Perte d'information?

Corpus de Gestion Publique

Lemmatisation singulier/pluriel

Laure Payret, Université de Picardie

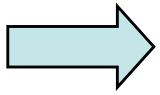
Systématique

Outil

-tronque: fréquence minimale, seuil …

-élague: mots-outils, hapax …

-regroupe: singulier - pluriel,
féminin –masculin,
type généralisé



lemmatisation
paramétrage

Types de corpus

Taille du corpus :

- questionnaires à réponses ouvertes 20, 30 observations 20 à 200 ko par questionnaire, corpus de 400 à 6 mo
- discours: 1 mo de 41 mo
- textes formatés: contrats de ville 16 mo, traité européen avec acte final 1,1 mo

⇒ Elagage: mots-outils, hapax …

Filtres

Parties:

- chapitre,
- thème,
- unité de temps
- variable signalétique

"l'objectif d'identification du genre serait un préalable à l'application de stratégies différenciées de la recherche d'informations dans le texte."Denise MALRIEU

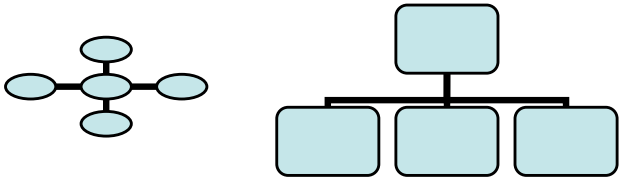
Spécificité

- Pathologies psychologiques: l'emploi du singulier stigmatise un comportement de retrait face au monde, (S.BRUNER JADT 2004)
- Groupe politique minoritaire au pouvoir: sous-utilisation des verbes et actions à court terme,
- Collectivités Territoriales et Culture: typologie issue de l'organisation, (C.LABBE, D.LABBE, D.MONIERE)
- Une typologie s'avère fondée, dans les différents genre du discours en Sciences de la Société,
- La spécificité du corpus est liée à la fonction de l'acteur.

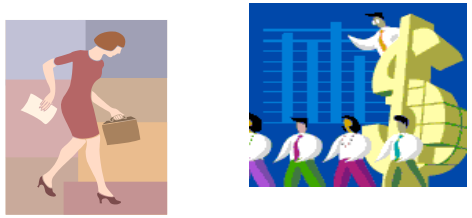
Ambiguïté

Genre du corpus

Arborescence textuelle



Type de variables retenues



Lemmatisation

Libre liberté libertés libéral libéralisme

lib

Plan

- *Méthodologie*
- *Discours*
  - *Discours pluriel au singulier*
  - *Concordances*
  - *Visualisation*

- *Réponses ouvertes:*
  - *Spécificités*
  - *Visualisation*

Méthodologie

1 Corpus primaire, dictionnaire
2 Mise en évidence formes signifiantes
3 La concordance de ces formes signifiantes
4 Calcul du coefficient d'implication réciproque (RD),
5 Pourcentage de ces formes signifiantes dans le corpus,
6 Spécificité
7 AFC
8 Reprise du processus avec le corpus réduit et lemmatisé.

corpus réduit,

fonction de la taille initiale

formes présentant un seuil de fréquence fixé,

une fois lemmatisé offre à certaines formes la fréquence nécessaire pour être retenue.

Pour que cela ait un sens "mesurable", il faudrait que la lemmatisation soit effectuée sur tous les substantifs

Discours présidentiels°

plus de 85% des signifiants sont des noms.

- Taille 63071
- Vocabulaire 7101
- maxfréq Signifiant /T 0,2822%
- maxfréq Signi/V 3,0418%
- fréqConservée/T 0,0342%
- fréqConservée/V 0,3042%

° Discours F.MITTERAND , 1981-1995, cd Fondation J.JAURES, filtre: laïcité

Discours

- 992 formes de fréquence supérieure ou égale à 5, signifiants sans mots-outils.
- 693 formes de fréquence supérieure ou égale à 5, sans les mots-outils avec les formes regroupées
  - *Concordances:*
- 215 formes signifiantes de fréquence supérieure ou égale à 14, dans le corpus sans les mots-outils.
- 225 formes signifiantes ayant une fréquence supérieure ou égale à 14, sans les mots-outils avec les formes regroupées (pluriel, singulier).
- Les concordances signifiantes sont extraites des 7 formes à gauche et à droite de la forme pôle.

Corpus signifiant non lemmatisé, concordance LIBERTE(S)

LIBERTES		CIR%
USAGE	*4*	*2,8571*
SOLIDARITE	1	*0,5102*
ORDRE	1	*0,2976*
NATIONALE	1	*0,1832*
SAUVEGARDE	1	*1,7857*
EPANOUISSEMENT	1	*1,4286*
BAFOUES	1	*3,5714*
DRAPEAU	1	
MENACEES	1	
BATAILLE	1	
FRANCE	1	
TIERS MONDE	1	
ECONOMIQUE	1	
CONSTRUCTION	1	

LS%	LS et L		L%
4,7619	***PRIVILEGES***	*3*	2,3810
0,8547	**DROITS**	2	0,5556
3,0303	**HOMME**	4	1,5504
6,0606	**PENSE**	4	2,0833
2,0833	**POUVOIR**	2	0,9009
6,6667	ENSEIGNER	2	1,2821
3,2258	**OBJECTIF**	2	1,0638

LIBERTE		CIR%
ENSEIGNEMENT	5	*0,1353*
LOIS	3	*0,0999*
ETAT	3	*0,0293*
CONQUETES	3	*0,6494*
BESOIN	3	*0,0248*
LIBERTES	2	*0,1855*
CREATRICES	2	*1,2987*
VIE	2	
RESPONSABILITE	2	
PUBLIC	2	
ORGANISATION	2	
EGALITE	2	
EXAMEN	2	
SAVOIR	2	
PENSE	2	

LIBERTES		CIR%	LS%		L%			CIR%

CIR Coefficient d'Implication Réciproque : rapport nombre de cooccurrences de 2 ou plusieurs formes au produit de leur fréquence

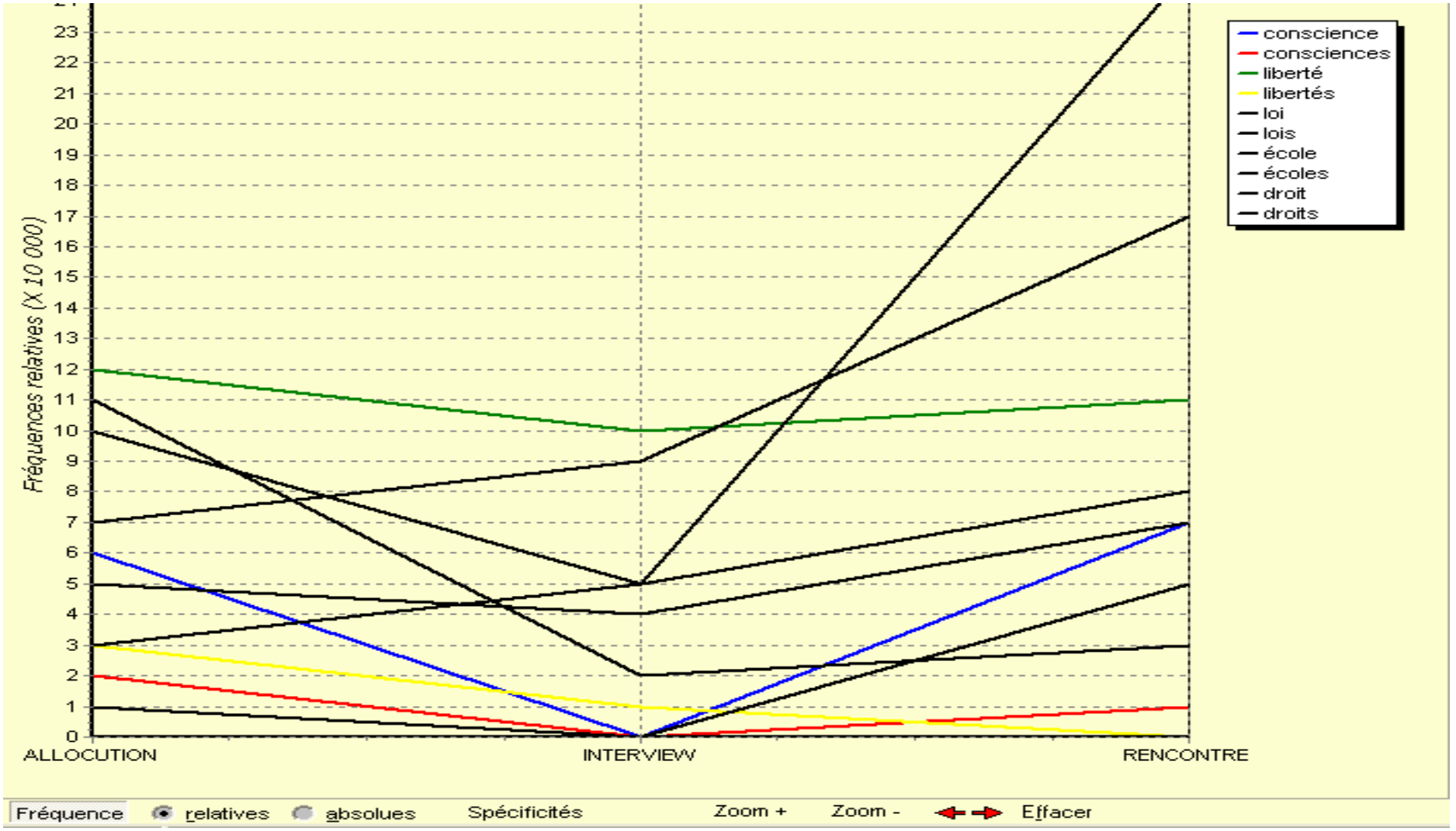
			LS et L			LIBERTE		
USAGE	*4*	*2,8571*	4,7619	*3*	2,3810			
			PRIVILEGES			ENSEIGNEMENT	5	*0,1353*
SOLIDARITE	1	*0,5102*	0,8547	2	0,5556			
			DROITS			LOIS	3	*0,0999*
ORDRE	1	*0,2976*	3,0303	4	1,5504			
			HOMME			ETAT	3	*0,0293*
NATIONALE	1	*0,1832*	6,0606	4	2,0833			
			PENSE			CONQUETES	3	*0,6494*
SAUVEGARDE	1	*1,7857*	2,0833	2	0,9009			
			POUVOIR			BESOIN	3	*0,0248*
EPANOUISSEMENT	1	*1,4286*	6,6667	2	1,2821			
			ENSEIGNER			LIBERTES	2	*0,1855*
BAFOUES	1	*3,5714*	3,2258	2	1,0638			
			OBJECTIF			*CREATRICES*	2	*1,2987*

ne classe pas selon le même ordre les cooccurrences communes à liberté et libertés. Enseigner,privilèges, objectif pour libertés;Privilèges, pense, homme pour liberté

Implication réciproque

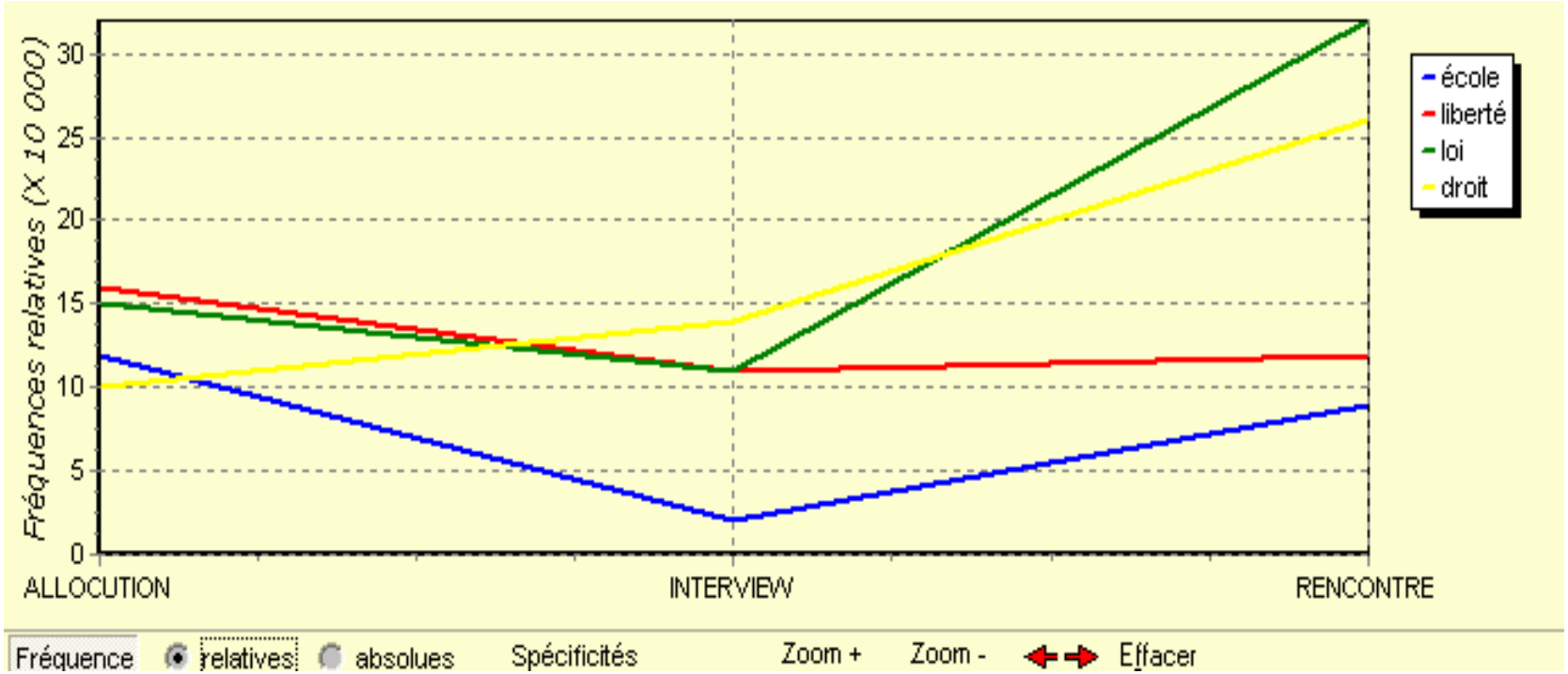
- Les termes en italiques n'appartiennent pas aux signifiants (fréquence trop faible).
- Les "libertés" sont liées fortement à l'"usage" alors que la "liberté" l'est avec "enseignement" puis "lois", "Etat", "conquêtes", "besoin".
- Cette différence pointe la séparation entre le pouvoir national, décisionnel, et le pouvoir local, opérationnel.
- Les formes communes relèvent du domaine conceptuel, du projet.

Spécificité avec formes singulier et pluriel

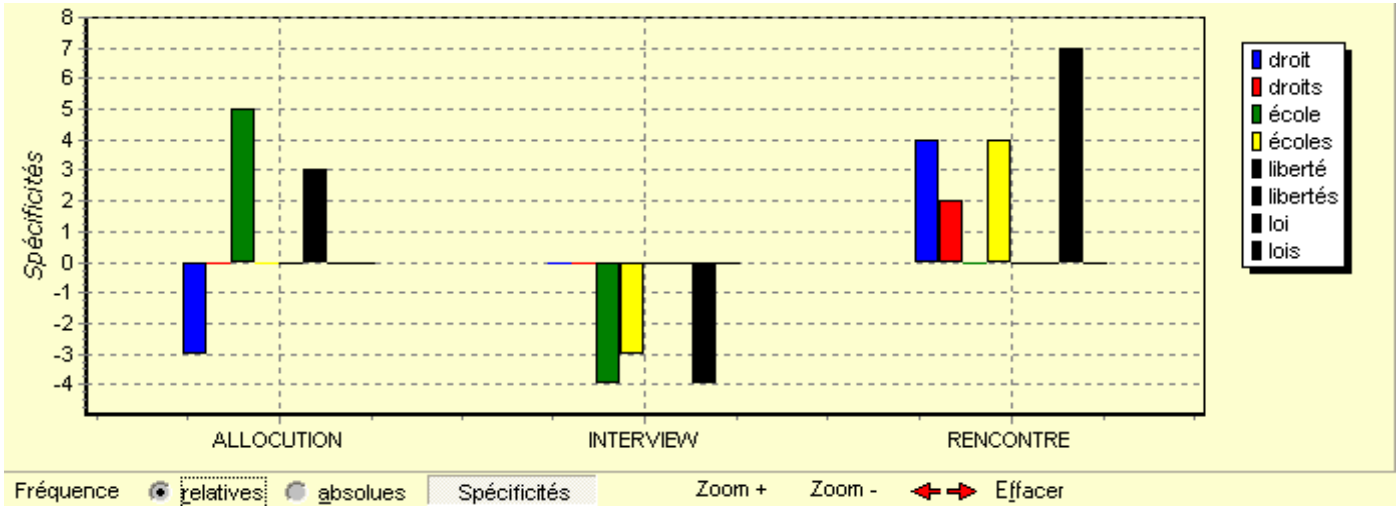


Les graphiques des spécificités ont été réalisés avec Lexico 3

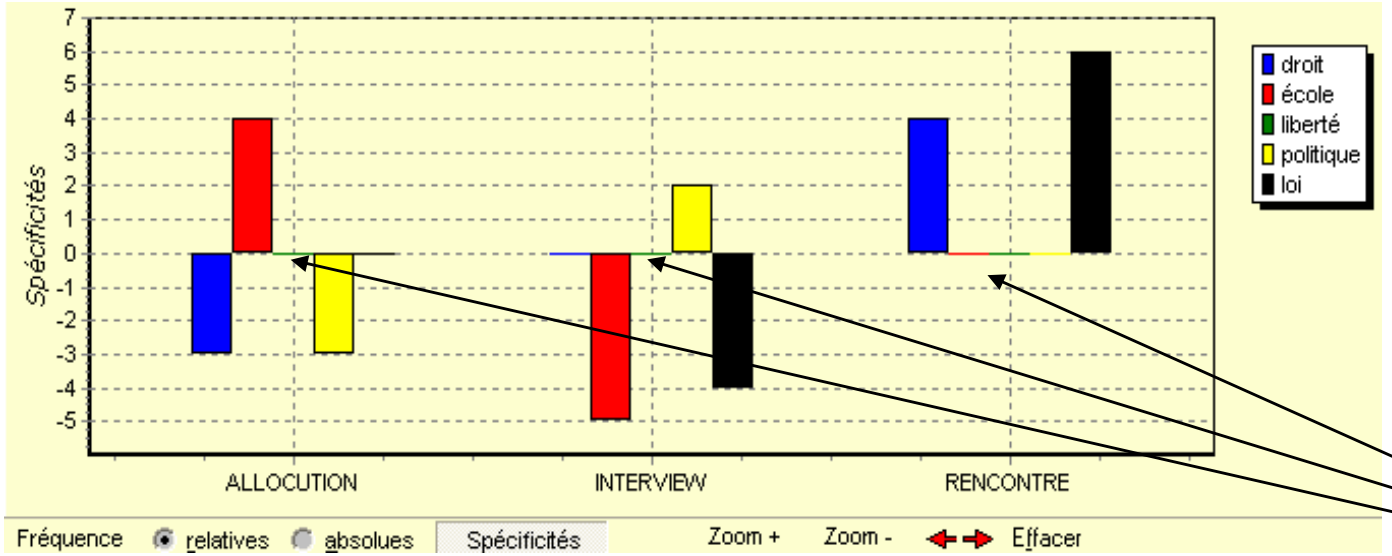
Formes ramenées au singulier



Spécificités



originales



lemmatisées

La forme liberté devient banale

Discours les pluriels au singulier

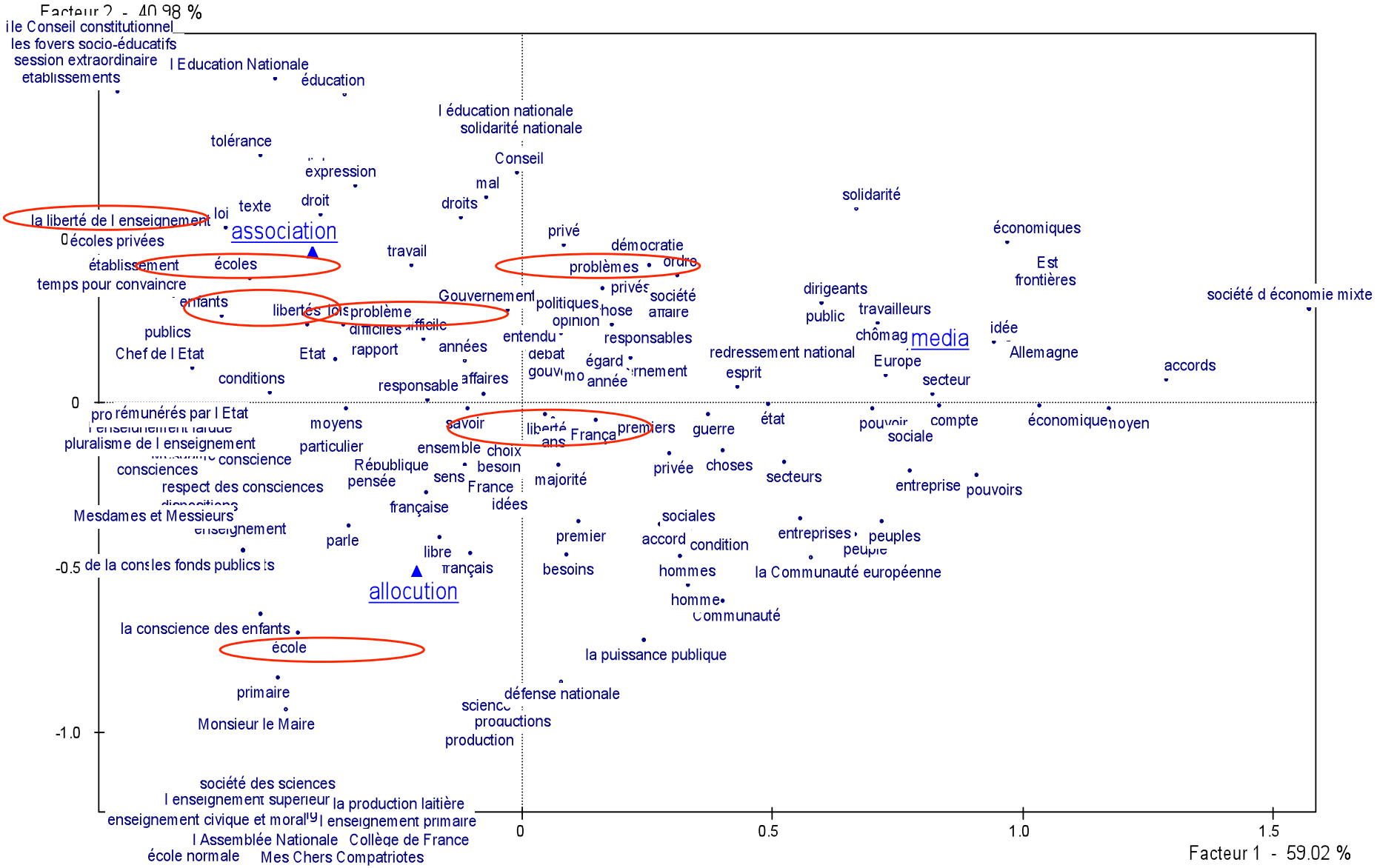
- Perte d'information avec la forme "liberté "
  qui devient banale, alors que
  au pluriel sens diffère en particulier dans cette thématique
  porteuse des libertés individuelles, "liberté de conscience« ,
  au singulier la personnification de la liberté,
- Ecole, forme très spécifique des types de discours,
  avec opposition entre les allocutions et les rencontres perd son sens "local" pour mettre en évidence uniquement le sens national,

En fonction du pourcentage de la forme au singulier, les spécificités de celui-ci l'emportent

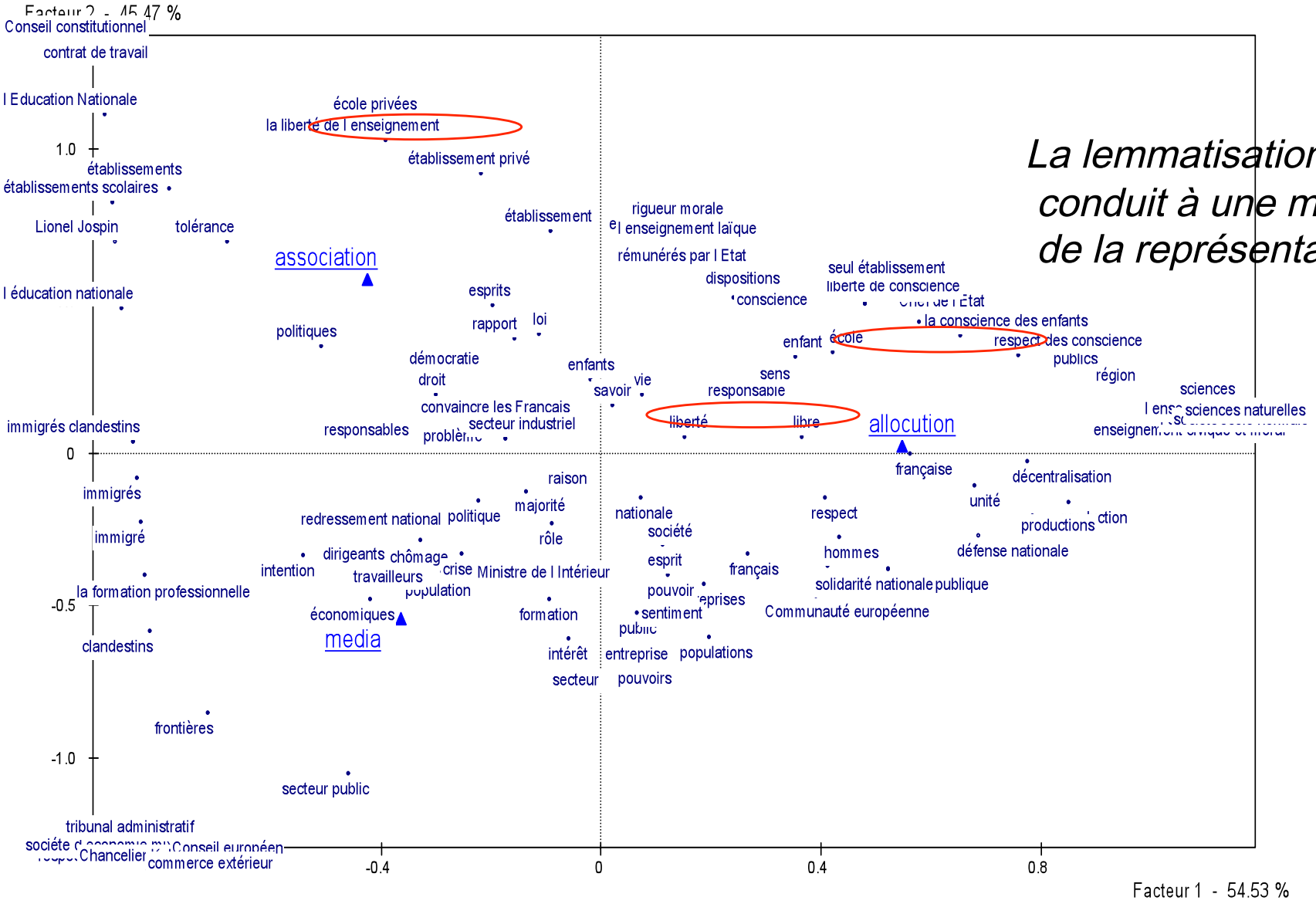
- Tous les termes retenus, ici, sauf "enfant" avaient un poids plus important au singulier qu'au pluriel,
- En revanche, dans le corpus lemmatisé, des termes
  INITIATIVE, ORGANISATION, CLANDESTIN, CITOYEN, DECISION, COMMUNISTE sont retenus.

Pour limiter les pertes de sens, il semble intéressant de regarder les contextes de ces formes avant de ramener au singulier.

Analyse Factorielle originale



Après lemmatisation partielle



La lemmatisation partielle conduit à une modification de la représentation,

Réponses ouvertes

- 16 acteurs de 3 CSP différentes, taille T=17822 formes, vocabulaire V=2989
- Sur les questionnaires à réponses ouvertes du CHU
- Taille 17822
- Vocabulaire 2989
- maxfréq signifiant /T 1,8180%
- maxfréq signifiant/V 10,8397%

- seuil 10%

Spécificités

Administratif

base	0,82	0,40	45	72	5,421	0,000
donnée	0,18	0,06	10	10	4,320	0,000
documents	*0,44*	*0,20*	*24*	*36*	*4,258*	*0,000*
l hôpital	0,35	0,15	19	27	4,026	0,000
pédiatrie	0,31	0,13	17	24	3,826	0,000
base données	0,15	0,04	8	8	3,767	0,000
projet	0,38	0,19	21	34	3,550	0,000
hôpital	0,35	0,17	19	30	3,488	0,000
service	0,93	0,62	51	110	3,322	0,000
service pédiatrie	0,20	0,08	11	15	3,126	0,001
services	0,31	0,16	17	28	3,078	0,001
préparation	0,13	0,04	7	8	2,957	0,002
hospitalisation	0,15	0,06	8	10	2,884	0,002
information	*0,15*	*0,06*	*8*	*10*	*2,884*	*0,002*
papier	*0,60*	*0,40*	*33*	*72*	*2,555*	*0,005*
temps	0,29	0,49	16	88	-2,565	0,005
infirmière	0,02	0,13	1	23	-2,833	0,002
cahier	*0,05*	*0,35*	*3*	*62*	*-4,928*	*0,000*

Spécificités

Infirmier

cahier	*0,92*	*0,35*	*49*	*62*	*7,881*	*0,000*
vert	0,34	0,10	18	18	6,164	0,000
cahiers	*0,26*	*0,09*	*14*	*16*	*4,541*	*0,000*
cahier vert	0,19	0,06	10	10	4,388	0,000
dossier médical	0,39	0,17	21	31	4,167	0,000
box	0,28	0,11	15	20	3,942	0,000
chambre	0,19	0,06	10	11	3,915	0,000
mauve	0,15	0,04	8	8	3,833	0,000
soin	*0,21*	*0,07*	*11*	*13*	*3,810*	*0,000*
note	*0,17*	*0,06*	*9*	*10*	*3,636*	*0,000*
dossier patient informatisé	0,19	0,07	10	12	3,540	0,000
alarme	0,13	0,04	7	7	3,524	0,000
violet	0,13	0,04	7	7	3,524	0,000
scope	0,13	0,04	7	7	3,524	0,000
feuilles	*0,30*	*0,13*	*16*	*24*	*3,523*	*0,000*
notes	*0,00*	*0,08*	*0*	*14*	*-2,461*	*0,007*
pédiatrie	0,02	0,13	1	24	-2,844	0,002
documents	*0,06*	*0,20*	*3*	*36*	*-2,893*	*0,002*

Spécificités

Médecin

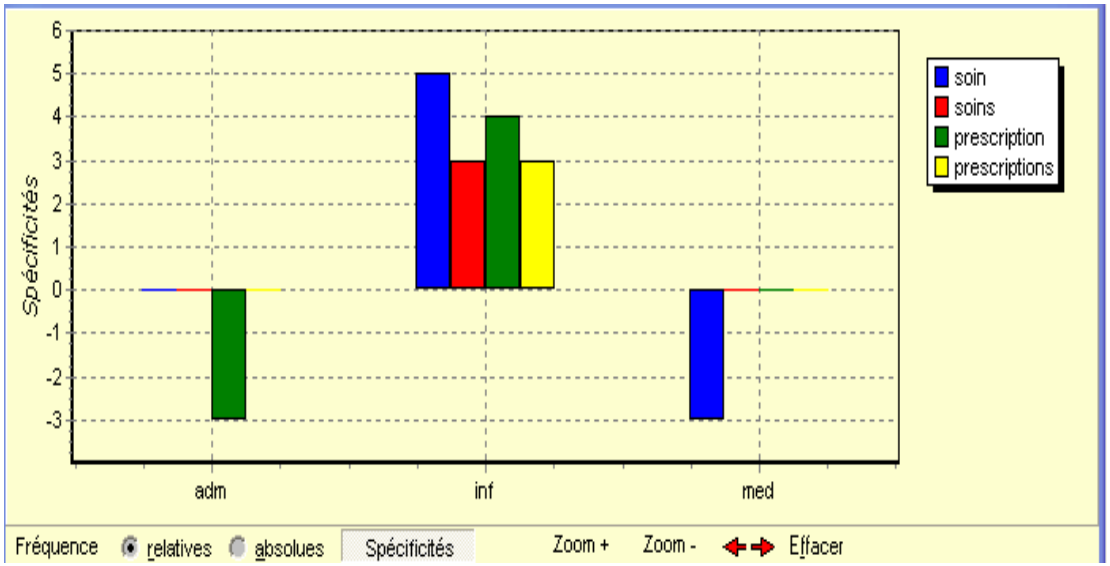
notes	***0,19***	***0,08***	***13***	***14***	***3,909***	***0,000***
rempli	0,14	0,06	10	10	3,756	0,000
activité	0,14	0,06	10	10	3,756	0,000
personnelles	0,13	0,05	9	9	3,514	0,000
notes personnelles	*0,13*	*0,05*	*9*	*9*	*3,513*	*0,000*
senior	0,11	0,04	8	8	3,258	0,001
terme	0,14	0,06	10	11	3,233	0,001
diagnostic	*0,10*	*0,04*	*7*	*7*	*2,982*	*0,001*
recherche	0,31	0,19	22	34	2,828	0,002
médecin	0,50	0,34	35	61	2,737	0,003
traitements	0,11	0,05	8	9	2,717	0,003
thérapeutiques	0,09	0,03	6	6	2,683	0,004
transmettre	0,09	0,03	6	6	2,683	0,004
exploration	0,09	0,03	6	6	2,683	0,004
soin	***0,01***	***0,07***	***1***	***13***	***-2,186***	***0,014***
hospitalisation	0,03	0,10	2	17	-2,188	0,014
feuille	***0,03***	***0,10***	***2***	***17***	***-2,188***	***0,014***
cahier vert	0,00	0,06	0	10	-2,464	0,007
donnée	0,00	0,06	0	10	-2,466	0,007
cahier	***0,14***	***0,35***	***10***	***62***	***-3,808***	***0,000***

Spécificités

• **Informations** 86,		
• **Information**	45	sur représentée chez les administratifs au singulier
• **Cahier**	82	sous-représentée chez les
• **Cahiers**	16	spécifiquement positif au singulier et au pluriel infirmiers, négatif médecins
• **Feuille**	17	spécificité négative des médecins
• **Feuilles**	24	spécificité positive des infirmiers
• **Note**	10	spécificité positive des infirmiers
• **Notes**	14	spécifique positive des médecins, négatives des infirmiers
• **Soins**	43	sur-représentée chez infirmier
• **Soin**	13	sur-représentée infirmier, sous

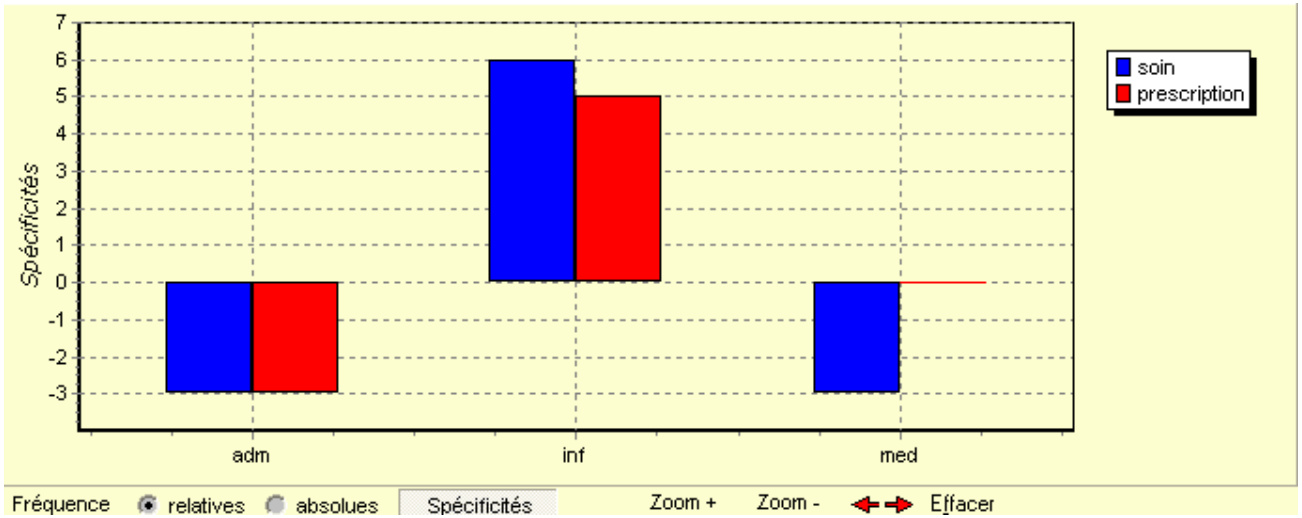
La fonction discrimine le vocabulaire et l'emploi de certaines formes au singulier ou pluriel

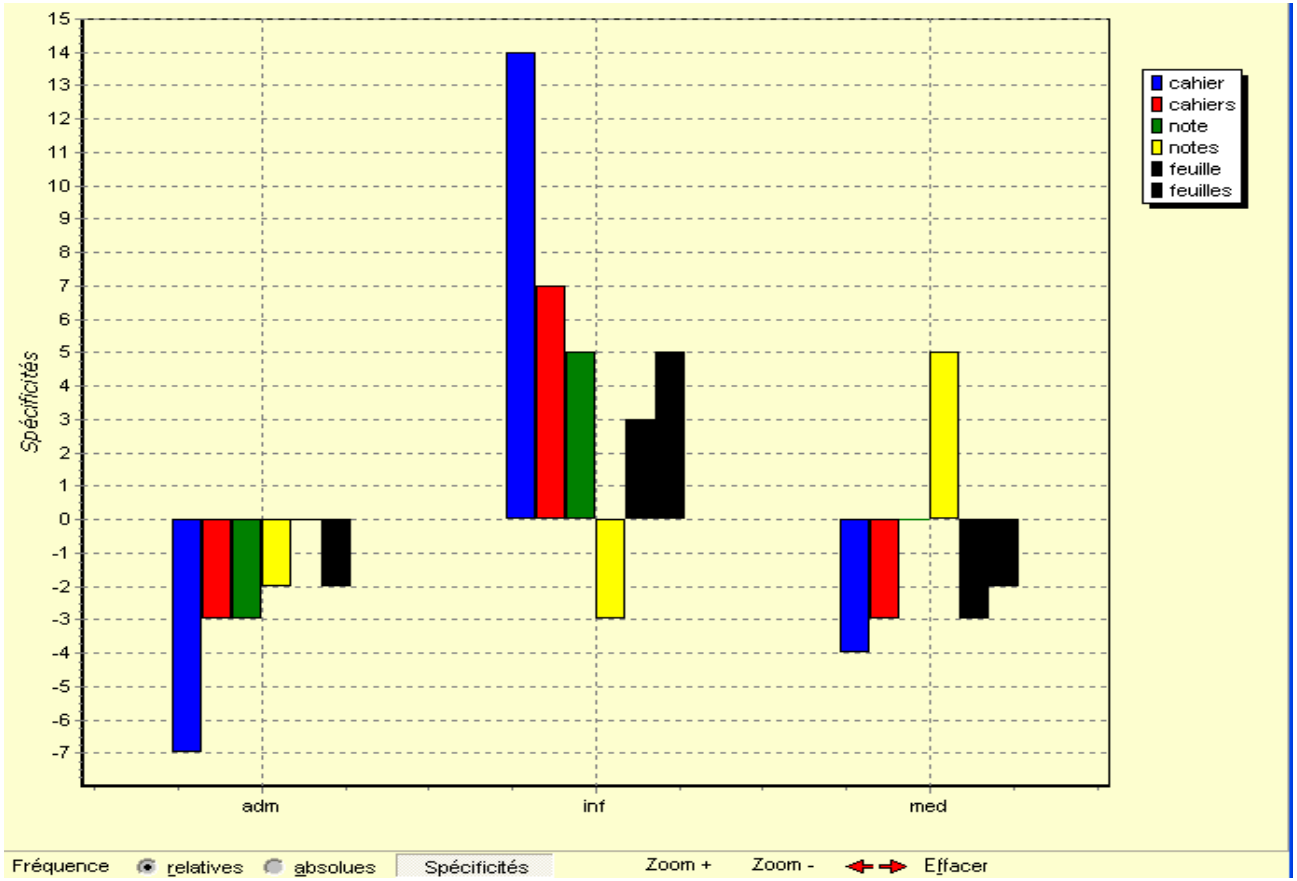
Soins/soin



Soin

Aplatissement des spécificités

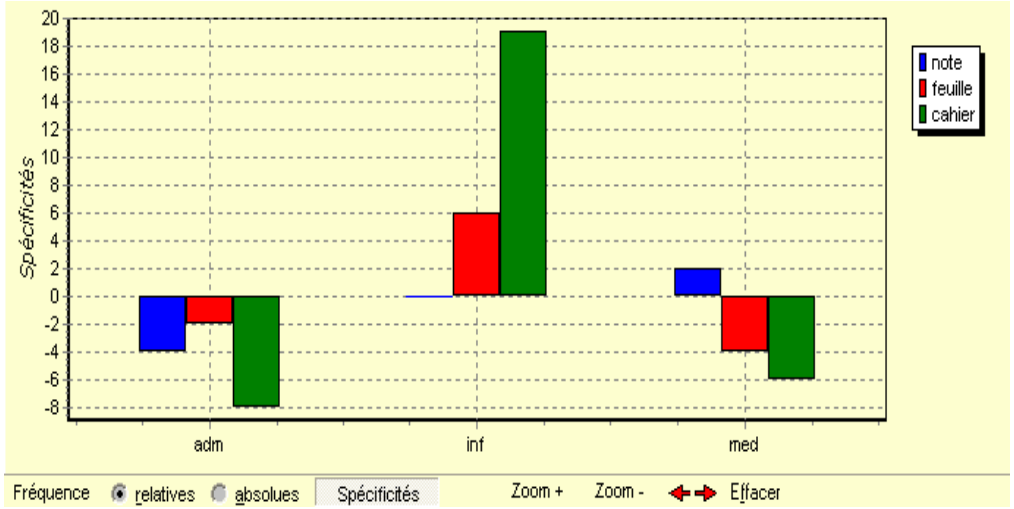


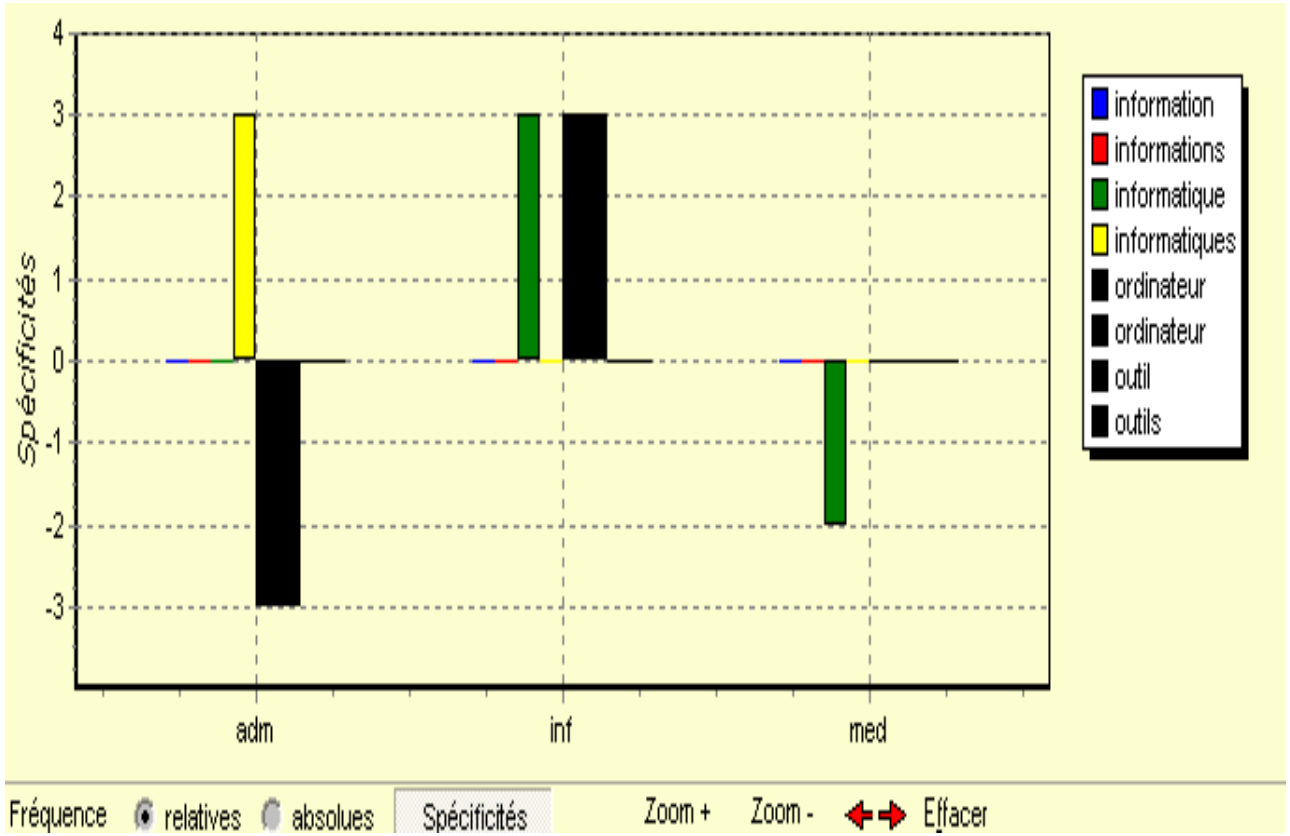


Cahiers/cahier

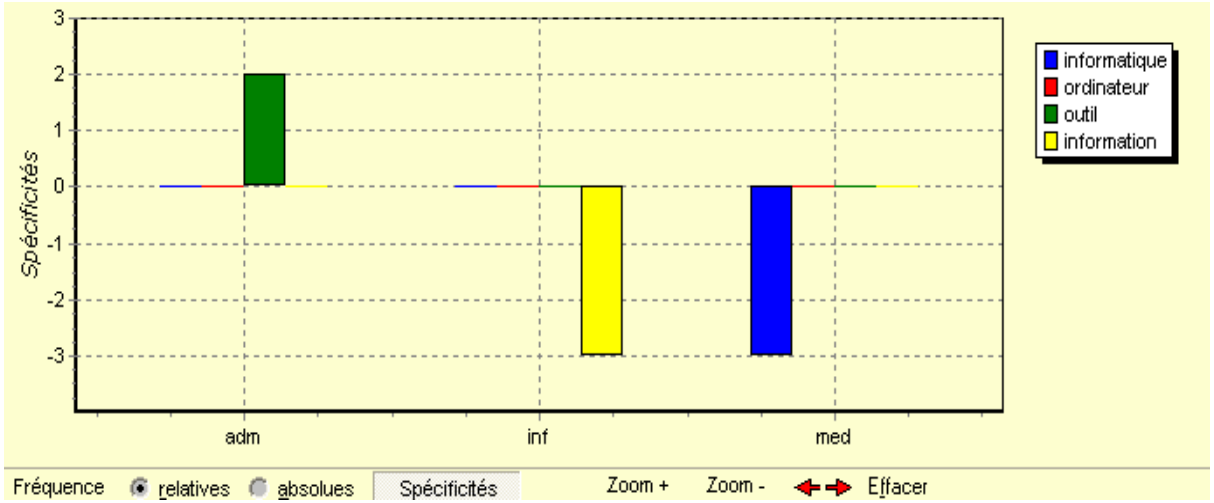
Cahier

Type de spécificité conservé à la limite

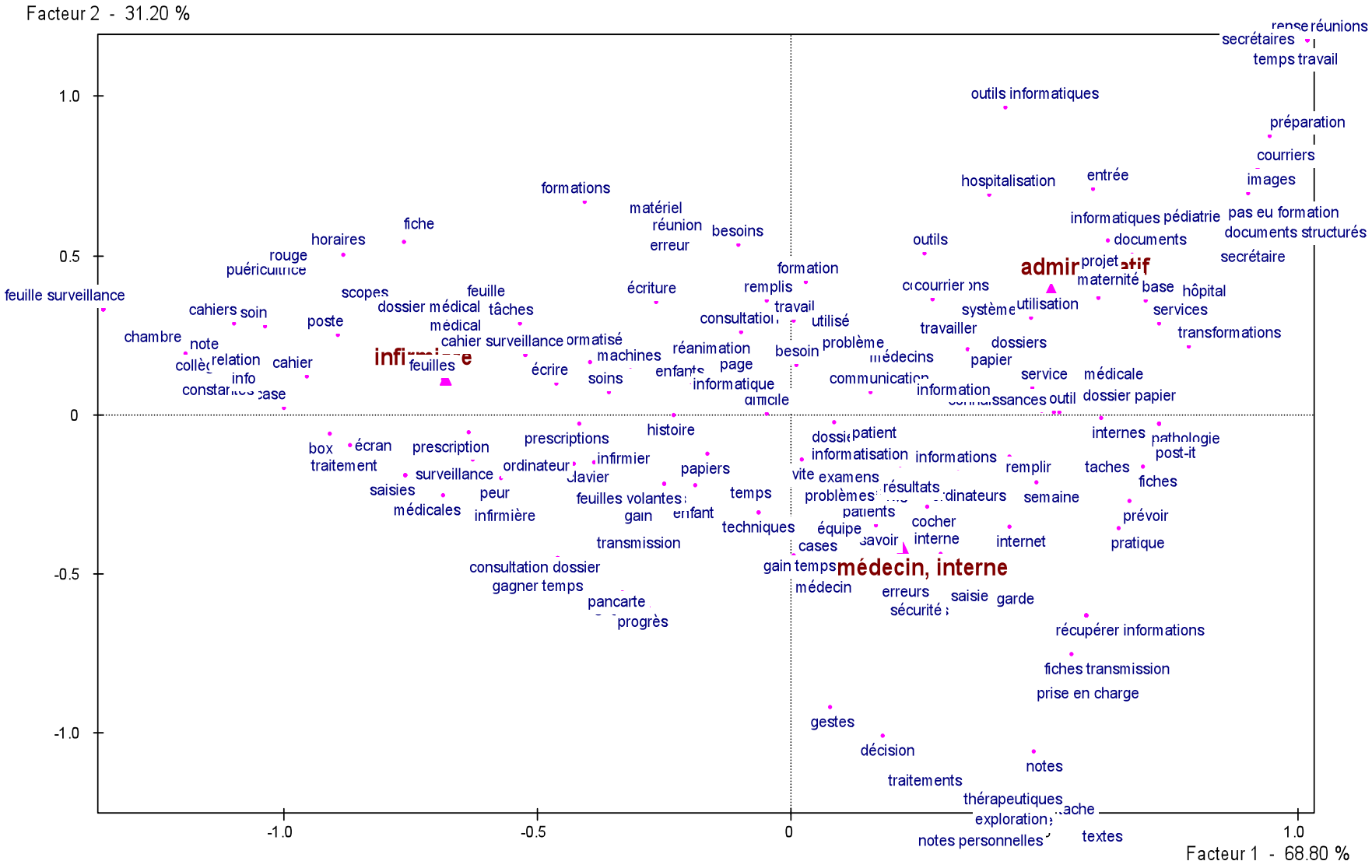




Retournement de situation

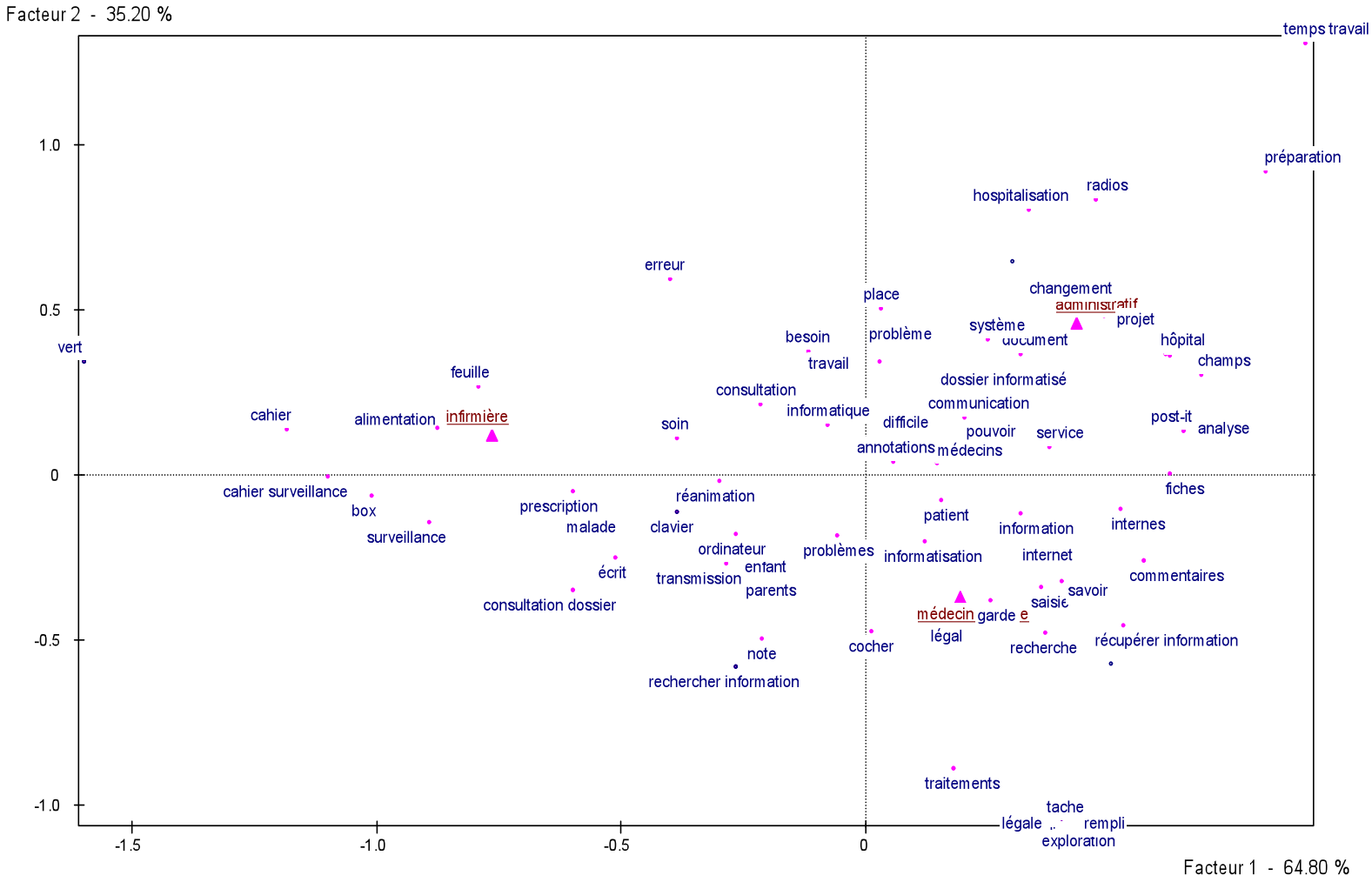


Réponses ouvertes: AFC



Les cartes factorielles ont été obtenues avec le logiciel SPAD 5.5

Réponses ouvertes: formes ramenées au singulier



Structure conservée avec réduction des spécificités.

Conclusion

- Perspective de Recherche documentaire, l'affichage multidimensionnel des formes, la création de dictionnaires par discipline? (Top-Down)

  formes au singulier et tronquées!
- La mise en évidence de spécificités du corpus, voire de typologies supposerait que dans un premier temps les formes aux singulier et au pluriel soient conservées
- Extraction de connaissances à partir des corpus,

  penser aux conclusions de L.LEBART! (Bottom-up)

Bibliographie articles…

AUBIN S., LELU A., " Vers un environnement complet de synthèse statistique de contenus textuels", Neuronav v2, *séminaire ADEST, 2005*

COURTIAL J.P., " Analysis of Social Representations in Action Based on Words Associated by Scientific Articles", European Review of Applied Psychology, 52, 2002, p.221-230

GOUADAIN D., " Les mots de la Gestion ", Gérer et Comprendre, n°66, 2001, p. 58-80, ESKA, Paris.

KRUSKAL J.B, " Multidimensional Scaling By Optimizing Goodness of Fit To a Nonmetric Hypothesis", Psychometrika Vol. 29, N°1, March 1964

LEBART L., " Validation des visualisations de données textuelles" , *Actes des JADT 2004*

REINERT M., " Approche statistique et problème du sens dans une enquête ouverte", *Journal de la Société Française de Statistique*, tome 142, vol 4, 2001

Bibliographie articles…

AUBIN S., LELU A., “ Vers un environnement complet de synthèse statistique de contenus textuels”, Neuronav v2, *séminaire ADEST, 2005*

BRUNER S., pathologies psychologiques: l’emploi du singulier stigmatise un comportement de retrait face au monde, JADT 2004

GOUADAIN D., “ Les mots de la Gestion ”, Gérer et Comprendre, n°66, 2001, p. 58-80, ESKA, Paris.

LEBART L., “ Validation des visualisations de données textuelles” , *Actes des JADT 2004*

PINCEMIN B., Lexicométrie sur corpus étiqueté, JADT 2004

REINERT M., “ Approche statistique et problème du sens dans une enquête ouverte”, *Journal de la Société Française de Statistique*, tome 142, vol 4, 2001

Bibliographie ouvrages

- LEBART L., SALEM A., Statistique textuelle, Dunod, 1994
- MULLER C., Principes et méthodes de statistique lexicale, Hachette, Collection Langues, Linguistique, 1970
- www.cavi.univ-paris3.fr/lexicometrica

  Actes des JADT 2002, 2004,
- Logiciel téléchargeable gratuit:

  LEXICO2