Apprentissage automatique et fouille de données
Institut Galilée, Université Paris 13, 27 avril 2006

Validité des visualisation de données textuelles

Ludovic Lebart,

CNRS, GET-ENST, 46 rue Barrault, 75013, Paris.
lebart@enst.fr

http://egsh.enst.fr/lebart/

PLAN

Partie 1

Visualisation en axes principaux

- Le bootstrap
- Les déclinaisons du bootstrap
- Les niveaux du bootstrap

Partie 2

Visualisation par cartes auto-organisées.

- L'analyse de contiguté
- Le « plan optimal »

1.Validation des visualisations en axes principaux

1.1 LE BOOTSTRAP

Raisons d'utiliser le "bootstrap" :

- Complexité de l'approche analytique
- Se libérer des hypothèses sur les distributions
- S'adapter à toutes les situations

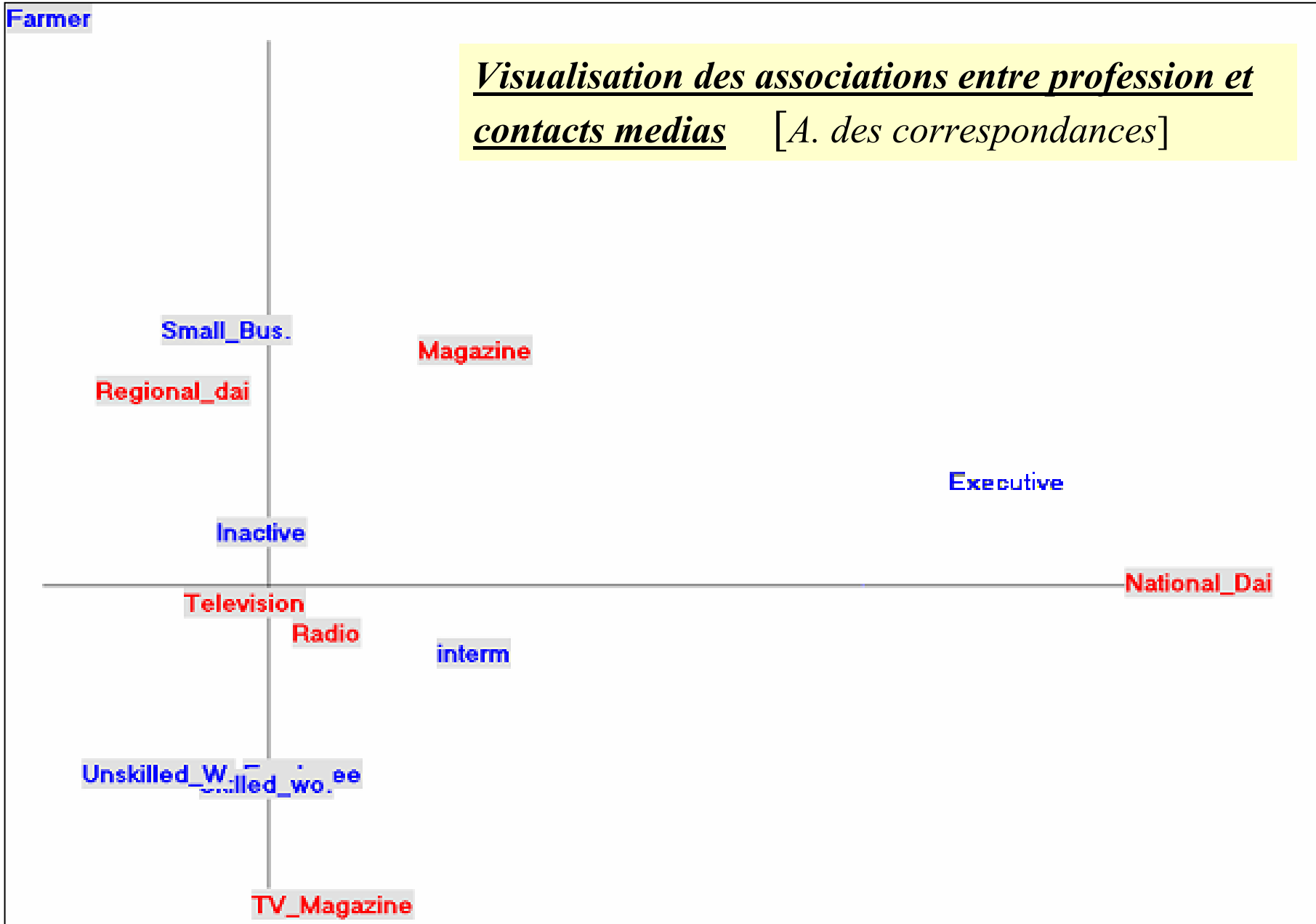
→ *Gifi (1981), Meulman (1982), Greenacre (1984) furent les premiers à proposer l'approche bootstrap dans ce cadre. Il reste cependant très difficile de procéder à des tests sur les valeurs propres. Il existe également plusieursmodalités d'applications selon les contextes.*

Rappel sur le Bootstrap

Exemple : Zones de confiances sur les visualisations.

- Exemple de table de contingence (CESP Multi-Media Survey, 1993).
- Dans chaque case: nombre de medias contactés la veille.
- Colonnes : **Media** [Radio, TV, National & Regional Daily N., Magazines].
- Lignes : **Catégories socio-professionnelles**.

	Radio	Tele	Nat.	Reg.	Maga	TV_Mag
Farmer	96.	118.	2.	71.	50.	17.
Small Business	122.	136.	11.	76.	49.	41.
Executive	193.	184.	74.	63.	103.	79.
Intermediate	360.	365.	63.	145.	141.	184.
Employee	511.	593.	57.	217.	172.	306.
Skilled worker	385.	457.	42.	174.	104.	220.
Unskilled worker	156.	185.	8.	69.	42.	85.
Housewives, Ret.	1474.	1931.	181.	852.	642.	782.


Visualisation des associations entre profession et contacts medias [A. des correspondances]
Farmer
Small_Bus.
Magazine
Regional_dai
Executive
Inactive
National_Dai
Television
Radio
interm
Unskilled_W
ee
TV_Magazine

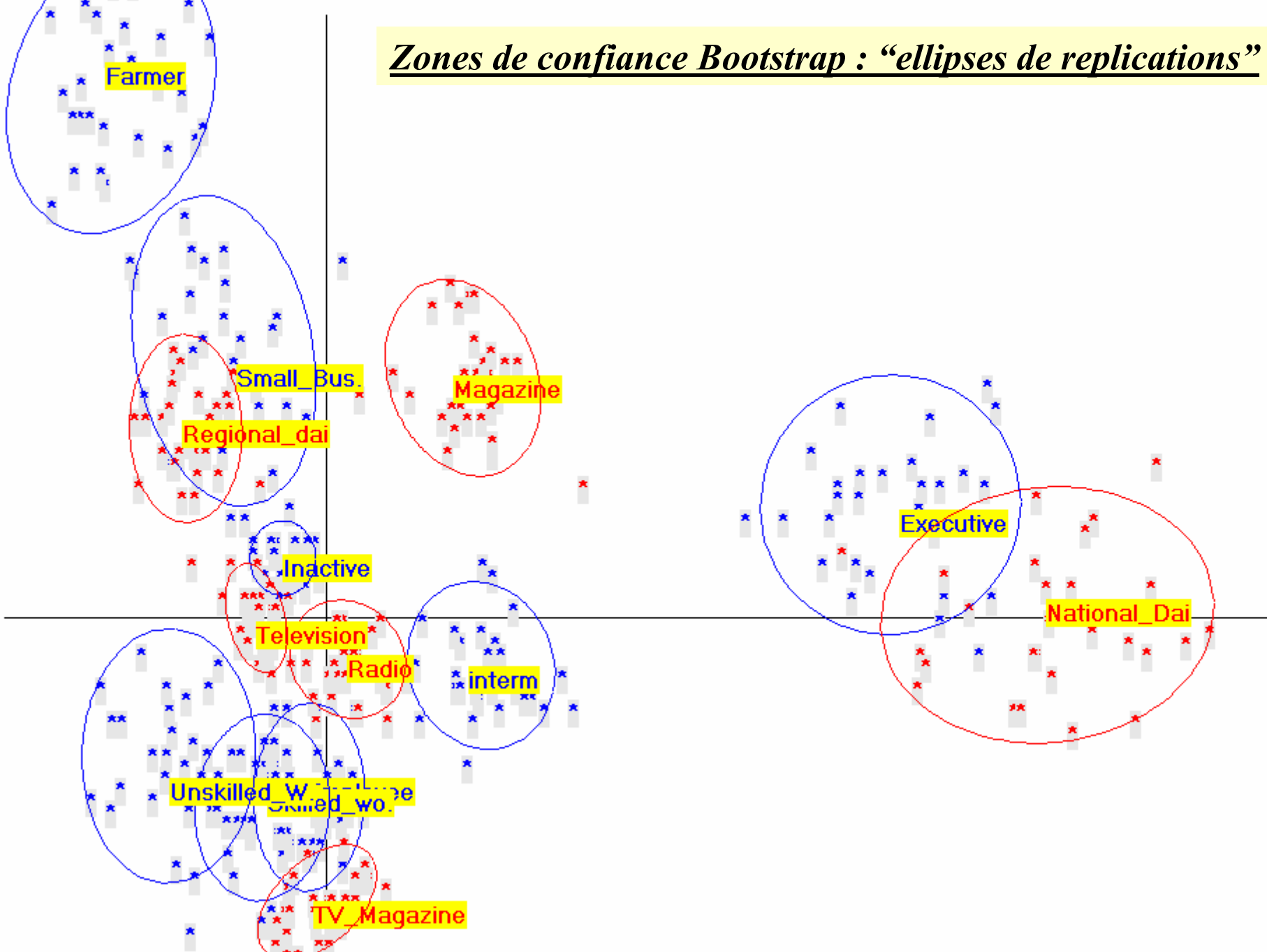
(Rappel sur le Bootstrap)

Exemple : Zones de confiances sur les visualisations

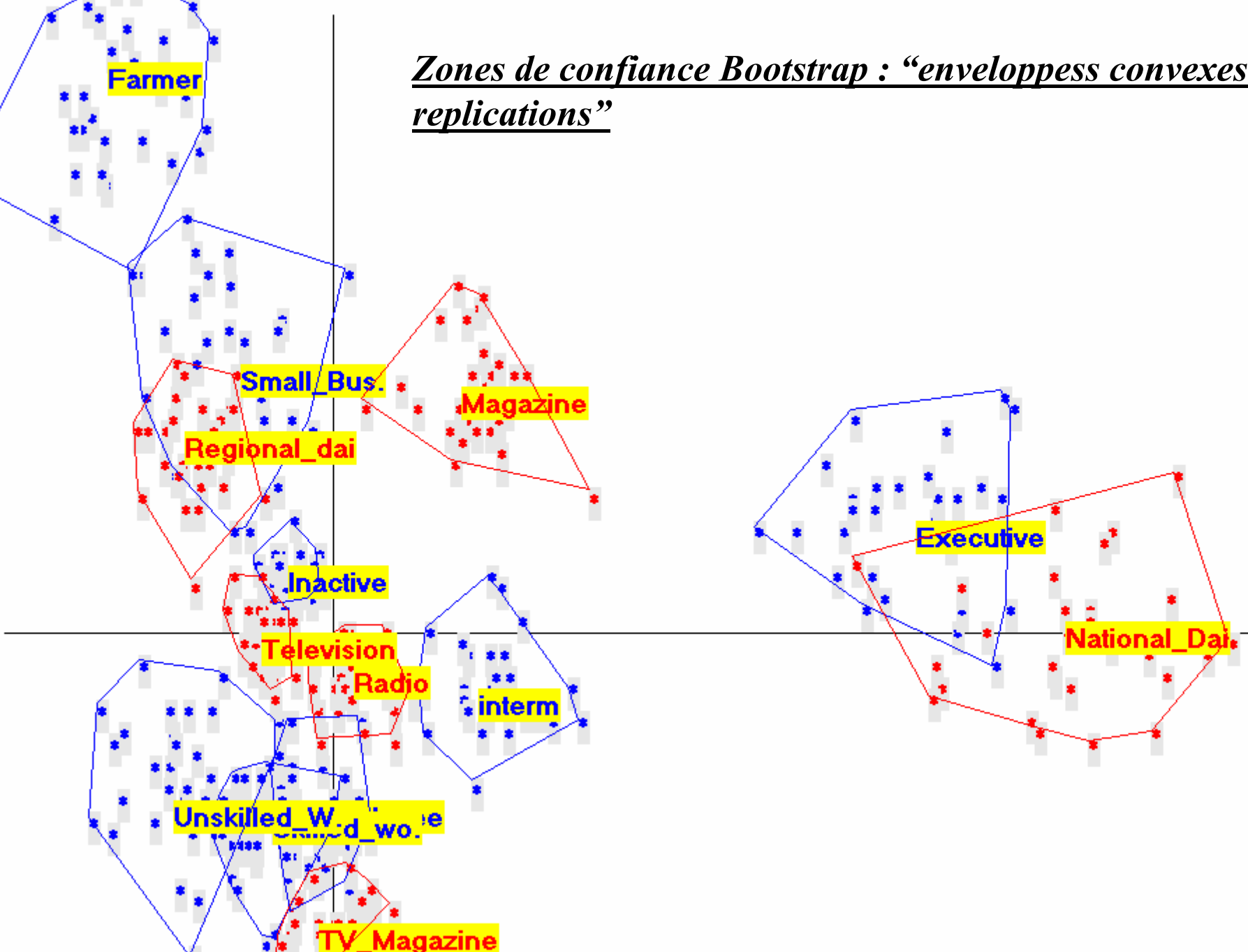
- Exemple de table répliquée

	Radio	Tele	Nat.	Reg.	Maga	TV_M
Farmer	109.	120.	1.	78.	48.	20.
Small Business	126.	142.	8.	76.	53.	30.
Executive	196.	181.	80.	77.	109.	72.
Intermediate	384.	365.	60.	133.	138.	203.
Employee	514.	596.	59.	228.	172.	316.
Skilled worker	378.	467.	33.	171.	100.	223.
Unskilled worker	169.	188.	8.	79.	38.	81.
Housewives, Ret.	1519.	1961.	158.	893.	632.	764.

(Tirage avec remise des 12 000 individus de cette table, dont les 48 cases définissent les 48 "couleurs" des boules l'urne)


Zones de confiance Bootstrap : “ellipses de replications”
Farmer
Small_Bus.
Regional_dai
Magazine
Executive
National_Dai
Inactive
Television
Radio
interm
TV_Magazine

Zones de confiance Bootstrap : "enveloppess convexes d
replications"



1.2 LES DECLINAISONS DU BOOTSTRAP

Le bootstrap partiel

La technique de bootstrap que l'on appellera *bootstrap partiel* (sans recalcul des valeurs propres) proposée notamment par Greenacre (1984) dans le cadre de l'analyse des correspondances, répond à plusieurs des préoccupations des utilisateurs dans le cas de l'analyse en composante principales.

Une réplication consiste en un tirage avec remise des n individus (vecteurs-observations), suivi du positionnement des p nouvelles variables ainsi obtenues en "variables supplémentaires" sur les q premiers axes de l'analyse de base.

Les procédures décrites ci-dessus peuvent être mises en oeuvre avec un programme classique de projection d'éléments supplémentaires.

On calcule donc les réplications de ce coefficient, ce qui revient à repondérer les individus avec les "poids Bootstrap" 0, 1, 2, ... qui caractérisent un tirage sans remise.

Trois types de « bootstrap total »

Bootstrap total de type 1 (épreuve sévère, très pessimiste) : simple changement (éventuel) de signes des axes homologues pour les réplications.
Il s'agit seulement de remédier au fait que les axes sont définis au signe près.
Un simple produit scalaire entre axes originaux et axes répliqués de mêmes rangs permet de rectifier le signe de ces derniers.

Bootstrap total de type 2 (épreuve assez sévère, plutôt pessimiste) : changement de signe et correction des interversions d'axes. Les axes répliqués sont affectés (séquentiellement, sans remise en cause d'affectations antérieures) du rang des axes originaux avec lesquels ils sont les plus corrélés en valeur absolue.
Puis on procède à un éventuel changement de signe des axes, comme en bootstrap de type 1.

Bootstrap total de type 3 (épreuve plutôt laxiste si on s'intéresse à la stabilité des axes, mais apte à décrire la stabilité des sous-espaces de dimension supérieure à 1) : une rotation dite procrustéenne (cf. Gower et Dijksterhuis, 2004) permet de rapprocher de façon optimale les système d'axes répliqués et les systèmes d'axes initiaux.

- Le bootstrap total de type 1 ignore les possibles interversions d'axes et rotations d'axes. Il permet de valider des structures stables et robustes. Chaque réplication doit produire les axes initiaux avec les mêmes rangs (ordre des valeurs propres).

- Le bootstrap total de type 2 est idéal si on veut valider des axes, c'est-à-dire des dimensions cachées, sans attacher une importance particulière aux rangs de celles-ci.

- Enfin le bootstrap de type 3 permet de valider globalement un sous-espace engendré par les axes principaux correspondant aux premières valeurs propres. Comme le bootstrap partiel, le bootstrap total de type 3 peut être qualifié de laxiste par les utilisateurs qui s'intéressent à l'individualité des axes, et pas seulement aux sous-espaces engendrés par plusieurs axes consécutifs.

Exemple (texte anglais)

Question ouverte :
"What is the single most important thing in life for you?"
Suivie par la relance : *"What other things are very important to you?".*

Question incluses dans une enquête internationale auprès de sept pays (Japon, France, Allemagne, Italie, Hollande, U K, USA) vers 1990 (Hayashi *et al*., 1992).

L 'exemple concerne le volet anglais de l 'enquête (taille d 'échantillon : 1043).

Le bilan de la première phase de codage numérique est :

Pour *1043* réponses, il y a *13 669* occurrences (*tokens*),

avec *1 413* mots distincts (*types*).

Si l 'on ne retient que les mots apparaissant au moins *16* fois, il reste *10 357* occurrences de ces mots (*tokens*), avec *135* mots distincts (*types*).

Exemple d'une table lexicale

Listage partiel d'une table lexicale croisant 135 mots apparaissant au moins 16 fois avec 9 catégories âge-education

	L-30	L-55	L+55	M-30	M-55	M+55	H-30	H-55	H+55
I	*2*	*46*	*92*	*30*	*25*	*19*	*11*	*21*	*2*
I'm	*2*	*5*	*9*	*3*	*2*	*1*	*0*	*0*	*0*
a	*10*	*56*	*66*	*54*	*44*	*19*	*20*	*22*	*7*
able	*1*	*9*	*16*	*9*	*7*	*4*	*4*	*5*	*0*
about	*0*	*3*	*13*	*7*	*1*	*2*	*4*	*1*	*0*
after	*1*	*8*	*11*	*3*	*1*	*2*	*0*	*0*	*0*
all	*1*	*24*	*19*	*8*	*18*	*6*	*3*	*5*	*2*
and	*8*	*89*	*148*	*86*	*73*	*30*	*25*	*32*	*13*
anything	*0*	*4*	*9*	*1*	*3*	*0*	*1*	*1*	*0*

• Les diapositives suivantes montrent le plan principal de l'analyse
•des correspondances de la table lexicale précédente.

• La proximité entre 2 points- categorie (colonnes) signifie similarité des profils lexicaux des 2 catégories.

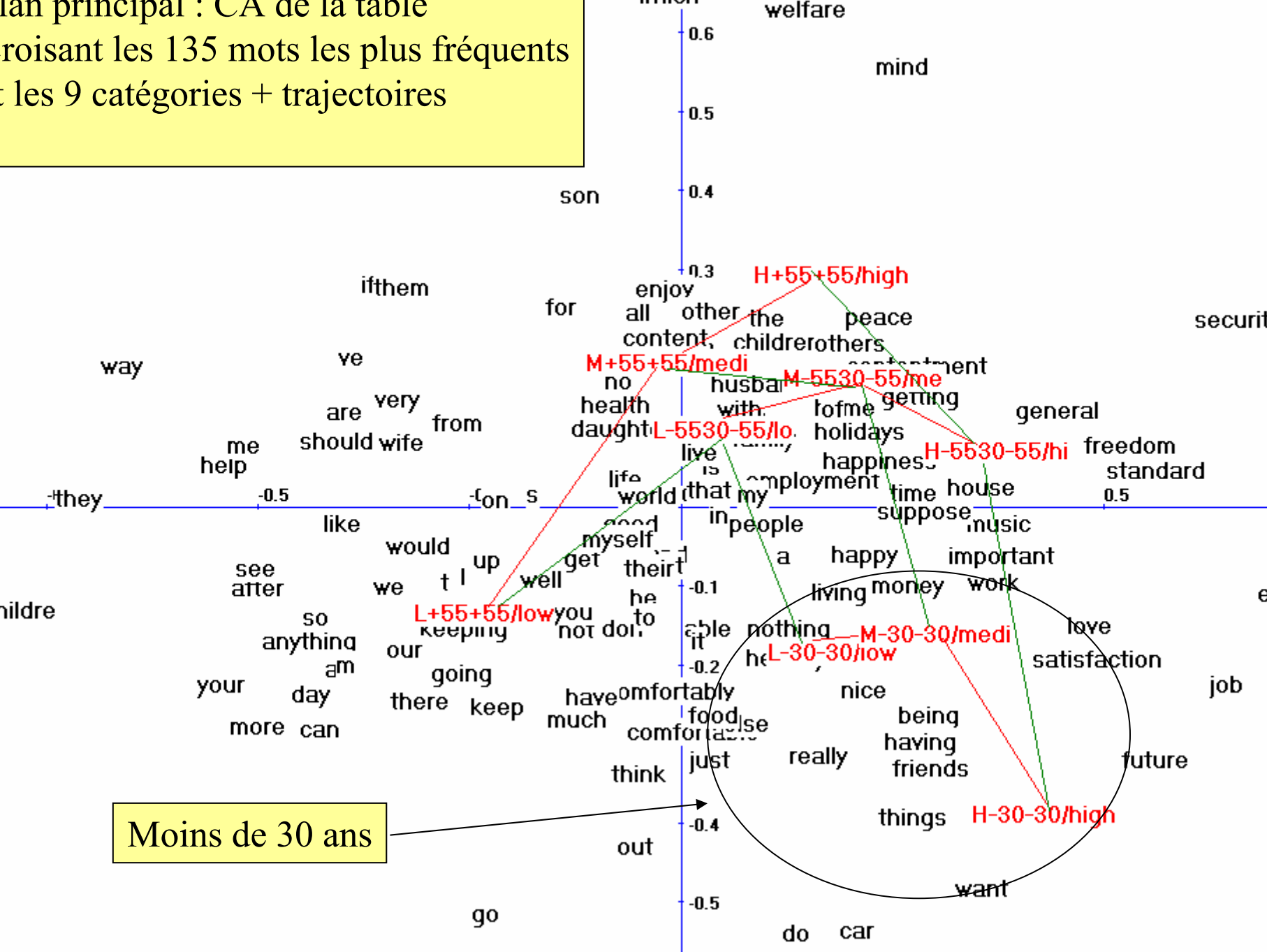
• La proximité entre 2 points- mots (lignes) signifie similarité des profils lexicaux de ces mots.

• Ellipses et enveloppes convexes décrivent l 'incertitude.

• 9 points categories, en rouge (toutes les categories, en fait)
• (L = low, M = medium, H = high)

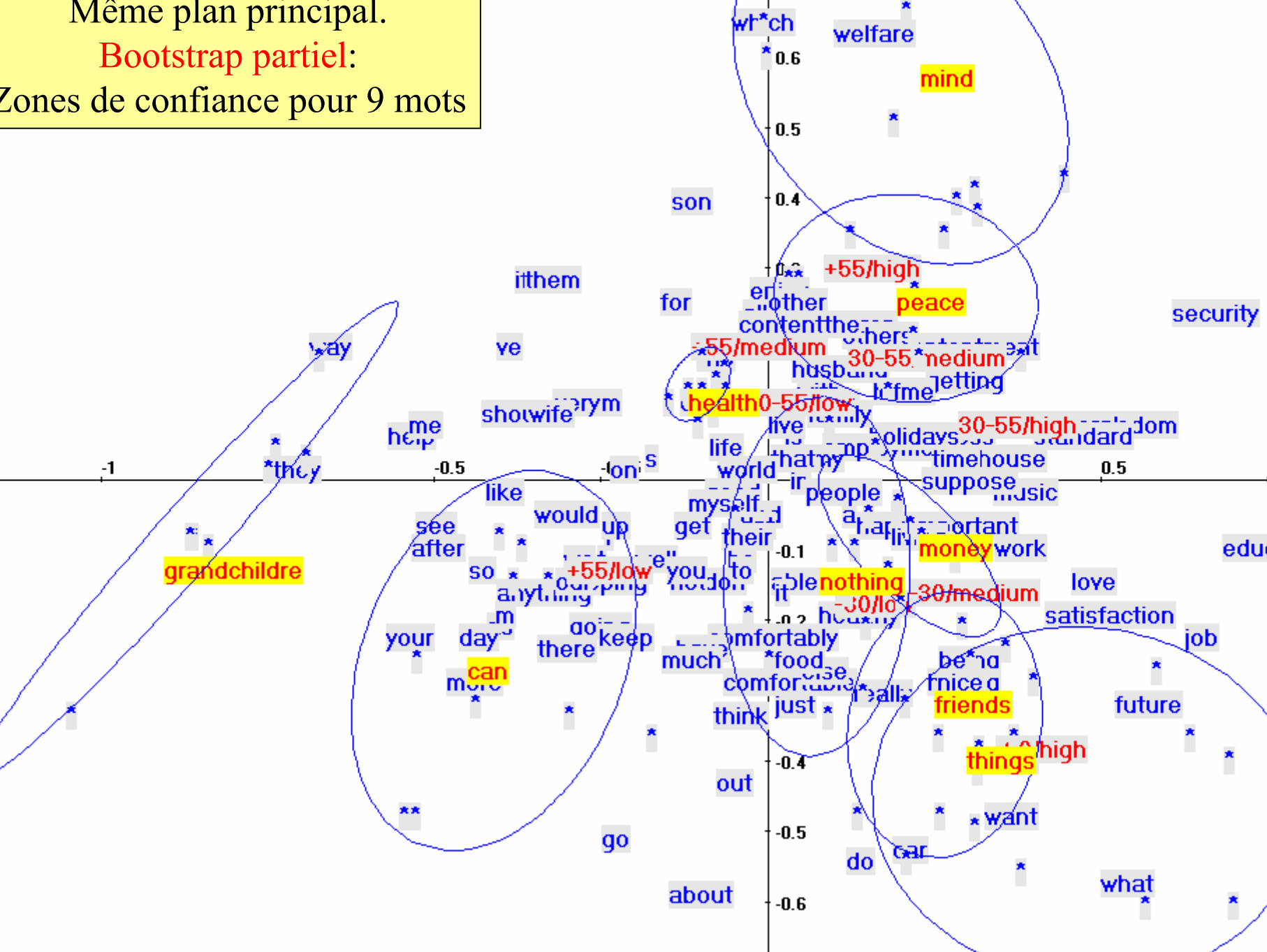
• 6 points mots (graphies), en bleu.

lan principal : CA de la table
roisant les 135 mots les plus fréquents
t les 9 catégories + trajectoires

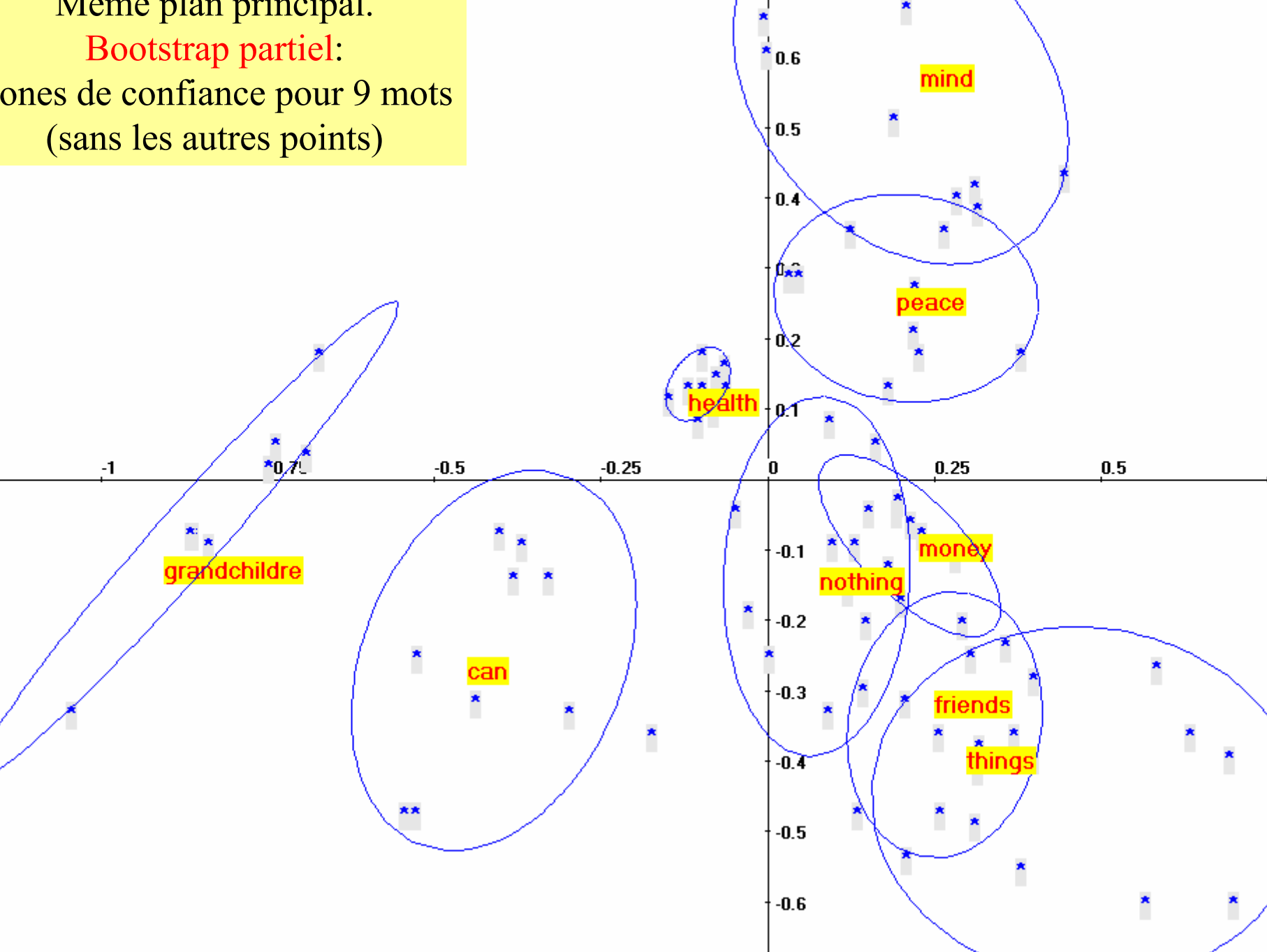


Moins de 30 ans

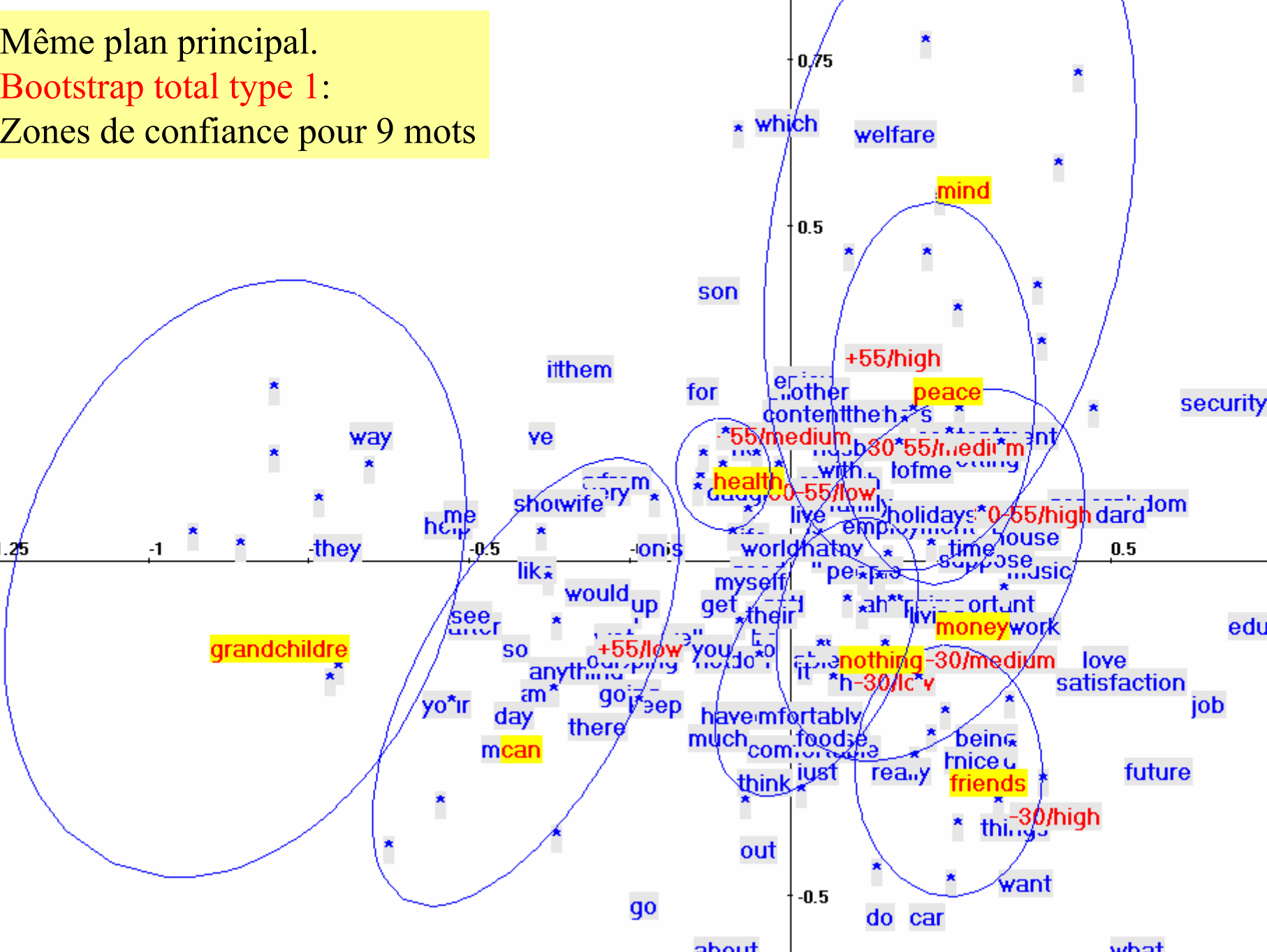
Même plan principal.
Bootstrap partiel:
Zones de confiance pour 9 mots



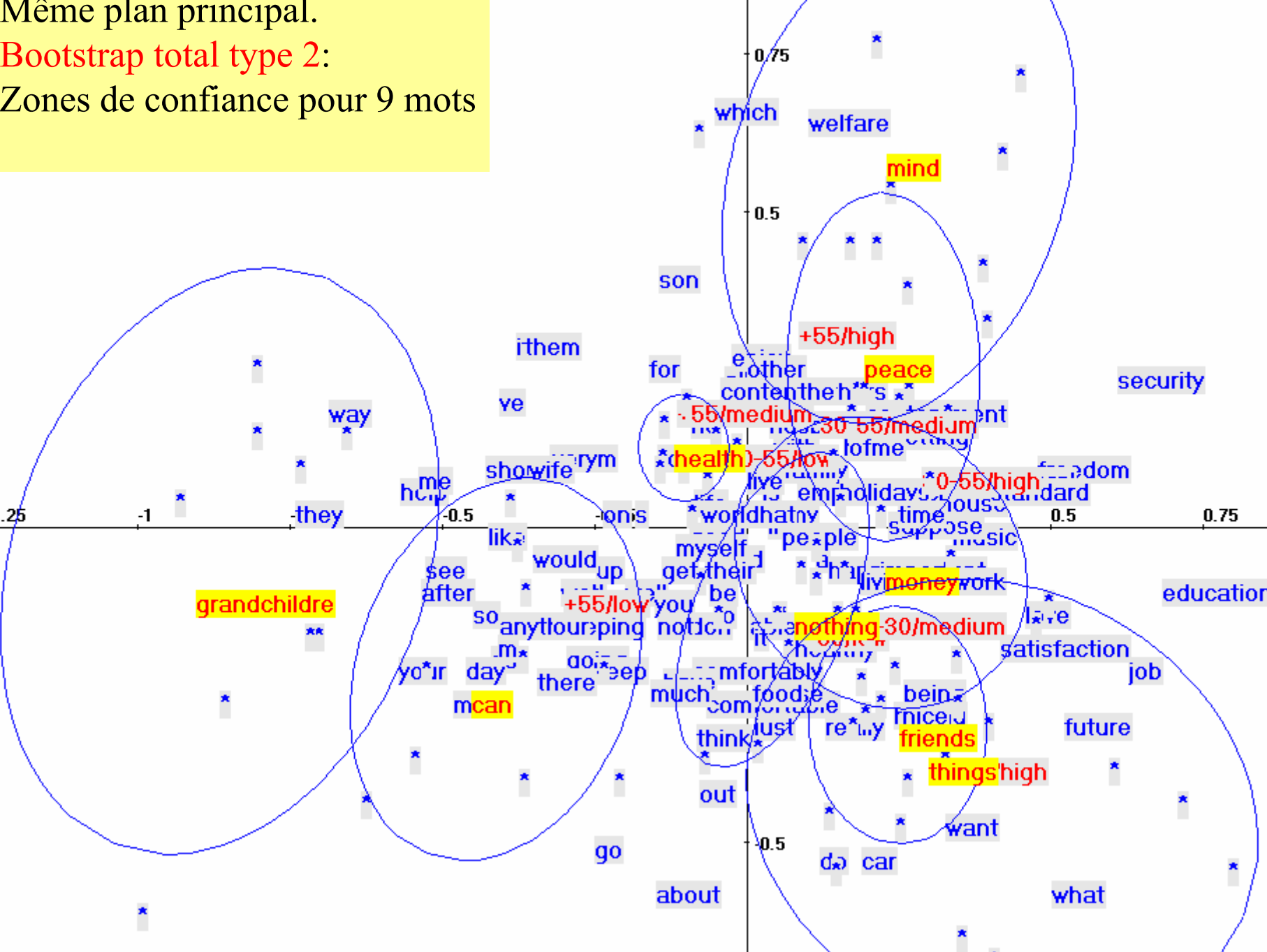
Même plan principal.
Bootstrap partiel:
ones de confiance pour 9 mots
(sans les autres points)



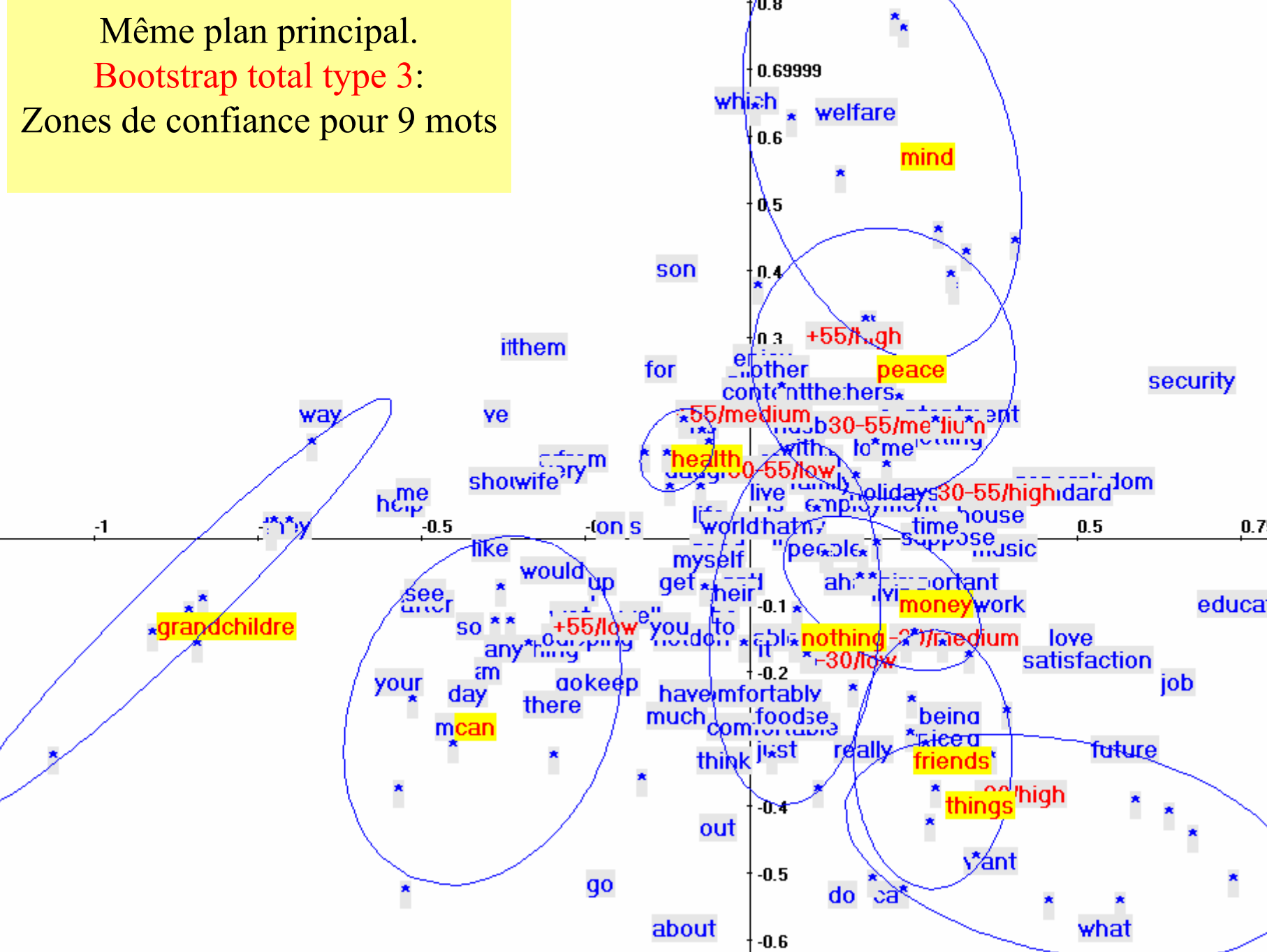
Même plan principal.
Bootstrap total type 1:
Zones de confiance pour 9 mots



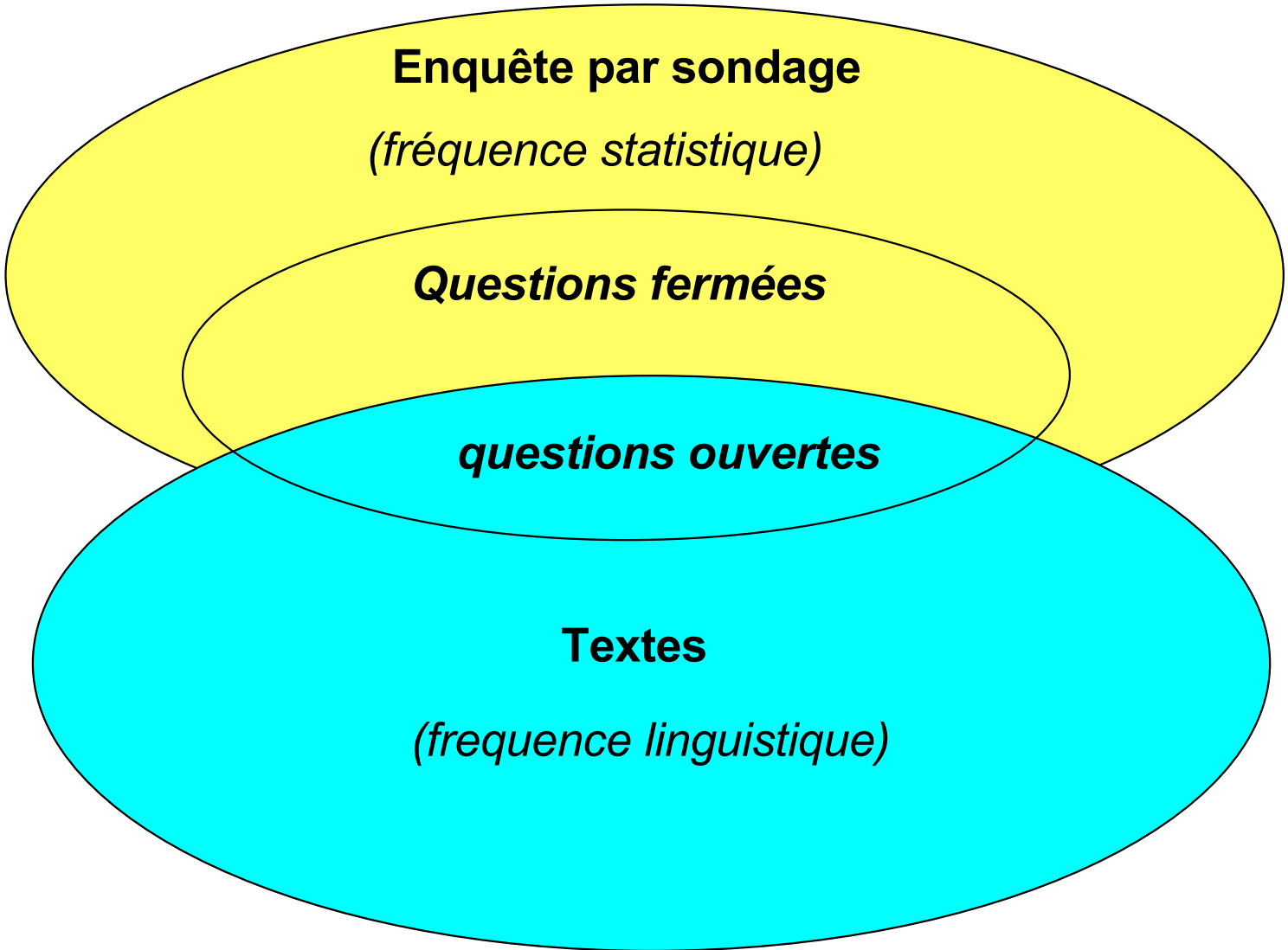
Même plan principal.
Bootstrap total type 2:
Zones de confiance pour 9 mots



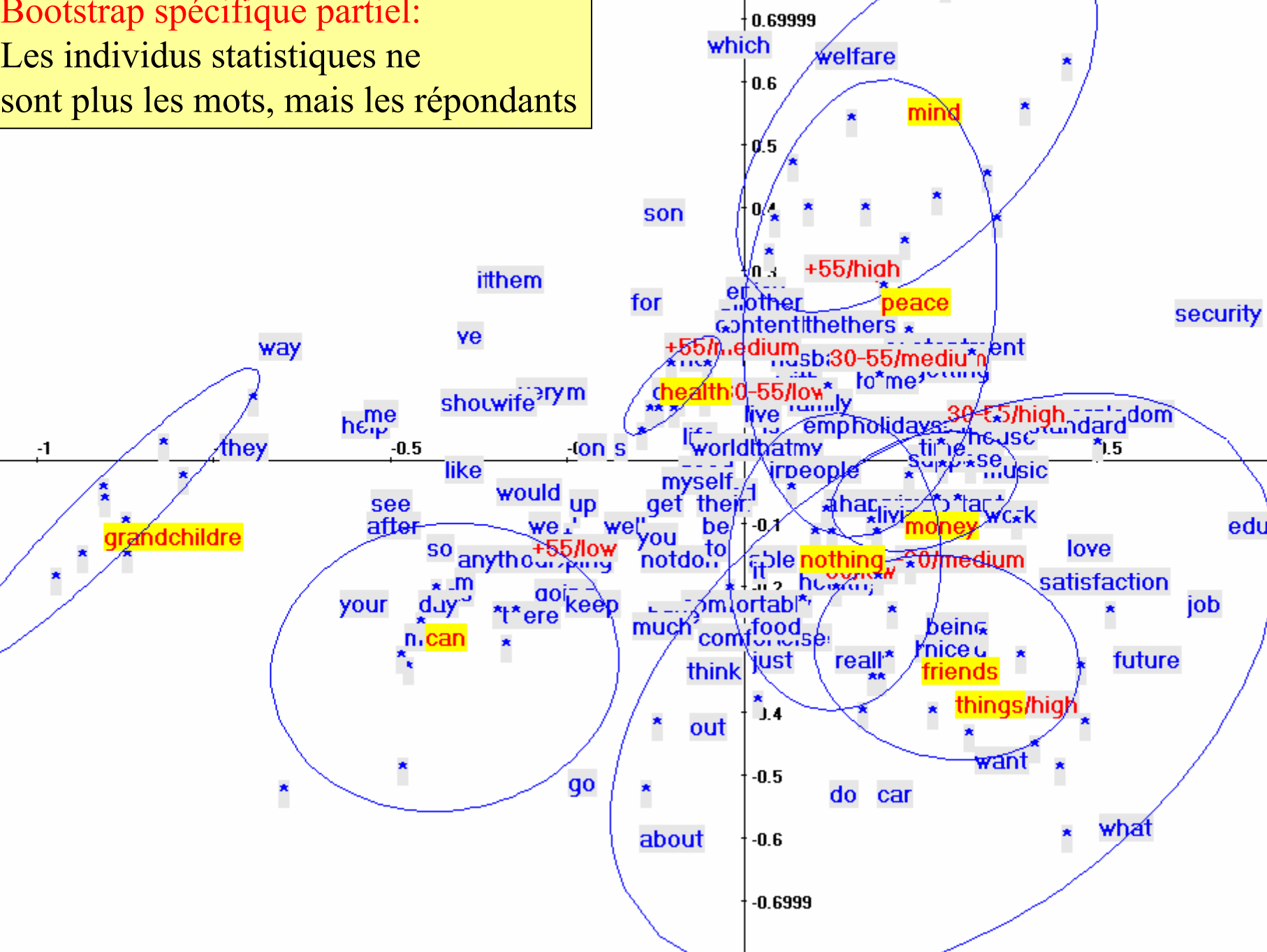
Même plan principal.
Bootstrap total type 3:
Zones de confiance pour 9 mots



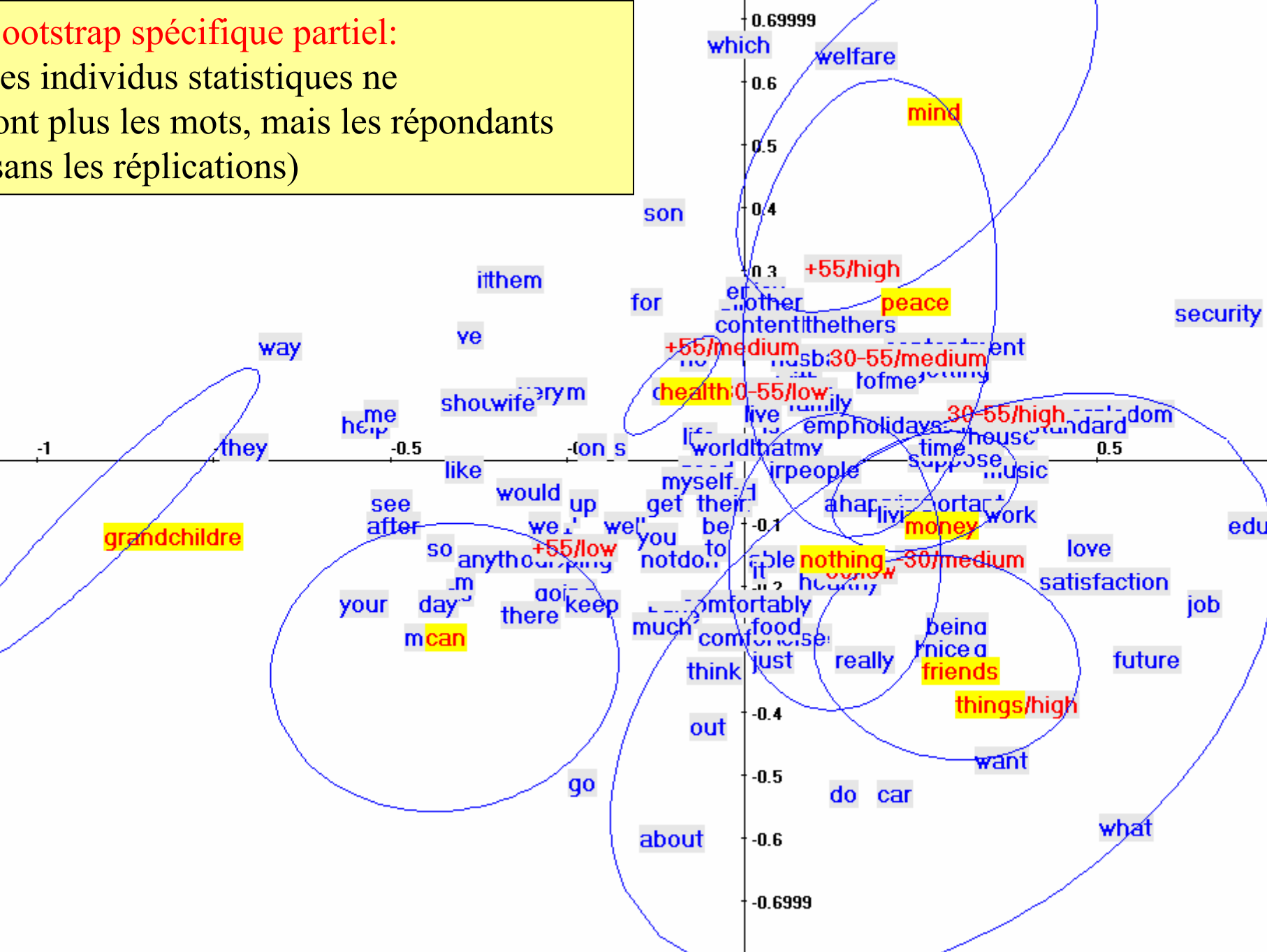
1.3 LES NIVEAUX DU BOOTSTRAP;
Fréquence statistique *versus* « fréquence linguistique »



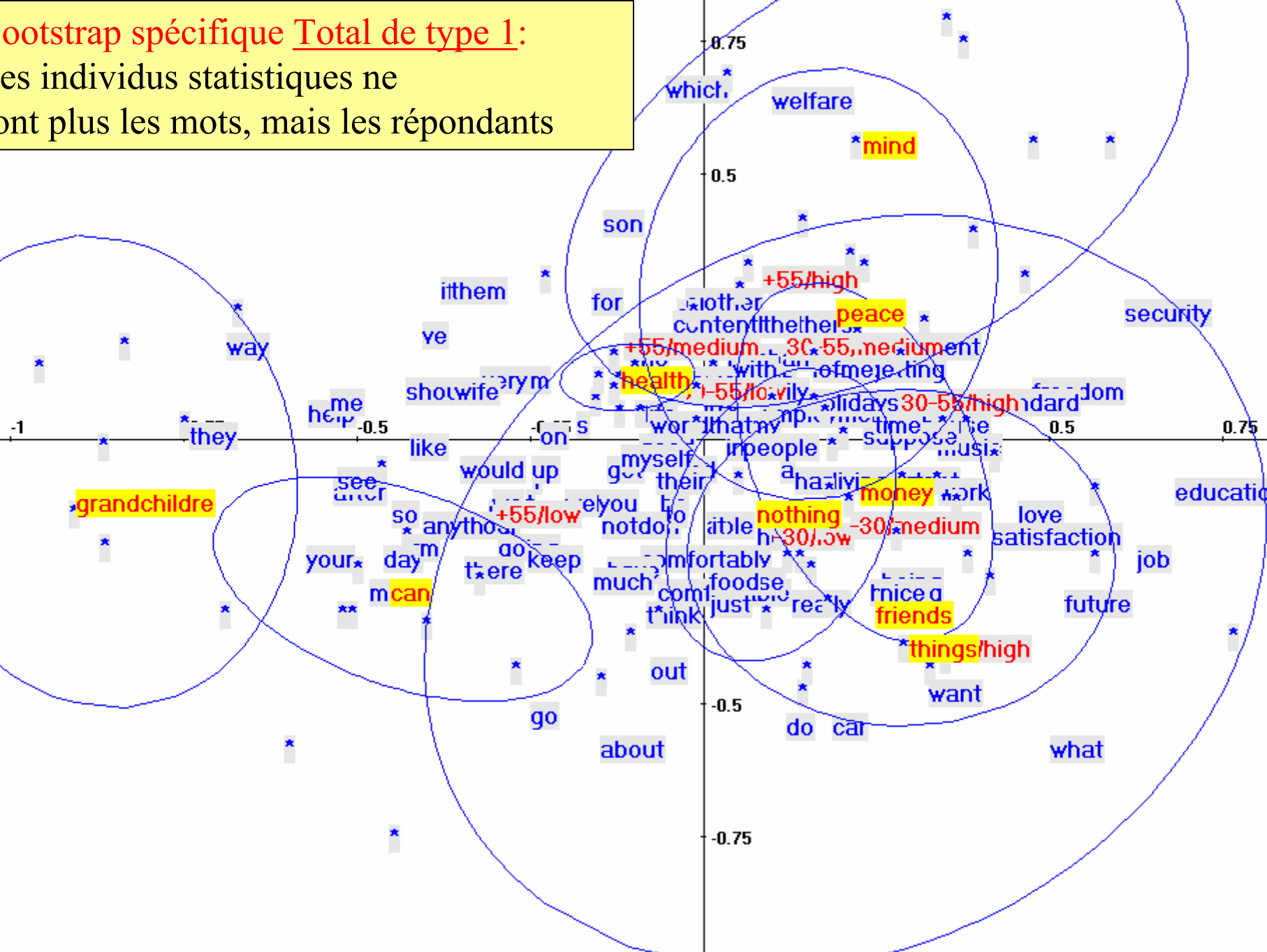
Bootstrap spécifique partiel:
Les individus statistiques ne sont plus les mots, mais les répondants



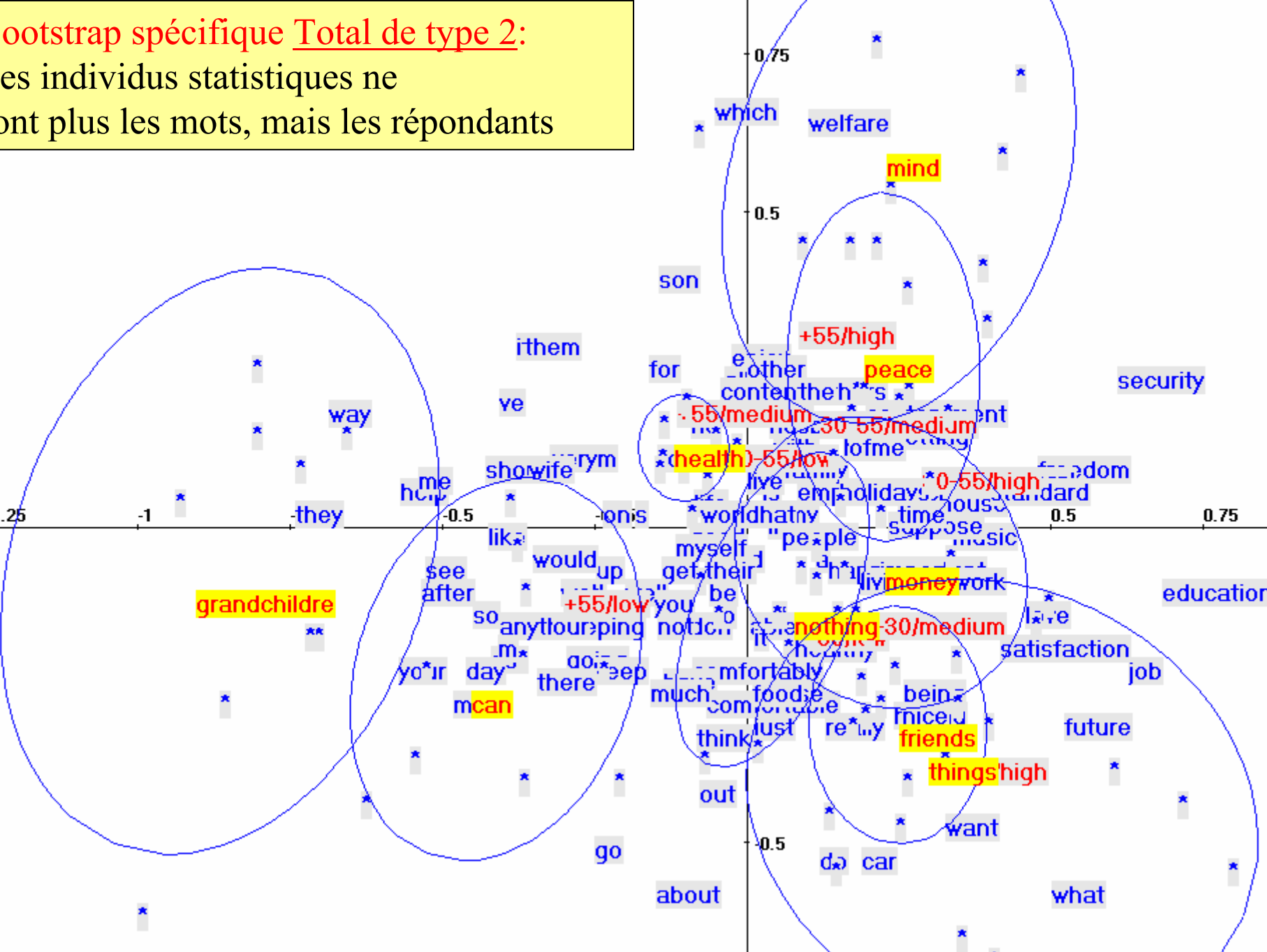
ootstrap spécifique partiel:
es individus statistiques ne
ont plus les mots, mais les répondants
ans les réplications)

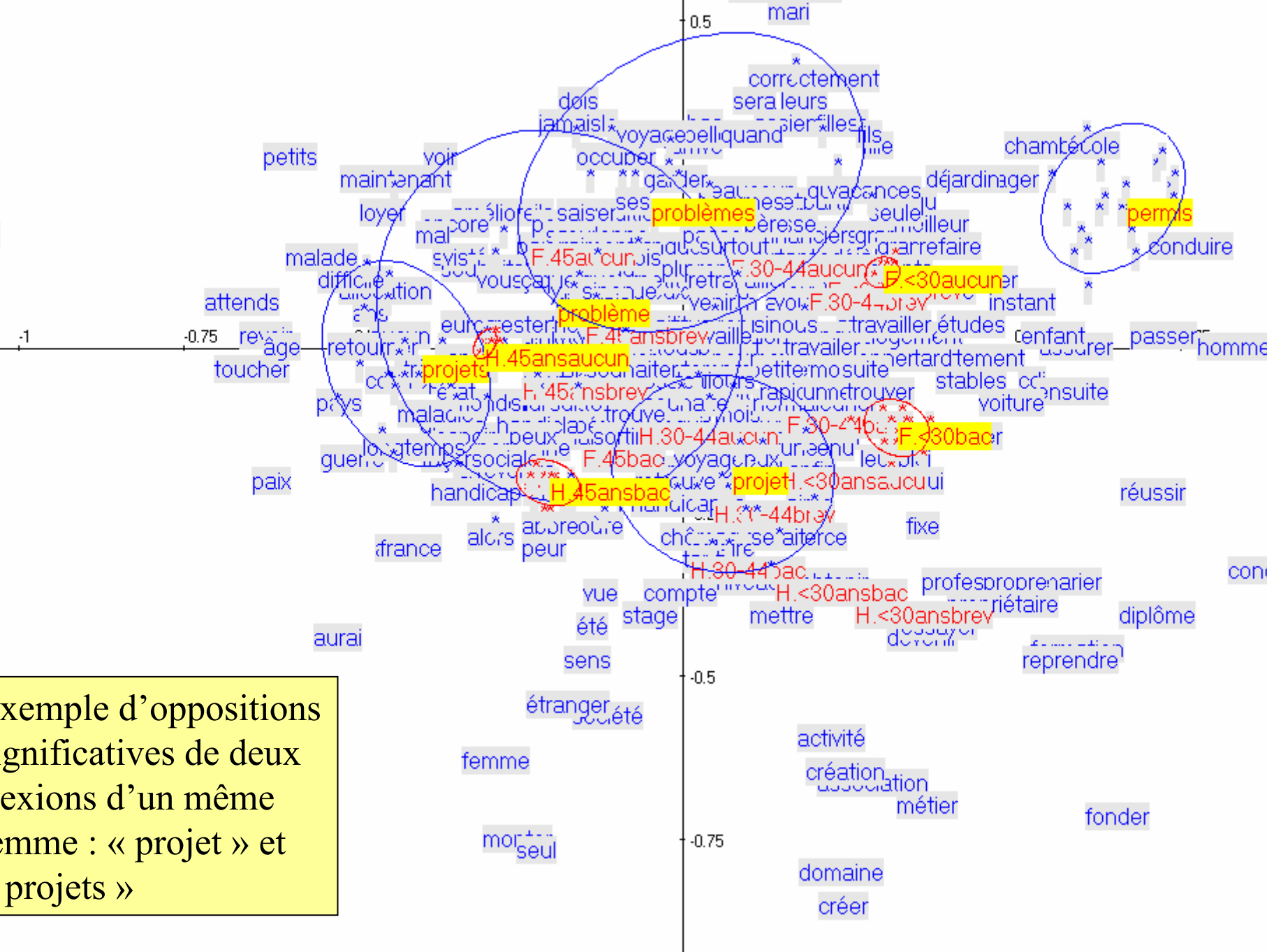


ootstrap spécifique Total de type 1:
es individus statistiques ne
ont plus les mots, mais les répondants



ootstrap spécifique Total de type 2:
es individus statistiques ne
ont plus les mots, mais les répondants





xemple d'oppositions
gnificatives de deux
exions d'un même
mme : « projet » et
projets »

Partie 2
Visualisation par cartes auto-organisées.
(Cartes de Kohonen ou SOM)

5 approches pour voir et classer…

(A) Construire la partition en s'efforçant d'optimiser un critère, puis, dans un second temps, représenter les classes dans un graphique plan d 'ACP (ou AC).

———————————

(B) Construire la partition en s'efforçant d'optimiser un critère, puis construire la représentation en tenant compte de la partition déjà trouvée (A. Discrim.)

———————————

(C) Construire simultanément la partition et la représentation, ce qui induit des contraintes sur la partition, mais peut conduire à une meilleure représentation. (Cartes auto organisées de Kohonen, ou: SOM)

———————————

(D) Une variante de l'approche précédente consiste à projeter les classes (ou leurs enveloppes convexes) dans le plan (1,2) d'une analyse de contiguïté faite à partir d'un k graphe des k plus proches voisins « symétrisé ».

———————————

(E) Réaliser une analyse de contiguité en prenant comme graphe externe une carte auto-organisée (SOM) de façon à approcher la grille par un plan…

———————————

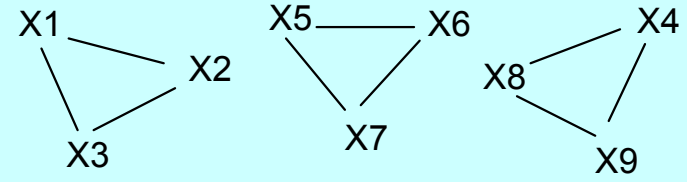
2.1 Rappel: Analyse de contiguïté

On considère, ***n* objets** décrits par ***p*** variables, conduisant à une matrice **Y**, dont les lignes ont une structure de graphe *a priori*.

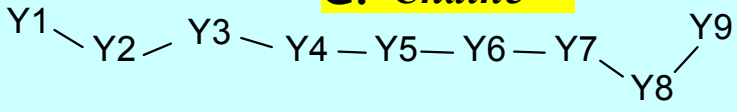
Les ***n* objets** sont les sommets d'un graphe symétrique G dont la matrice (***n***, ***n***) associée est **M** .

($\mathbf{m}_{ii'} = 1$ si les sommets i et i' sont joints par une arête, $\mathbf{m}_{ii'} = 0$ sinon).

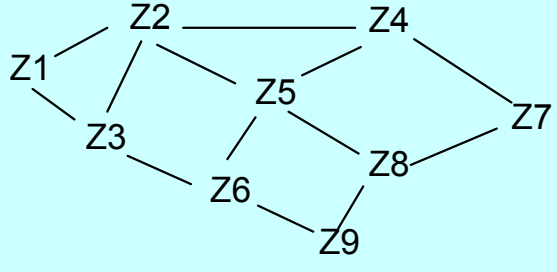
G: ***Partition: 3 cliques***



G: ***Chaîne***



G: ***Cas Général : relation binaire***



Les n **objets** (lignes de Y) sont les sommets d 'un graphe symétrique G dont la matrice associée (n, n) est **M** .

$m_{ii'} = 1$ si les sommets i et i' sont joints par une arête, $m_{ii'} = 0$ sinon.

$m = \sum m_{ii'}$ (nombre d 'arêtes du graphe G)

Variance locale $v^c(y) = (1/2m) \sum m_{ii'} (y_i - y_{i'})^2$

Variance globale $v(y) = (1/2n(n-1)) \sum (y_i - y_{i'})^2$

(La variance empirique est un cas particulier de la variance locale losque le graphe est complet, i.e.: $m_{ii'} = 1$ pour tout i et i ')

Le coefficient de contiguïté (Geary, 1954; après Moran et Von Neumann)

$$c(y) = v^*(y) / v(y)$$

« Correction » de la définition de la variance locale, nouveau coefficient de contiguïté

Nouvelle variance locale

$$m_i^* = (1/n_i)\sum_k m_{ik} y_k$$

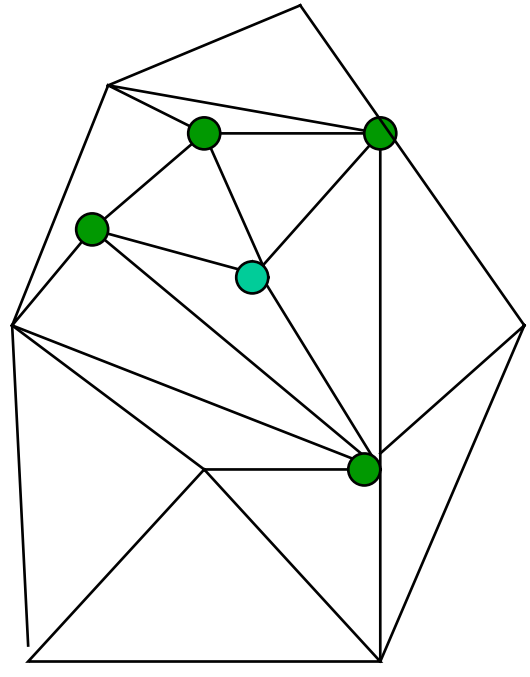
$$v^*(y) = (1/n)\sum (y_i - m_i^*)^2$$

Nouveau coefficient de contiguïté

$$c(y) = v^*(y) / v(y)$$

Avec, comme d'habitude :

$$v(y) = 1/n \sum_{i=1}^{n} (y_i - m\ )^2$$



La matrice diagonale **N** (matrice des degrés) est telle que :

$$n_i = \Sigma_k m_{ik}$$

c(y) s ’écrit, en notations matricielles (**U** = matrice associée au graphe complet):

$$c(y) = \mathbf{y'}\,(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'\,(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\,\mathbf{y} \;/\; \mathbf{y'}\,(\mathbf{I} - (1/n)\mathbf{U})\,\mathbf{y}$$

La *(p, p)* matrice de covariance locale **V*** est définie comme :

$$\mathbf{V^*} = (1/n)\,\mathbf{Y'}(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})'\,(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})\,\mathbf{Y}$$

Cette matrice définit un puissant outil de mesure de corrélation partielles, si le tableau Y est n tableau de variables instrumentales.

Généralisation aux observations multivariées

Soit **Y'u** le vecteur des *n* valeurs de la combinaison linéaire *u* des *p* variables.

Son coefficient de contiguïté vaut alors :

$$c(u) = \mathbf{u'\,V^*\,u} \;/\; \mathbf{u'\,V\,u}$$

… où : $\mathbf{V^*} = (1/n)\,\mathbf{Y'(I - N^{-1}M)'(I - N^{-1}M)Y}$

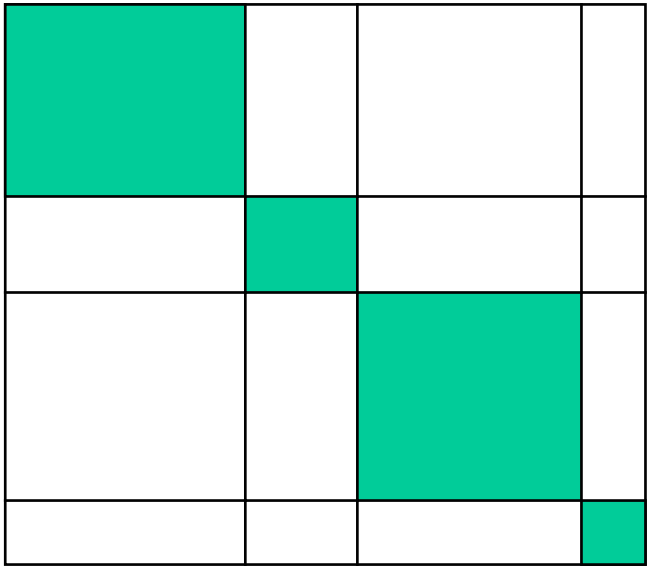
est la matrice *(p, p)* de covariance locale.

L ’analyse de contiguïté est la recherche du minimum de *c(u):*

$$c(u) = \mathbf{u'\,V^{*}\,u \;/\; u'\,V\,u}$$

Elle se réduit à une Analyse Discriminante de Fisher quand G est associé au graphe d’une partition.

M =



= 1

= 0

C(u) nous permet de travailler avec des classes empiétantes, des partitions floues.

Développement à partir d'un exemple :
Visualisation en Sémiométrie

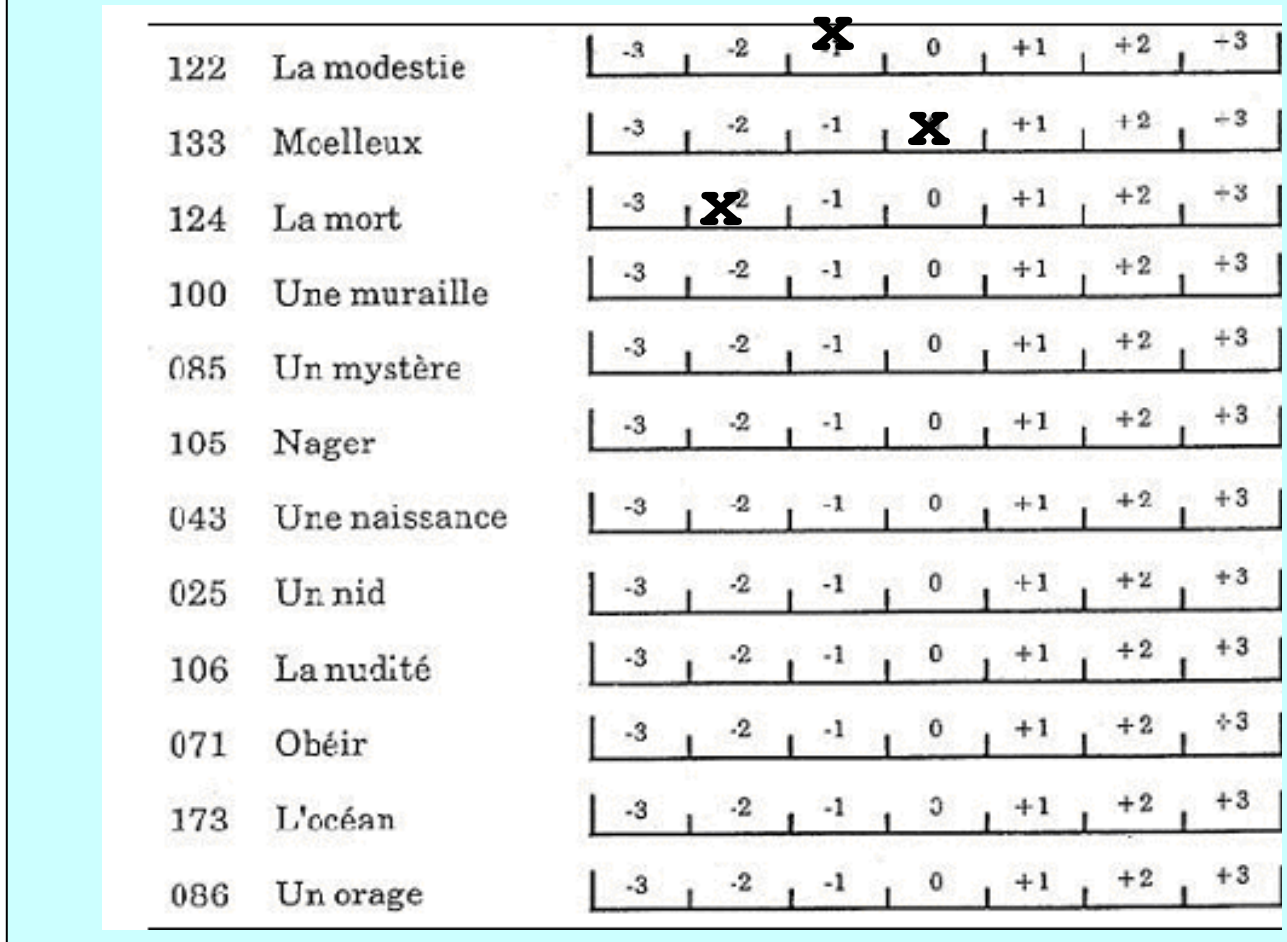
- Idée de base:
Questionnaire de 210 mots, version abrégée 70 mots.

Notes de 1 à 7

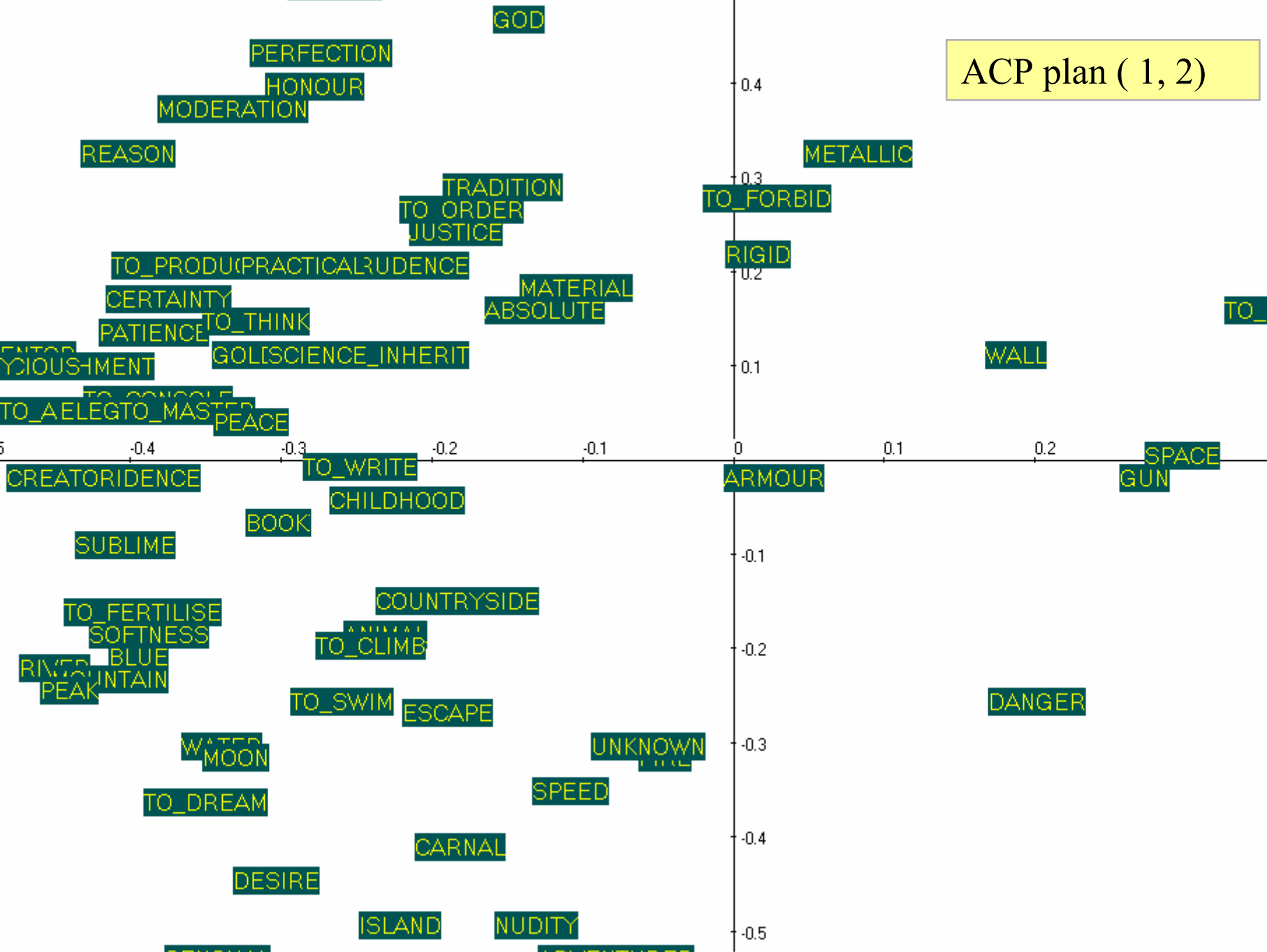
(Très agréable à très désagréable)

Questionnaires en 5 langues

FRENCH	ENGLISH	GERMAN	SPANISH	ITALIAN
l'absolu	absolute	absolut	el absoluto	l'assoluto
l'acharnement	persistence	hartnaeckig	el empeno	l'accanimento
acheter	to buy	kaufen	comprar	comprare
admirer	to admire	bewundern	admirar	ammirare
adorer	to love	anbeten	adorar	adorare
l'ambition	ambition	der ehrgeiz	la ambicion	l'ambizione
l'âme	soul	die seele	el alma	l'anima
l'amitié	friendship	die freundschaft	la amistad	l'amicizia
l'angoisse	anguish	die angst	la angustia	l'angoscia
un animal	animal	ein tier	un animal	un animale
un arbre	tree	ein baum	un arbol	un albero
l'argent	silver	das geld	el dinero	il denaro
une armure	armour	die ruestung	una armadura	un'armatura
l'art	art	die kunst	el arte	l'arte

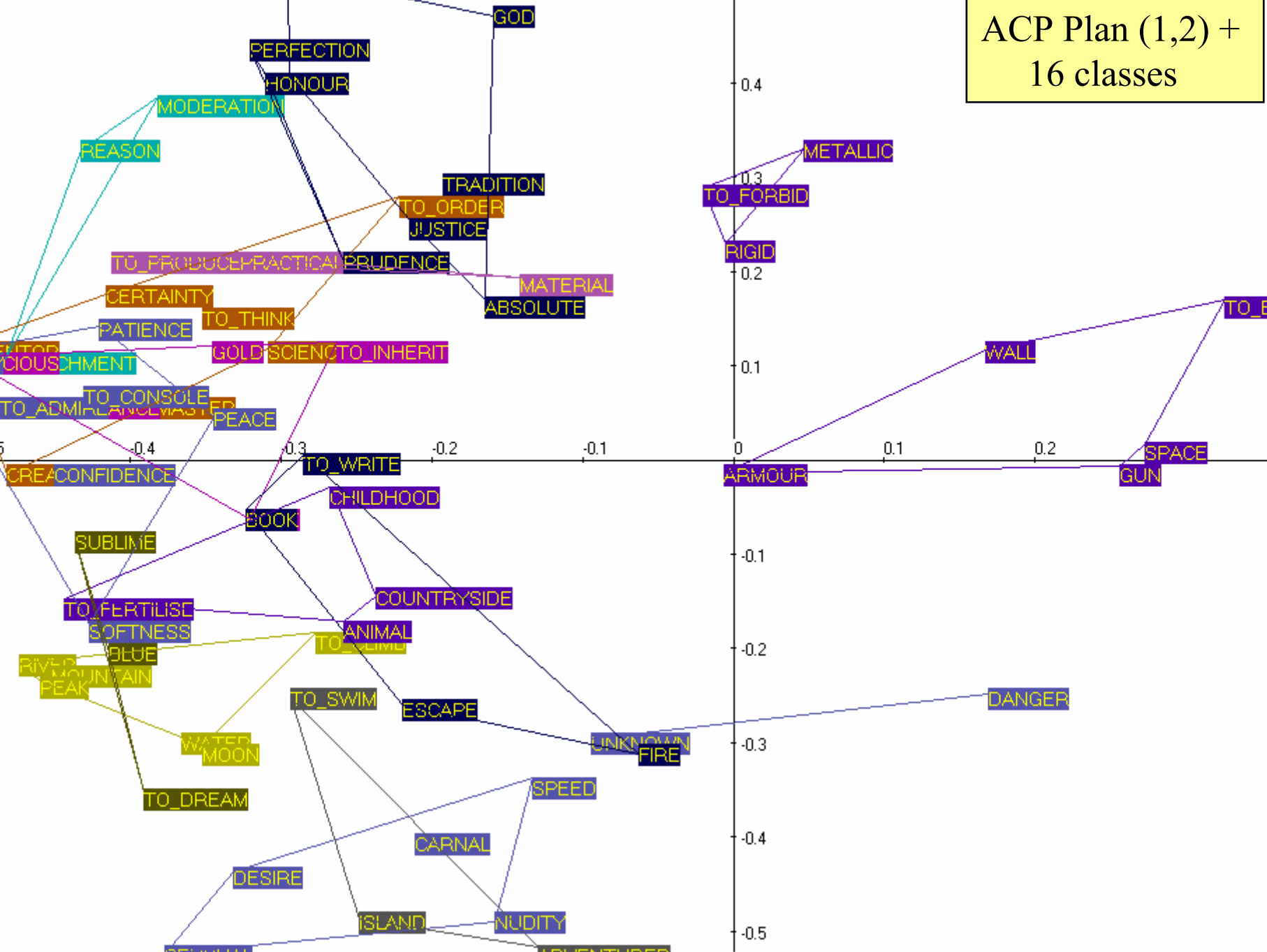


Facsimile d'un questionnaire


ACP plan (1, 2)
GOD
PERFECTION
HONOUR
MODERATION
REASON
METALLIC
TRADITION
TO_ORDER
JUSTICE
TO_FORBID
RIGID
TO_PRODU(PRACTICALRUDENCE
CERTAINTY
MATERIAL
ABSOLUTE
PATIENCE
TO_THINK
GOLISCIENCE_INHERIT
WALL
PEACE
TO_WRITE
CREATORIDENCE
ARMOUR
SPACE
GUN
CHILDHOOD
BOOK
SUBLIME
COUNTRYSIDE
TO_FERTILISE
SOFTNESS
TO_CLIMB
BLUE
PEAK
TO_SWIM
ESCAPE
DANGER
MOON
UNKNOWN
SPEED
TO_DREAM
CARNAL
DESIRE
ISLAND
NUDITY
0.4
0.3
0.2
0.1
-0.1
-0.2
-0.3
-0.4
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2

S.O.M. + Axe 1

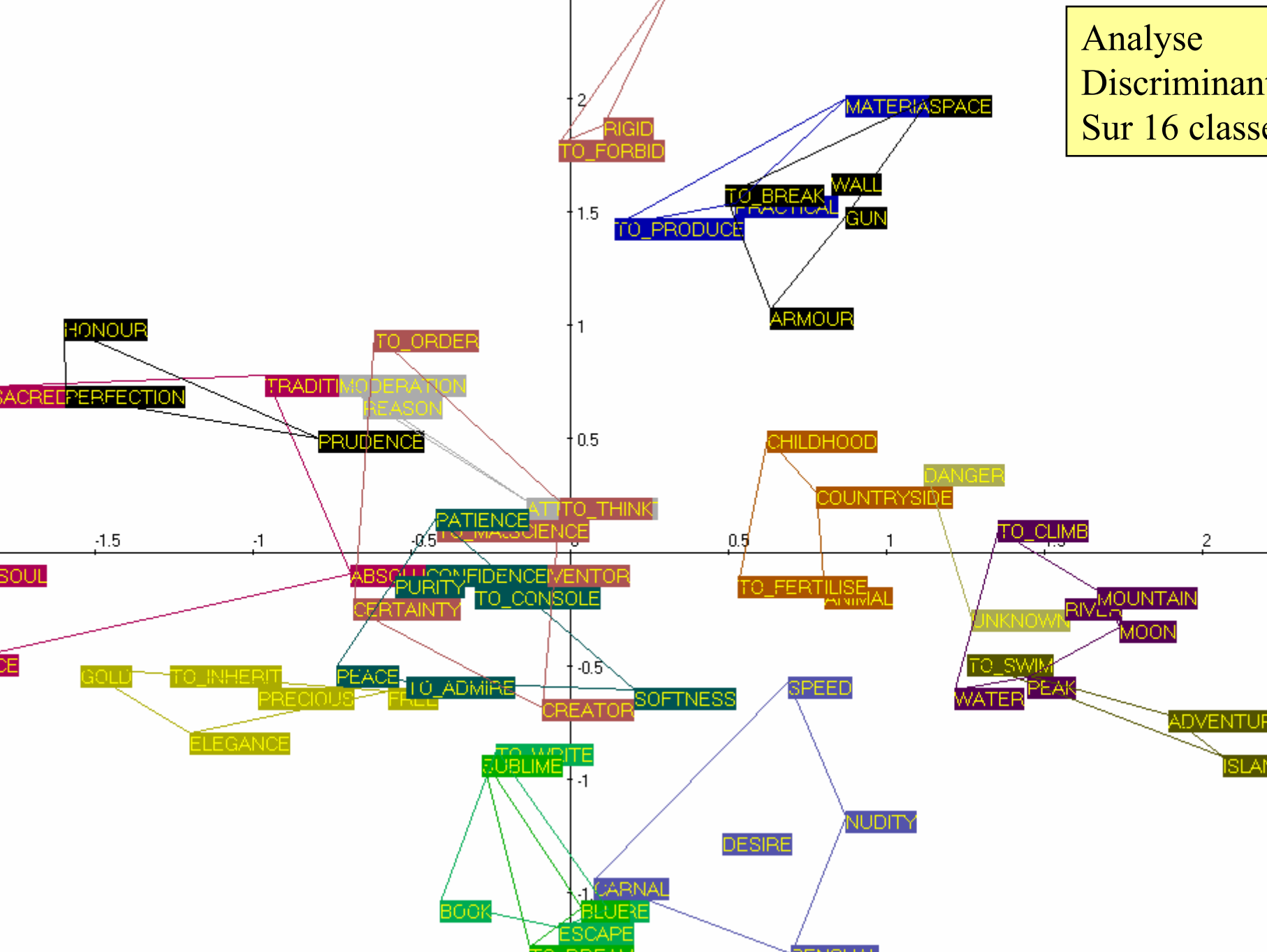
_DREAM			WALL
NSUAL			TO_BREAK
DITY			SPACE
SIRE	UNKNOWN	TO_FORBID	RIGID
RNAL	SPEED	MATERIAL	METALLIC
VENTURER	FIRE	DANGER	GUN
			ARMOUR
TER			
ON	SUBLIME	TO_ORDER	
AND	ESCAPE	PRUDENCE	
_SWIM			TRADITION
_CLIMB			SACRED
ER			PERFECTION
AK	TO_ADMIRE	TO_MASTER	NOBLE
UNTAIN	PRECIOUS	TO_INHERIT	HONOUR
E	FREE	GOLD	GOD
MAL	ELEGANCE	CERTAINTY	ABSOLUTE
	TO_PRODUCE		
	TO_CONSOLE	TO_THINK	
	PURITY	SCIENCE	
	PRACTICAL	REASON	
_FERTILISE	PEACE	MODERATION	
FTNESS	PATIENCE	INVENTOR	TO_WRITE
UNTRYSIDE	CONFIDENCE	CREATOR	SOUL
LDHOOD	ATTACHMENT	BOOK	JUSTICE


ACP Plan (1,2) + 16 classes
GOD
PERFECTION
HONOUR
MODERATION
REASON
TRADITION
TO_ORDER
JUSTICE
PRUDENCE
MATERIAL
ABSOLUTE
CERTAINTY
TO_THINK
PATIENCE
GOLD
SCIENC
TO_INHERIT
CIOUS
CHMENT
TO_CONSOLE
PEACE
CONFIDENCE
TO_WRITE
CHILDHOOD
BOOK
SUBLIME
COUNTRYSIDE
TO_FERTILISE
SOFTNESS
ANIMAL
BLUE
PEAK
TO_SWIM
ESCAPE
DANGER
FIRE
MOON
TO_DREAM
SPEED
CARNAL
DESIRE
ISLAND
NUDITY
METALLIC
TO_FORBID
RIGID
WALL
SPACE
GUN
ARMOUR
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.4
-0.3
-0.2
-0.1
0.1
0.2

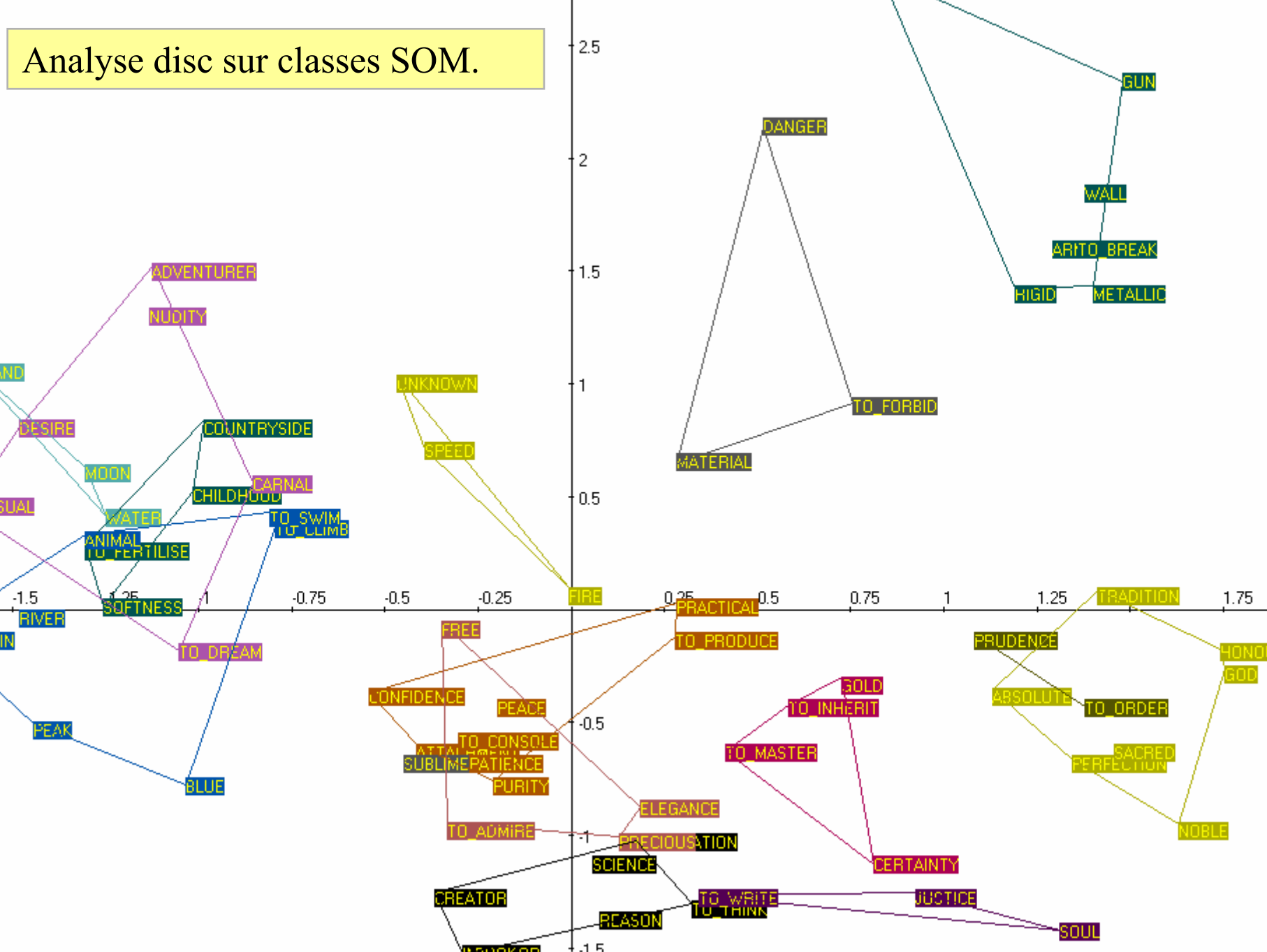
Matrice associé au graphe de l'analyse discriminante (9 classes)

1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Analyse
Discriminant
Sur 16 classe



Analyse disc sur classes SOM.



Matrice associé à une carte de Kohonen carrée (9 x 9)

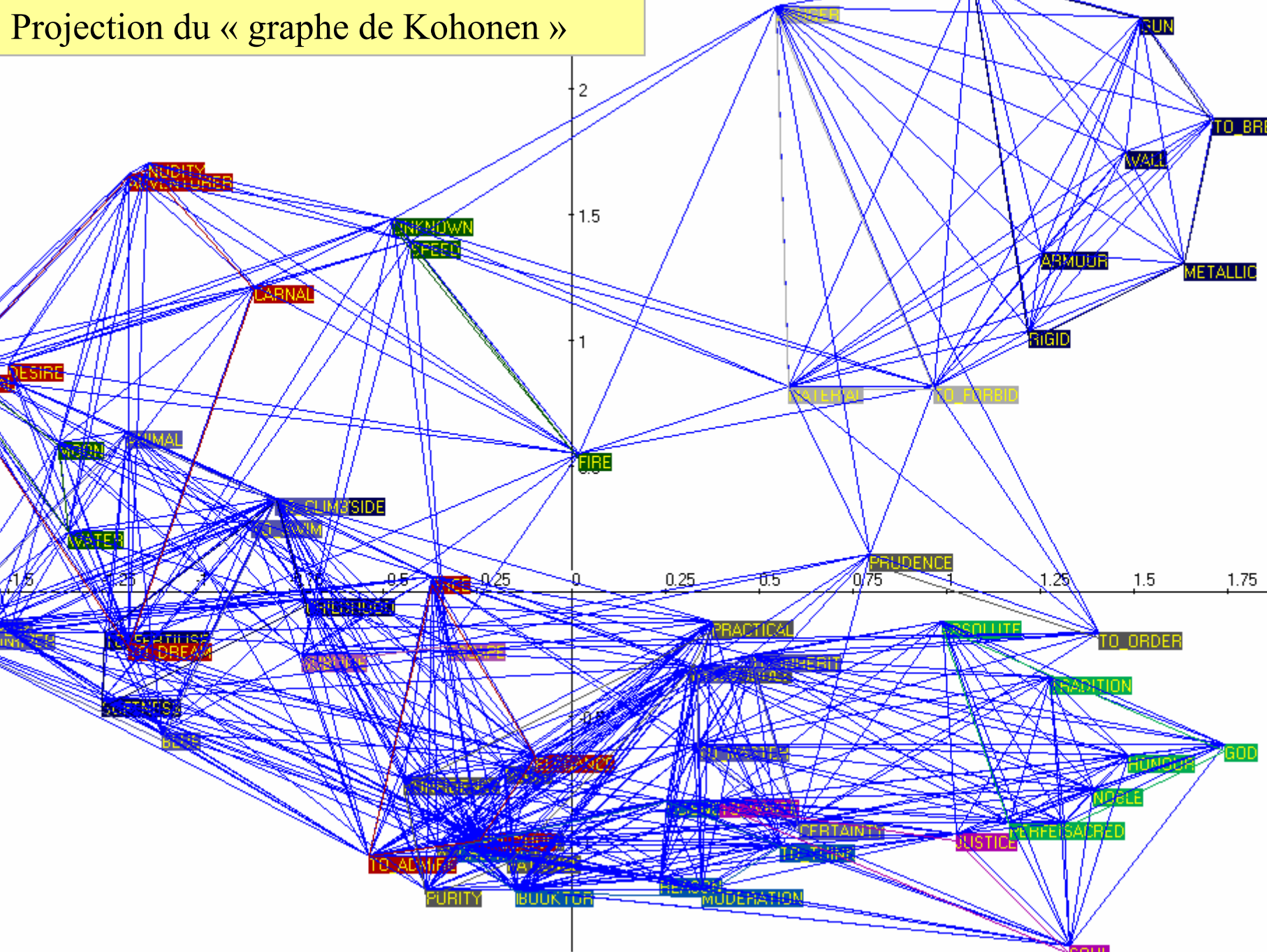
1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
7	0	0	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0
8	0	0	0	1	1	1	1	1	0	0	1	1	1	1	0	0	0	0
9	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	1	0	0
10	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	1	0	0
11	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
12	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0
13	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0	0	0	1	1
14	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0	0	0	1	1
15	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0
16	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	1	0	0
17	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1
18	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1
19	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1
20	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	1

1
2
3
4
5
6
7
8
9

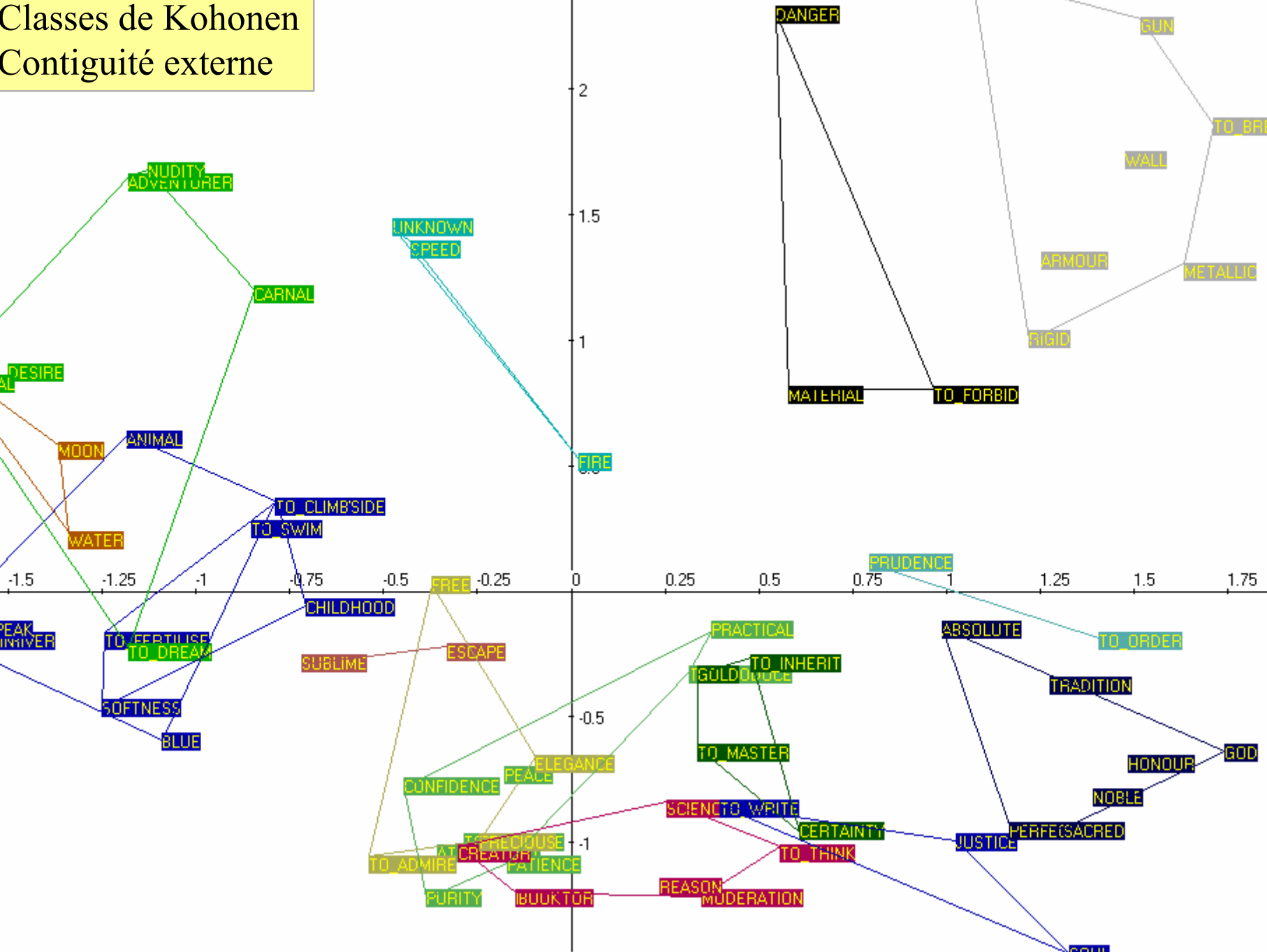
7	8	9
4	5	6
1	2	3

Locations of the 9 clusters in the SOM

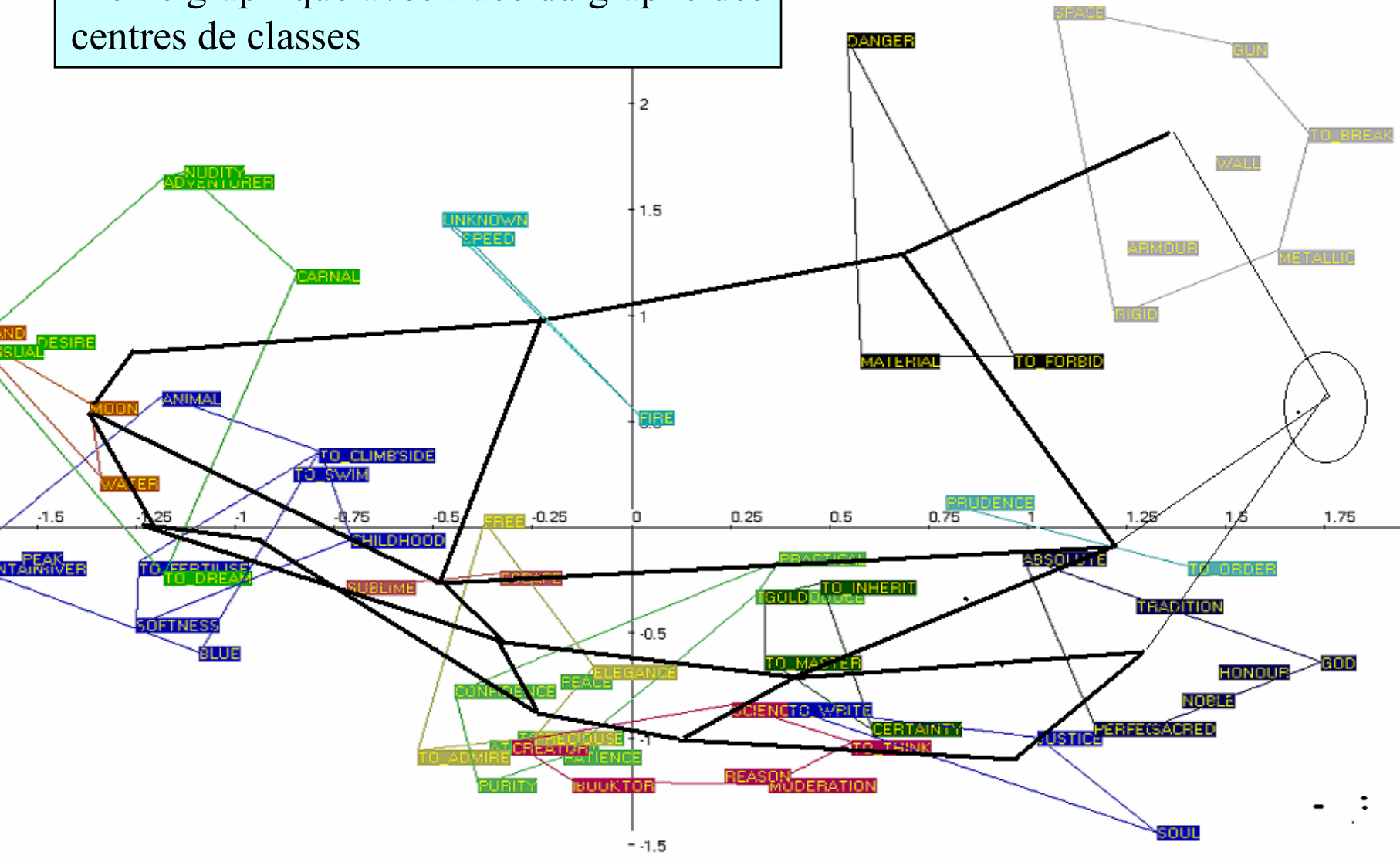
Projection du « graphe de Kohonen »



Classes de Kohonen
Contiguité externe



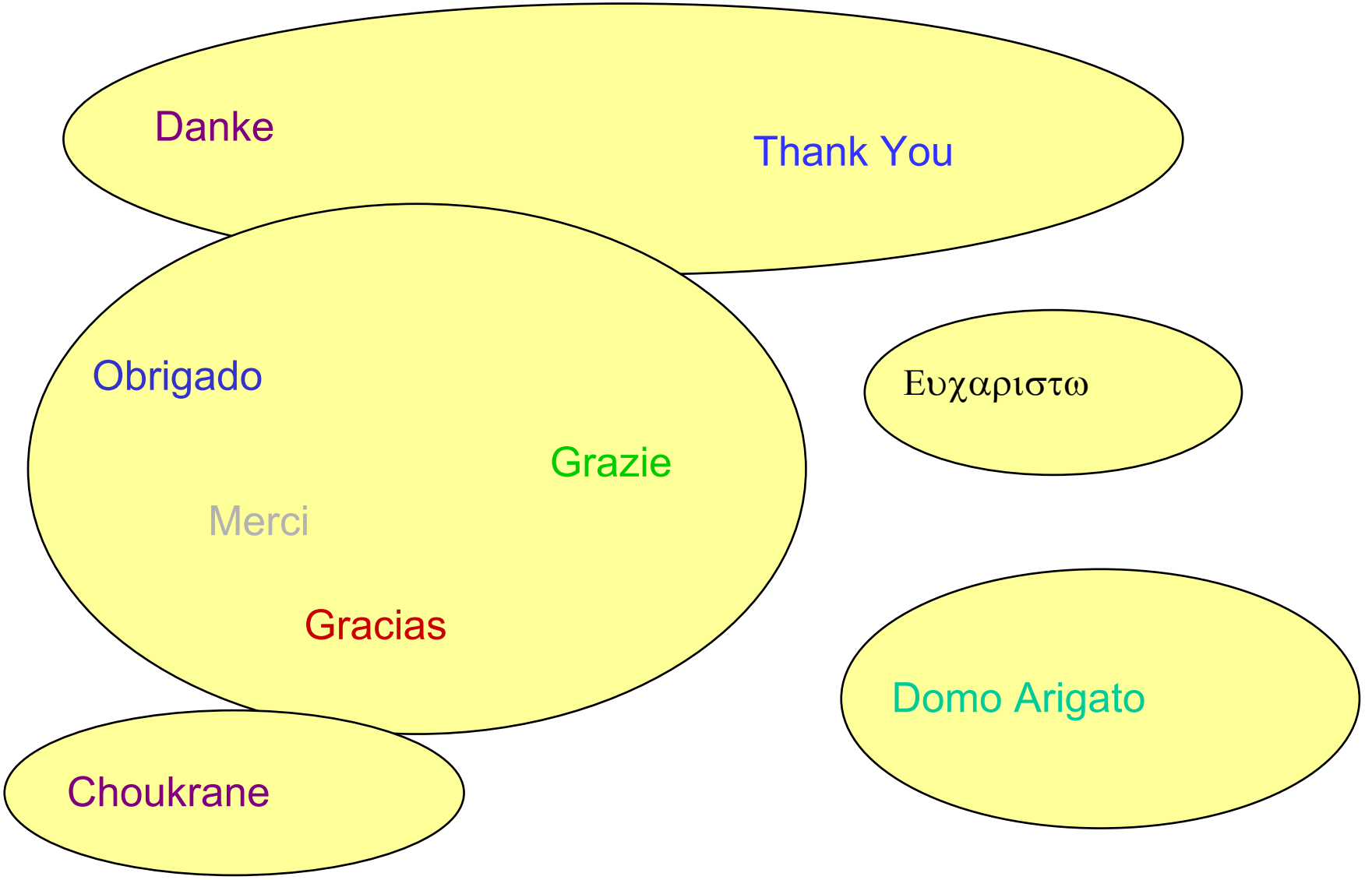
Même graphique avec tracé du graphe des centres de classes



La forme, le contour, la structure interne des classes sont maintenant visibles

Conclusion

- Outils variés, mais stratégie complexe
- Implémentation interactive nécessaire
- Prix à payer pour un statut scientifique des visualisations ?
- A suivre …


Danke
Thank You
Obrigado
Grazie
Merci
Gracias
Ευχαριστω
Domo Arigato
Choukrane