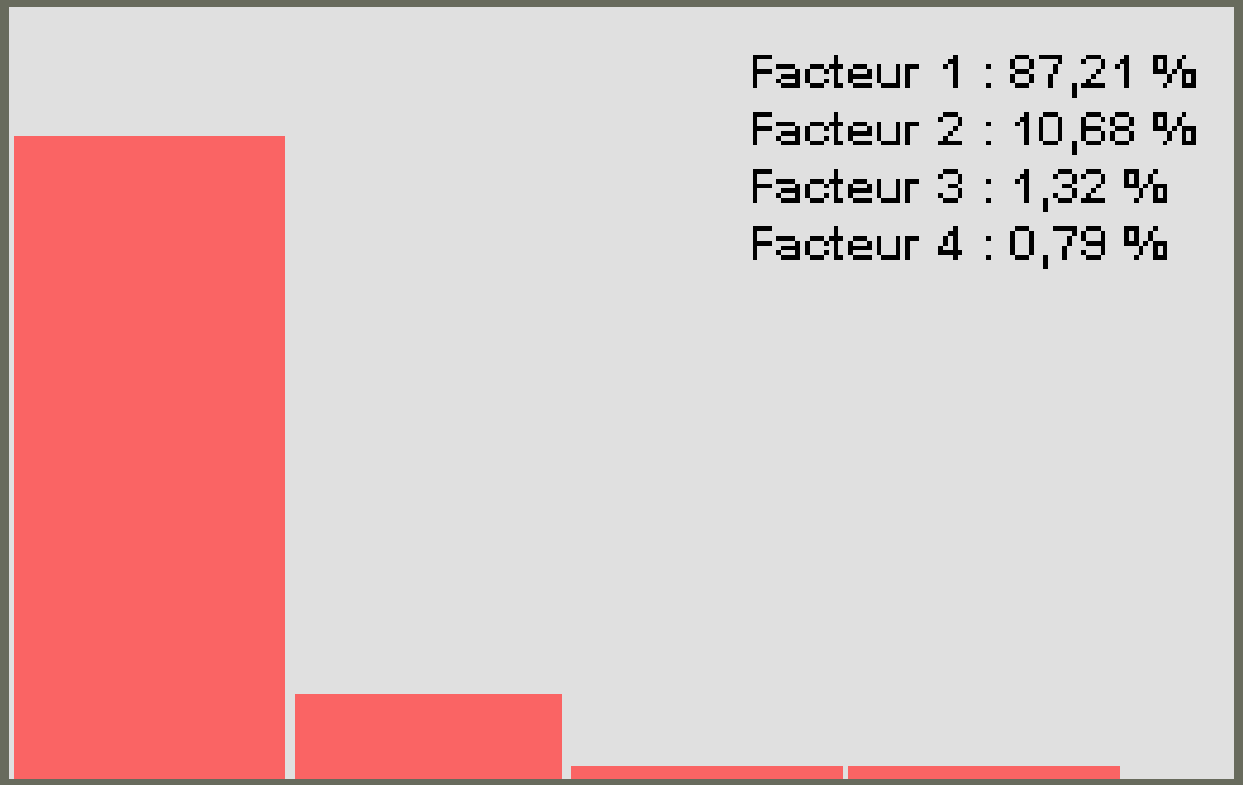
Un système de coordonnées sur un plan de 2 « facteurs »



L’analyse factorielle des correspondances.

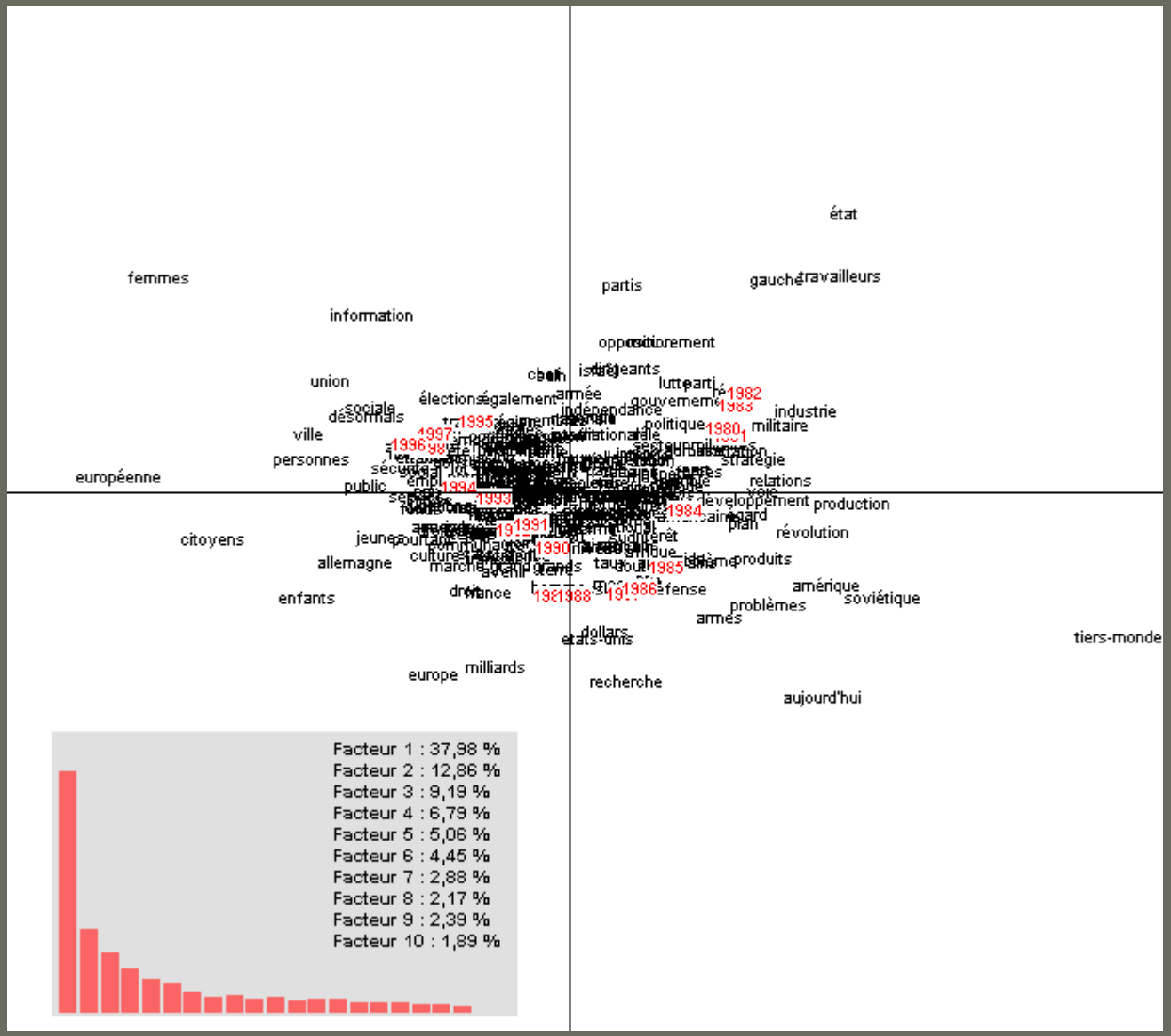
Autant d’itérations (de « facteurs ») que le tableau comporte de colonnes (ou de lignes, s’il y a moins de lignes que de colonnes) –
en comptant le Facteur 0, qui correspond au tableau T0 de départ.

Chaque itération extrait une part décroissante de l’information totale du tableau d’origine (le total des « pourcentages d’inertie » est 100%.



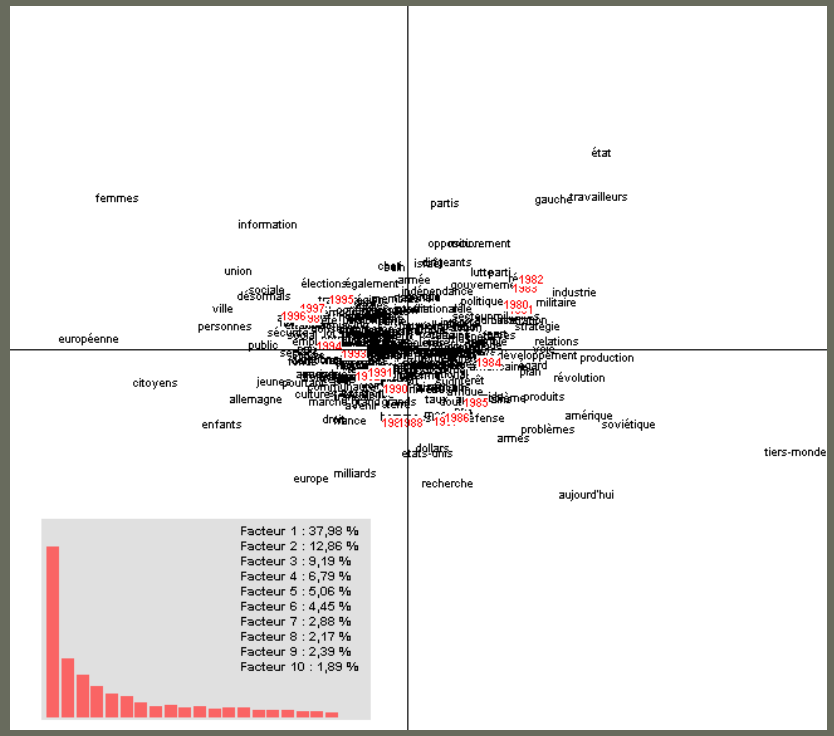
L’analyse factorielle des correspondances.

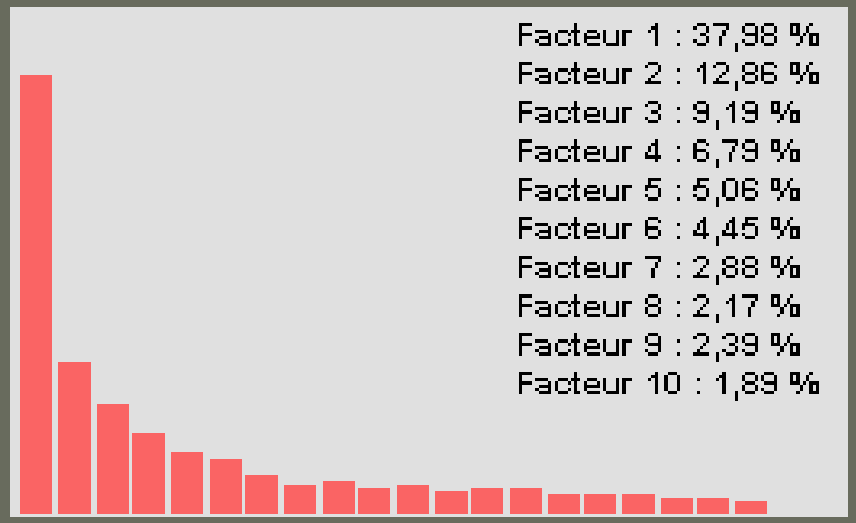
Un système de coordonnées sur un plan de 2 « facteurs »



L’analyse factorielle des correspondances.

Un système de coordonnées sur un plan de 2 « facteurs »
Mais ici, avec 244 lignes et 21 colonnes, il y a 20 facteurs, dont l’histogramme indique la « décroissance » en % d’information.
Le 3ème facteur mériterait d’être visualisé.
Nous en reparlerons plus loin.

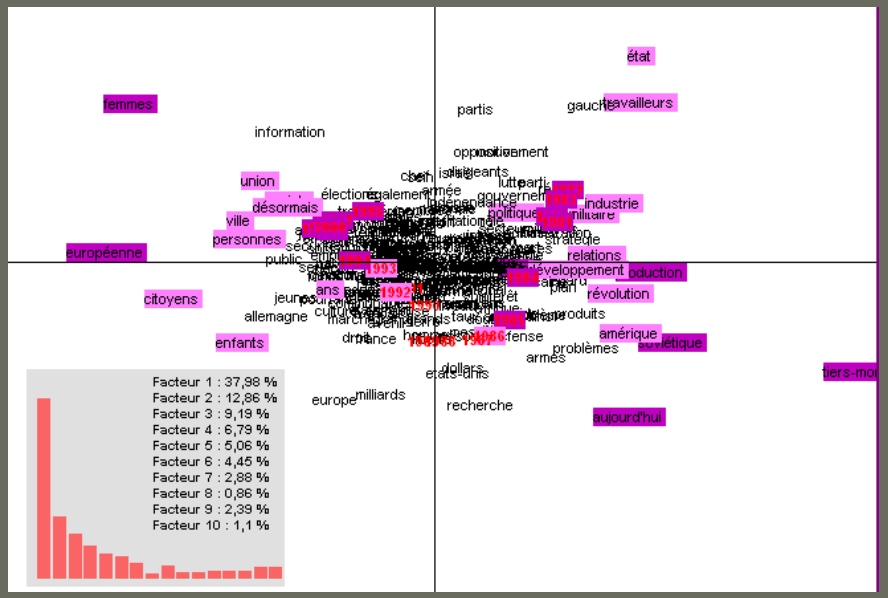


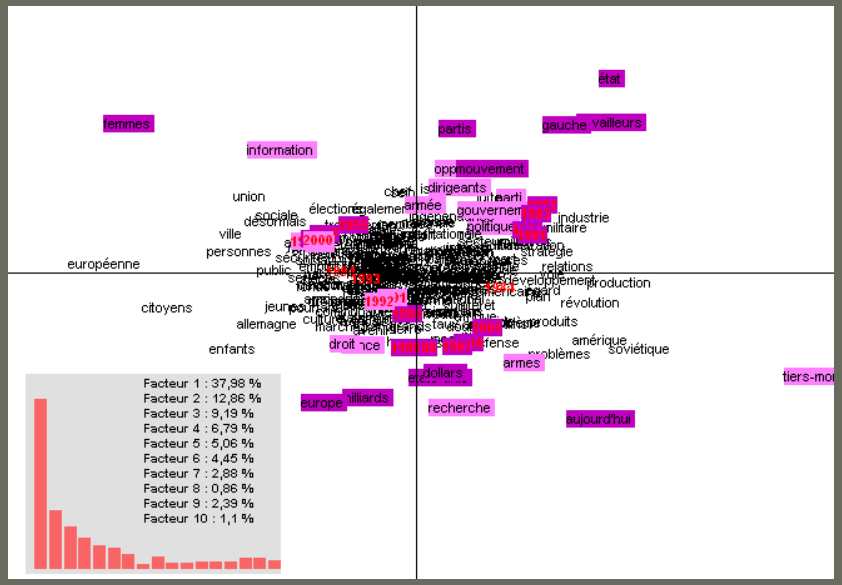


L’analyse factorielle des correspondances.

D’autres indicateurs méritent d’être accessibles.

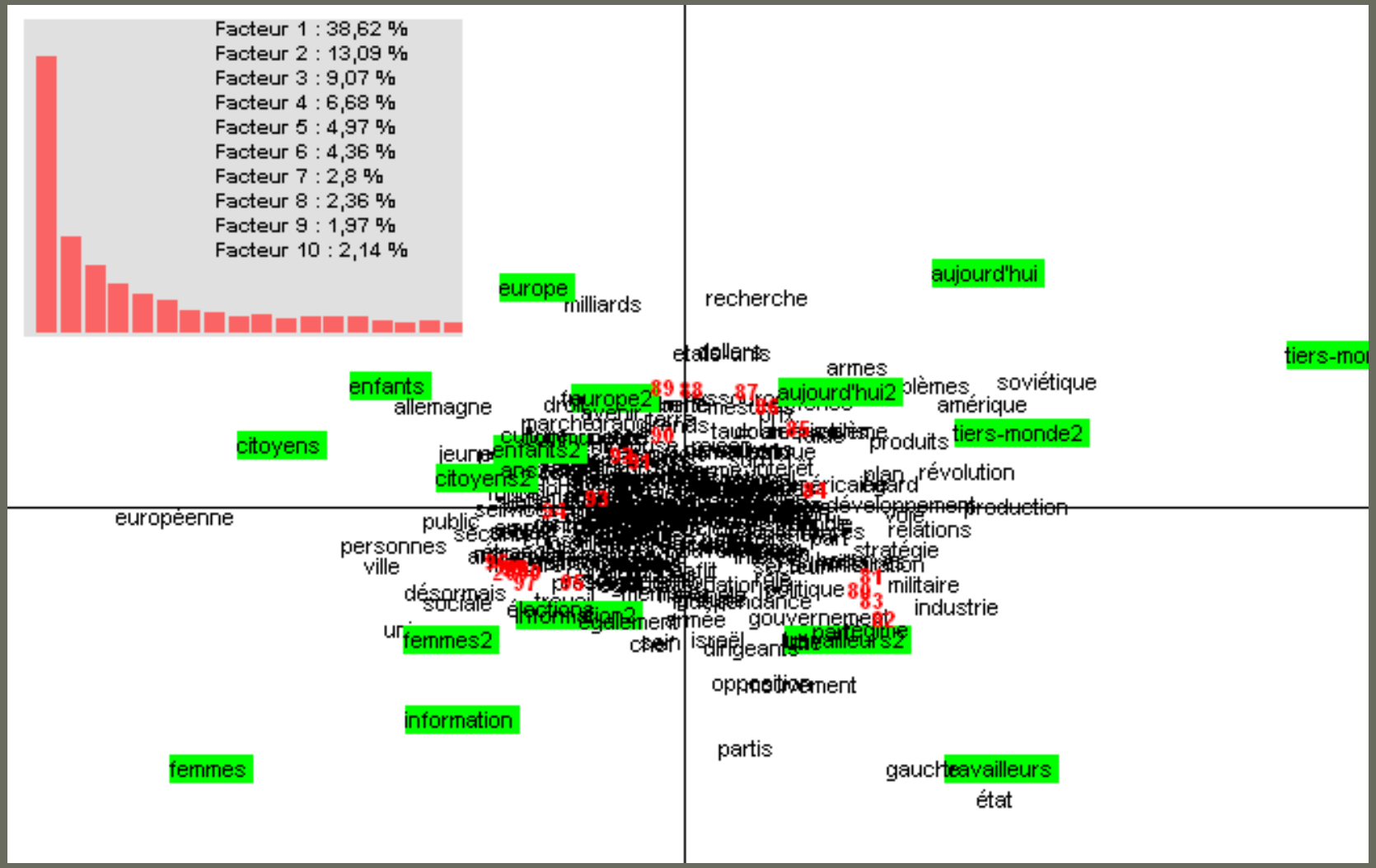
Exemple : la contribution aux « axes »





L’analyse factorielle des correspondances.

IMPORTANCE DE LA POSITION <u>ANGULAIRE</u> SUR LE GRAPHE (profils +/- accentués)

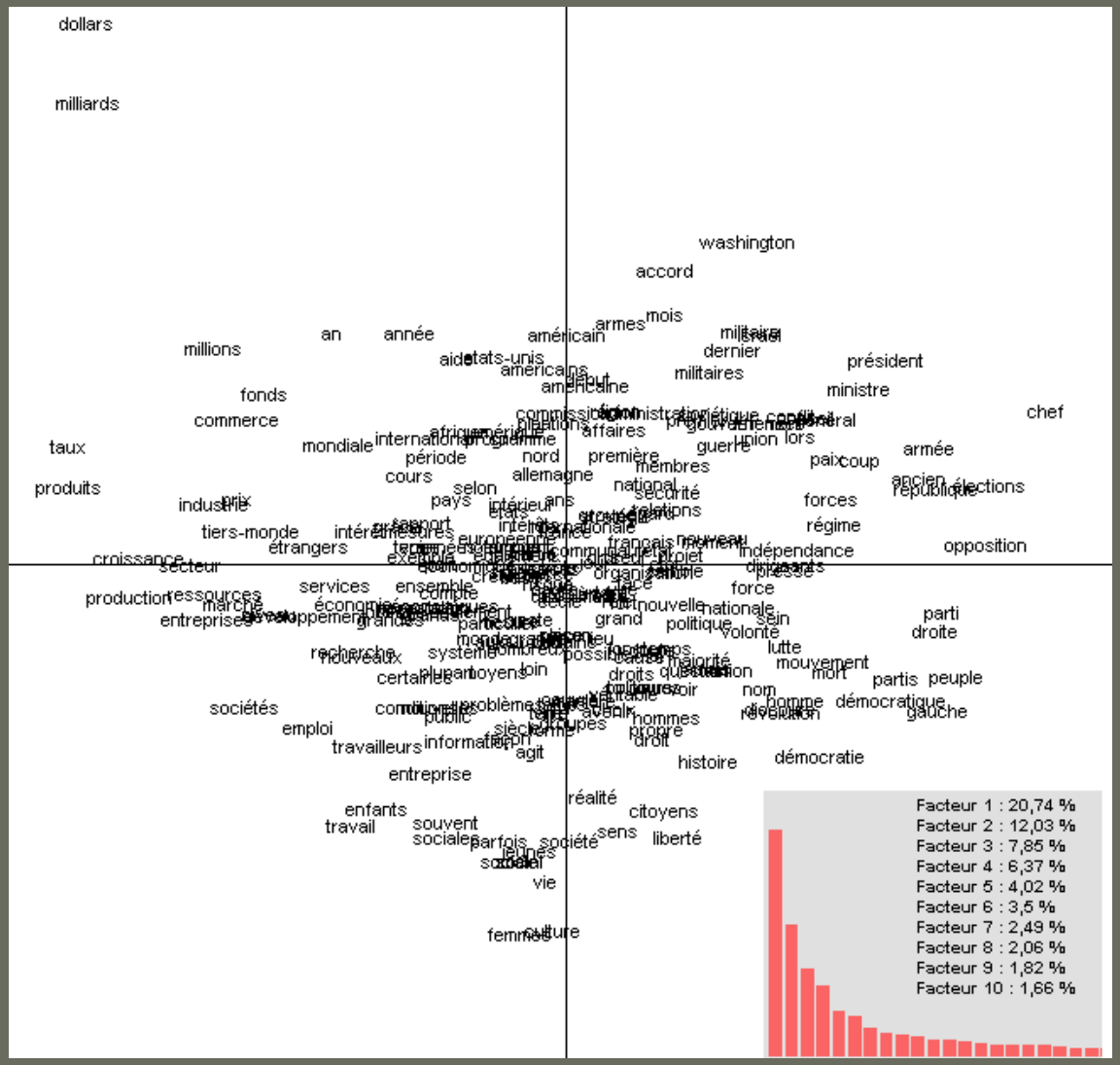


L’analyse factorielle des correspondances.

Tout tableau de données comportant

- **des lignes décrivant des classes d’occurrences du corpus**
- **des colonnes décrivant des variables recouvrant une partie du corpus**
- **des cellules « intersections » rendant compte d’effectifs**

Cooccurrence généralisée
Q-occurrence (Massonie 1993)
Isotropie (Viprey 1998)



Un exemple de mise en œuvre de ® MacSpin

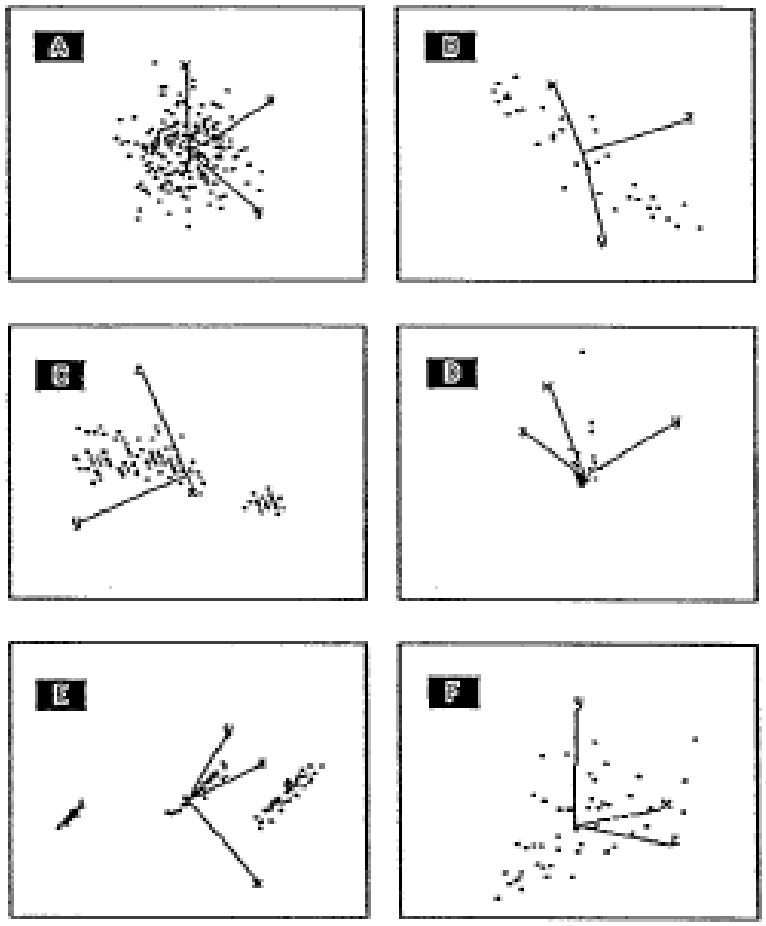
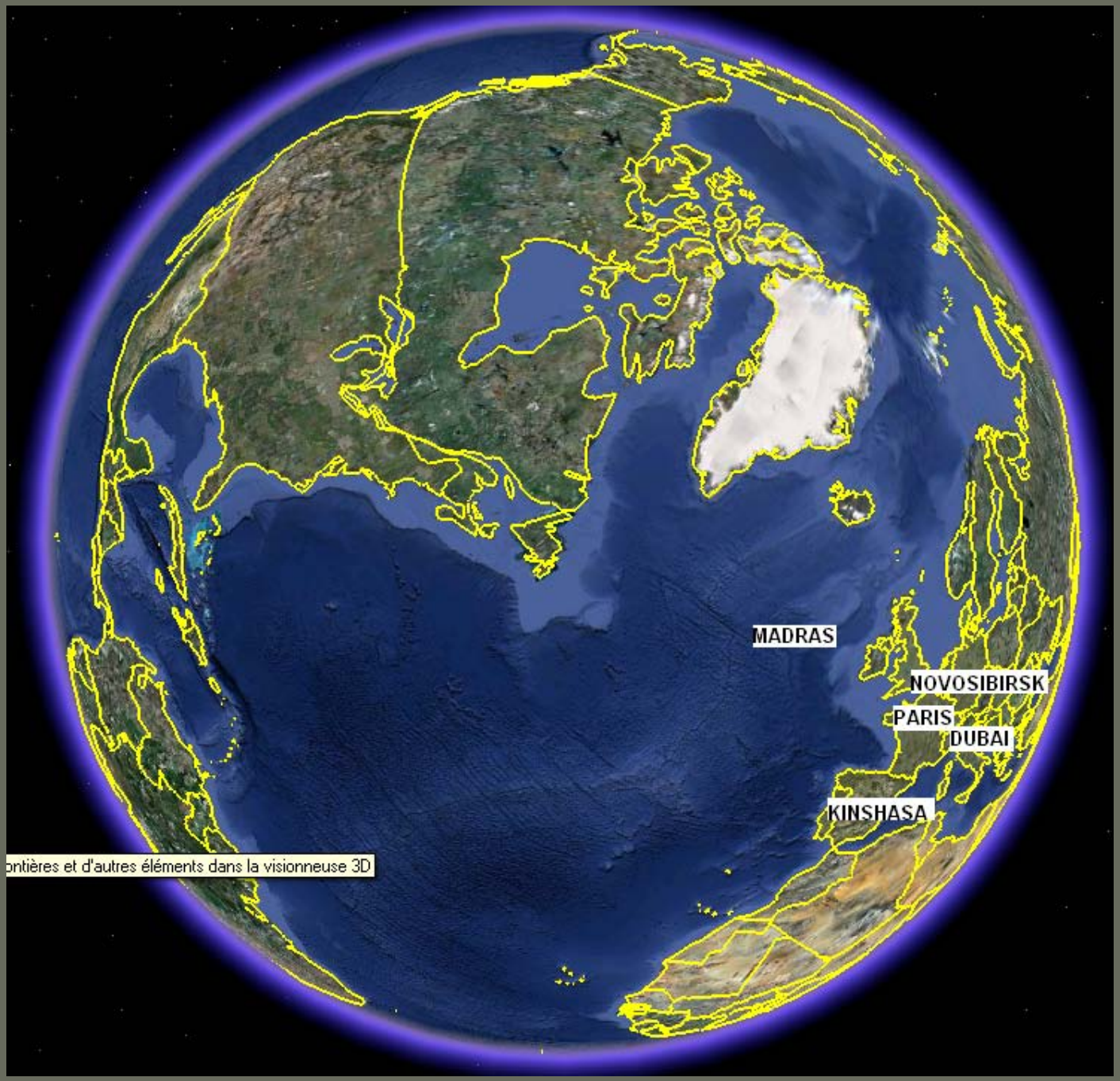
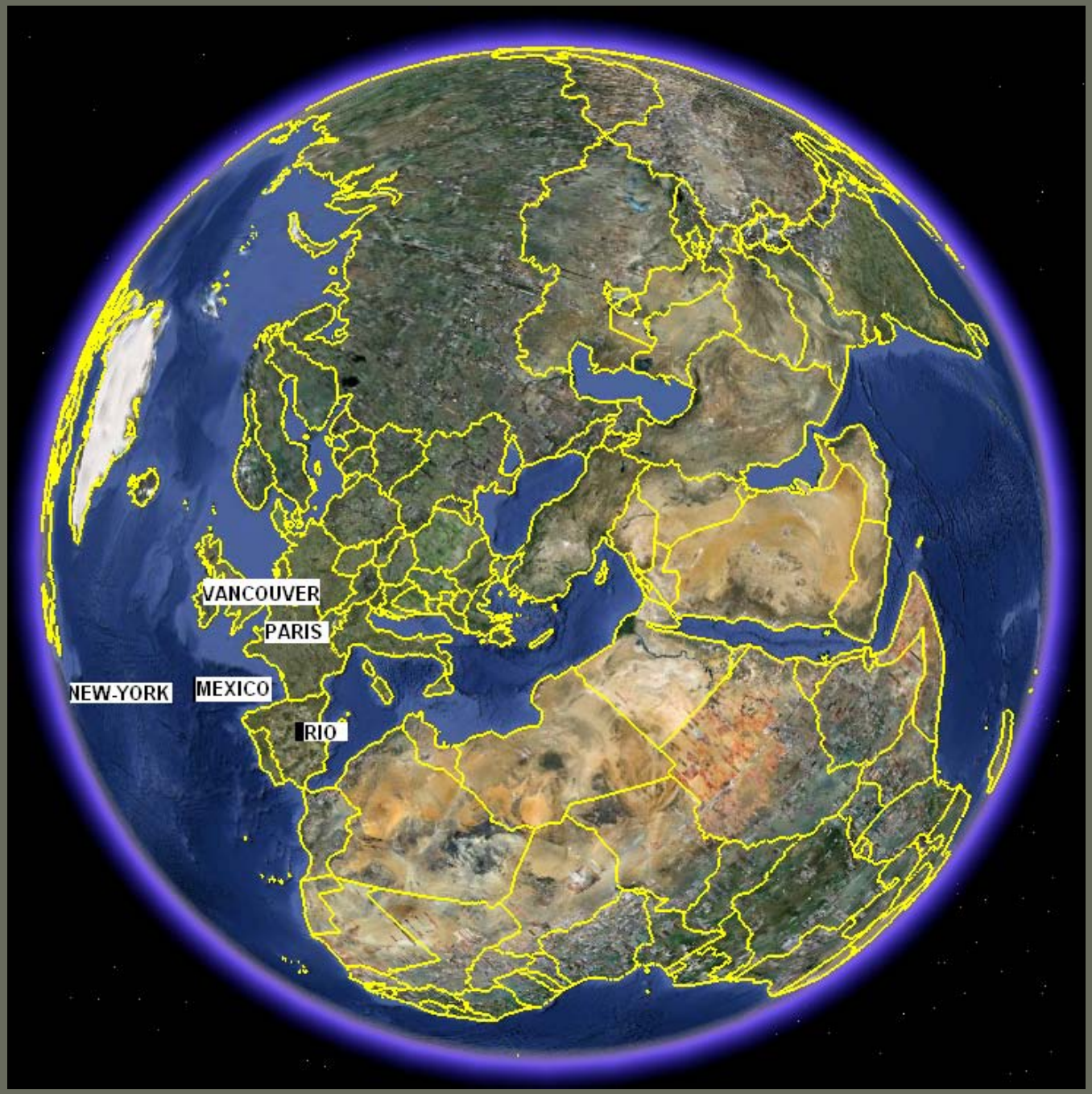


Figure 1: Exemples de formes "significatives" de nuages de points.

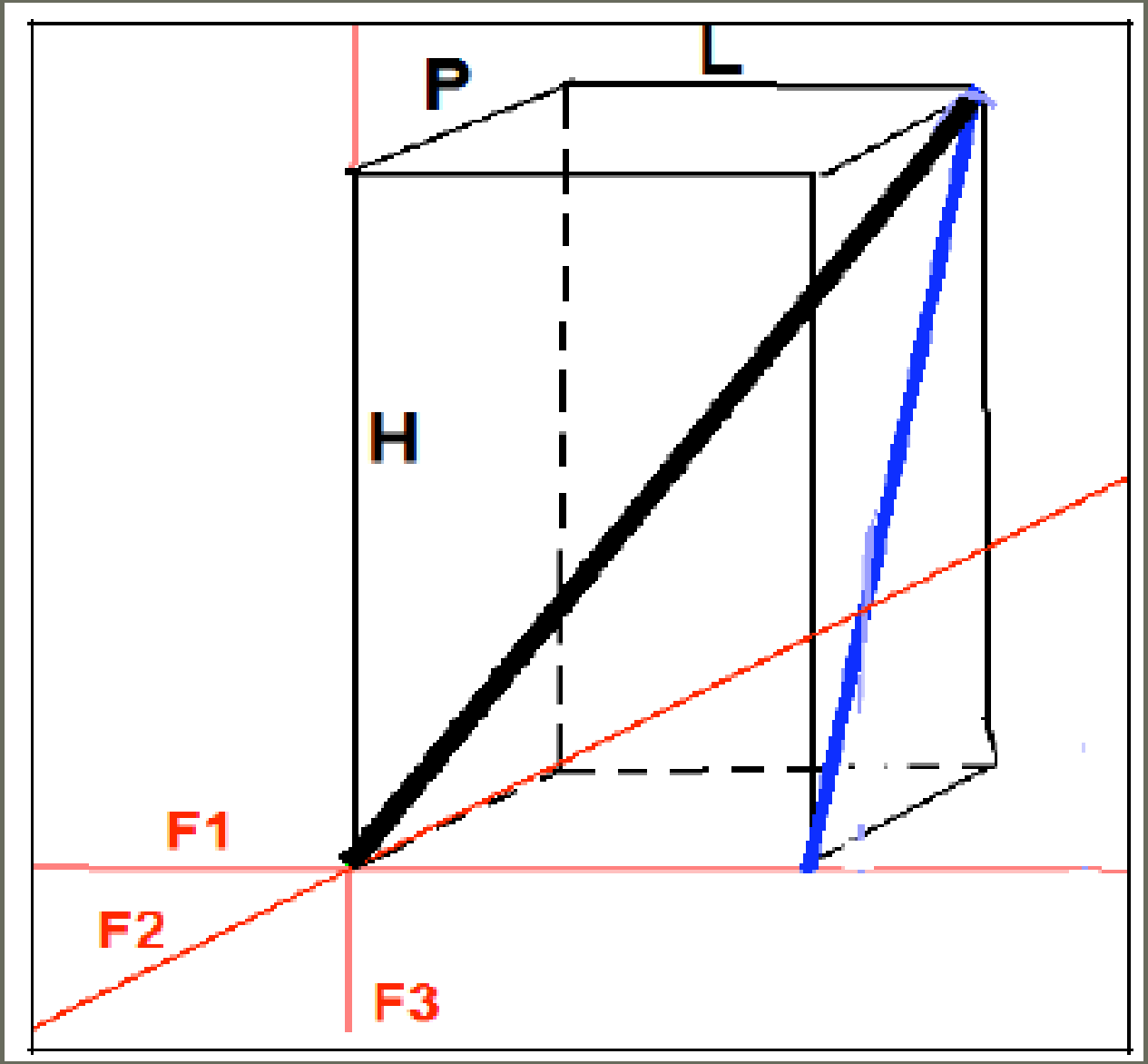
« L'analyse exploratoire des données : une approche interactive nouvelle de l'information statistique »
Philippe Waniez *Séminfor V*. Montpellier, Ed. de l'ORSTOM, Col. Colloques et séminaires, 13p

PARIS
MELBOURNE
WELLINGTON
PAPEETE
NOUMEA

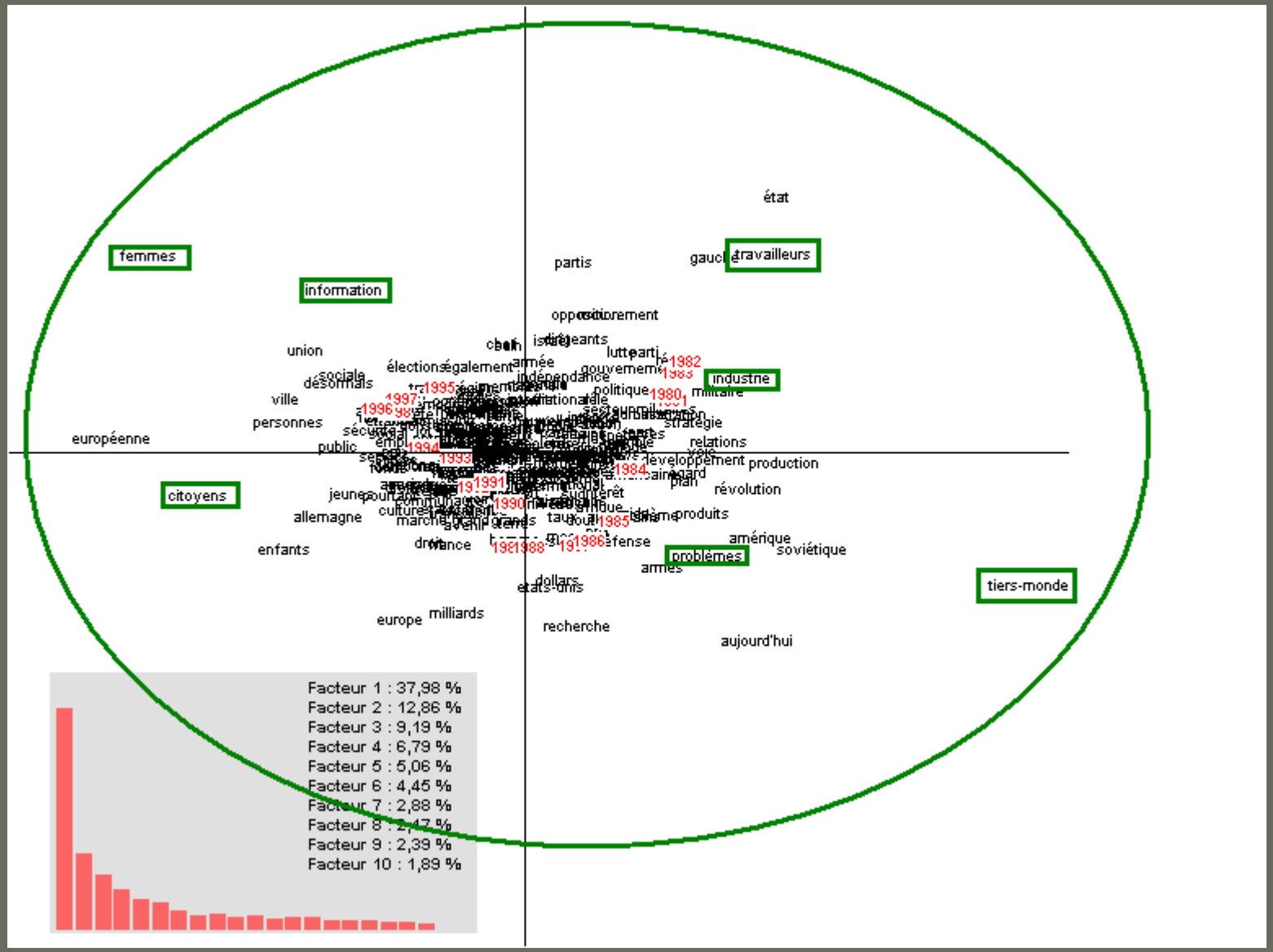

MADRAS
NOVOSIBIRSK
PARIS
DUBAI
KINSHASA
ontières et d'autres éléments dans la visionneuse 3D


VANCOUVER
PARIS
NEW-YORK
MEXICO
RIO

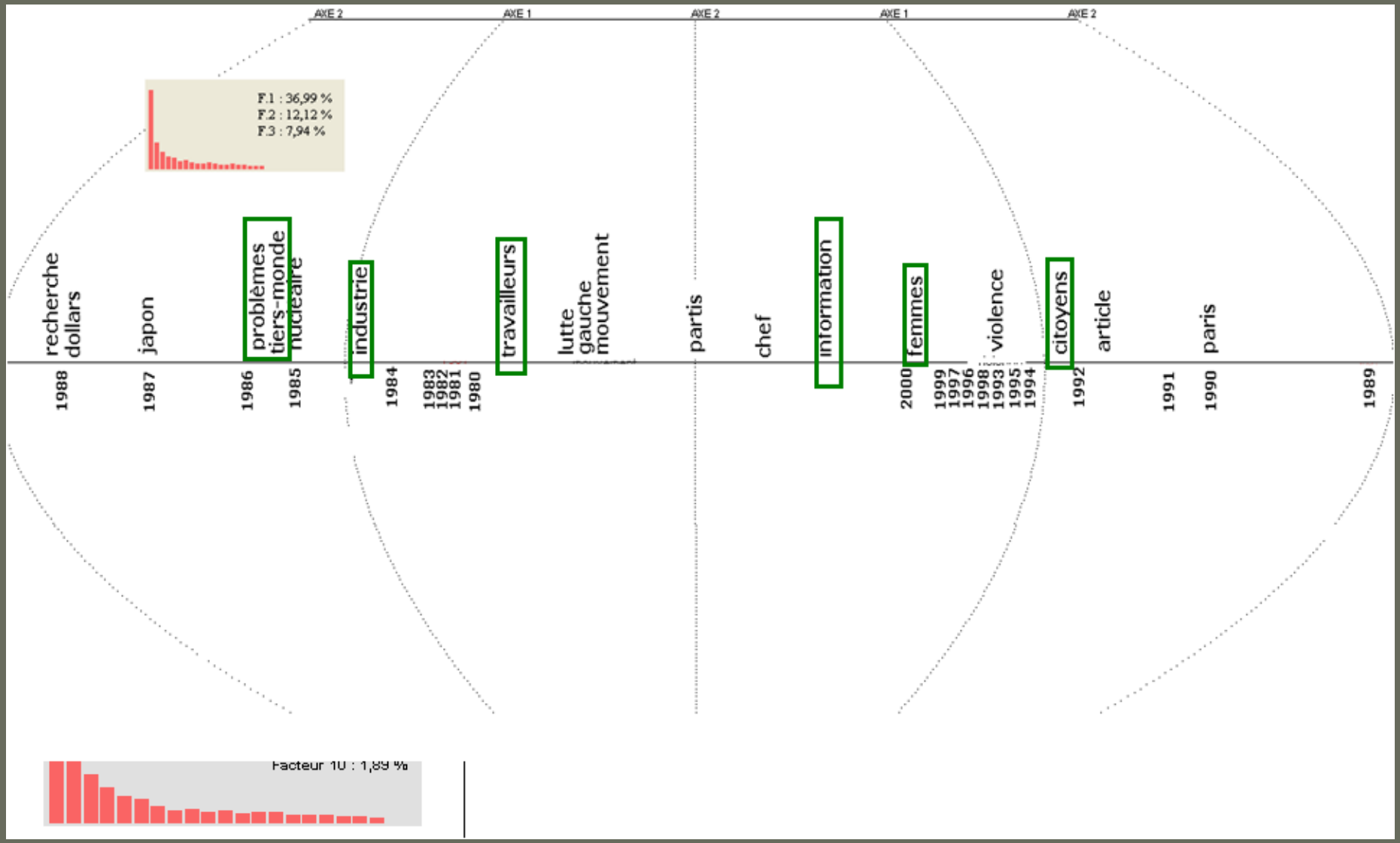
Parallélipipède calé dans un « huitième » de la boule



Plan des 2 premiers facteurs : un « horizon équatorial »



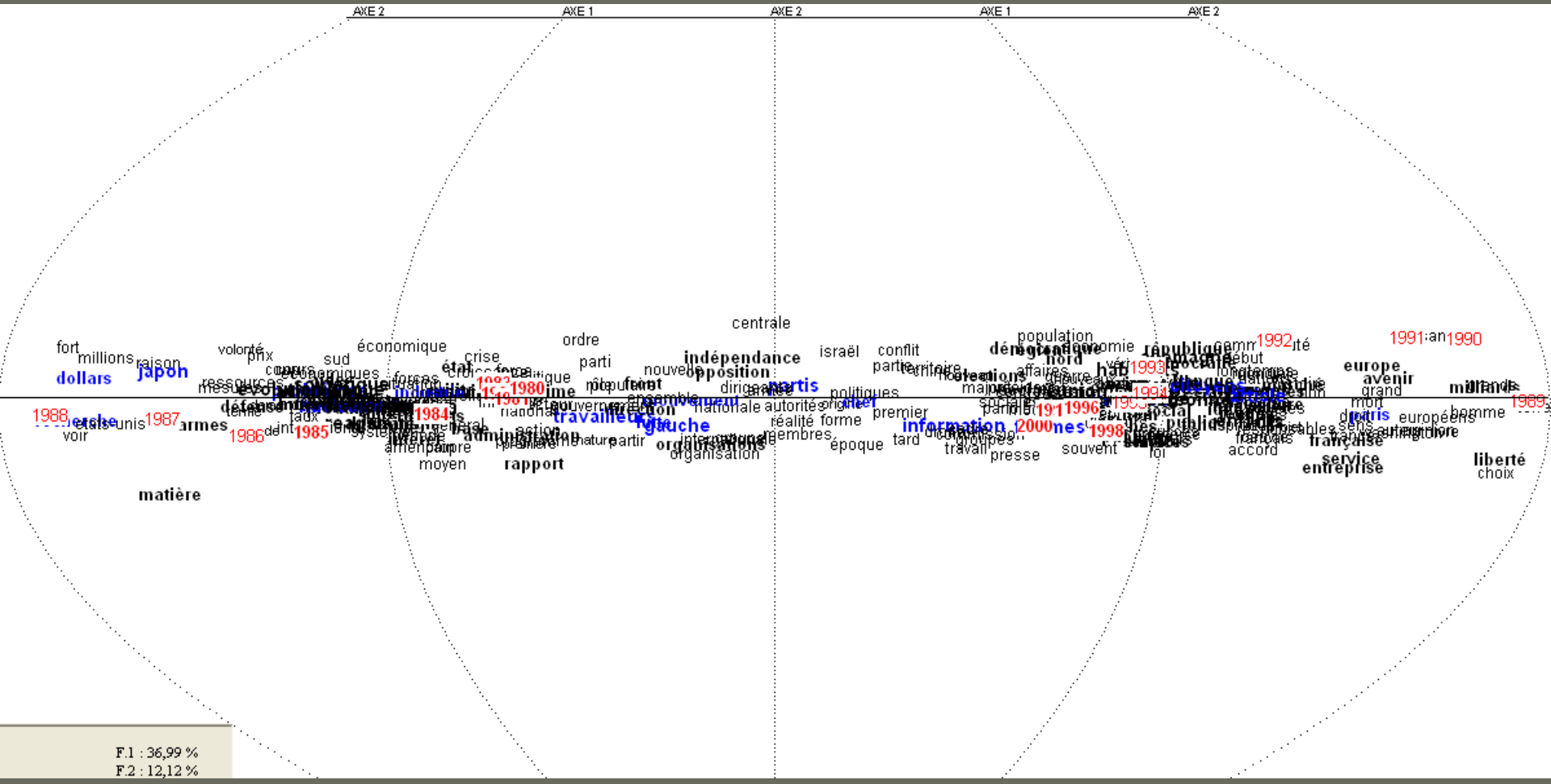
Les mêmes, sur l' « équateur », dans l'hypothèse d'un 3ème facteur entièrement neutre (tous à zéro).



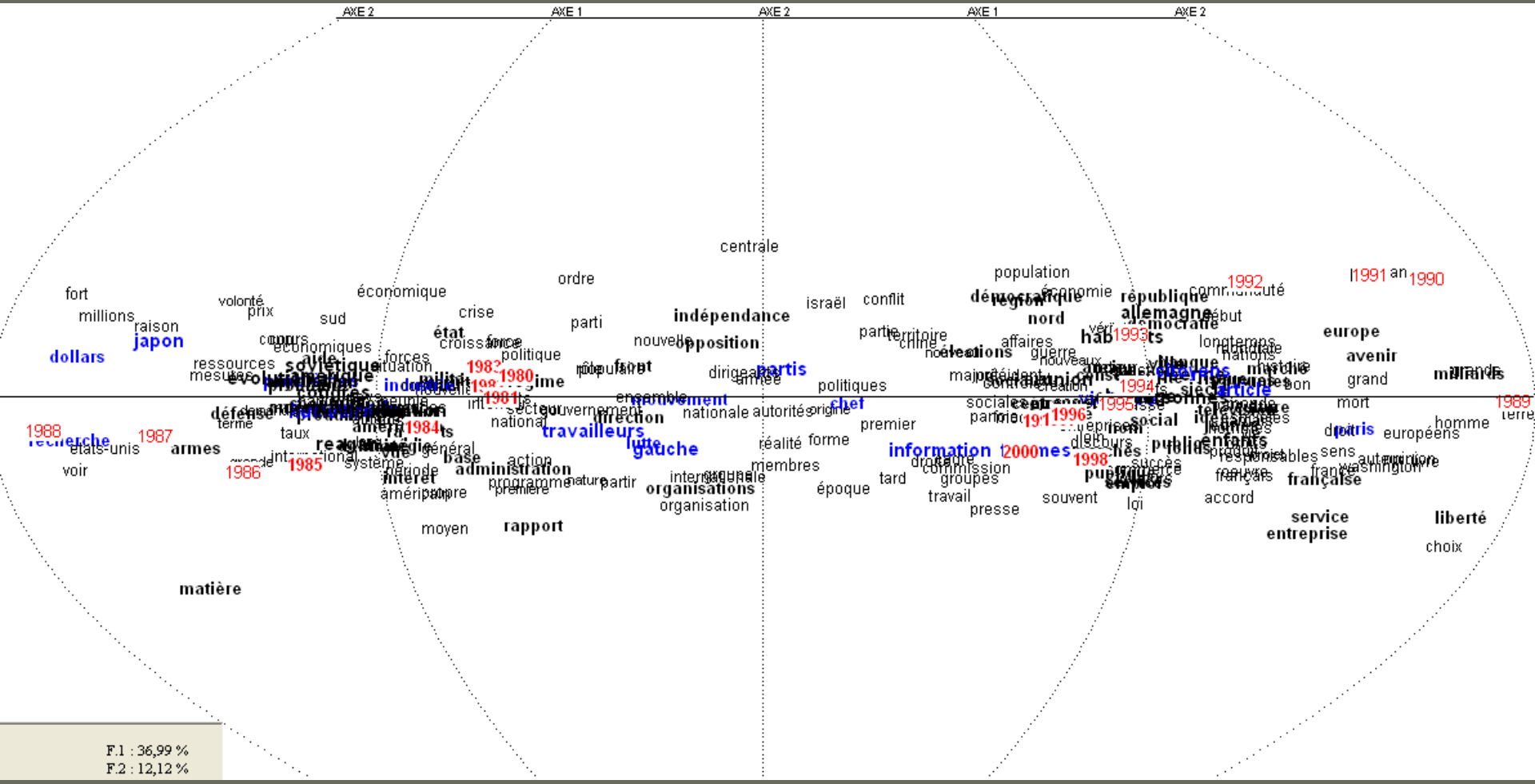
Coordonnées sur les 3 premiers facteurs

gauche	0,2072	-0,2111	-0,1782
femmes	-0,4087	-0,2129	-0,1731
travail	-0,1022	-0,0697	-0,1281
travailleurs	0,2703	-0,2136	-0,1203
internationale	0,0314	-0,0577	-0,1165
etats-unis	0,028	0,1467	-0,1113
afrique	0,0823	0,0605	-0,102
information	-0,1957	-0,1758	-0,101
sociétés	0,0209	-0,0095	-0,097
droits	-0,1302	0,0384	-0,0849
droit	-0,1036	0,0988	-0,0826
production	0,2824	0,0122	-0,0795
international	0,0412	0,0349	-0,079
loi	-0,1092	-0,0239	-0,0781
entreprise	-0,0541	0,0639	-0,0777
armes	0,1492	0,1254	0,0671
fort	0,0027	0,043	0,0699
passé	-0,0501	-0,0089	0,0723
milliards	-0,0731	0,1742	0,0758
communauté	-0,0924	0,0533	0,0764
washington	-0,0518	0,0244	0,0783
ville	-0,2595	-0,0583	0,0813
population	-0,0553	-0,0211	0,0822
conflit	0,0037	-0,058	0,0849
soviétique	0,3118	0,1053	0,0969
guerre	-0,071	-0,0067	0,1033
israël	0,0292	-0,1205	0,106
démocratie	-0,1232	0,0163	0,1128
europe	-0,1348	0,1819	0,1186
allemagne	-0,2129	0,0712	0,1862

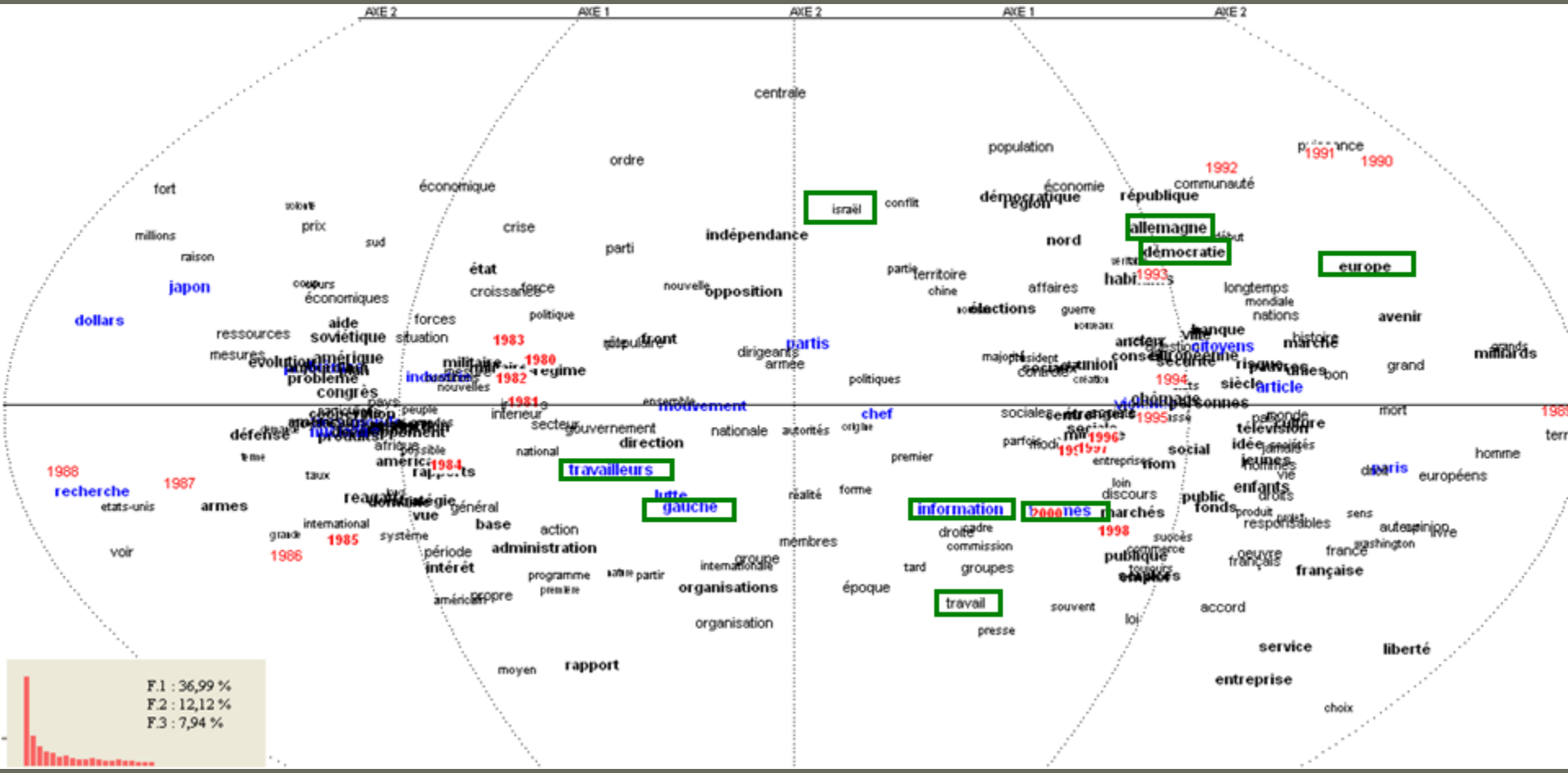
Visualisation des 3 premiers facteurs



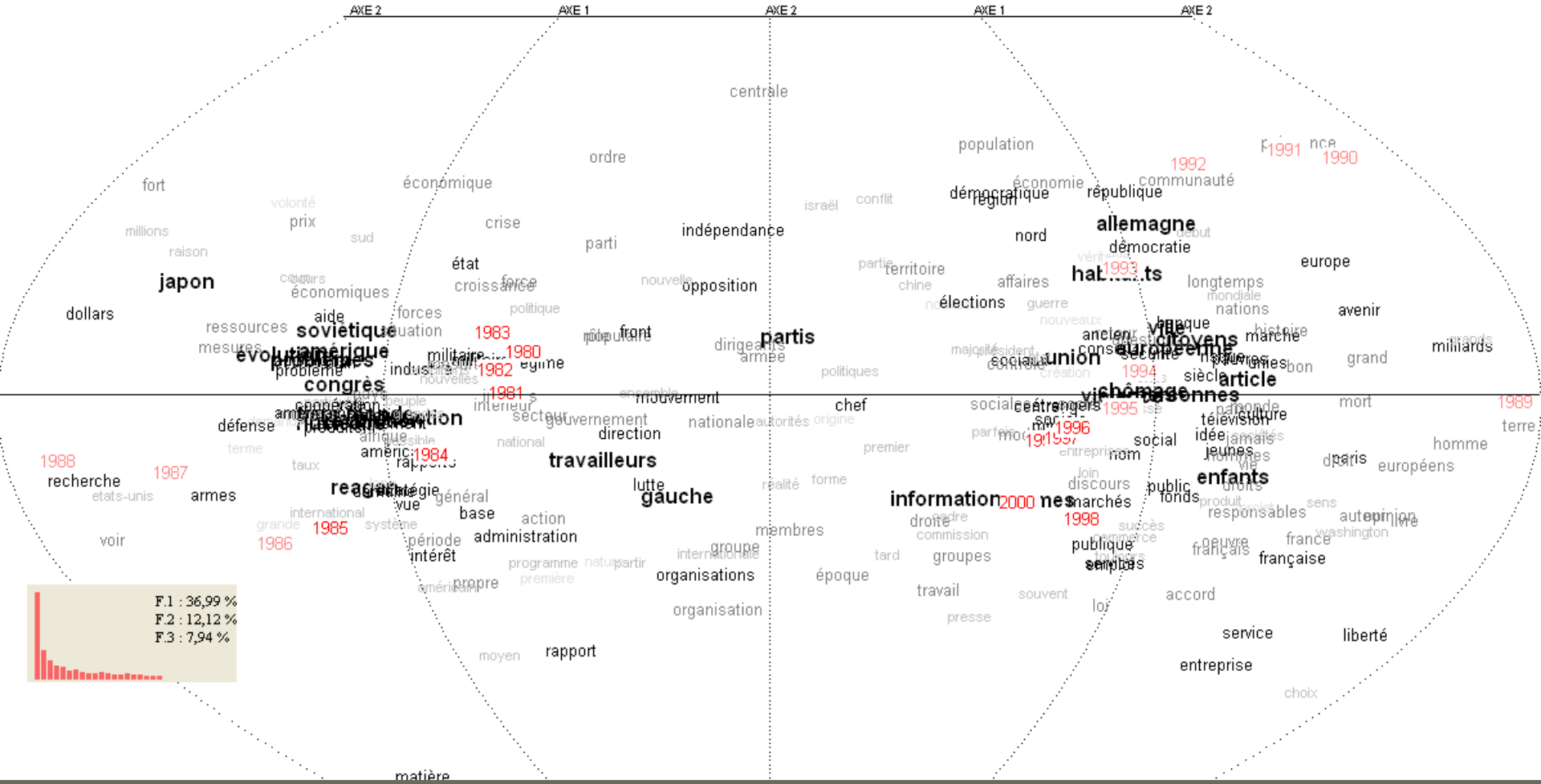
Visualisation des 3 premiers facteurs



Visualisation des 3 premiers facteurs

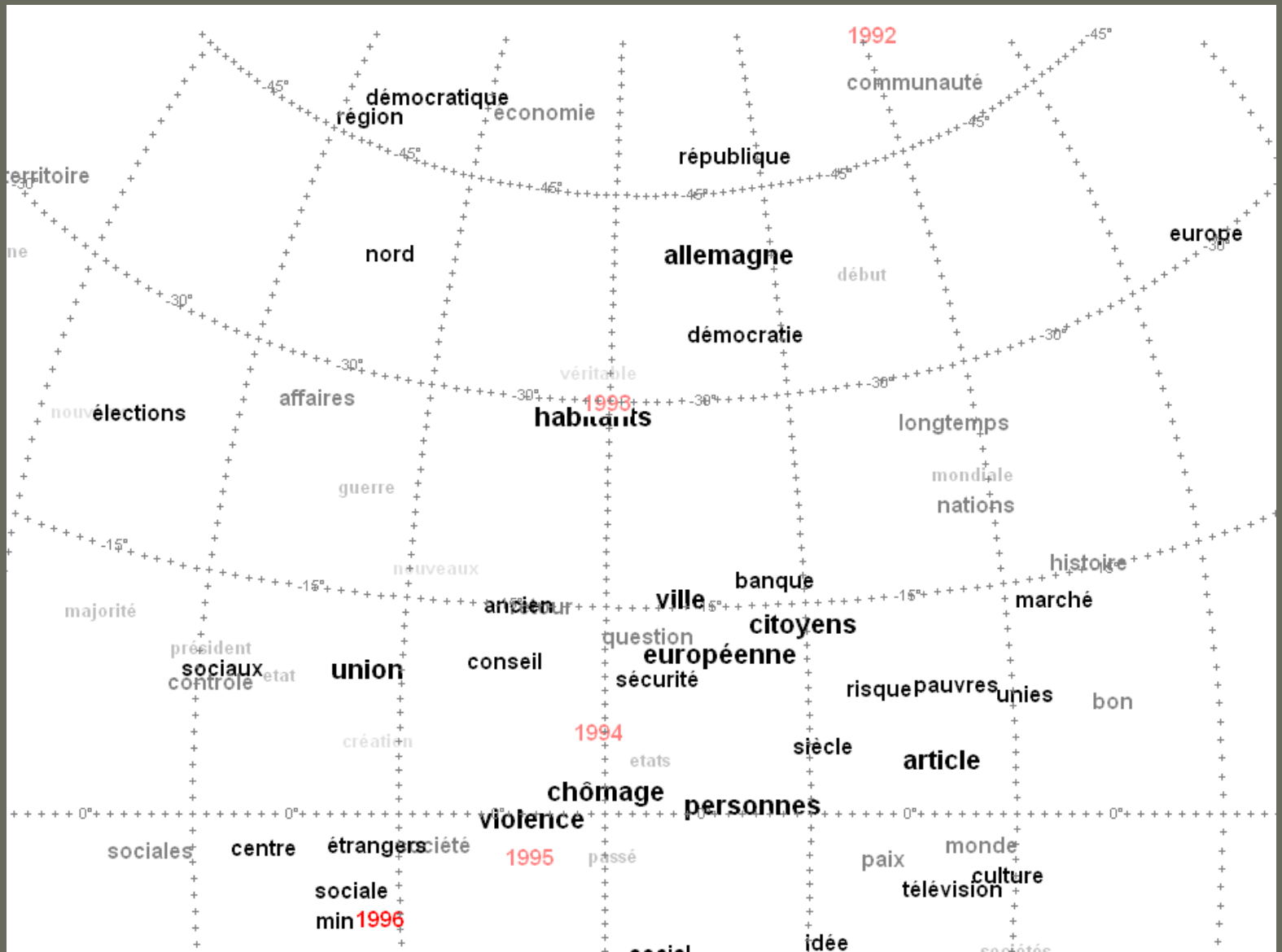


Visualisation des 3 premiers facteurs



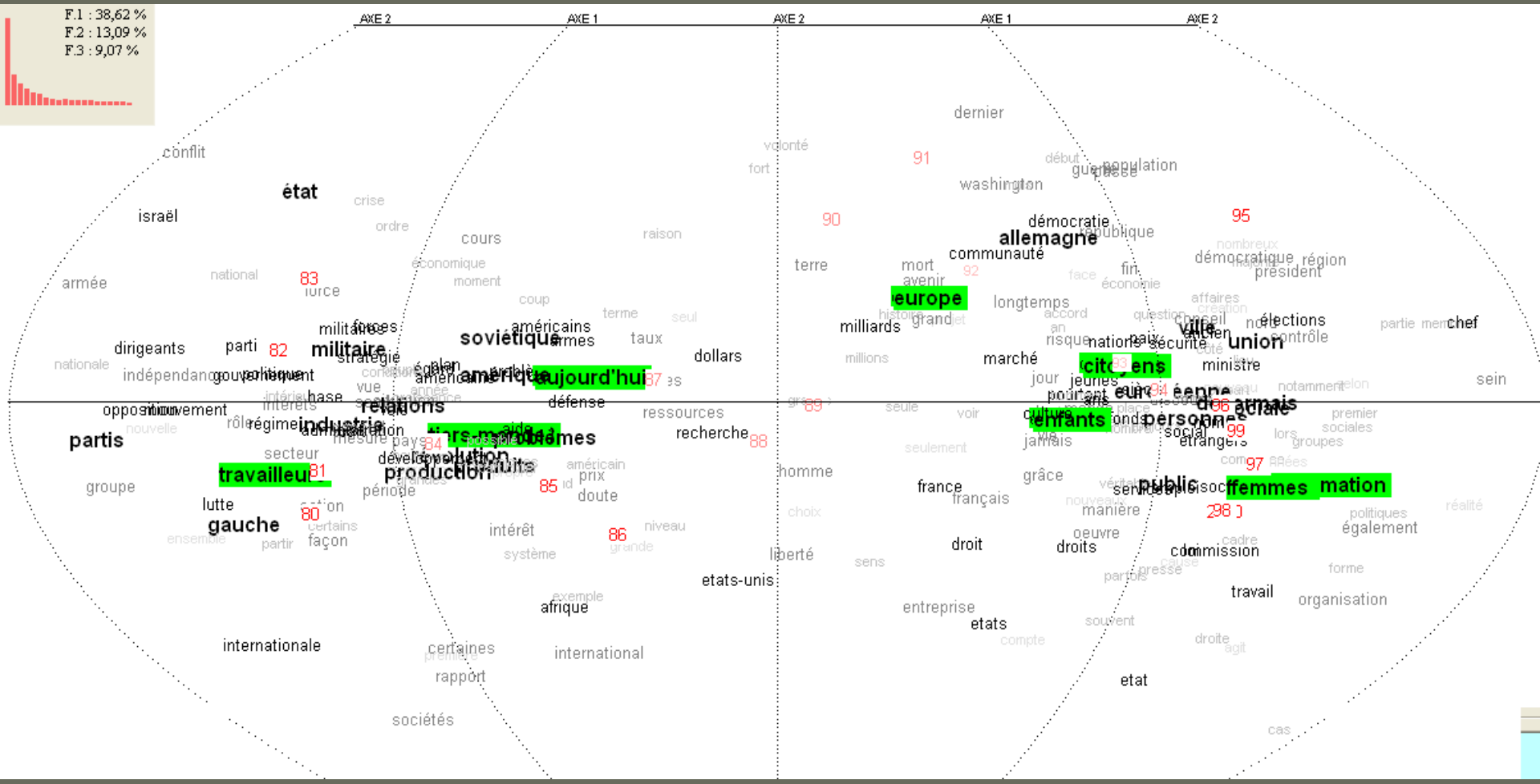
Visualisation des 3 premiers facteurs

Zoom « régional »



EQUIVALENCE DES POSITIONS ANGULAIRES :

Les points sont superposés (leur « gras » seul est différent)



AFC des cooccurrences

Exemple : 12 formes fréquentes dans *Mondiplo*

culture	5194
dollars	9691
élections	4402
fonds	3991
liberté	4385
libertés	1176
parti	12814
production	8199
ressources	3557
social	5197
société	10837
sociétés	4714

Balayage de l'ensemble du corpus à la recherche des cooccurrences entre ces 12 formes, dans les limites de 15 mots à gauche et à droite, et dans les limites de la phrase (ponctuation forte)

	culture	dollars	élections	fonds	liberté	libertés	parti	partis	production	ressources	social	société	sociétés
culture	358	7	1	15	32	11	18	15	42	13	39	129	32
dollars	7	1598	4	181	4	1	15	11	91	46	12	57	57
élections	1	4	84	6	18	12	357	132	1	5	6	12	1
fonds	15	181	6	154	4	0	21	10	26	63	45	24	32
liberté	32	4	18	4	376	38	55	18	4	11	21	55	18
libertés	11	1	12	0	38	44	11	5	0	0	12	18	2
parti	18	15	357	21	55	11	1566	242	18	17	114	141	5
partis	15	11	132	10	18	5	242	146	6	4	26	54	4
production	42	91	1	26	4	0	18	6	306	62	49	87	77
ressources	13	46	5	63	11	0	17	4	62	60	19	37	26
social	39	12	6	45	21	12	114	26	49	19	82	90	40
société	129	57	12	24	55	18	141	54	87	37	90	426	56
sociétés	32	57	1	32	18	2	5	4	77	26	40	56	158

AFC des cooccurrences

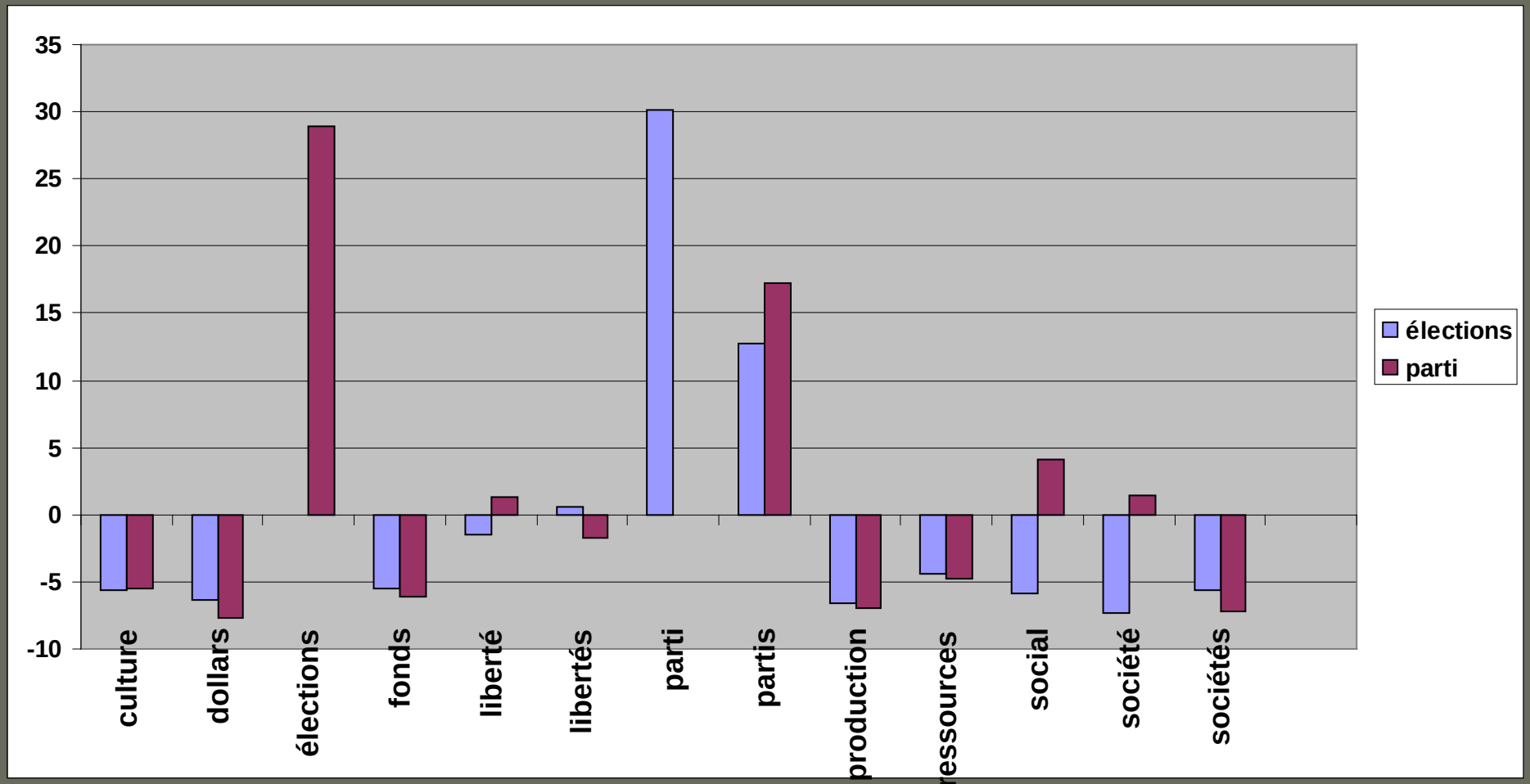
Ecarts à l’indépendance (tableau T1)

La diagonale est neutralisée

	culture	dollars	élections	fonds	liberté	libertés	parti	partis	production	ressources	social	société	sociétés
culture	0,0000	-4,1619	-5,7677	-2,0373	4,0435	1,8438	-5,8313	-2,9481	3,0365	-1,1214	2,2954	13,6626	2,6713
dollars	-4,1138	0,0000	-6,3436	26,1306	-3,9473	-2,6466	-8,0158	-5,0032	9,2681	4,5638	-4,3563	-0,4877	5,6788
élections	-5,6659	-6,3045	0,0000	-5,4651	-1,4844	0,6354	30,1856	12,6993	-6,5914	-4,4091	-5,8780	-7,3452	-5,6298
fonds	-2,0243	26,2676	-5,5278	0,0000	-3,5873	-2,8003	-6,4969	-4,6319	-1,1713	9,3082	2,1514	-4,2788	1,5607
liberté	4,0702	-4,0198	-1,5210	-3,6341	0,0000	14,8675	1,4158	-1,2846	-3,8726	-0,7754	-0,1248	3,6982	0,5285
libertés	1,8826	-2,7338	0,6604	-2,8774	15,0805	0,0000	-1,8658	-1,5283	-3,0058	-2,3978	1,2372	1,2400	-1,7676
parti	-5,4862	-7,6296	28,9093	-6,1516	1,3233	-1,7193	0,0000	17,2584	-6,9917	-4,8229	4,1535	1,3945	-7,1811
partis	-2,9033	-4,9849	12,7313	-4,5909	-1,2568	-1,4742	18,0658	0,0000	-5,5923	-4,4464	-2,4210	-1,5378	-4,9145
production	3,0076	9,2870	-6,6459	-1,1676	-3,8106	-2,9159	-7,3607	-5,6243	0,0000	8,3426	2,2761	4,1252	10,0786
ressources	-1,1263	4,6376	-4,5082	9,4094	-0,7737	-2,3589	-5,1490	-4,5349	8,4602	0,0000	-0,9655	-0,1306	2,1281
social	2,2715	-4,3614	-5,9213	2,1427	-0,1227	1,1992	4,3689	-2,4327	2,2741	-0,9513	0,0000	4,3256	2,5427
société	13,1711	-0,4757	-7,2081	-4,1513	3,5418	1,1708	1,4289	-1,5053	4,0151	-0,1254	4,2139	0,0000	1,9331
sociétés	2,6722	5,7472	-5,7329	1,5712	0,5252	-1,7319	-7,6355	-4,9920	10,1791	2,1194	2,5703	2,0059	0,0000

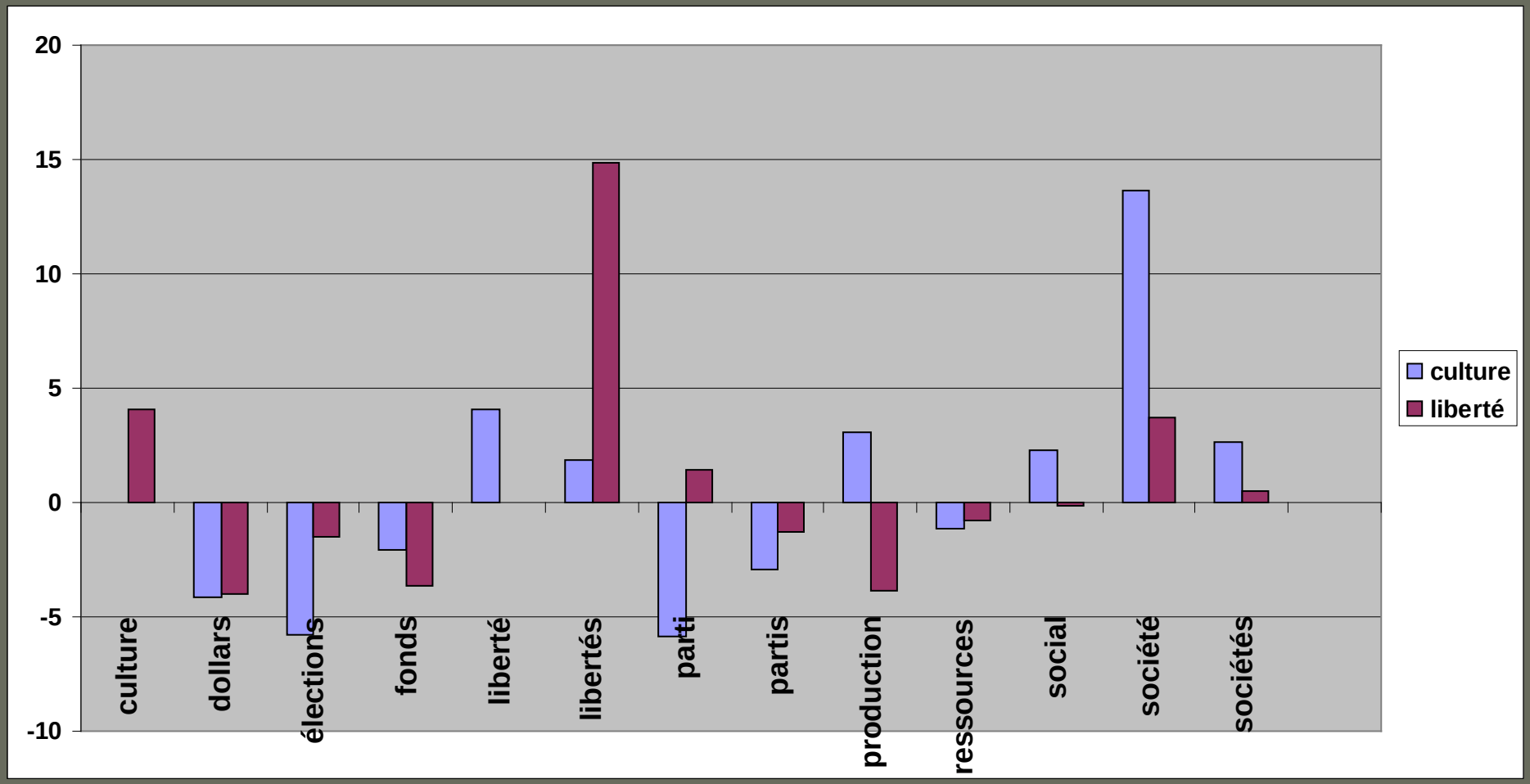
AFC des cooccurrences

Comparaison des profils cooccurrentiels de *élections* et *parti*



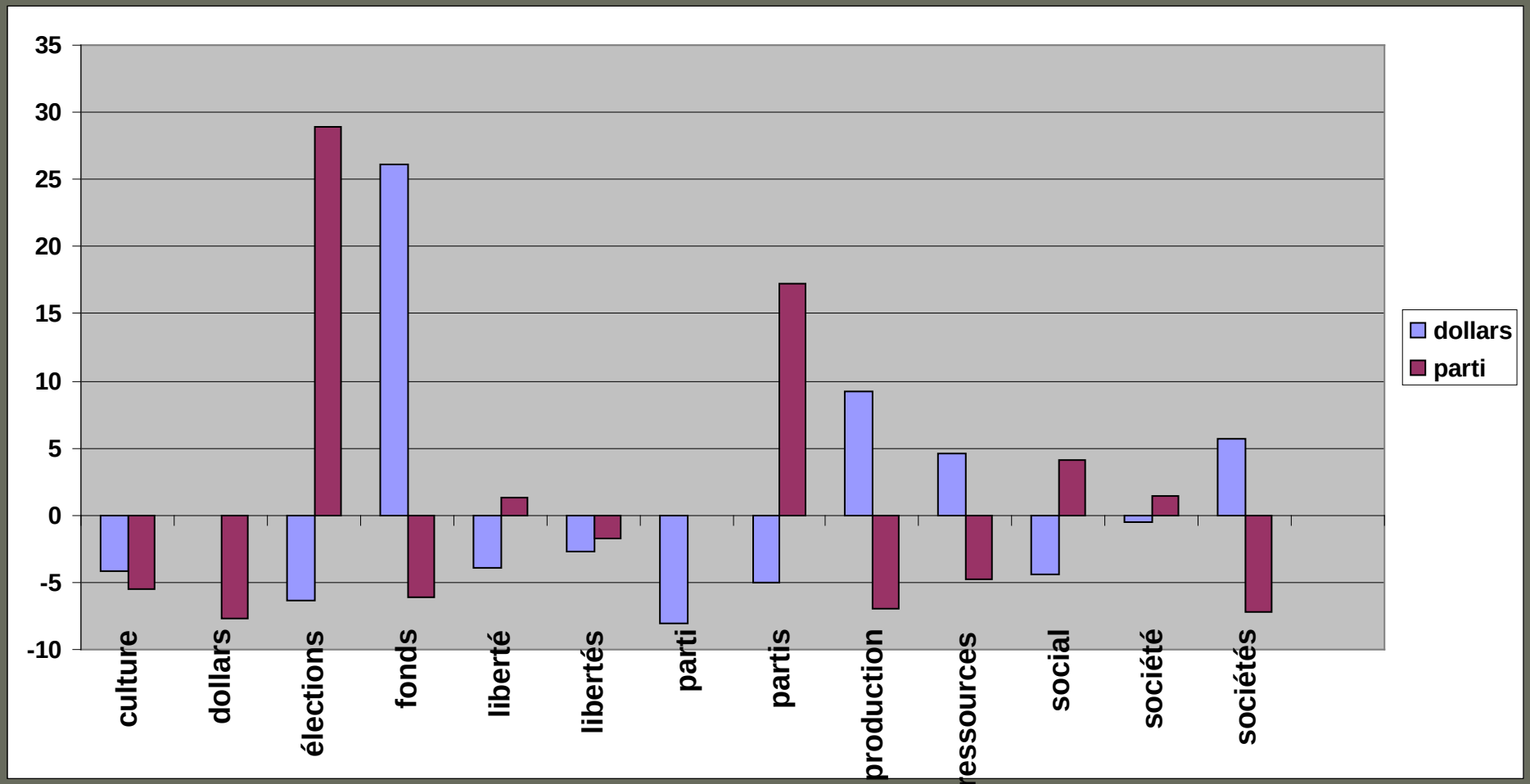
AFC des cooccurrences

Comparaison des profils cooccurrentiels de *liberté* et *culture*



AFC des cooccurrences

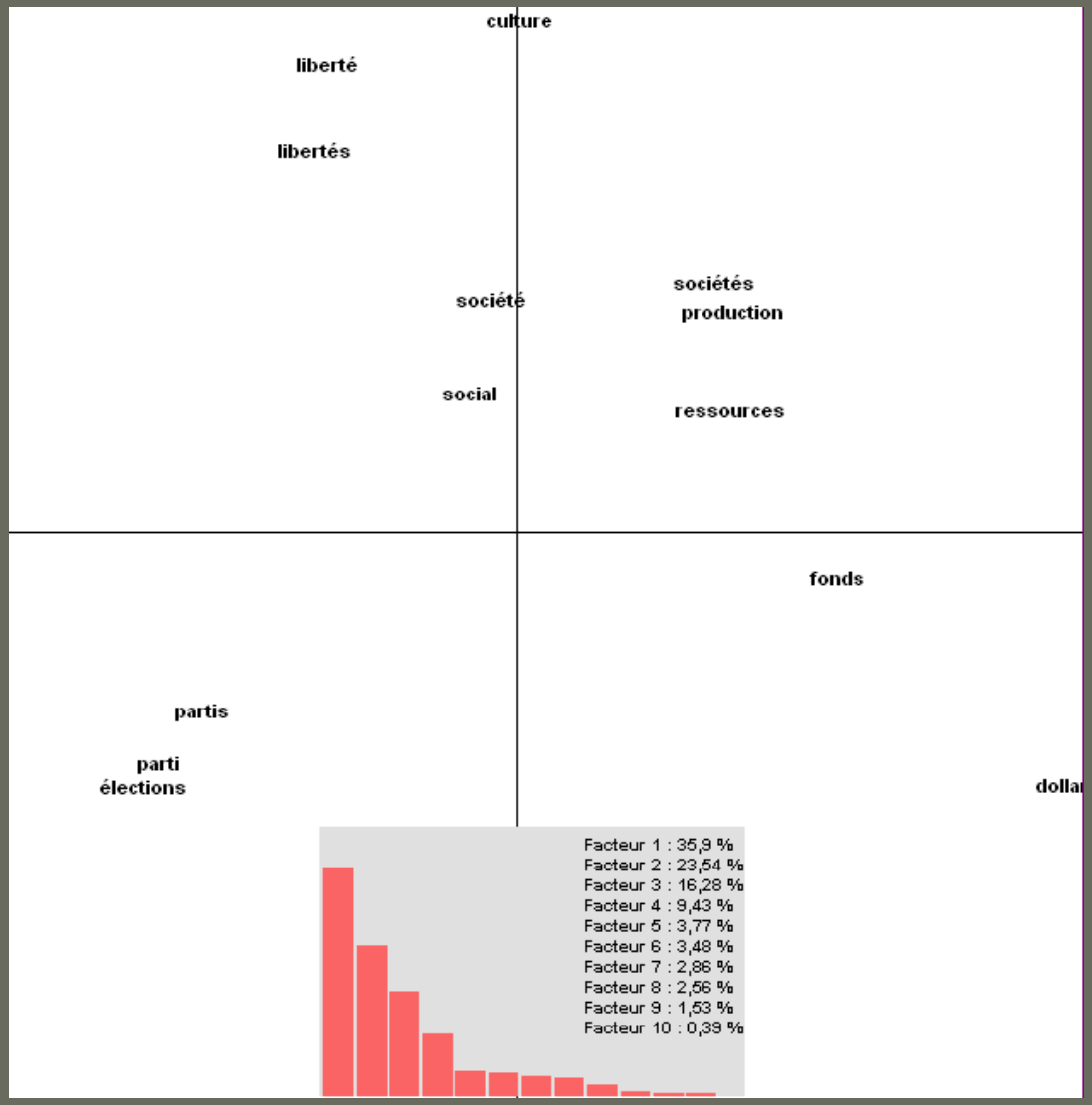
Comparaison des profils cooccurrentiels de *dollars* et *parti*



AFC des cooccurrences

Plan des 2 premiers facteurs

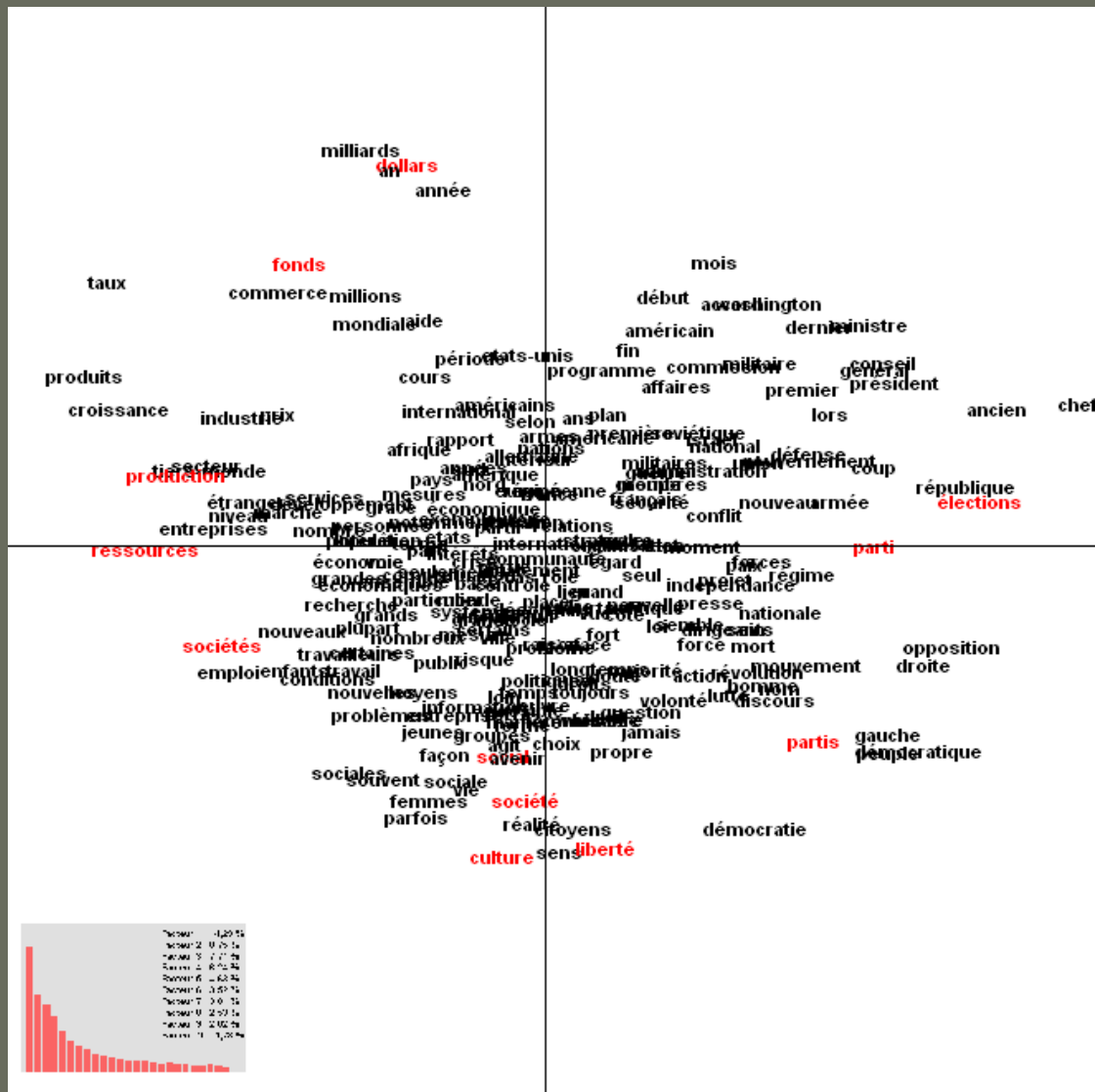
	culture	dollars	élections	fonds	liberté	libertés	parti	partis	production	ressources	social	société	sociétés	TOTAL
culture	358	7	1	15	32	11	18	15	42	13	39	129	32	712
dollars	7	1598	4	181	4	1	15	11	91	46	12	57	57	2084
élections	1	4	84	6	18	12	357	132	1	5	6	12	1	639
fonds	15	181	6	154	4	0	21	10	26	63	45	24	32	581
liberté	32	4	18	4	376	38	55	18	4	11	21	55	18	654
libertés	11	1	12	0	38	44	11	5	0	0	12	18	2	154
parti	18	15	357	21	55	11	1566	242	18	17	114	141	5	2580
partis	15	11	132	10	18	5	242	146	6	4	26	54	4	673
production	42	91	1	26	4	0	18	6	306	62	49	87	77	769
ressources	13	46	5	63	11	0	17	4	62	60	19	37	26	363
social	39	12	6	45	21	12	114	26	49	19	82	90	40	555
société	129	57	12	24	55	18	141	54	87	37	90	426	56	1186
sociétés	32	57	1	32	18	2	5	4	77	26	40	56	158	508
TOTAL	712	2084	639	581	654	154	2580	673	769	363	555	1186	508	11458



AFC des cooccurrences

244 formes lexicales les plus fréquentes, empan 15g/15d limite de phrase.

Plan des 2 premiers facteurs (23 % de l'inertie).



AFC des cooccurrences

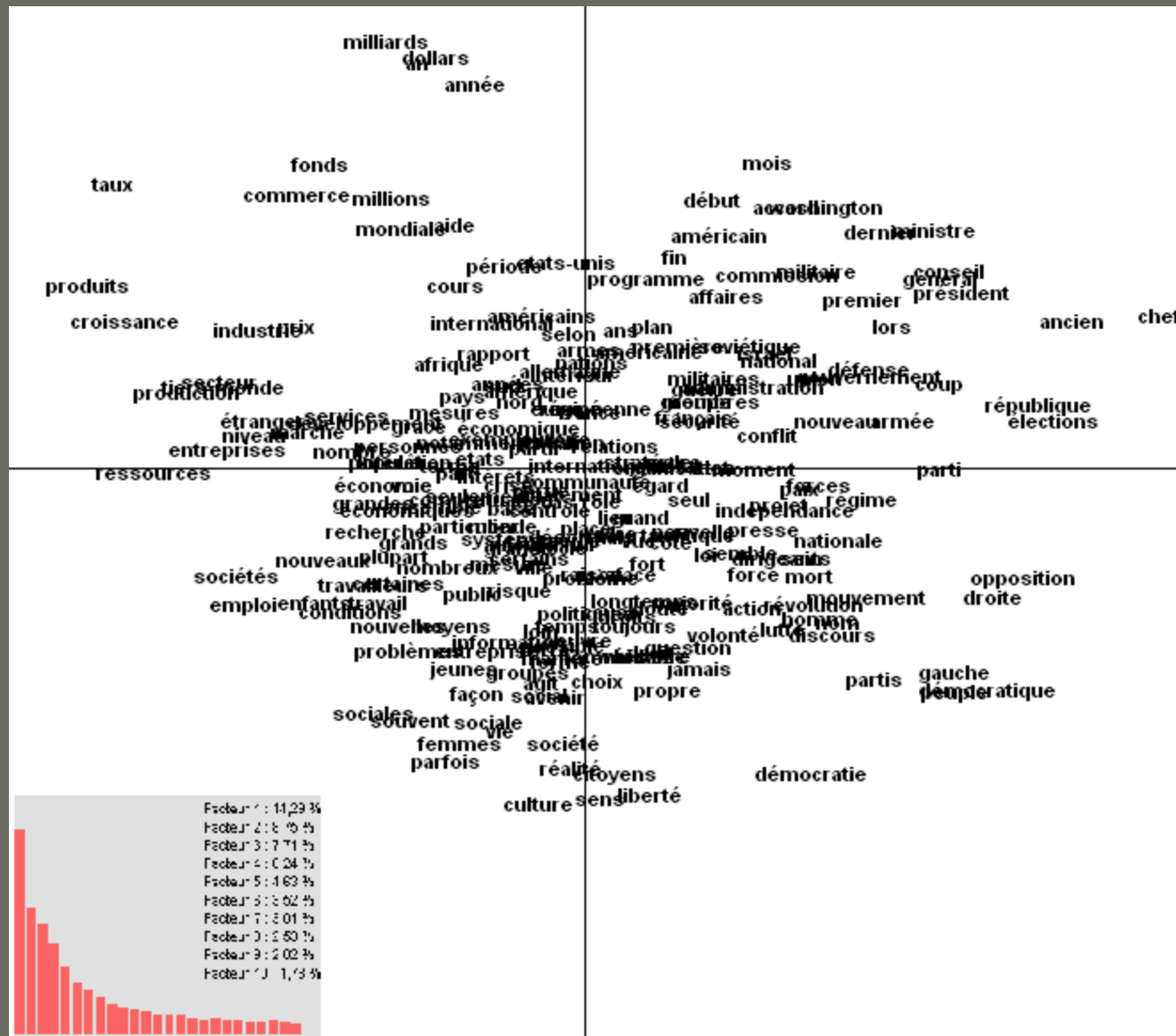
244 formes lexicales les plus fréquentes, empan 15g/15d limite de phrase

Inerties cumulées des facteurs 1 et 2 : 23 %

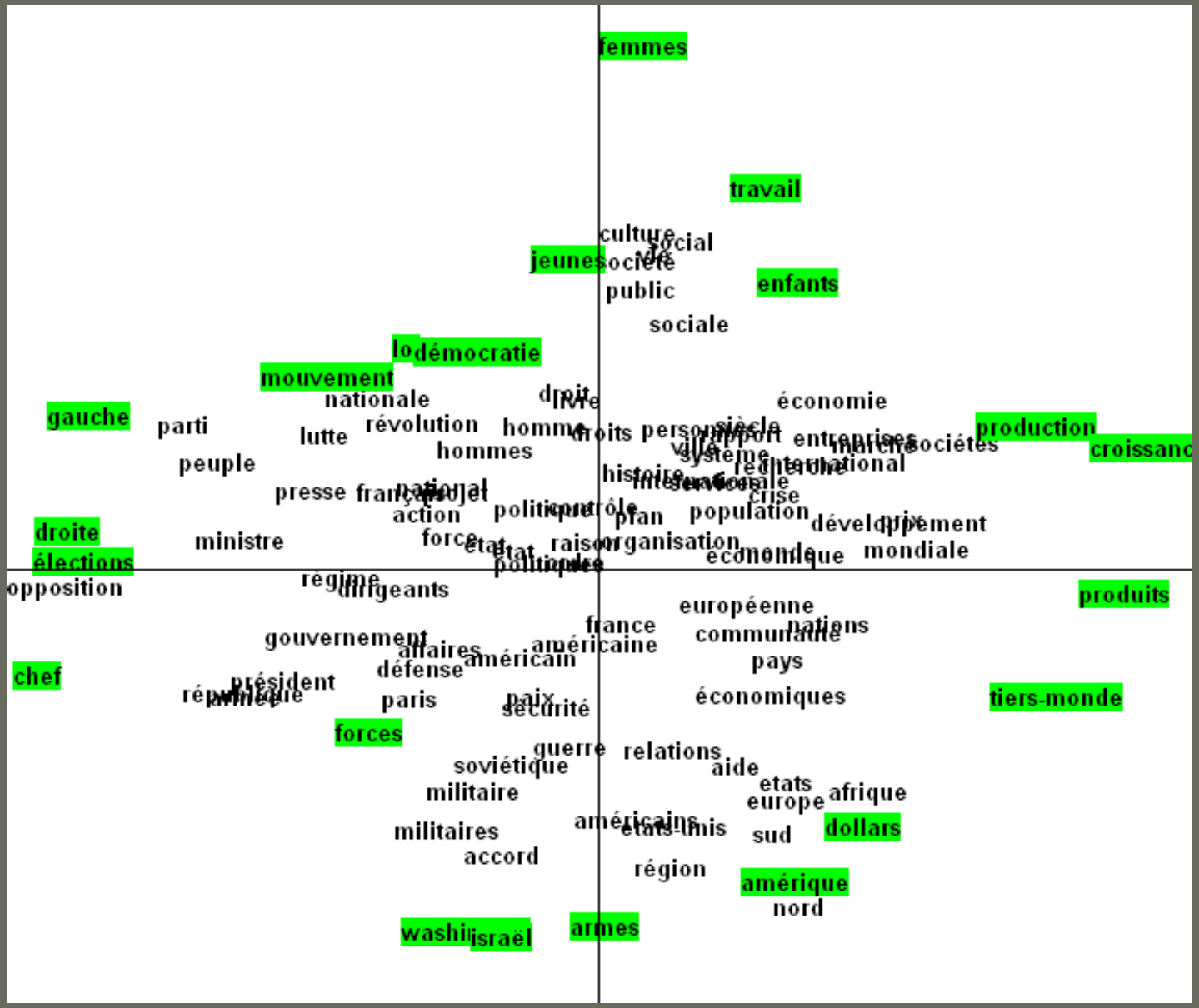
Visualiser l'ensemble des 3 premiers facteurs donnerait une meilleure approximation (Cibois 1994:85)

Nuage très enchevêtré

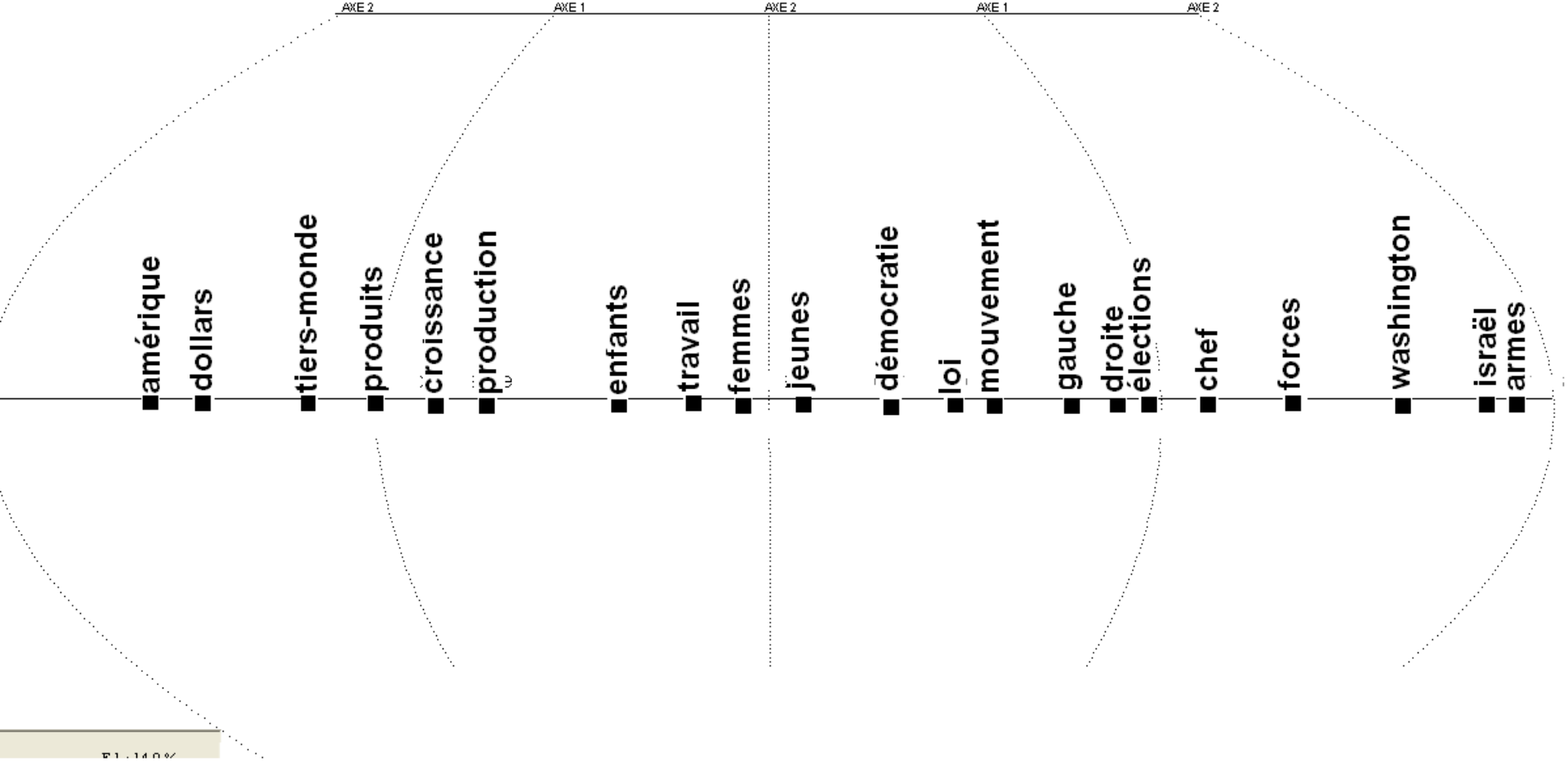
Centre du nuage encombré de points



Items distants du « centre » (saillants)



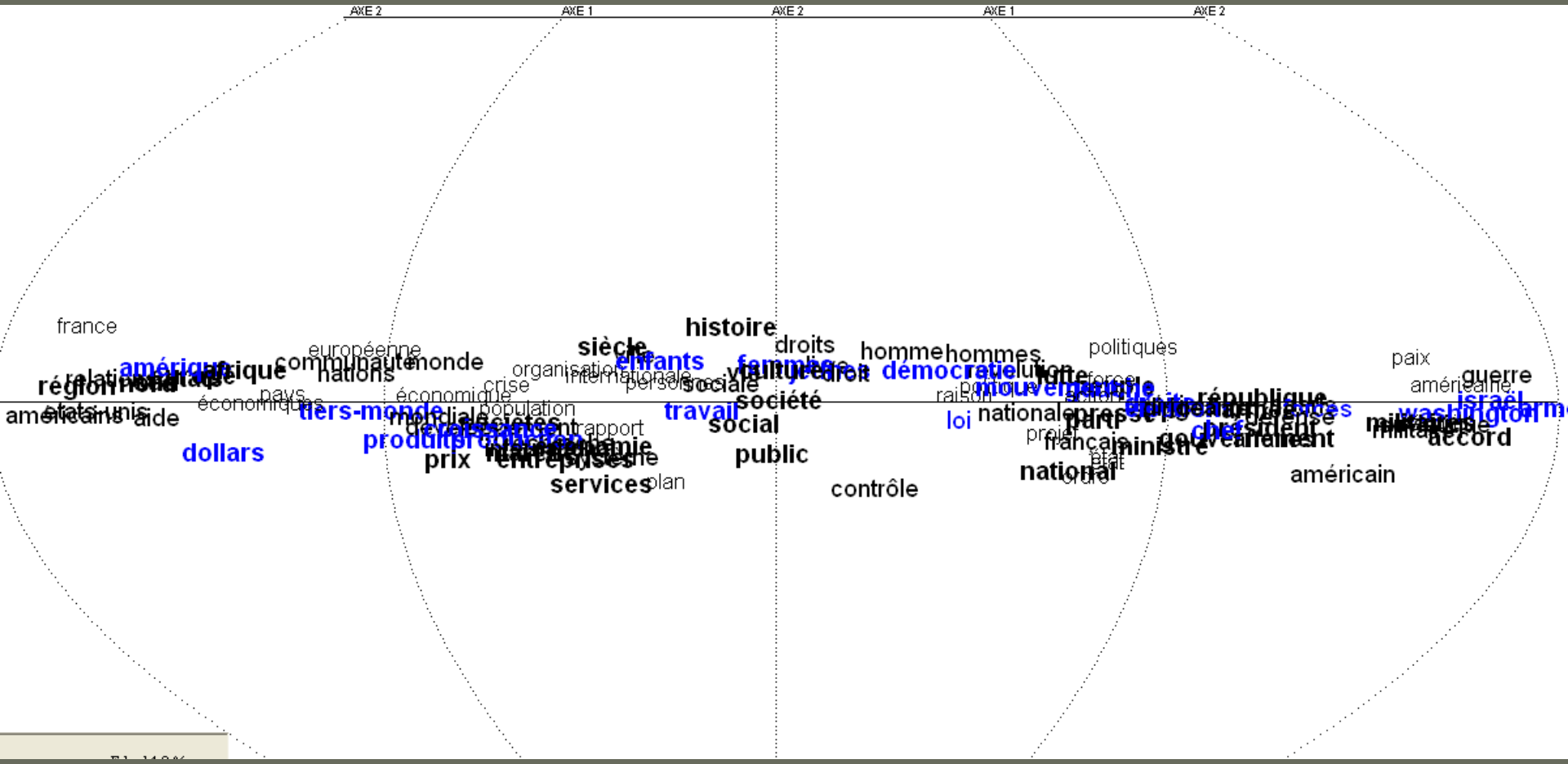
Les mêmes, sur l' « équateur », dans l'hypothèse d'un 3ème facteur entièrement neutre (tous à zéro).



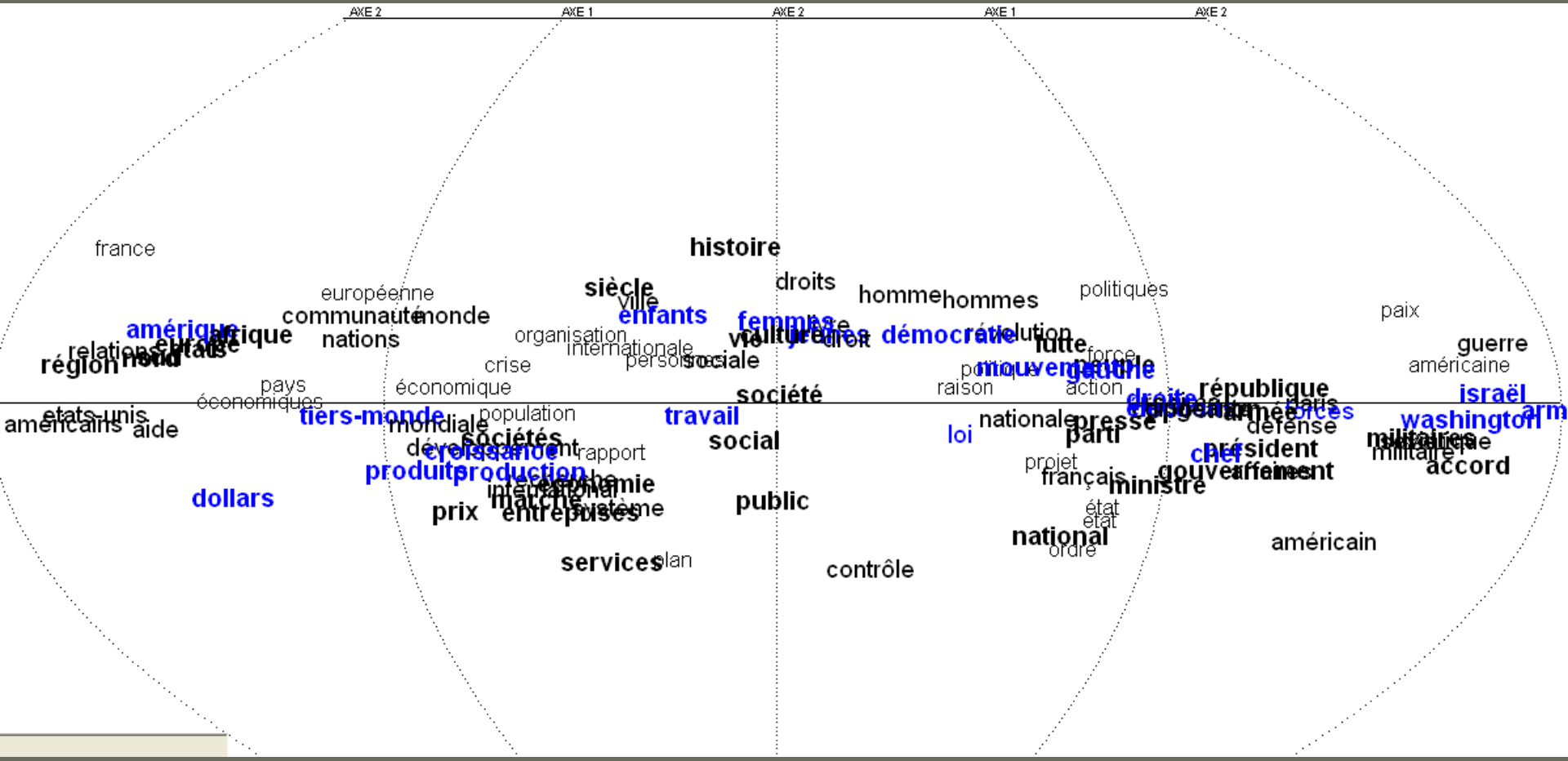
Coordonnées sur les 3 premiers facteurs

	1	2	3
amérique	0,1865	0,3273	-0,2777
armes	-0,0071	0,3748	0,0063
chef	-0,6045	0,1093	0,2223
croissance	0,5626	-0,1327	0,1969
démocratie	-0,1532	-0,2346	-0,1903
dollars	0,2638	0,2705	0,3282
droite	-0,5768	-0,0439	-0,0523
élections	-0,5662	-0,0116	-0,0117
enfants	0,1984	-0,3093	-0,3425
femmes	0,0321	-0,5603	-0,4768
forces	-0,2548	0,1697	0,0048
gauche	-0,5576	-0,1657	-0,1919
israël	-0,1153	0,3859	-0,0491
jeunes	-0,0456	-0,3322	-0,2245
loi	-0,2109	-0,2388	0,0646
mouvement	-0,3115	-0,2075	-0,1289
production	0,441	-0,1551	0,2592
produits	0,5411	0,0219	0,2833
tiers-monde	0,4599	0,1314	0,0284
travail	0,1629	-0,4085	0,0234
washington	-0,1658	0,3803	0,0355

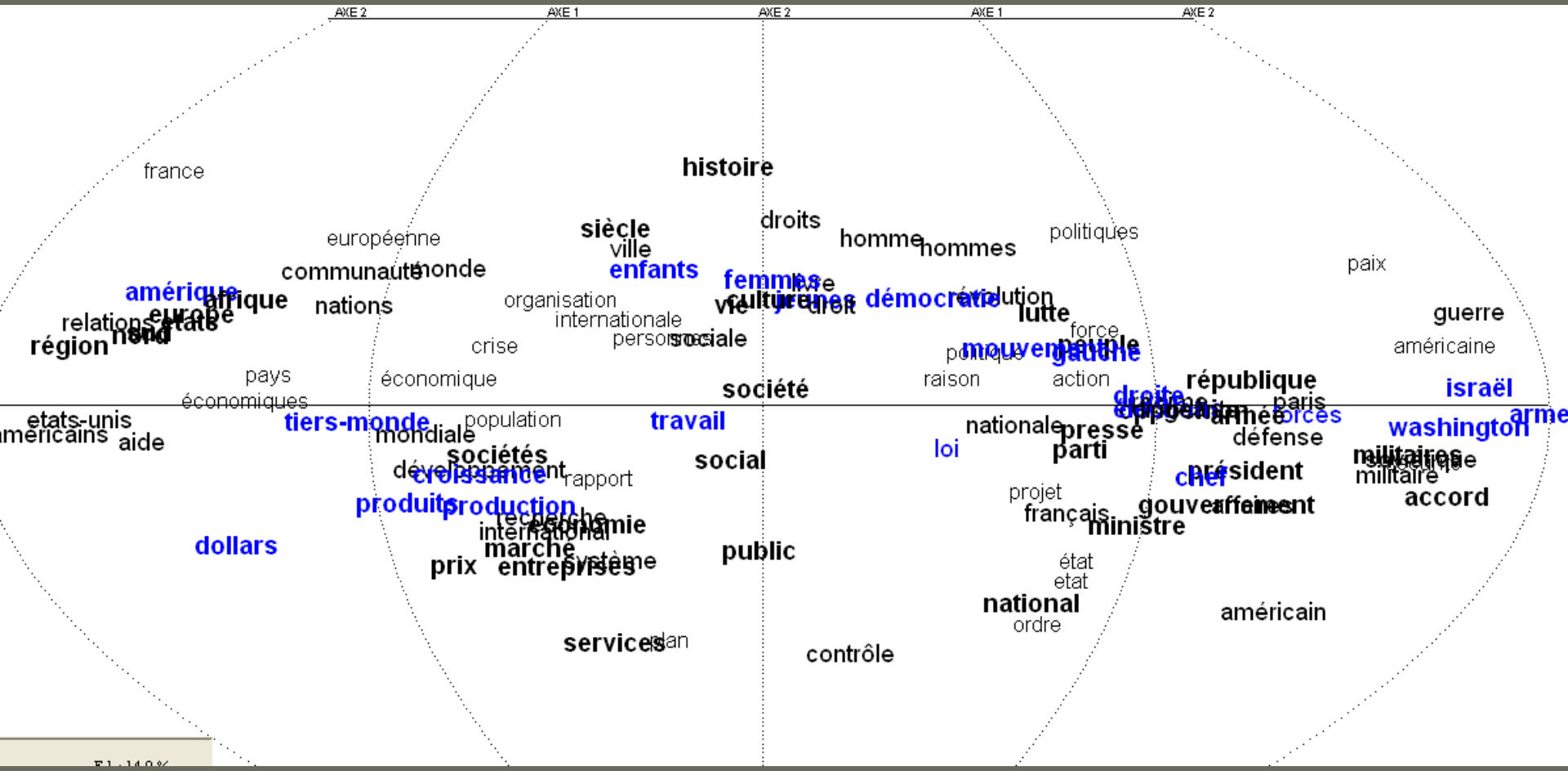
Visualisation des 3 premiers facteurs



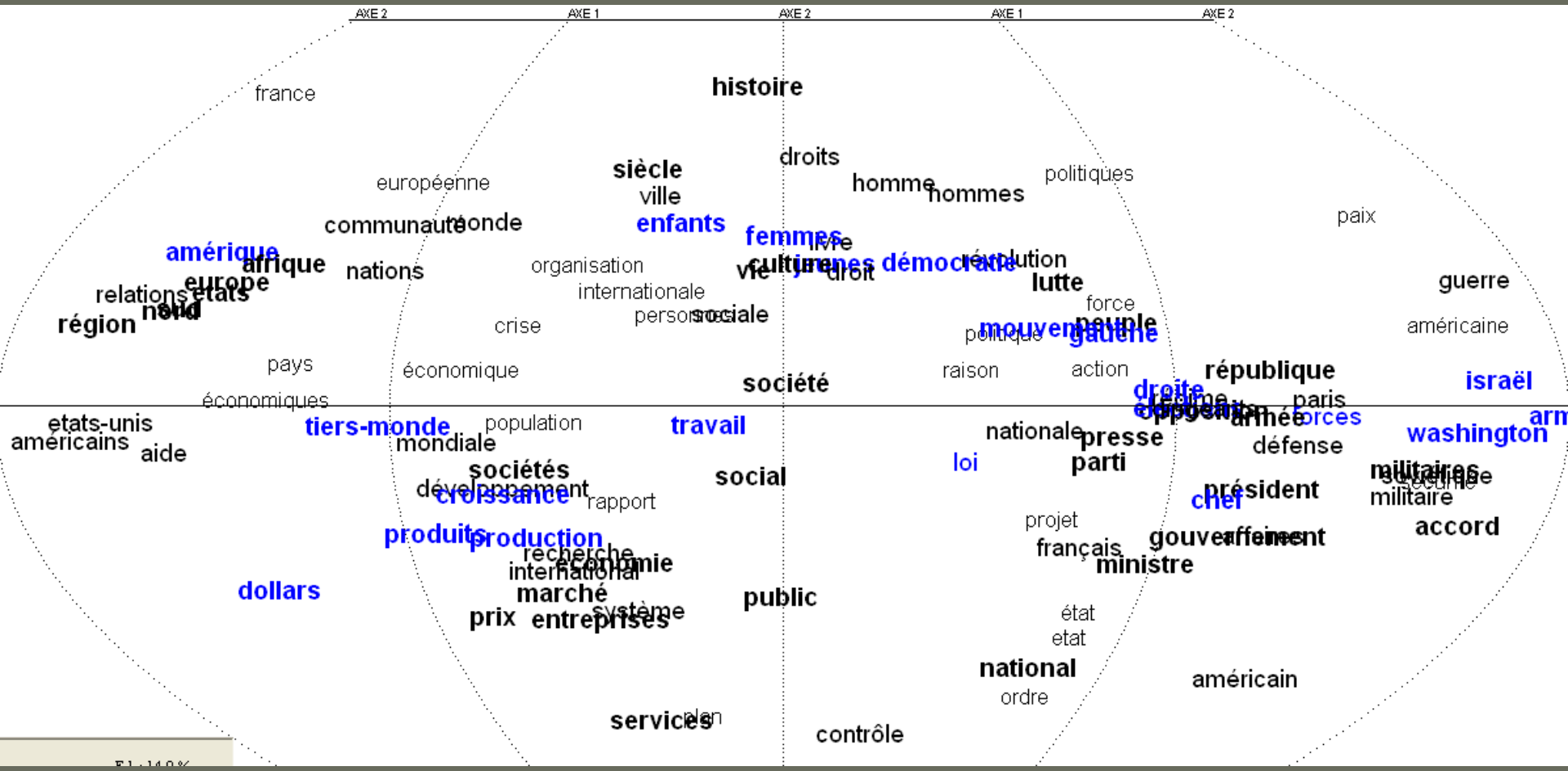
Visualisation des 3 premiers facteurs



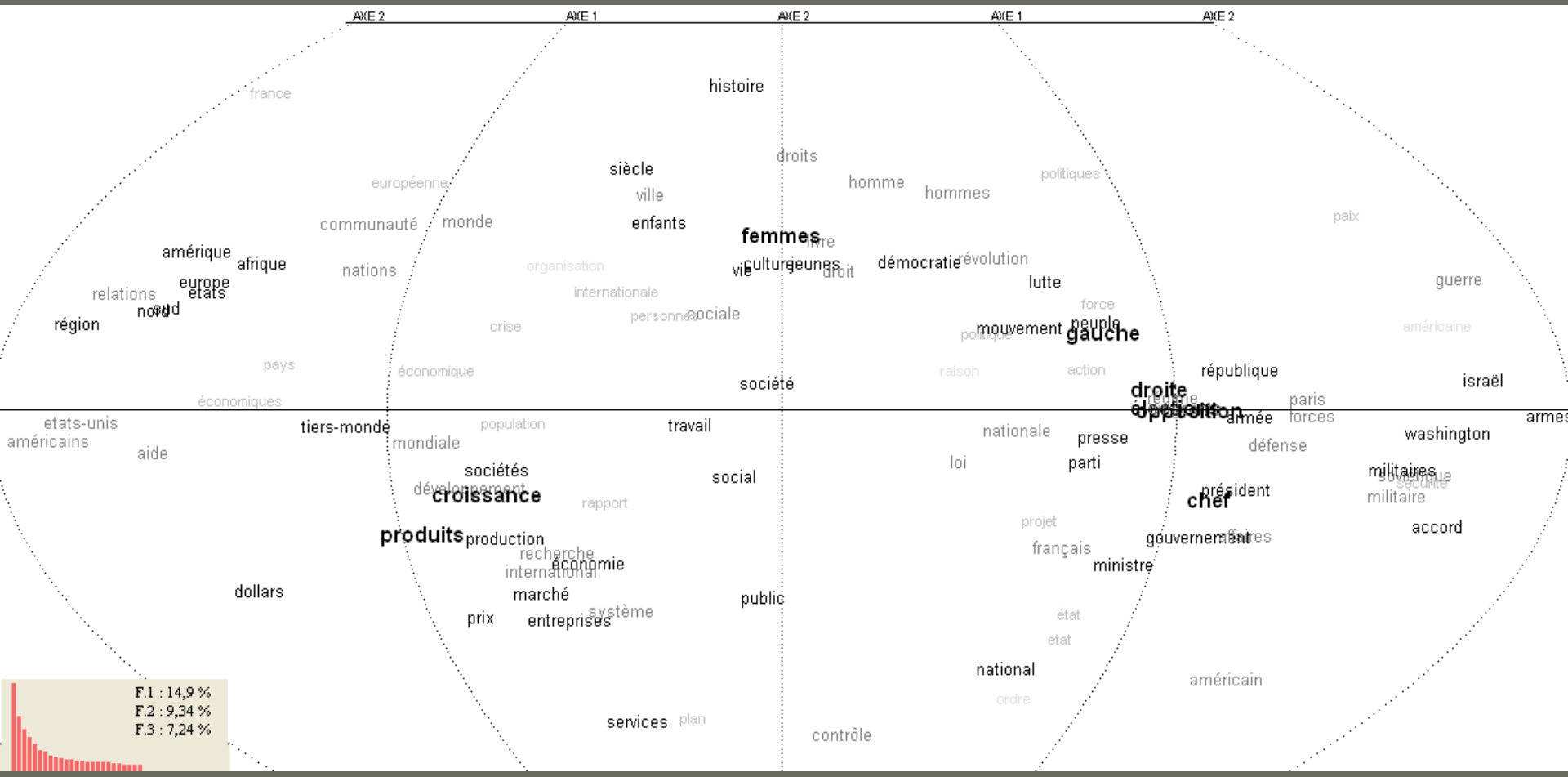
Visualisation des 3 premiers facteurs

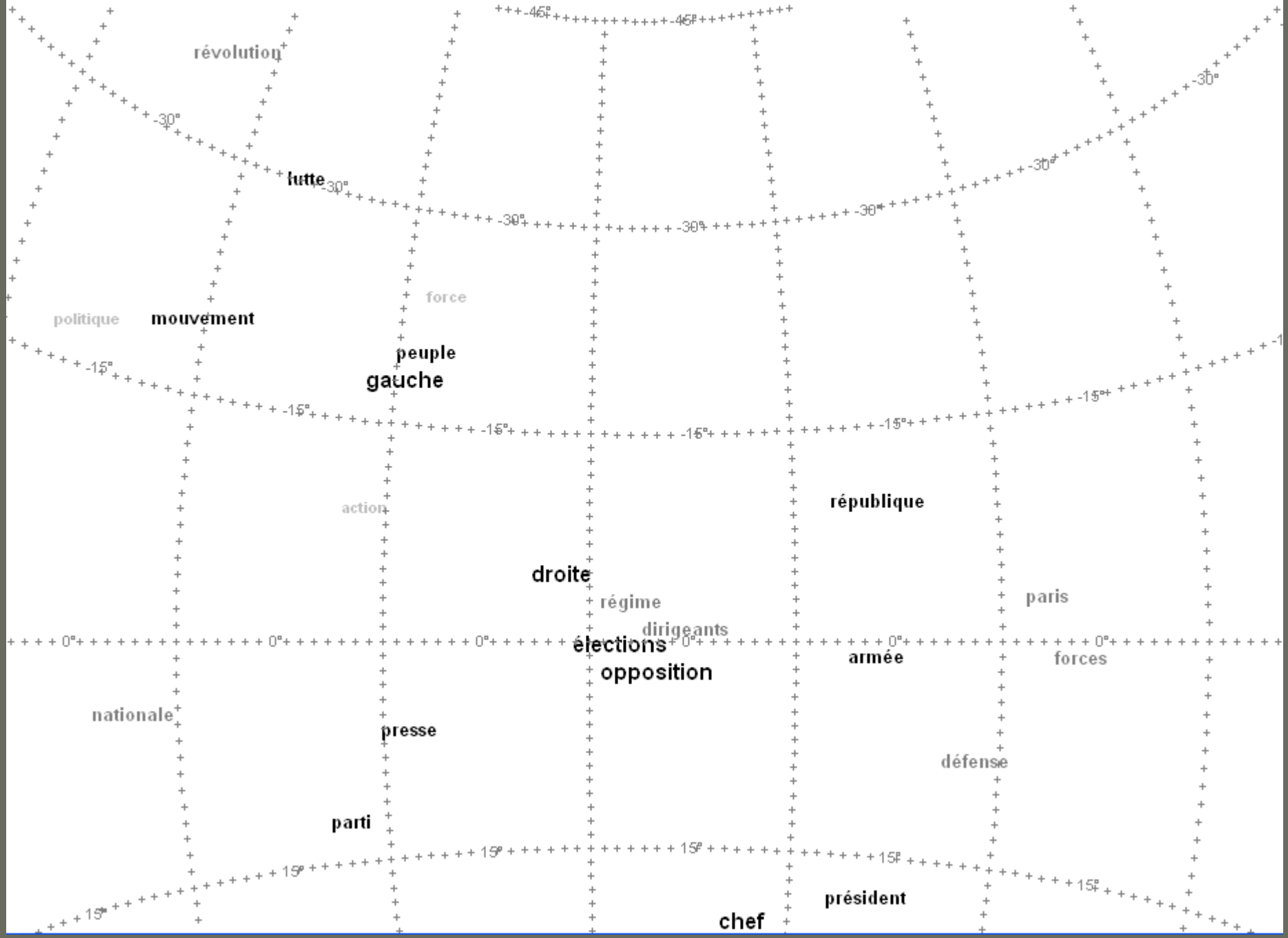


Visualisation des 3 premiers facteurs



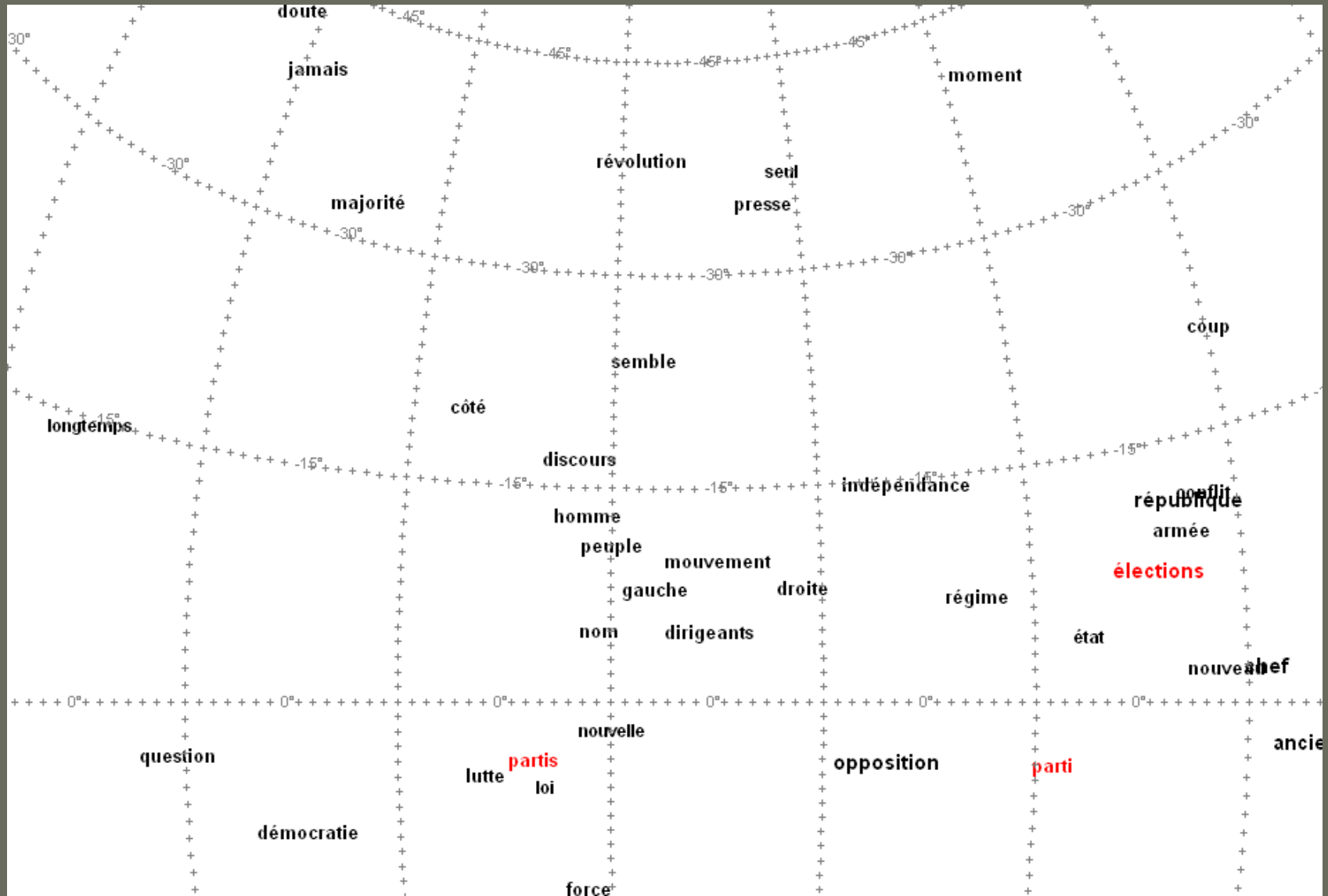
Visualisation des 3 premiers facteurs




révolution
lutte
force
politique
mouvement
peuple
gauche
action
république
droite
régime
paris
dirigeants
élections
opposition
armée
forces
nationale
presse
défense
parti
président
chef

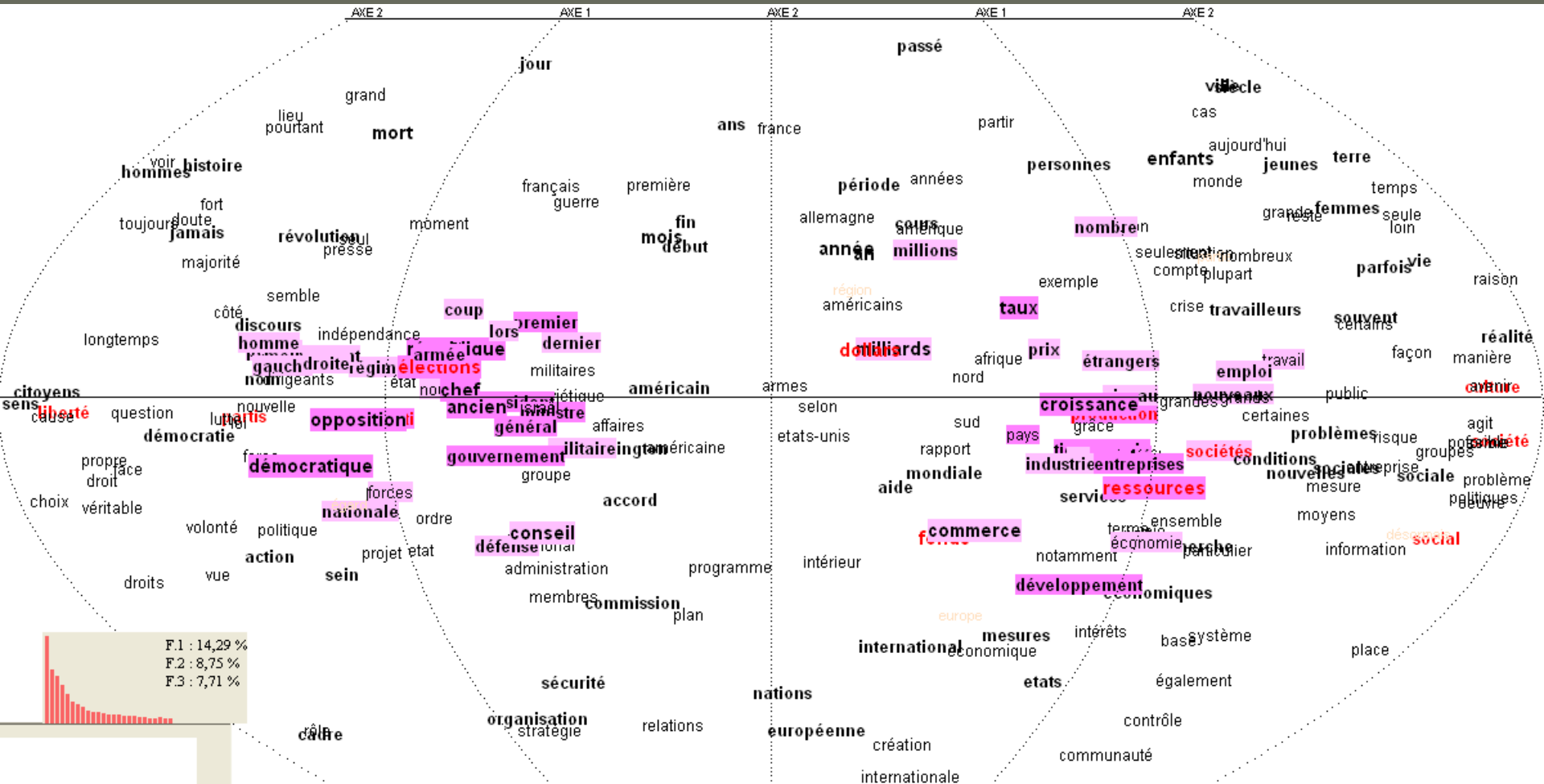
Visualisation des 3 premiers facteurs

Zoom « régional »



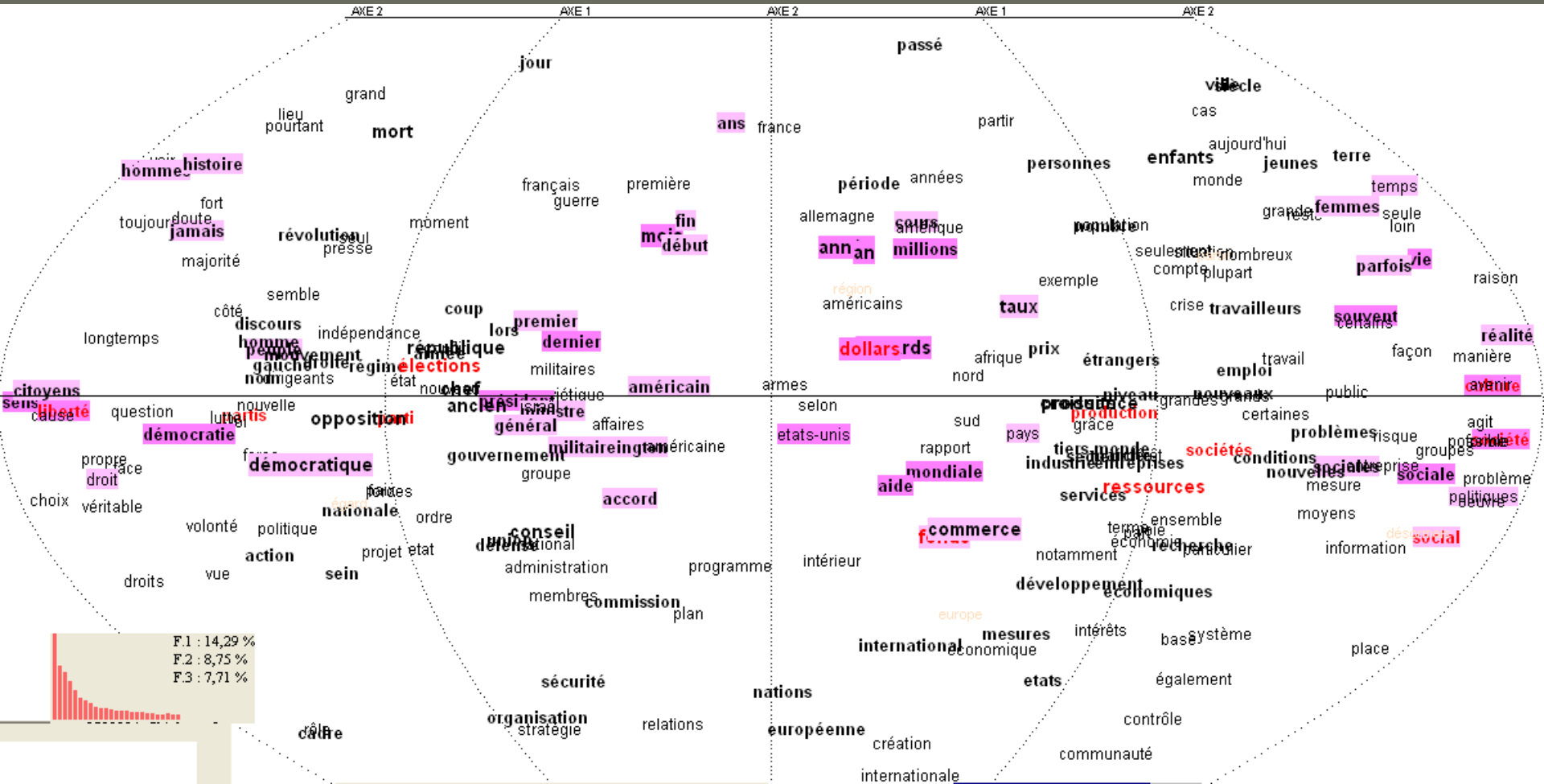
Visualisation des 3 premiers facteurs

Les fortes contributions au facteur 1 :



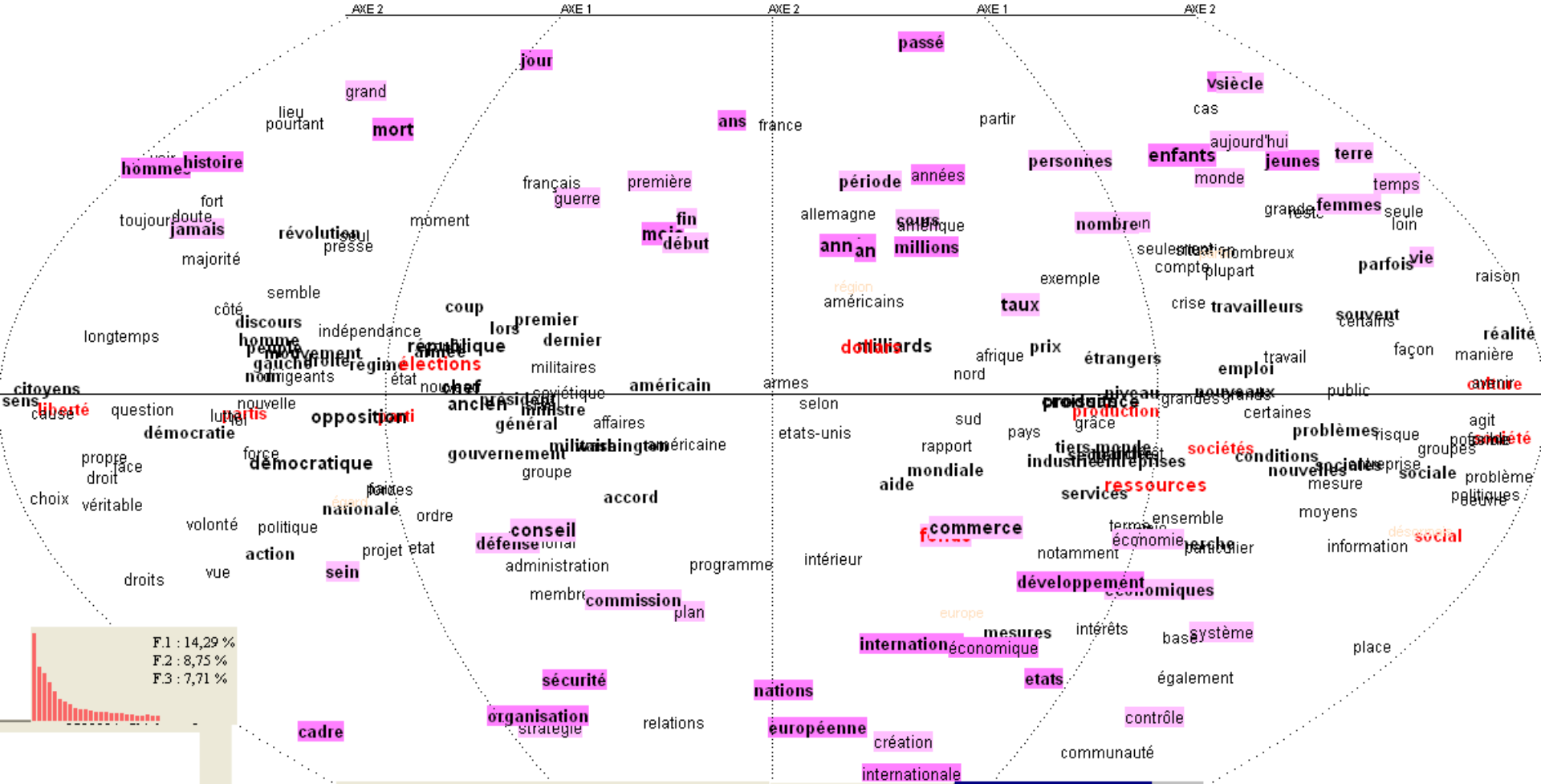
Visualisation des 3 premiers facteurs

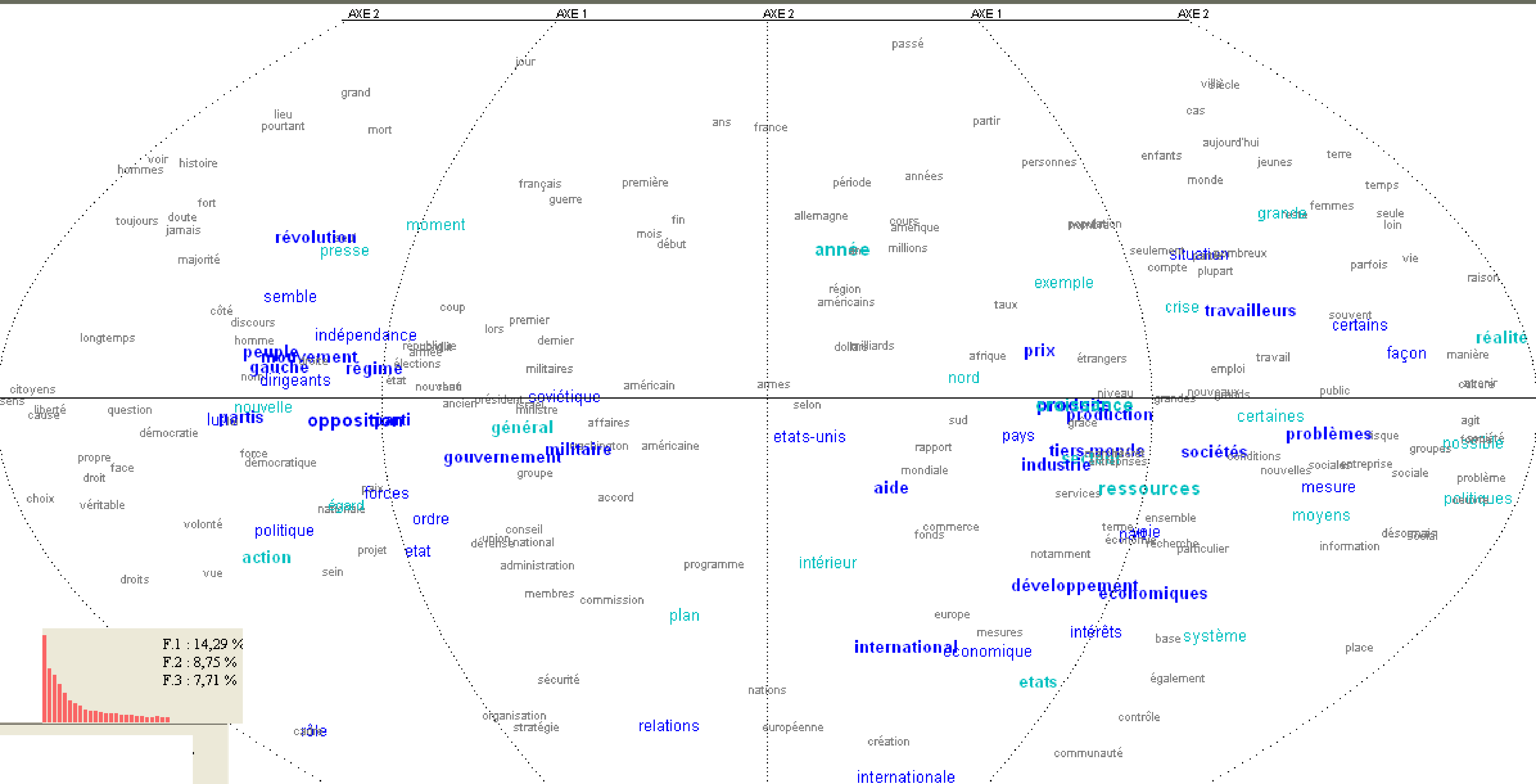
Les fortes contributions au facteur 2 :



Visualisation des 3 premiers facteurs

Les fortes contributions au facteur 3 :



Visualisation des 3 premiers facteurs

Marquage des formes excédentaires en 1980 :