Une cartographie des genres et des styles fondée sur la distribution des quantifieurs

Michel DELARCHE (EILA/LANSAD)

Présentation au séminaire CLILLAC
(19 novembre 2012)

"Certaines personnes ont peur des chiffres parce qu'elles s'imaginent, à tort, que la connaissance des chiffres aliène leur liberté" (É. BOREL)

"I only believe in statistics I doctored myself" (W. CHURCHILL (?))

"Statistics are like bikinis: what they reveal is suggestive but what they conceal is vital " (A. LEVENSTEIN)

"All models are wrong, but some are useful" (G. BOX)

Michel DELARCHE
Université Paris Diderot, Sorbonne Paris Cité, Labo CLILLAC-ARP
UFR EILA, case 7002, 75205 Paris cedex 13
delarche@eila.univ-paris-diderot.fr

université PARIS DIDEROT PARIS 7

Une cartographie des genres et des styles fondée sur la distribution des quantifieurs

Programme de recherche

Concepts, méthodes et outils

Résultats et consolidations

Interprétations

Conclusions

Autres travaux en cours

Cartographie des quantifieurs : Programme de recherche (1/2)

2010-2012: modélisation et exploration

- construction d'un modèle algébrique des quantifieurs (anneau commutatif) complétant les outils classiques de la sémantique formelle (théorie des types, Théorie Générale des Quantifieurs (GQT en anglais)...)
- construction de grammaires syntaxiques (outil NooJ) pour extraire et trier les quantifieurs naturels non-numéraux
- première vérification de la stabilité auctoriale de la distribution de ces quantifieurs sur des corpus réduits (6 romans d'Austen, 6 romans de Maupassant etc.)

2012-2013: étude statistique du genre et du style

- construction d'un premier corpus de validation
- étude statistique multidimensionnelle de ce corpus par une ACP (Analyse en Composantes Principales)
- *applications de la démarche à d'autres domaines*

Cartographie des quantifieurs: Programme de recherche (2/2)

6 hypothèses étudiées sur un corpus monolingue (anglais):

H1 : la distribution des quantifieurs est un indicateur du sexe de l'auteur
« Les femmes n'écrivent pas / ne parlent pas comme les hommes »

H2 : la distribution des quantifieurs est un indicateur de genre littéraire
Des discours de genres différents montreront des distributions différentes

H3 : la distribution des quantifieurs est un indicateur de style
Pour un auteur donné, la distribution des quantifieurs est stable

H4 : vis-à-vis des quantifieurs, le genre littéraire domine le style
Au sein d'un genre donné, la variabilité d'un auteur à l'autre est secondaire

H5 : vis-à-vis des quantifieurs, le style domine le genre littéraire
Pour un auteur donné, la variabilité liée au genre est secondaire

H6 : la distribution des quantifieurs est un indicateur de nationalité
« Les Océaniens n'écrivent pas / ne parlent pas comme les Estasiens »

Cartographie des quantifieurs : Concepts, méthodes et outils (1/15)

Qu'est-ce qu'un quantifieur ?

Un quantifieur naturel est un modifieur nominal (le nom modifié peut être sous-entendu) qui joue un rôle prédicatif (NB1: morpho-syntaxiquement, ce peut être un adjectif: "de nombreux” + <N+p> = "beaucoup de" + <N+p>)

Il y a une demi-douzaine de quantifieurs qui semblent exister dans toutes les langues (pas de, peu de, un peu de, un certain nombre de, beaucoup de, tout)

NB2: j'ai développé une catégorisation inter-langue des quantifieurs distinguant des quantifieurs absolus (sans référence explicite à la totalité) et relatifs:

Absolu: « Many people » / Relatif: « Many of them »

Un de mes objectifs secondaires est de voir si tout ou partie de cette distinction formelle est statistiquement pertinente

Cartographie des quantifieurs : Concepts, méthodes et outils (2/15)

Qu'est-ce qu'un genre littéraire ?

Cette notion est très (trop?) élastique et peut se raffiner en sous-genres à l'infini, mais on peut identifier deux schémas principaux de catégorisation (pas complètement orthogonaux entre eux) :

- des classifications dérivées des propriétés structurales des discours (le sonnet, l'article scientifique, la pièce de théâtre, le récit en prose...) indépendamment de la sémantique des contenus
- des typologies s'appuyant sur le contenu et/ou le registre (la poésie épique, le roman de chevalerie, le roman policier, la tragédie classique...)

Cartographie des quantifieurs : Concepts, méthodes et outils (3/15)

Qu'est-ce que le style d'un auteur ?

Le style est la combinaison de différentes propriétés lexicales et morpho-syntaxiques du discours évaluables statistiquement et permettant de caractériser un auteur en relation avec un (/des) genre(s) (idéalement, de manière unique et certaine)

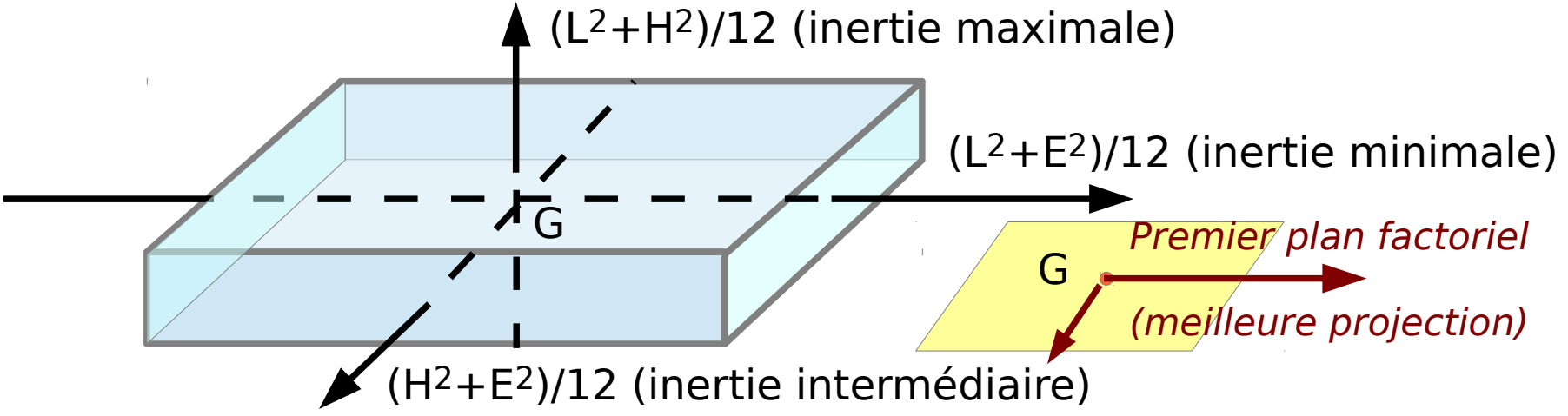
La stylométrie recourt à des métriques de distribution statistique (longueur des mots, longueur des phrases, fréquence des mots, collocations les plus fréquentes, formes spécifiques de certains mots)
pour résoudre des problèmes de “paternité littéraire” (attribution, identification de contributeurs multiples, détection d'interpolations etc.)

Cartographie des quantifieurs : Concepts, méthodes et outils (4/15)

Qu'est-ce qu'une ACP ? (1/3)

Une Analyse en Composante Principale est une technique statistique d'analyse de données permettant de représenter un nuage de n points (les individus) dans un espace à p dimensions (les variables) par des projections à 2 dimensions (plans factoriels) définis par les axes dits « principaux » d'inertie du nuage (axes qui maximisent/minimisent les sommes des carrés des distances des points à ces axes, orthogonaux deux à deux entre eux et se croisant au centre de gravité G du nuage)

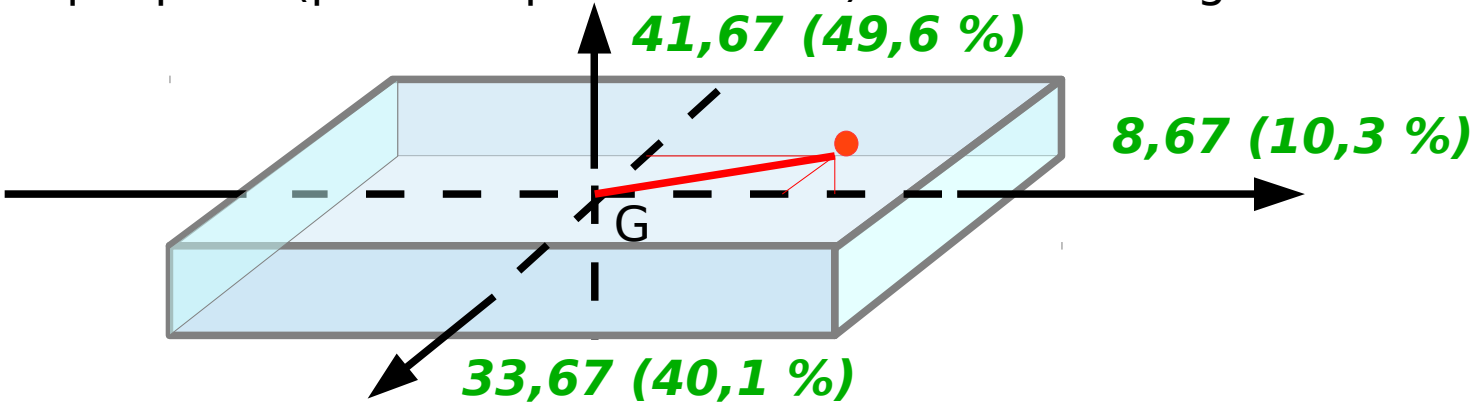
Axes principaux d'inertie d'un livre de hauteur H > largeur L > épaisseur E



Cartographie des quantifieurs : Concepts, méthodes et outils (5/15)

Qu'est-ce qu'une ACP ? (2/3)

L'inertie totale d'un nuage de points est la somme des carrés des distances de chaque point (pondéré par sa masse) au centre de gravité du nuage



Le carré de la distance au centre de gravité est la somme des carrés des distances aux 3 axes orthogonaux et l'inertie totale est donc la somme des inerties par rapport aux 3 axes soit: $(H^2+L^2+E^2)/6$

Pour H=20, L=10 et E=2, l'inertie totale vaut: (400+100+4)/6 = 84

L'inertie maximale vaut 500/12 = 41,67 et l'intermédiaire 404/12=33,67

L'écart entre le moment d'inertie associé au plan principal et l'inertie totale se normalise en pourcentage de l'inertie totale et on dit alors que « ce plan explique N% de l'inertie du nuage de point» (ici, 75,33/84= 89,7%)

Cartographie des quantifieurs : Concepts, méthodes et outils (6/15)

Qu'est-ce qu'une ACP ? (3/3)

Techniquement, les axes cherchés correspondent aux vecteurs propres de la matrice de variance-covariance (ou de la matrice de corrélation) parce qu'il s'agit d'un problème de maximisation sous contrainte: toute solution $\vec{v}$ annule la dérivée du lagrangien: $M\vec{v} - \lambda\vec{v}=0$

Intuitivement, le premier axe est celui qui traverse le nuage dans sa plus grande longueur (combinaison de variance maximale), le second axe est orthogonal au premier dans le sens de la plus grande largeur du nuage (combinaison des variables donnant la deuxième plus grande variance) etc. pour obtenir la meilleure projection bidimensionnelle possible (cf. l'exemple précédent du livre faisant 20 x 10 x 2 cm^3)

Si par extraordinaire notre nuage de points se révélait homogène et parfaitement sphérique, tous les axes se vaudraient et l'on ne pourrait extraire aucune information (les variables seraient interchangeables)

information ≡ anisotropie

(isotropie = entropie maximale = absence d'information)

Cartographie des quantifieurs : Concepts, méthodes et outils (7/15)

Quelques précautions techniques recommandables pour l'ACP:

1°) mener une analyse de sensibilité :

l'introduction du carré des distances dans les calculs d'inertie donne un poids élevé aux éléments les plus excentrés du nuage de point

=> refaire l'ACP sans les points isolés pour vérifier la stabilité des différentes projections obtenues

2°) limiter le nombre de dimensions :

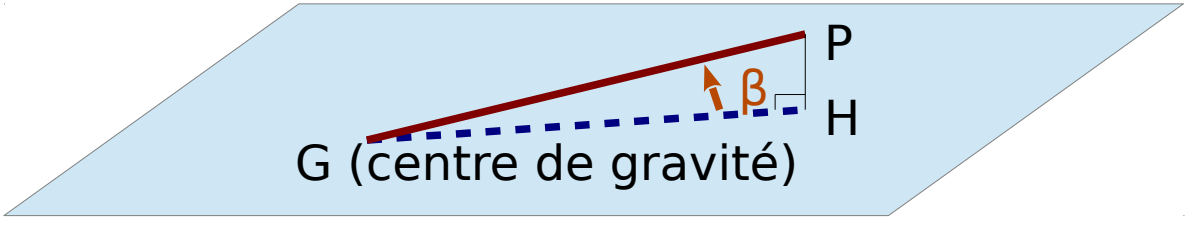
=> supprimer les facteurs ayant des contributions très faibles

(taux d'occurrence < 1/100ème des facteurs principaux)

=> fusionner éventuellement les facteurs trop parfaitement corrélés

3°)vérifier au besoin la qualité des projections:

=> très bonne: $\cos^2(\beta)>0,8$ (ie $\beta<25°$) ; bonne: $\cos^2(\beta)>0,65$ ($\beta<35°$)



Deux points P1 et P2 dont les projections H1 et H2 sont proches peuvent être en fait très éloignés : éviter les conclusions hâtives !

Cartographie des quantifieurs : Concepts, méthodes et outils (8/15)

Contraintes techniques sur l'étude statistique:

1°) les taux d'occurrence des quantifieurs sont faibles

0,5 à 1,5 % des mots d'un texte sont des quantifieurs

Les taux d'occurrence des plus rares varient de 10 à 100 ppm

=> traiter des échantillons importants (>> 10 000 mots)

2°) l'objectif est d'évaluer des styles et des genres

=> construire un corpus comportant assez d'auteurs et de genres différents, chacun autant que possible représenté par plusieurs œuvres

=> limiter l'hétérogénéité de l'échantillon, donc la variabilité vis-à-vis d'autres paramètres explicatifs des variations (une seule langue, empan temporel pas trop vaste, genre commun analysable en sous-genres afin de bien tester les limites de sensibilité de l'analyse)

Contraintes pragmatiques sur le matériau utilisé:

=> textes disponibles en version électronique maniable (texte brut non formaté en pages) et libres de droits

Cartographie des quantifieurs : Concepts, méthodes et outils (8/15)

Contraintes de la modélisation :

→ Les différents quantifieurs étudiés définissent un espace de représentation comportant une dizaine de dimensions, il est donc recommandé d'avoir un nuage de points de taille suffisante

=> règle heuristique pour garantir la robustesse : nombre de points = 5-10 x nombre de dimensions)

→ Notre analyse des quantifieurs est morpho-syntaxique

=> comparer les résultats obtenus avec ce que donne un simple comptage lexical, si l'on veut démontrer une réelle valeur ajoutée vis-à-vis de la stylométrie basique (outils du type Signature ou JGAAP)

Cartographie des quantifieurs : Concepts, méthodes et **outils** (9/15)

Corpus de validation exploratoire: 60 romans anglais du 18ème siècle

« the time of the Georges » (dixit Clarissa Dalloway) : 1720-1830

23 auteurs représentant tous les sous-genres du roman de l'époque (picaresque, féministe, philosophique, gothique, historique, érotique...)
9 femmes : Austen, Brunton, Burney, Edgeworth, Lamb, Lennox, Radcliffe, Reeve, Wollstonecraft-Shelley
14 hommes : Cleland, Defoe, Fielding, Godwin, Golsmith, Johnson, Lewis, Maturin, Polidori, Richardson, Scott, Smollett, Sterne, Walpole
3 nationalités : irlandaise (Edgeworth, Goldsmith, Maturin, Sterne) écossaise (Brunton, Scott, Smollett) et anglaise (les 16 autres)
De 1 à 7 œuvres max par auteur (éviter une surdose de Scott!)

Taille totale du corpus : près de **10 000 000 mots**
Des textes variant de **16 000 mots** (Shamela de Fielding)
à **975 000 mots** (Clarissa de Richardson)
Taille moyenne d'ouvrage : **165 000 mots**

Cartographie des quantifieurs : Concepts, méthodes et outils (10/15)

Données brutes toutes positives

=> recentrage des données sur la moyenne (centre de gravité)
simple translation ne changeant pas la géométrie du nuage

Textes de longueurs très variables

prélever des « tranches » calibrées sur la taille minimale ?
    => problème potentiel de représentativité narrative
normaliser les compteurs pour définir des taux de quantifieurs ?
    => construction d'une matrice de variance-covariance

Taux variables de quantifieurs: 0,5 à 1,5 % des mots

deuxième échelon de normalisation : exprimer le taux d'occurrence de chaque quantifieur sur les seuls quantifieurs (et pas sur tous les mots)
    => construction de la matrice de corrélation

Les deux modes d'analyse peuvent (et doivent) être utilisés :
Il ne faut pas se limiter a priori aux covariances ou aux corrélations

Cartographie des quantifieurs : Concepts, méthodes et outils (11/15)

Données brutes (comptage d'occurrences) :

Auteur	Œuvre	Taille (mots)	Code	RQT	AQN
Austen	Northanger Abbey	78287	JA1	256	224
Austen	Sense and Sensibility	121085	JA2	449	391
Austen	Pride and Prejudice	122525	JA3	456	321

Données normalisées (taux d'occurrence):

Auteur	Œuvre	Taille (mots)	Code	RQT	AQN
Austen	Northanger Abbey	78287	JA1	3270	2861
Austen	Sense and Sensibility	121085	JA2	3708	3229
Austen	Pride and Prejudice	122525	JA3	3722	2620

Les taux ci-dessus sont donnés en ppm (parties par million) 1000 ppm = 0,1 %

Cartographie des quantifieurs : Concepts, méthodes et outils (12/15)

Annotations des quantifieurs développées à partir de :
NooJ version 3.2 www.nooj4nlp.net
(remerciements au passage à Max Zylberstein)

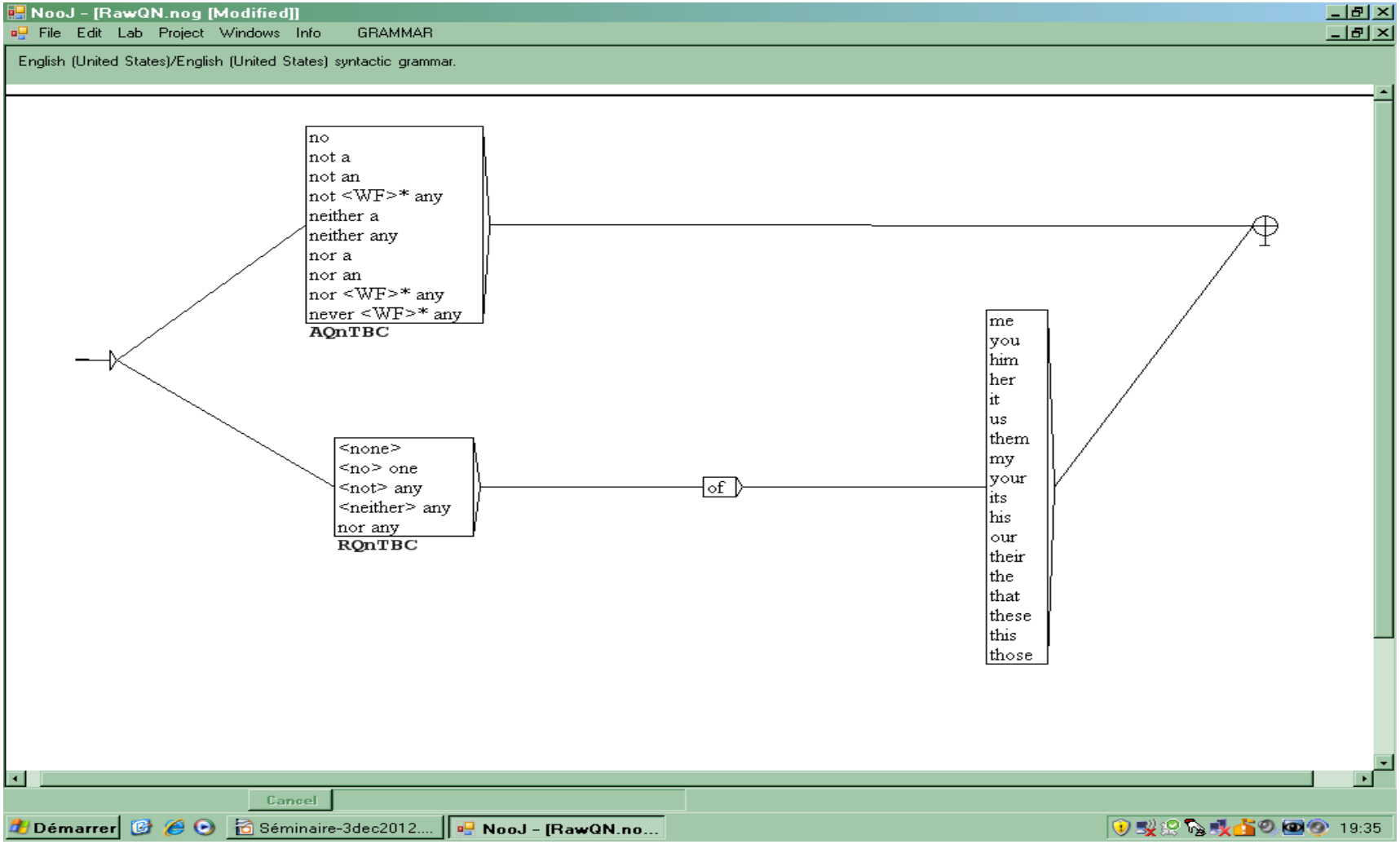
Analyses multivariées conduites avec l'outil inclus dans:
Jean-Pierre Georgin
Analyse interactive des données (ACP, AFC) avec Excel 2000
Théorie et pratique (2ème édition) Presses Univ. de Rennes (2007)

Outil limité à 100 points et 10 dimensions, mais très complet (calcul automatique de tous les indicateurs de qualité) hautement maniable (prise en compte interactive des modifications) et ne demandant aucun apprentissage (il suffit de copier-coller sa matrice de données sur la feuille de données initiales et d'aller étudier les résultats)

Sources électroniques du corpus:
The Gutenberg Project (90% des ouvrages)
complété par quelques sites universitaires américains et australiens

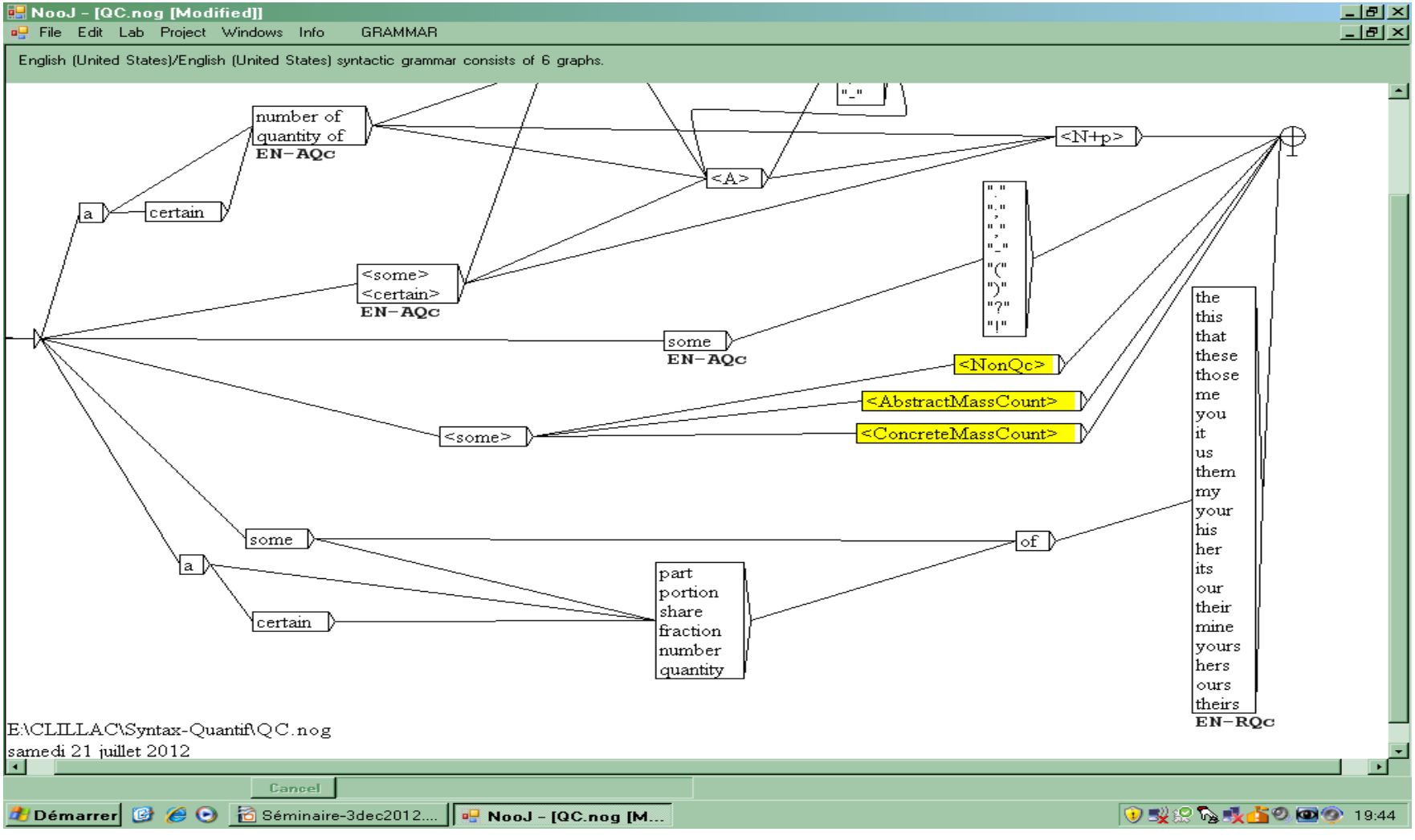
Cartographie des quantifieurs : Concepts, méthodes et outils (13/15)

Exemple de grammaire NooJ d'annotation:



Cartographie des quantifieurs : Concepts, méthodes et outils (14/15)

Exemple de grammaire NooJ d'annotation:

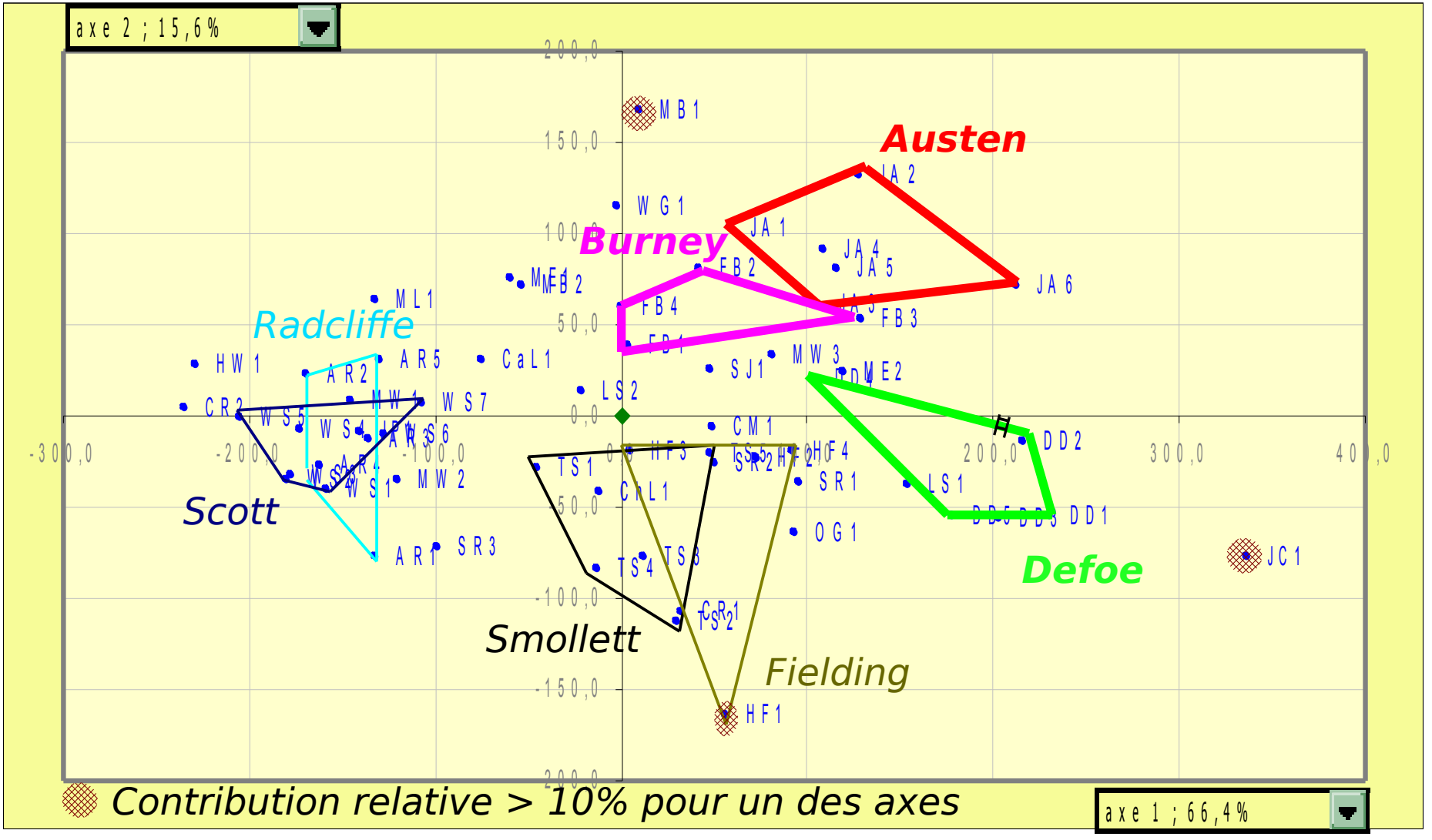


Cartographie des quantifieurs :
Concepts, méthodes et outils (15/15)
Incertitudes sur les comptages réalisés :
=> ambigüité du repérage des quantifieurs

Some one, some people, some danger
Some time or other, it took them some time
A little girl, a little time, a little hesitation
A lot of land, a lot of wind, a lot of jewels
I am no judge, there is no solution, No way!
They are few, few are here
He was all pride, you could not all go

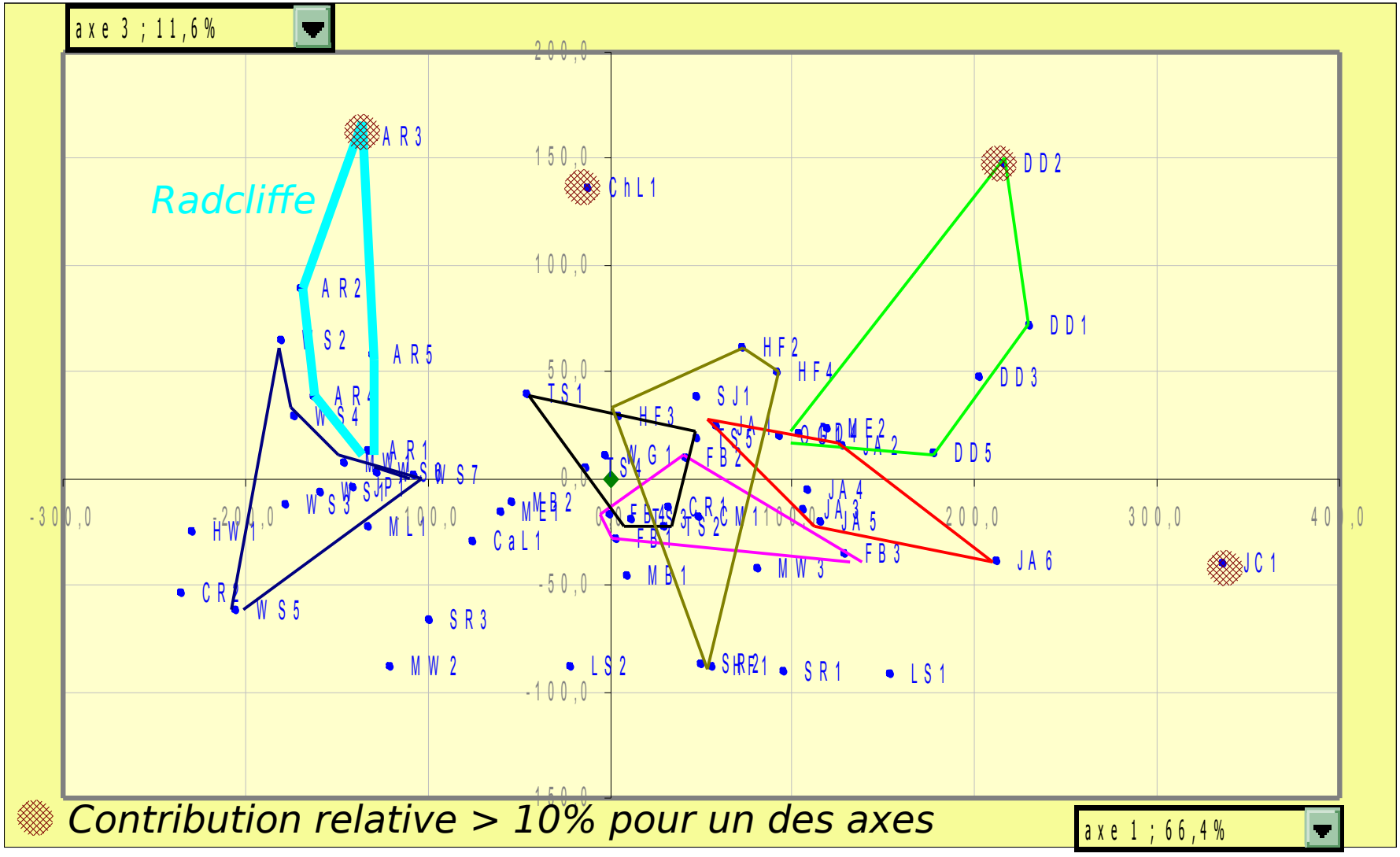
Même en utilisant des critères syntaxiques, le comptage doit être considéré comme défini à seulement ±5 voire 10%

Cartographie des quantifieurs : Résultats et consolidations (1/15)



Cartographie globale (N=60 V=9, M=VarCov, PF1) (bonne projection: 82%)

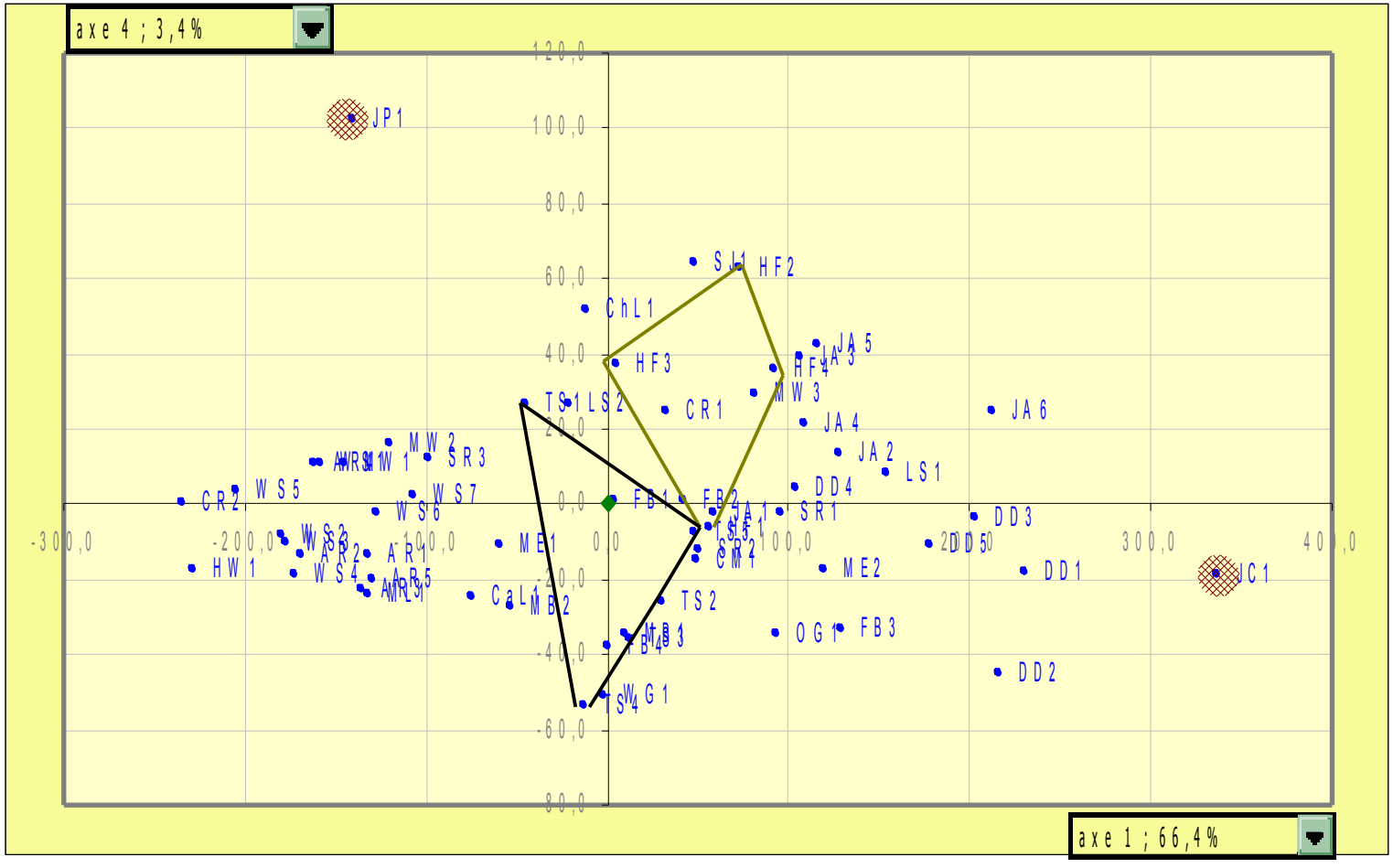
Cartographie des quantifieurs : **Résultats** et consolidations (2/15)



Cartographie globale des auteurs (N=60 V=9, M=VarCov, PF2)

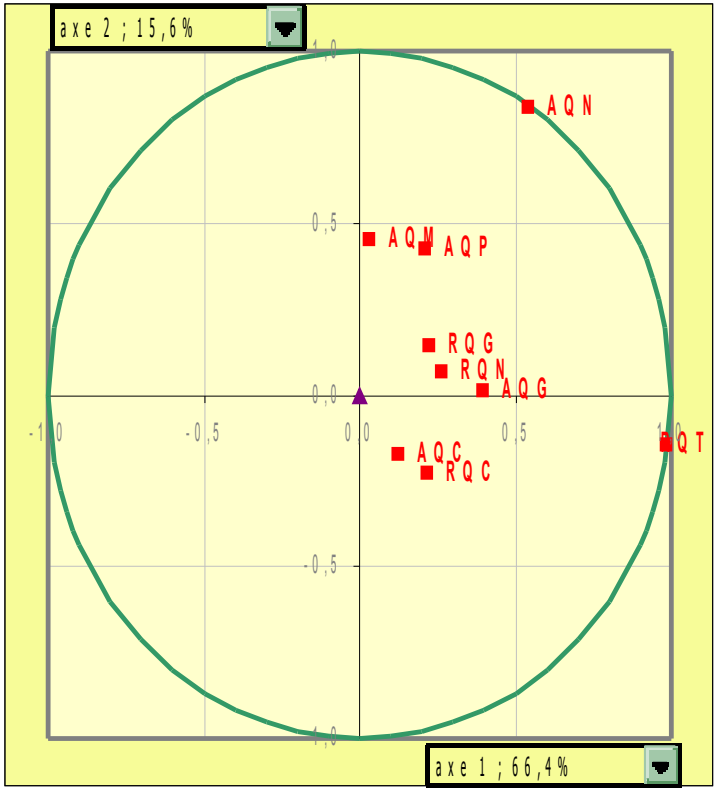
Cartographie des quantifieurs : Résultats et consolidations (3/15)

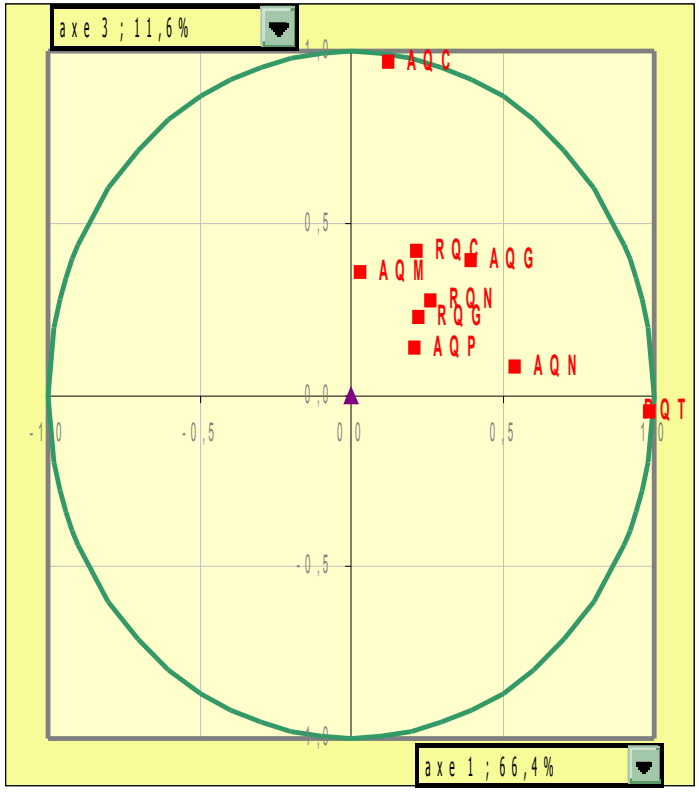
Le quatrième axe apporte un peu d'information nouvelle:
→ un 7ème élément marginal (20% de contribution à l'axe 4!)
→ un angle de vue séparant Fielding de Smollett



Cartographie des quantifieurs : Résultats et consolidations (4/15)

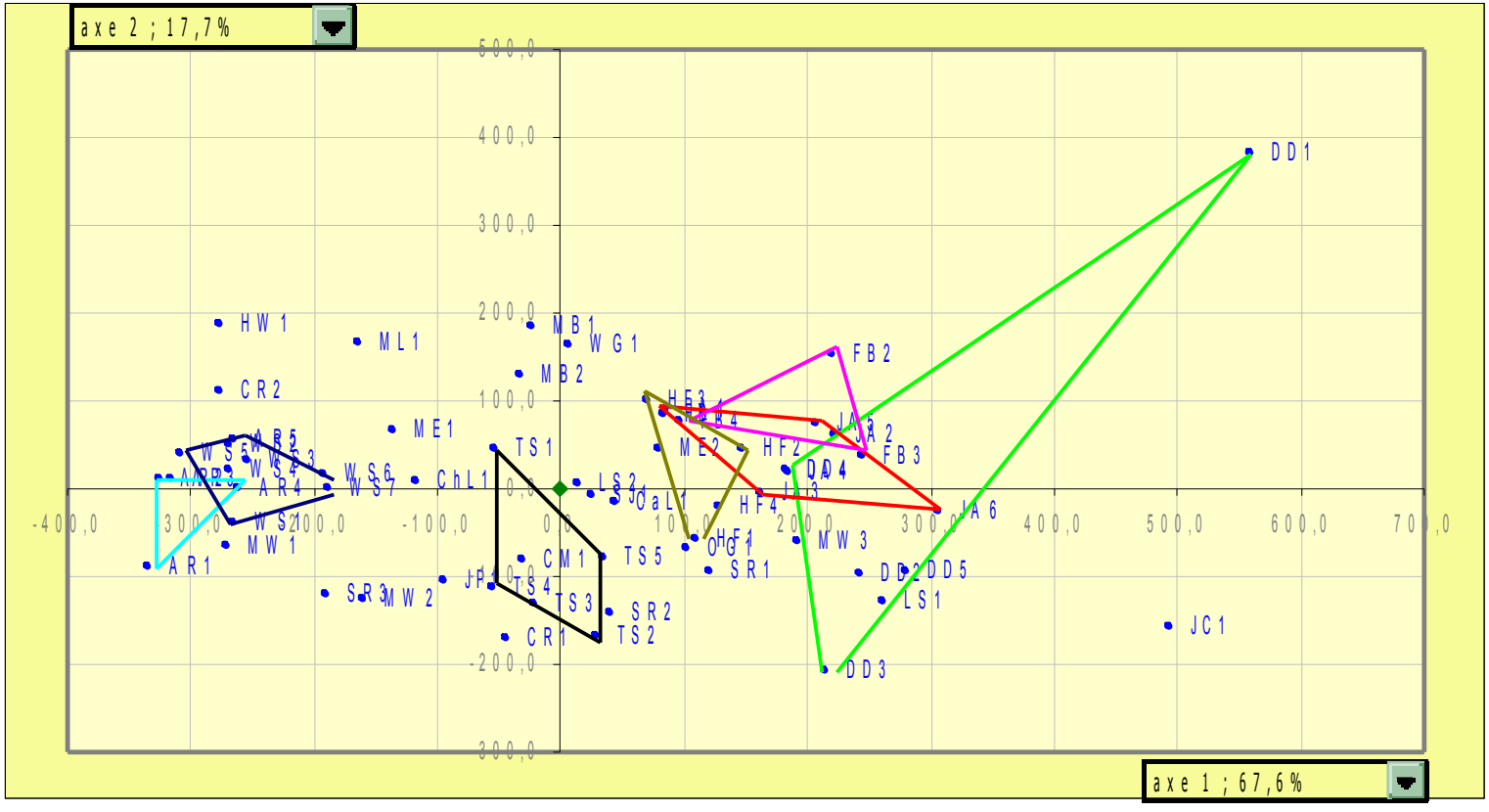
Position des variables par rapport au cercle unité des plans principaux





L'axe 1 est défini par RQT (all/whole) et un peu par AQN (no), l'axe 2 par AQN (no) et dans une moindre mesure par AQM (little/few) et AQP (a little/a few) l'axe 3 par AQC (some/a certain quantity) et un peu par AQM et RQC l'axe 4 est défini essentiellement par AQG (much/many)

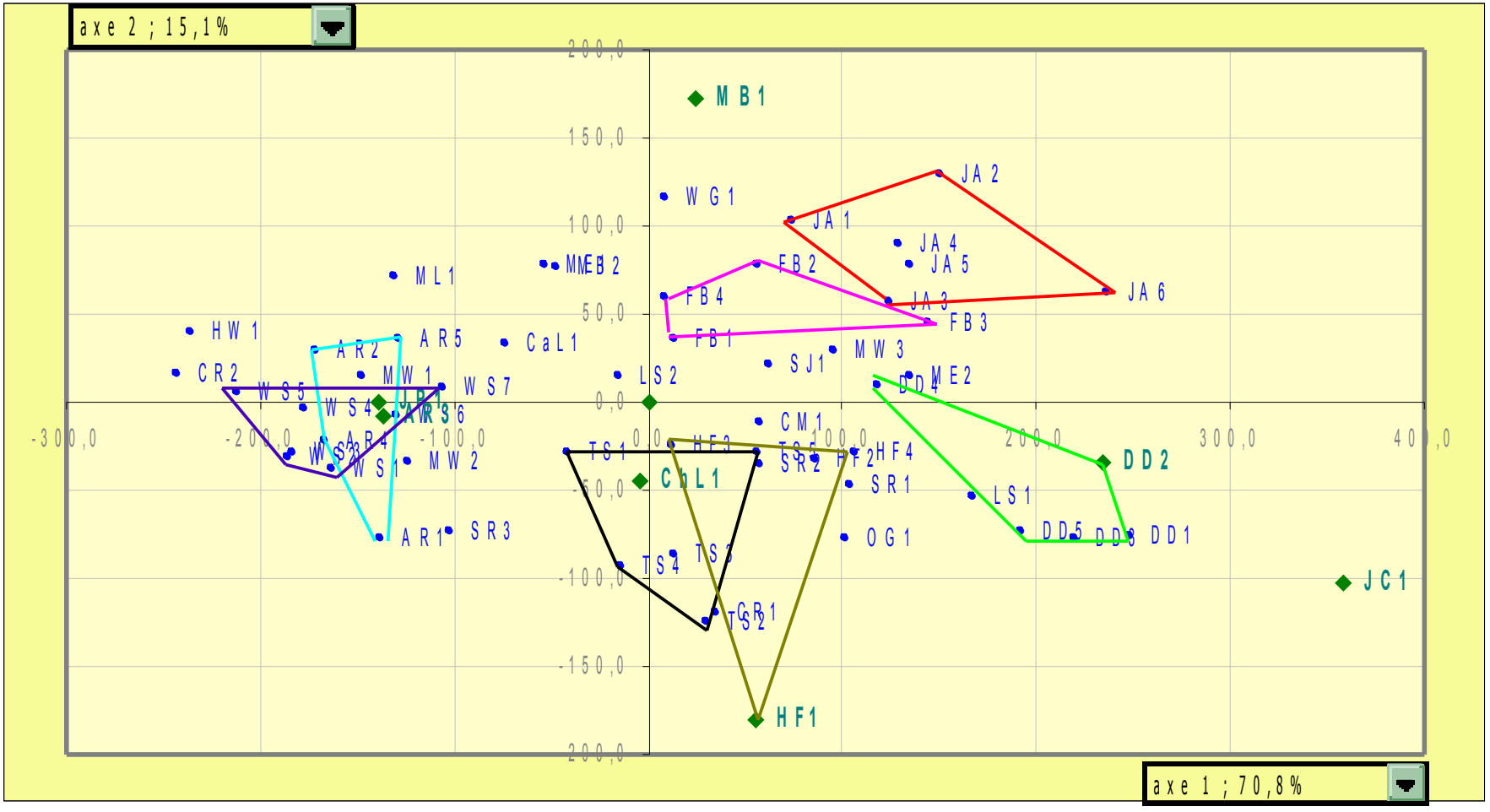
Cartographie des quantifieurs : Résultats et consolidations (5/15)



Plan factoriel principal pour la matrice de variance-covariance obtenue sans aucun filtrage syntaxique (comptage brut des mots-quantifieurs)
la cartographie paraît beaucoup plus bruitée (recouvrements multiples)
→ les quantifieurs sont une structure du discours, pas du pur lexique
→ des petits polygones auctoriaux subsistent (homogénéité stylistique)

Cartographie des quantifieurs
Résultats et consolidations (6/15)

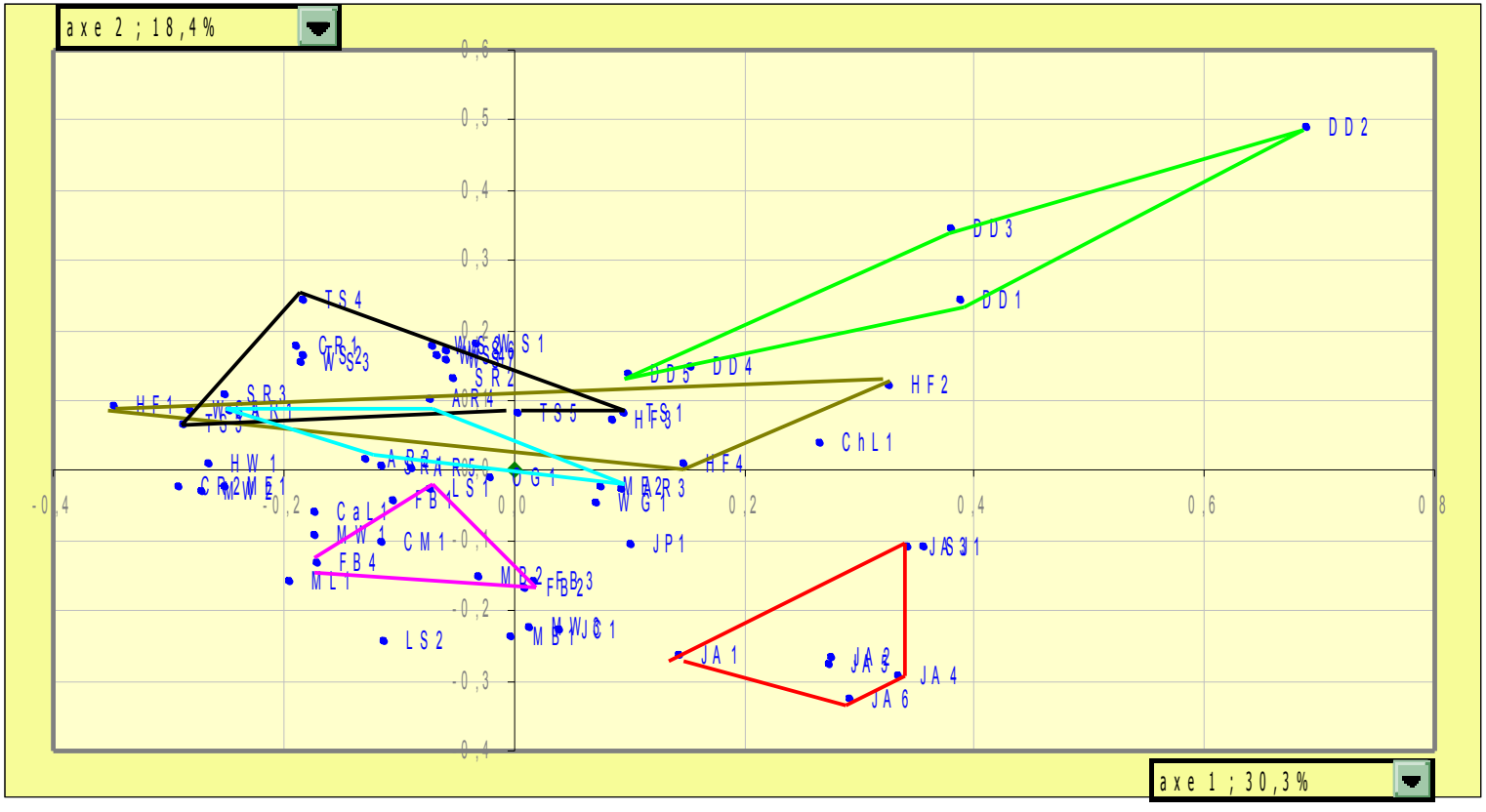
Traitement des points marginaux (7 ouvrages très contributifs aux axes)



NB: ici, ces 7 points sont visualisés mais ils ne participent pas au calcul (N=53 V=9, M=VarCov, PF1) La projection s'améliore encore (82% → 86%)

Cartographie des quantifieurs
Résultats et consolidations (7/15)

Analyse complémentaire par la matrice des corrélations

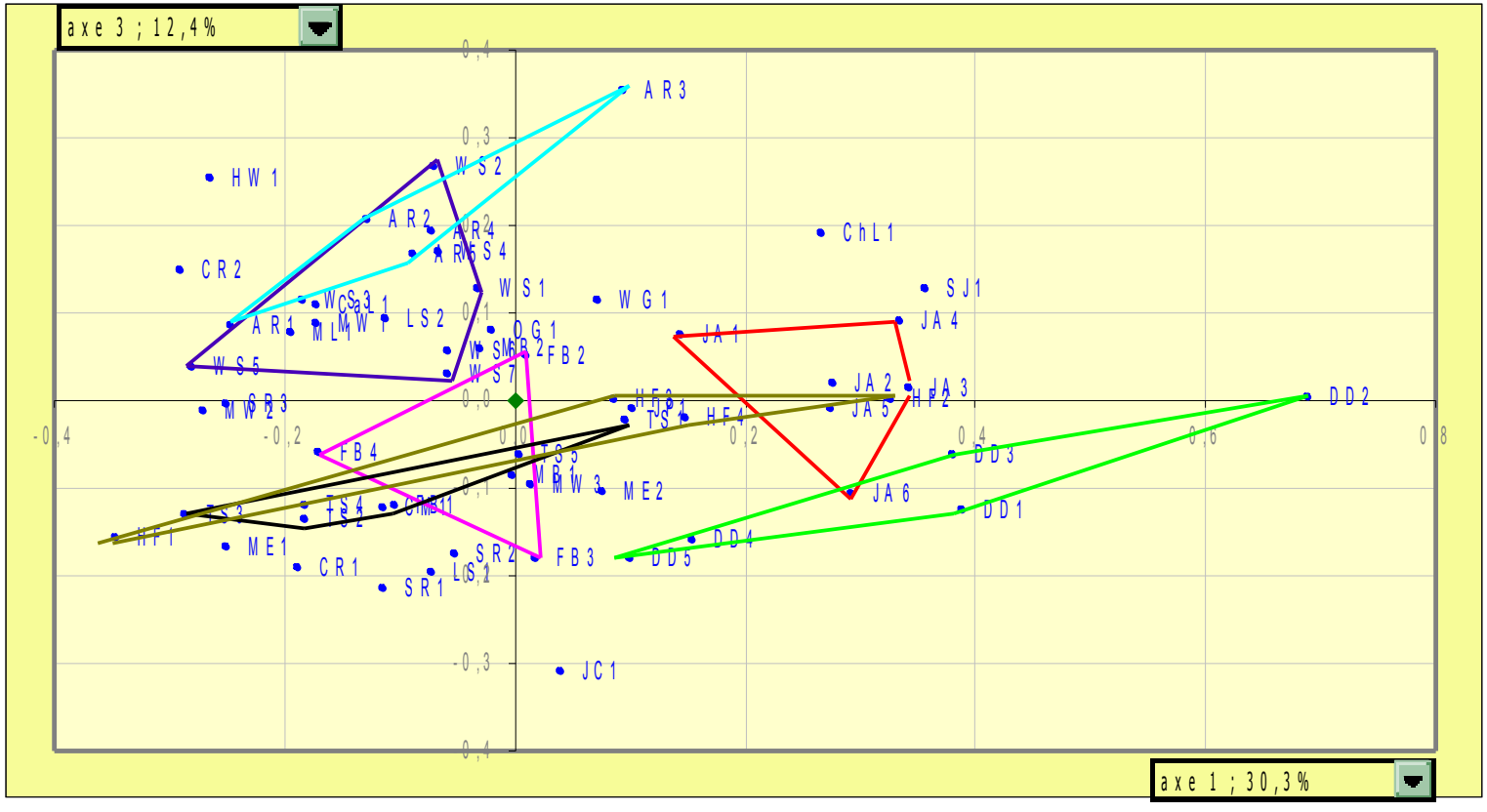


La projection obtenue est moins bonne (<50% de l'inertie au niveau du plan principal) mais on obtient quelques résultats nouveaux: séparation nette entre Austen et Burney, Defoe aligné sur un axe...

(N=60 V=9, M=Correl, PF1)

Cartographie des quantifieurs : Résultats et consolidations (8/15)

Analyse complémentaire par la matrice des corrélations

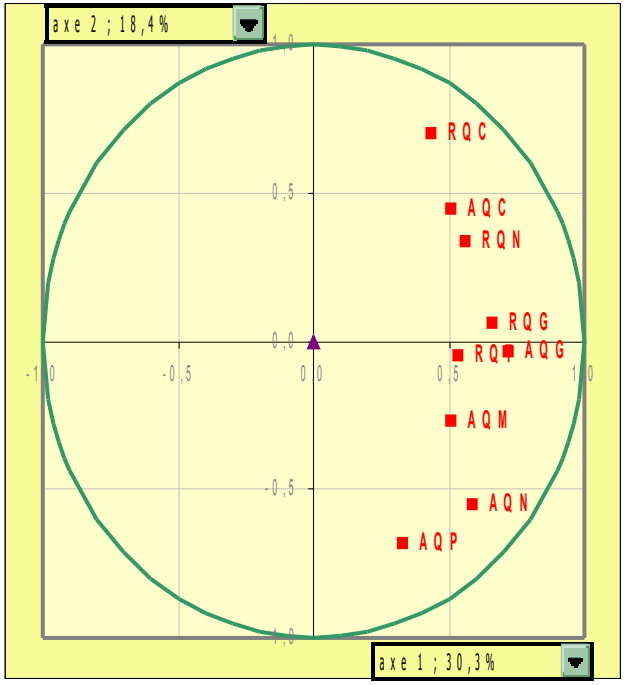


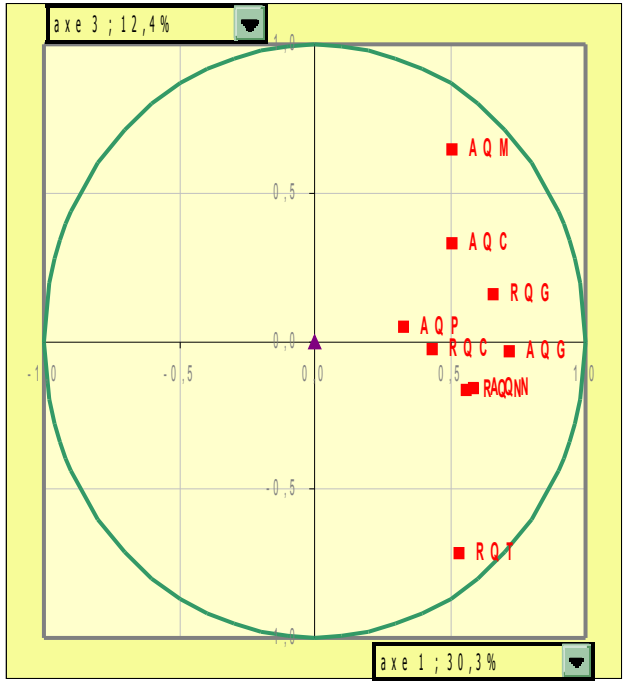
Les polygones très aplatis obtenus pour certains auteurs font apparaître certaines corrélations spécifiques entre les parts relatives des quantifieurs (→ tests statistiques d'attribution)

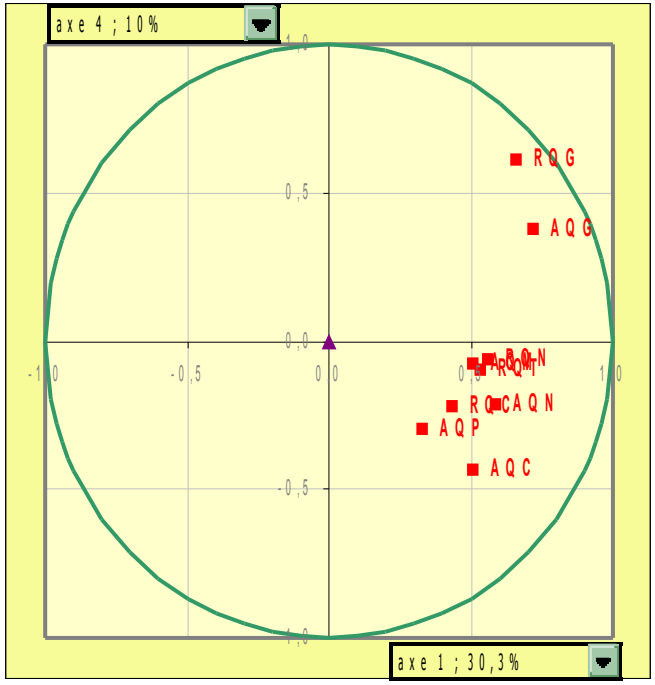
(N=60 V=9, M=Correl, PF2)

Cartographie des quantifieurs : Résultats et consolidations (9/15)

Analyse complémentaire par la matrice des corrélations



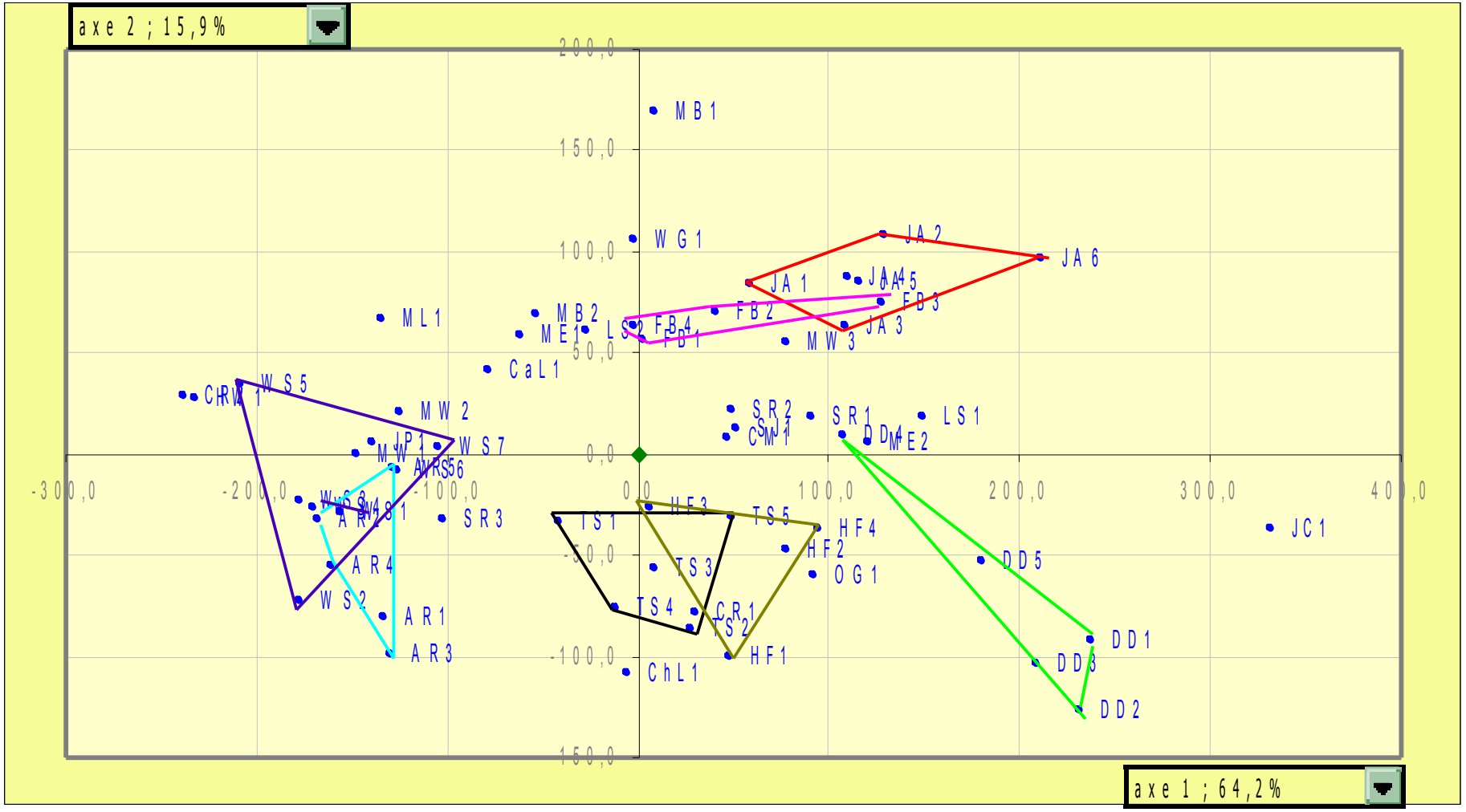




L'axe principal est défini positivement par AQG et RQG et le second axe est défini positivement par RQC et négativement par AQP.
Le troisième axe est défini positivement par AQM et négativement par RQT.
Le quatrième axe est défini positivement par RQG et négativement par AQP et AQC

Cartographie des quantifieurs
Résultats et consolidations (10/15)

Regroupement des quantifieurs absolus et relatifs (réduction à 6 variables)

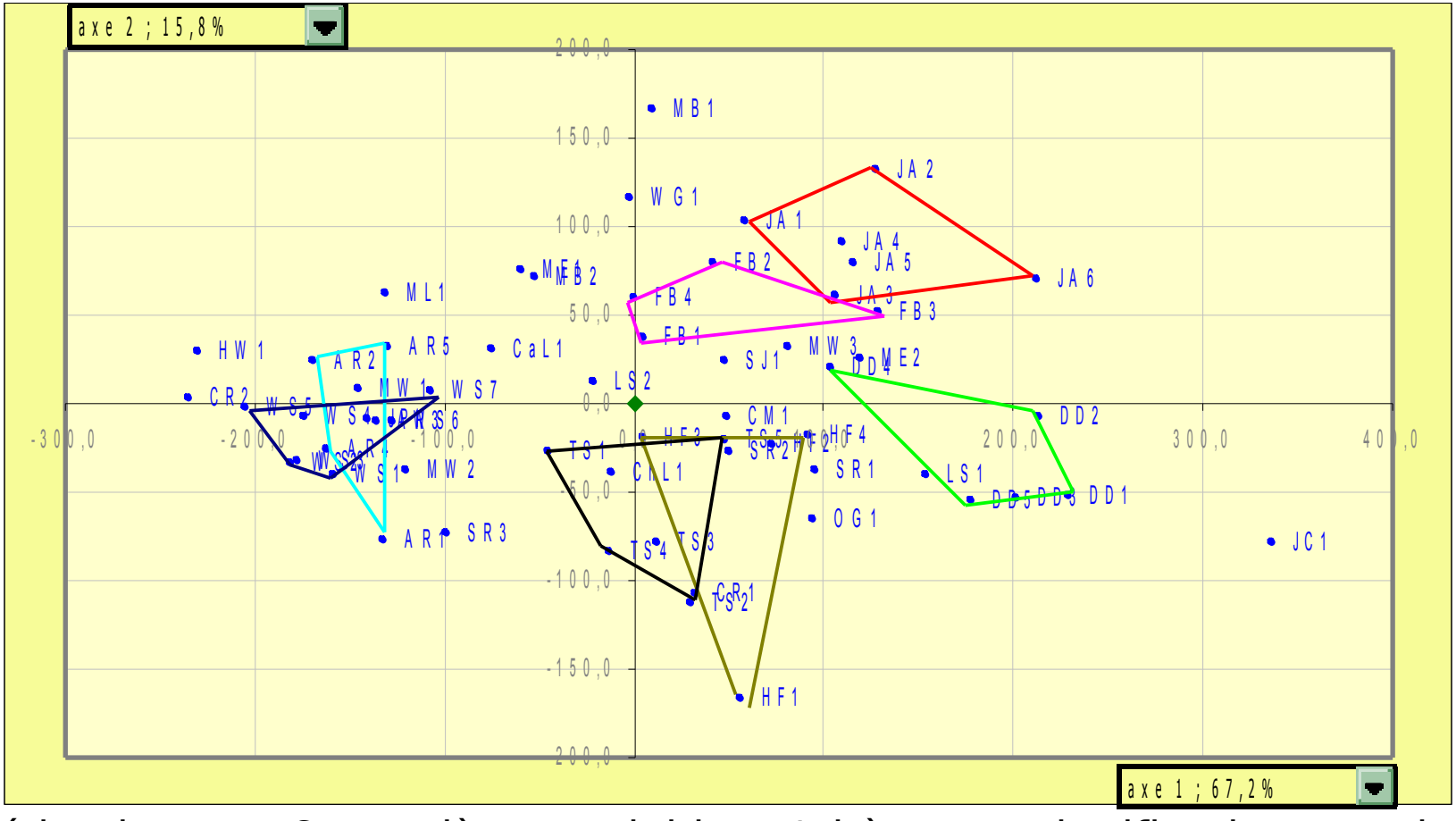


(N=60 V=6, M=VarCov, PF1) Cartographie un peu moins précise (JA vs FB)

université PARIS DIDEROT PARIS 7

Cartographie des quantifieurs : Résultats et consolidations (11/15)

Réduction à 6 du nombre de variables prises en compte dans l'ACP

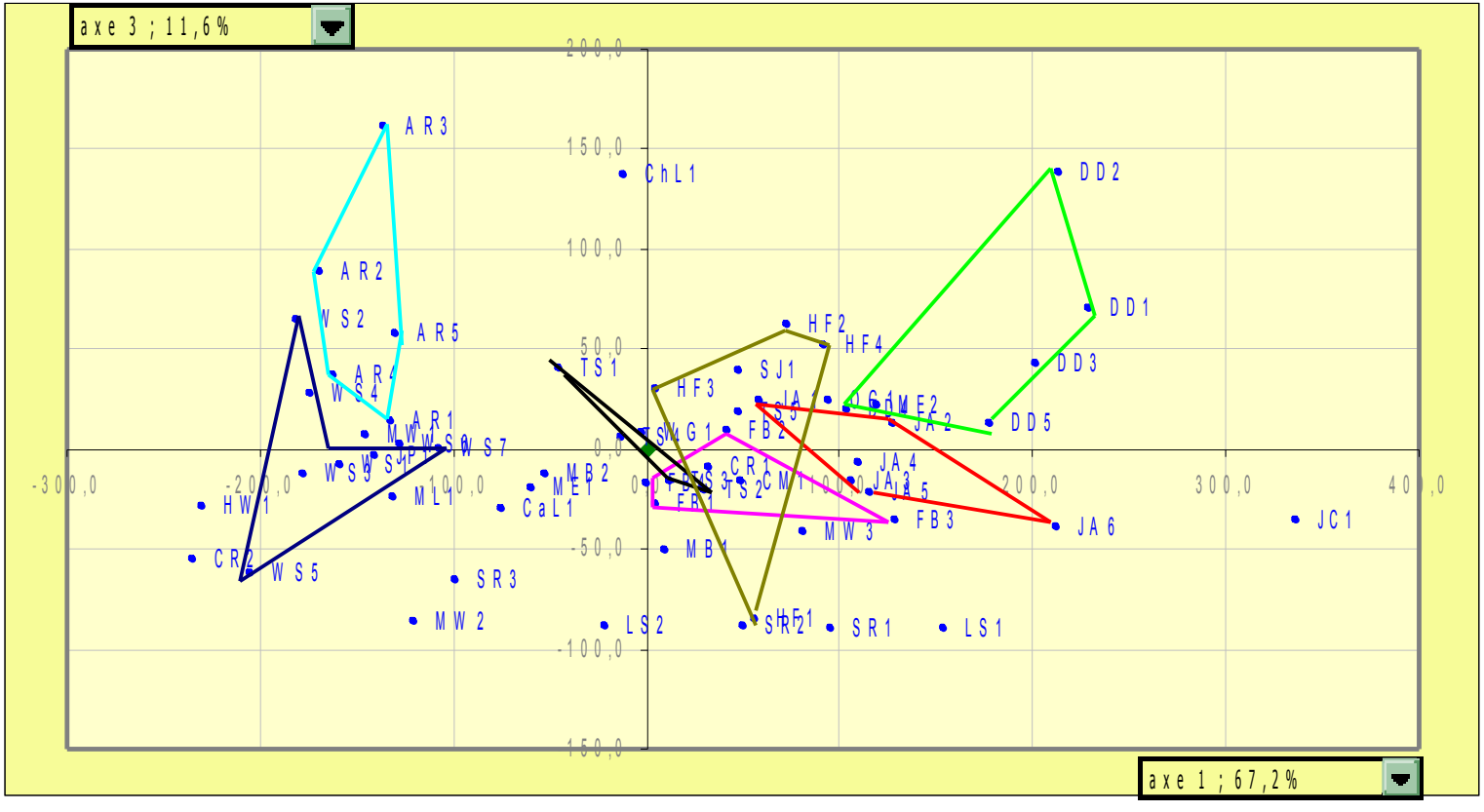


La réduction aux 6 premières variables n'altère pas significativement la cartographie obtenue précédemment avec 9 variables

(N=60 V=6, M=VarCov, PF1)

Cartographie des quantifieurs : Résultats et consolidations (12/15)

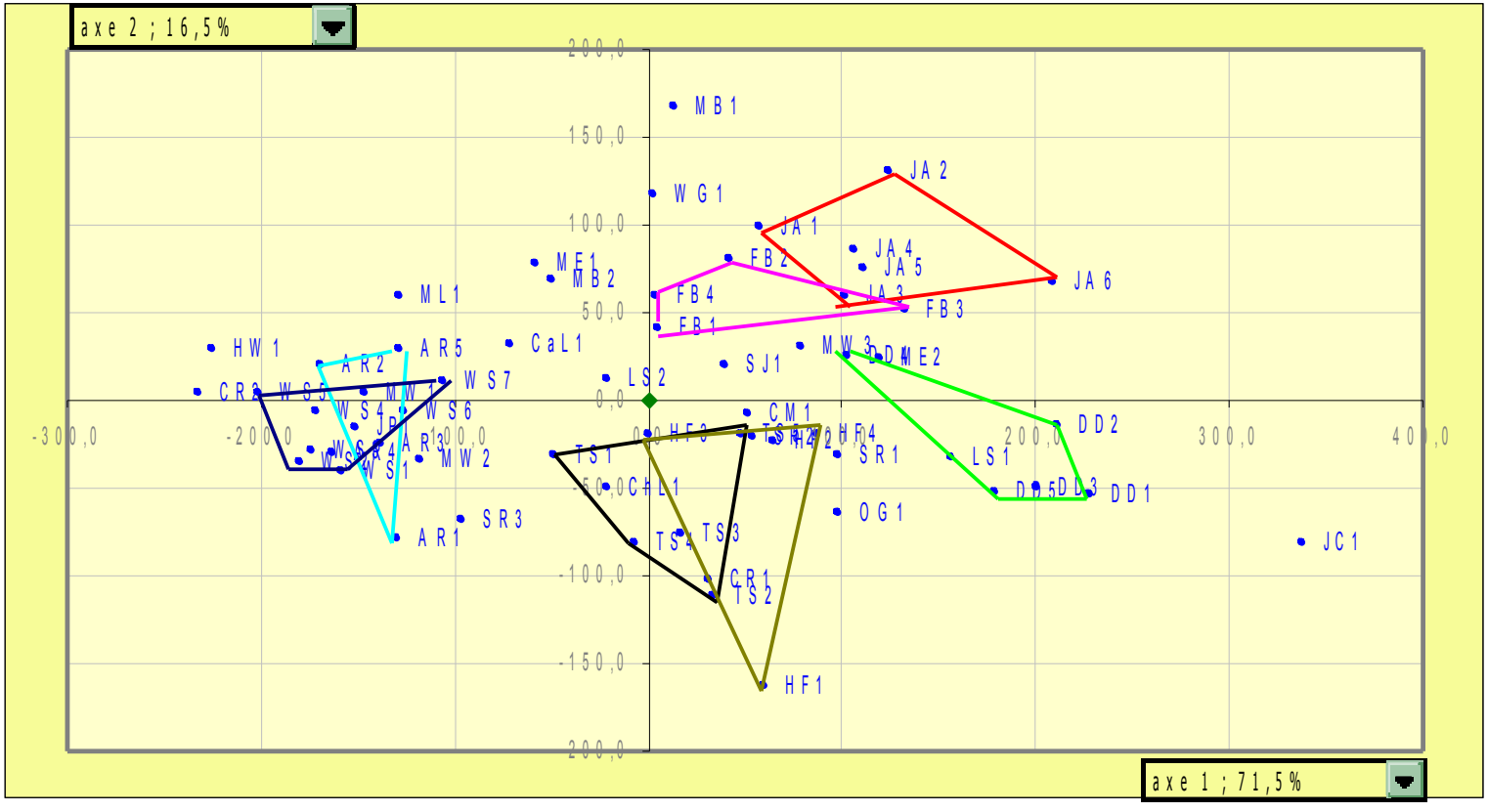
Réduction à 6 du nombre de variables prises en compte dans l'ACP



(N=60 V=6, M=VarCov, PF2)

Cartographie des quantifieurs : Résultats et consolidations (13/15)

Réduction à 3 du nombre de variables prises en compte dans l'ACP



On retrouve l'essentiel de la cartographie des auteurs mais sans les dimensions permettant d'améliorer la séparabilité

(N=60 V=3, M=VarCov, PF1)

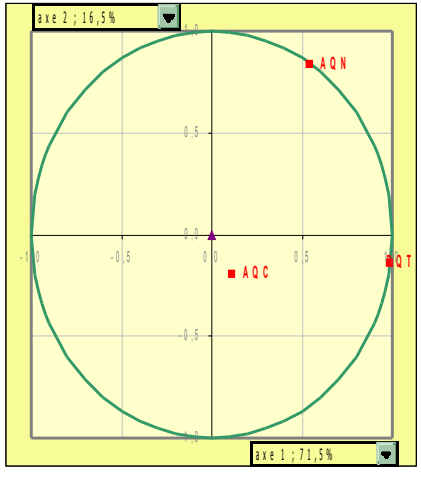
Cartographie des quantifieurs : Résultats et consolidations (14/15)

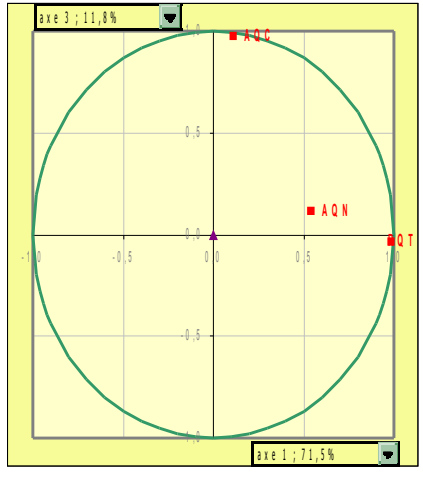
Réduction à 3 du nombre de variables prises en compte dans l'ACP

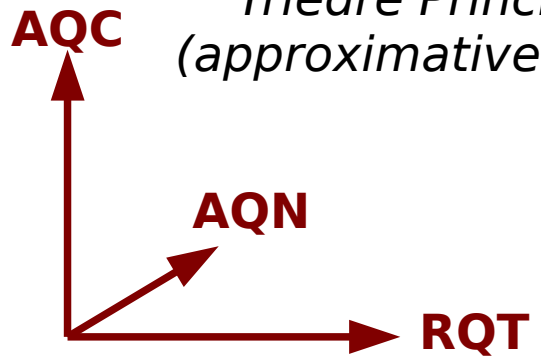
(N=60 V=3, M=VarCov, PF2)

Cartographie des quantifieurs : Résultats et consolidations (15/15)

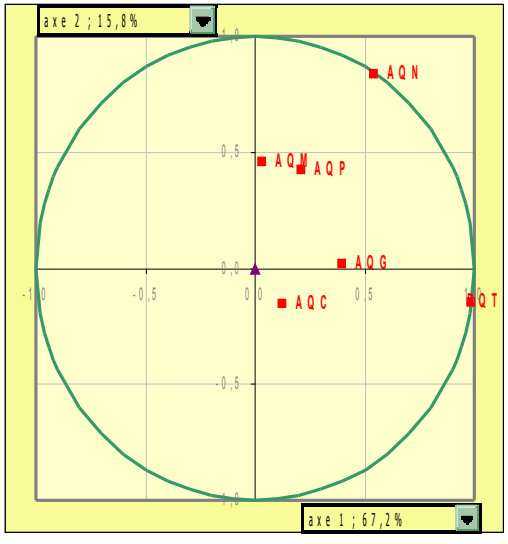
Identification des axes de l'analyse réduite

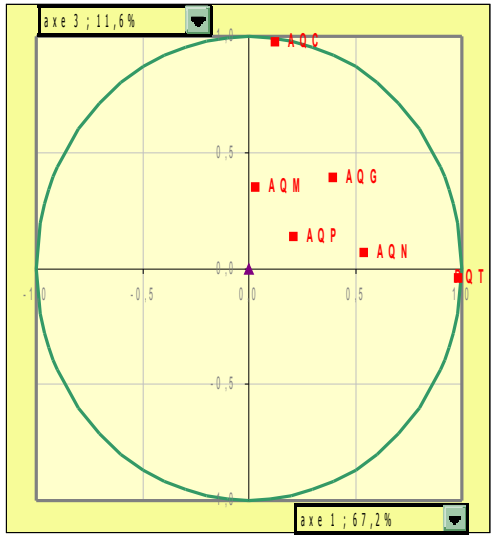


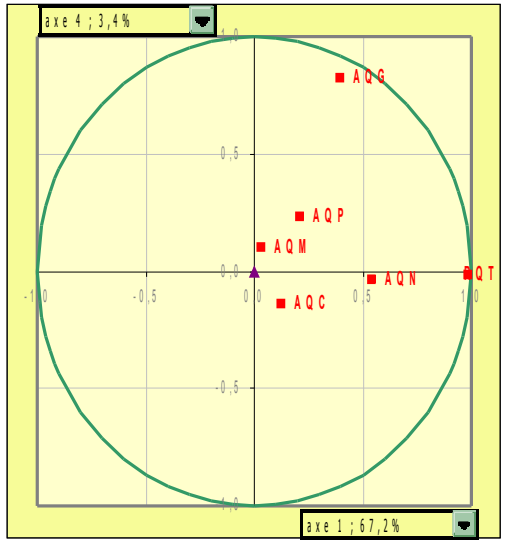




Trièdre Principal (approximativement)

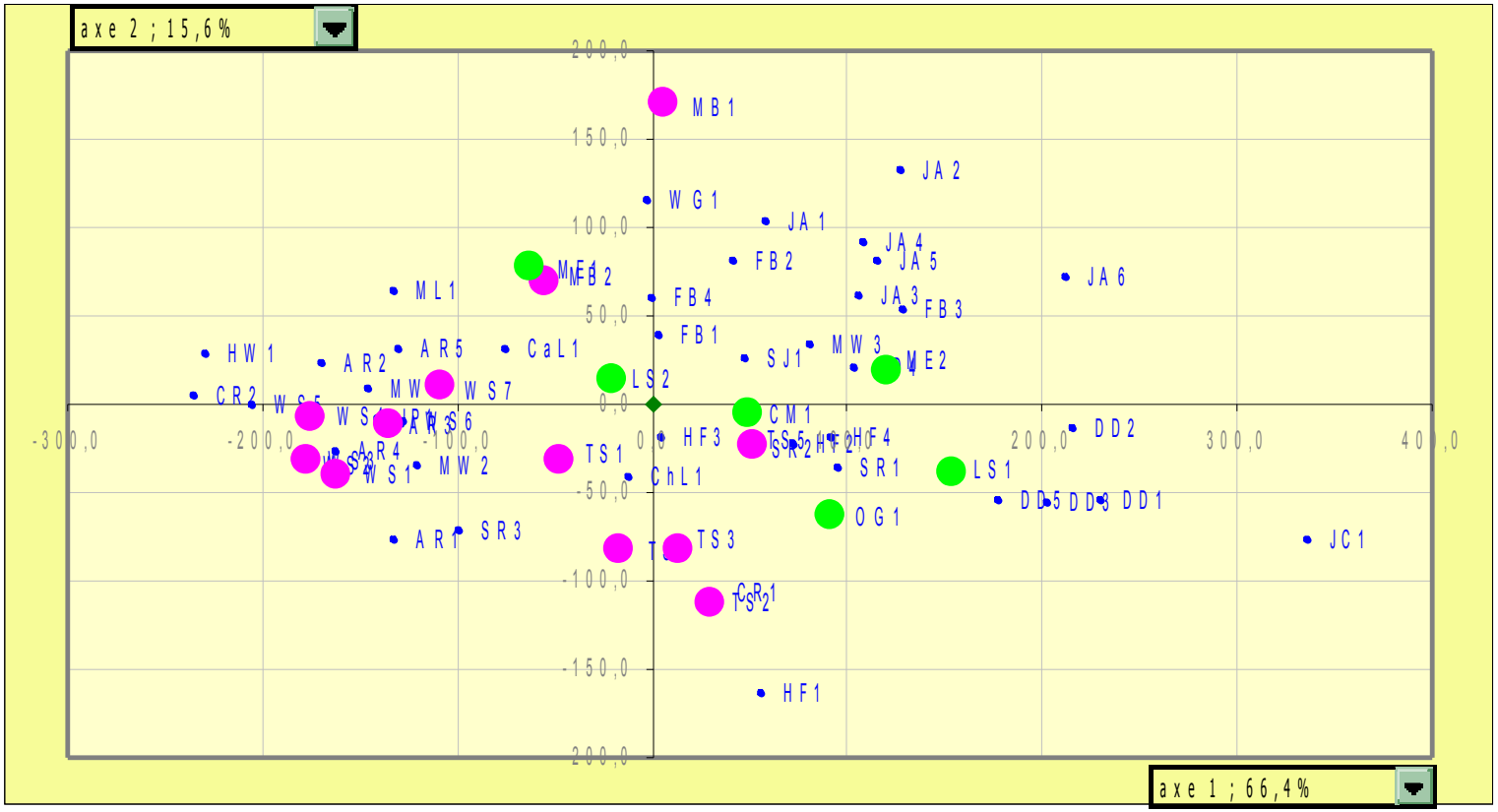






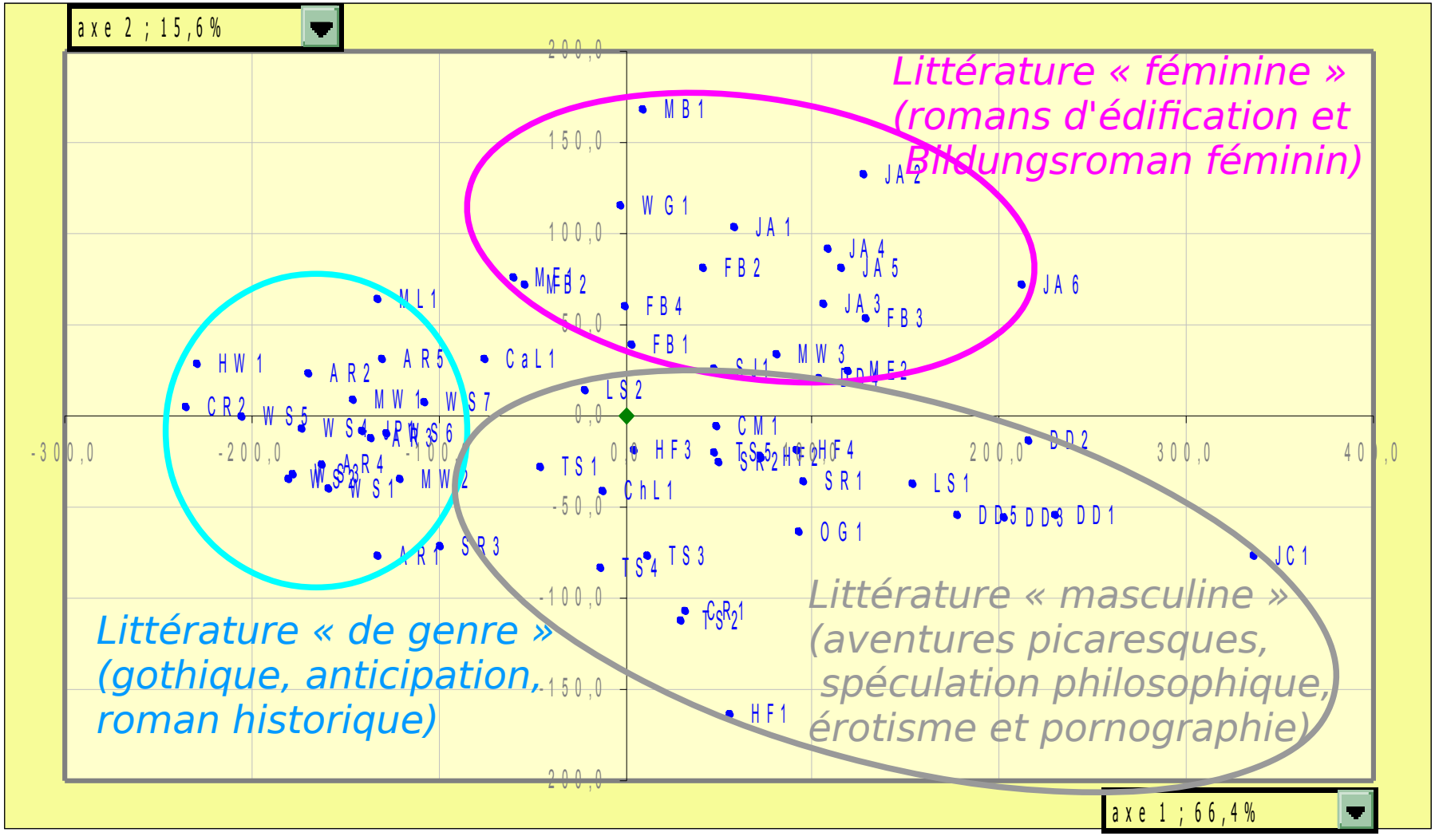
université PARIS DIDEROT PARIS 7

Cartographie des quantifieurs : Interprétations (1/9)



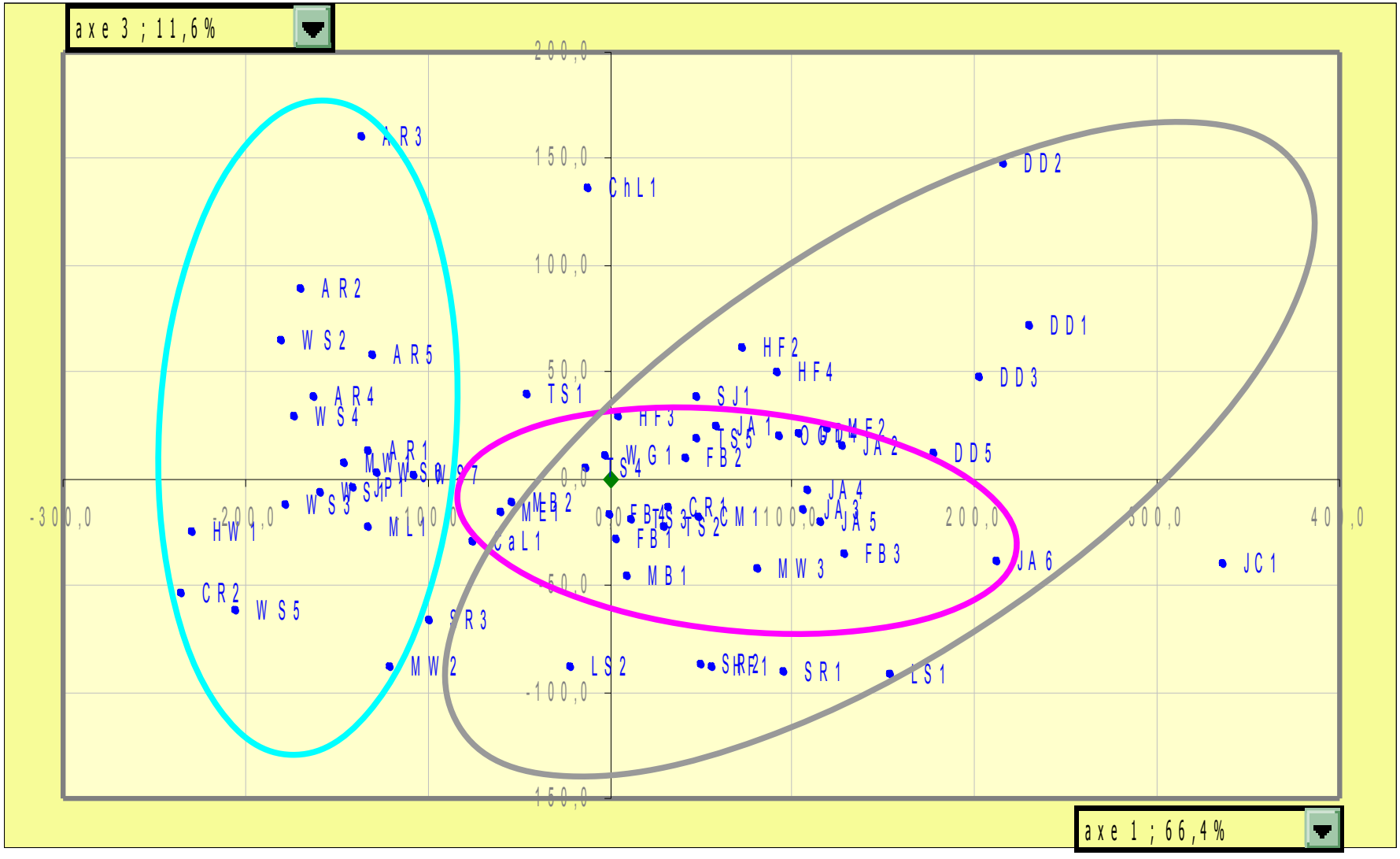
Marquage des nationalités (N=60 V=9, M=VarCov, PF1)
Ecossais, Irlandais: aucune séparabilité discernable sur cette base

Cartographie des quantifieurs : Interprétations (2/9)

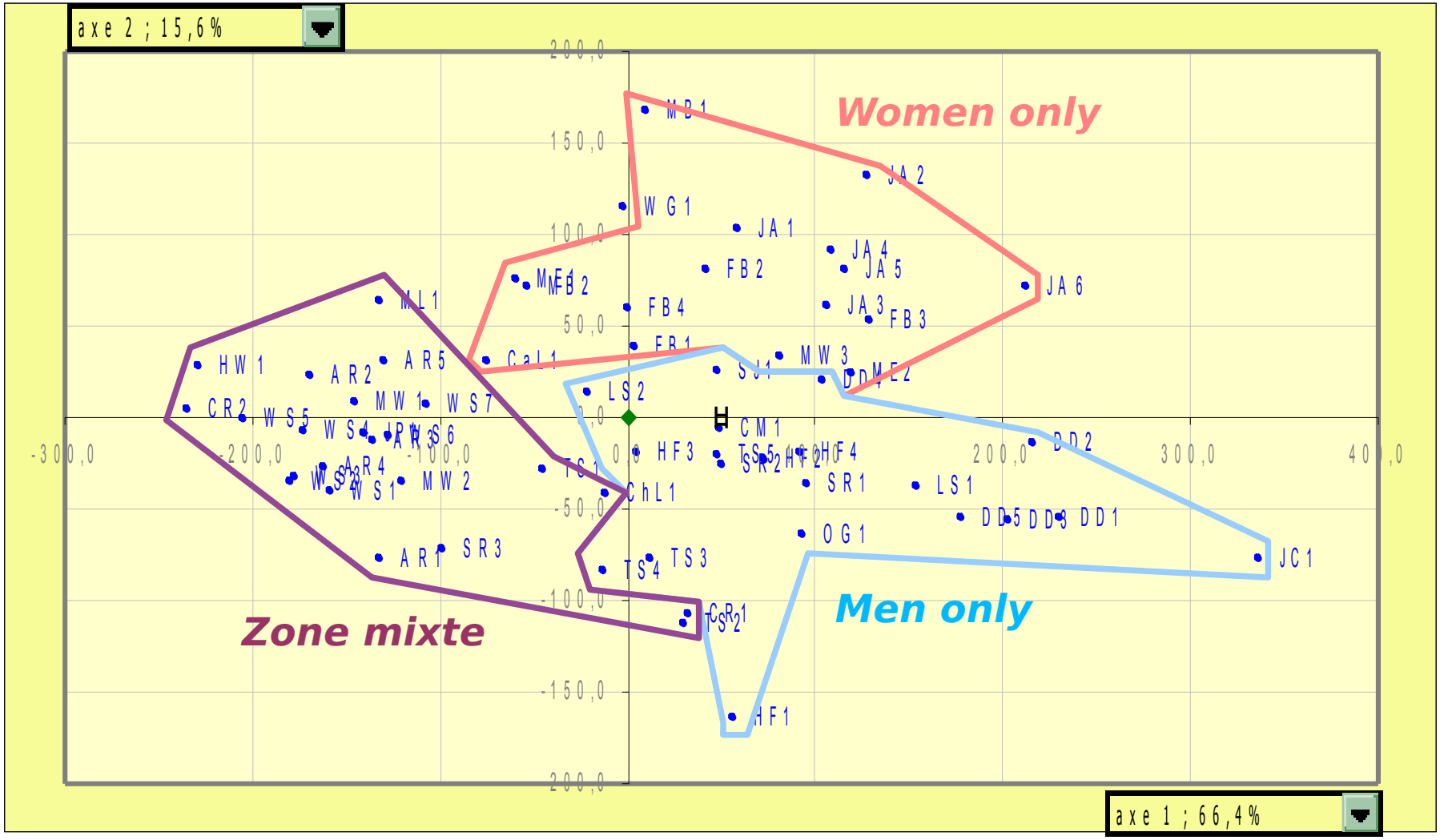


Cartographie globale des genres (N=60 V=9, M=VarCov, PF2)

Cartographie des quantifieurs : Interprétations (3/9)

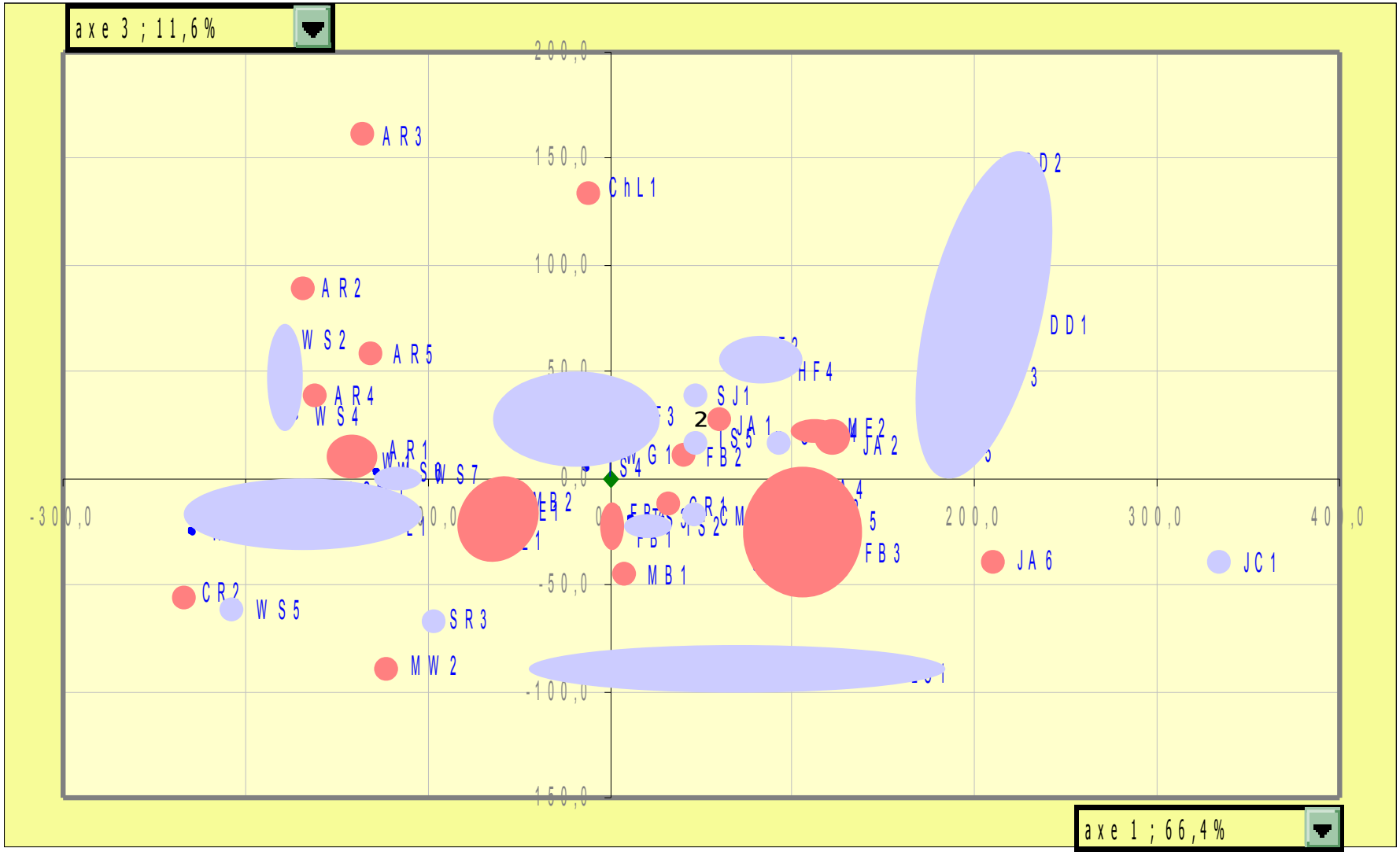


Cartographie des quantifieurs : Interprétations (4/9)



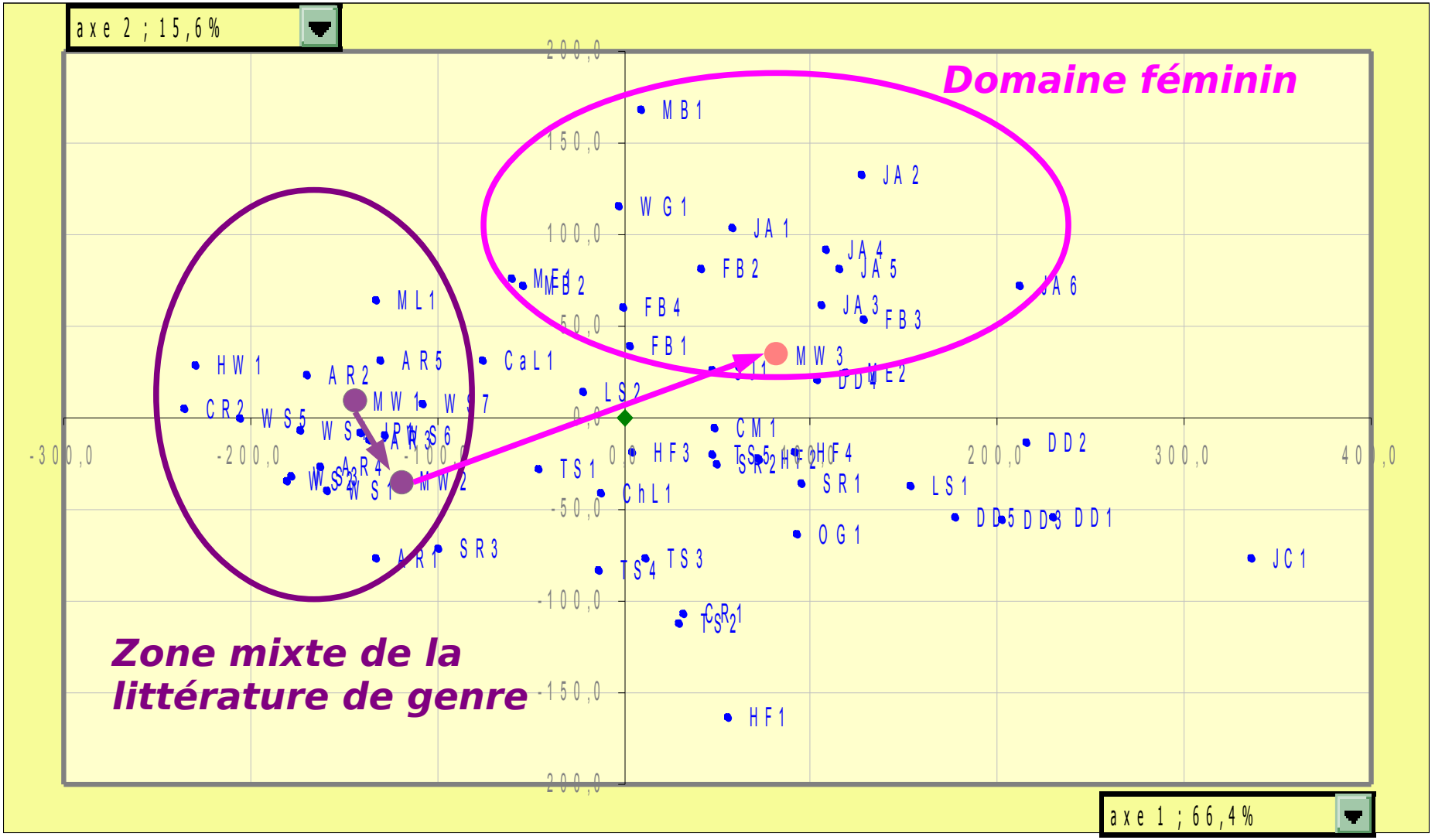
Territoires des bleus et des roses (N=60 V=9, M=VarCov, PF1)

Cartographie des quantifieurs : Interprétations (5/9)



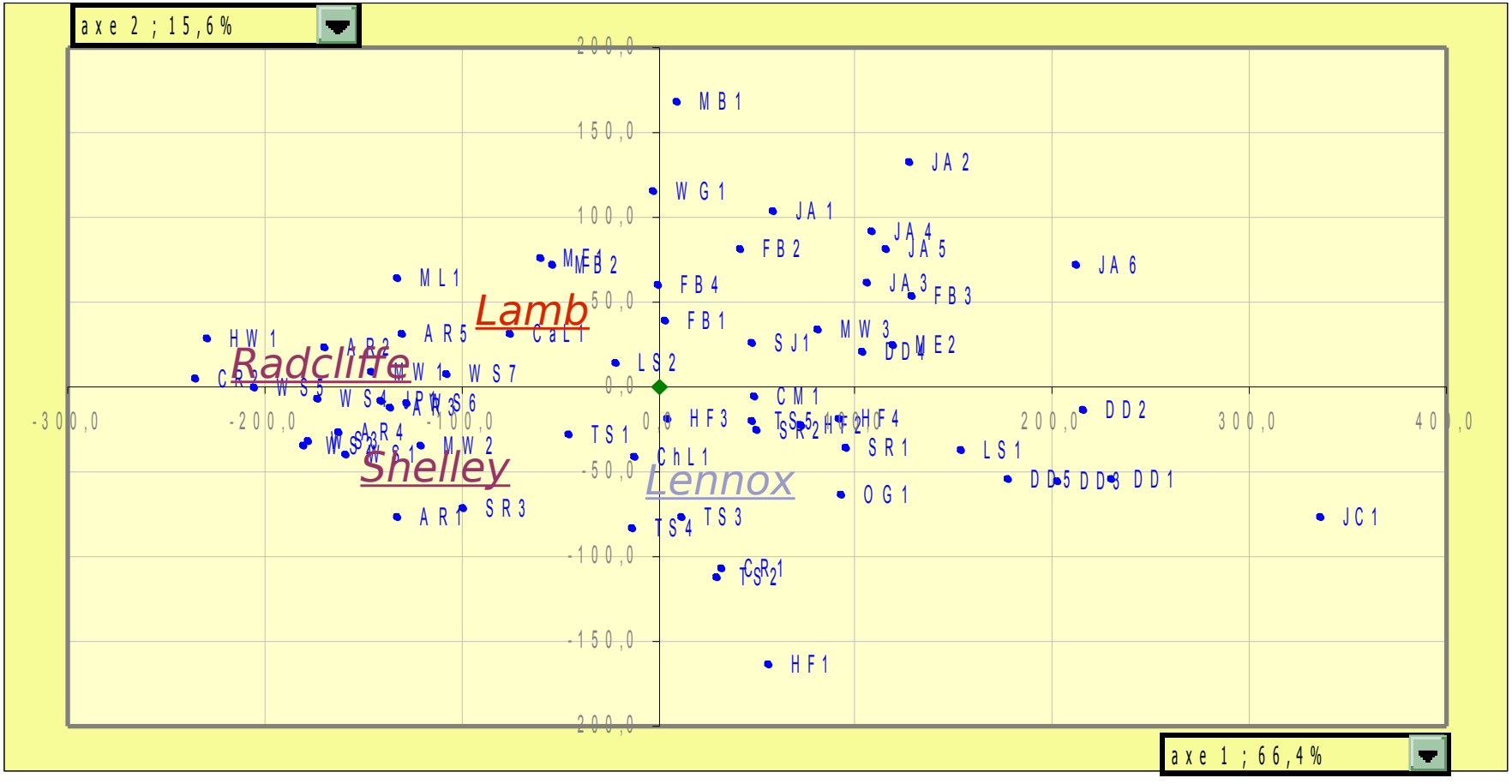
Territoires masculins et féminins (N=60 V=9, M=VarCov, PF2)

Cartographie des quantifieurs : Interprétations (6/9)



Quand Mary Shelley change de genre: de Frankenstein à Mathilda

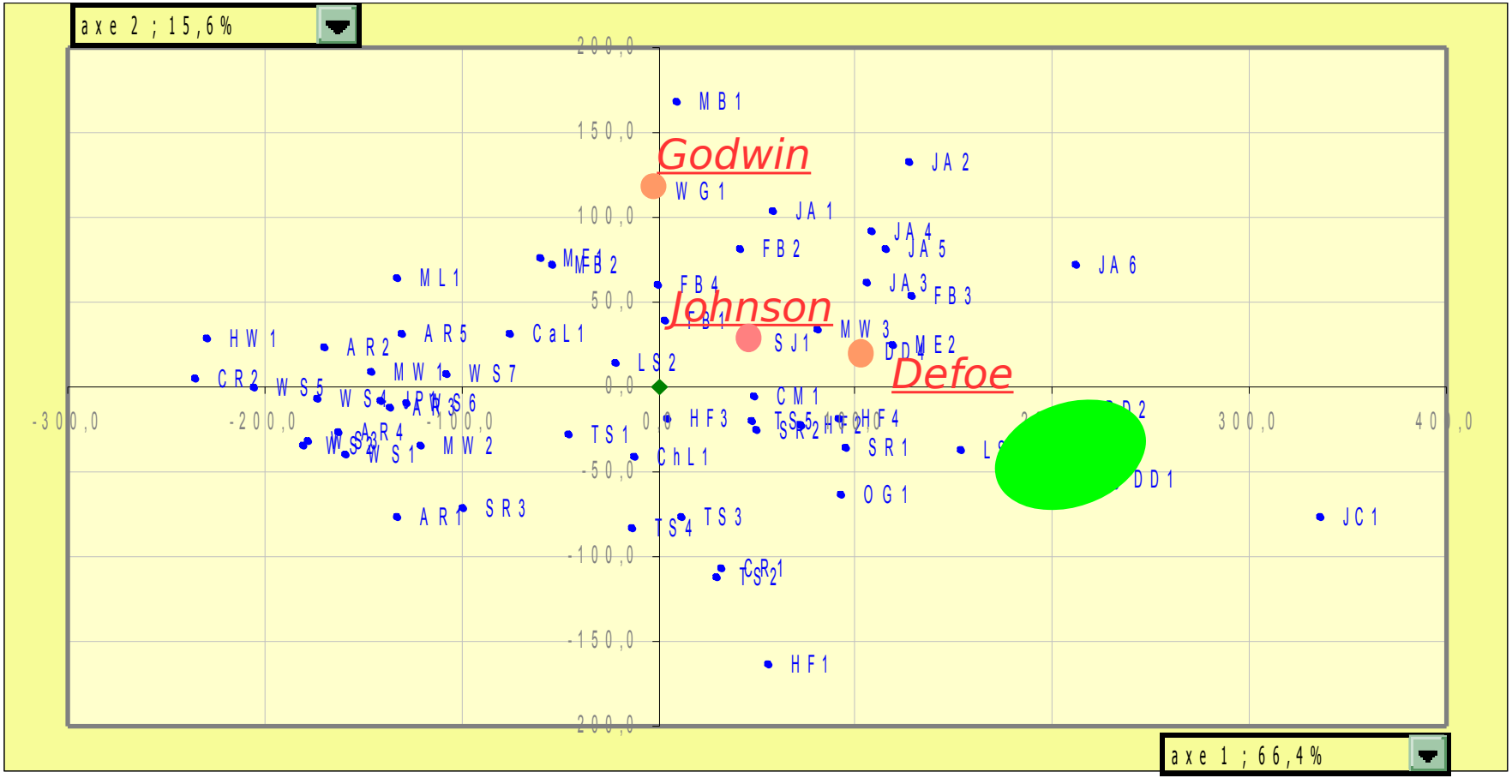
Cartographie des quantifieurs
Interprétations (7/9)



De fortes personnalités: le(s) mauvais genre(s) des « transgressives »
A. Radcliffe (reine du gothique) Ch. Lennox (''The Female Quixote'')
Lady Caroline Lamb (aristocrate à scandales, maîtresse de Byron)
M. Wollstonecraft-Shelley (féminisme + romans-catastrophes)

université PARIS DIDEROT PARIS 7

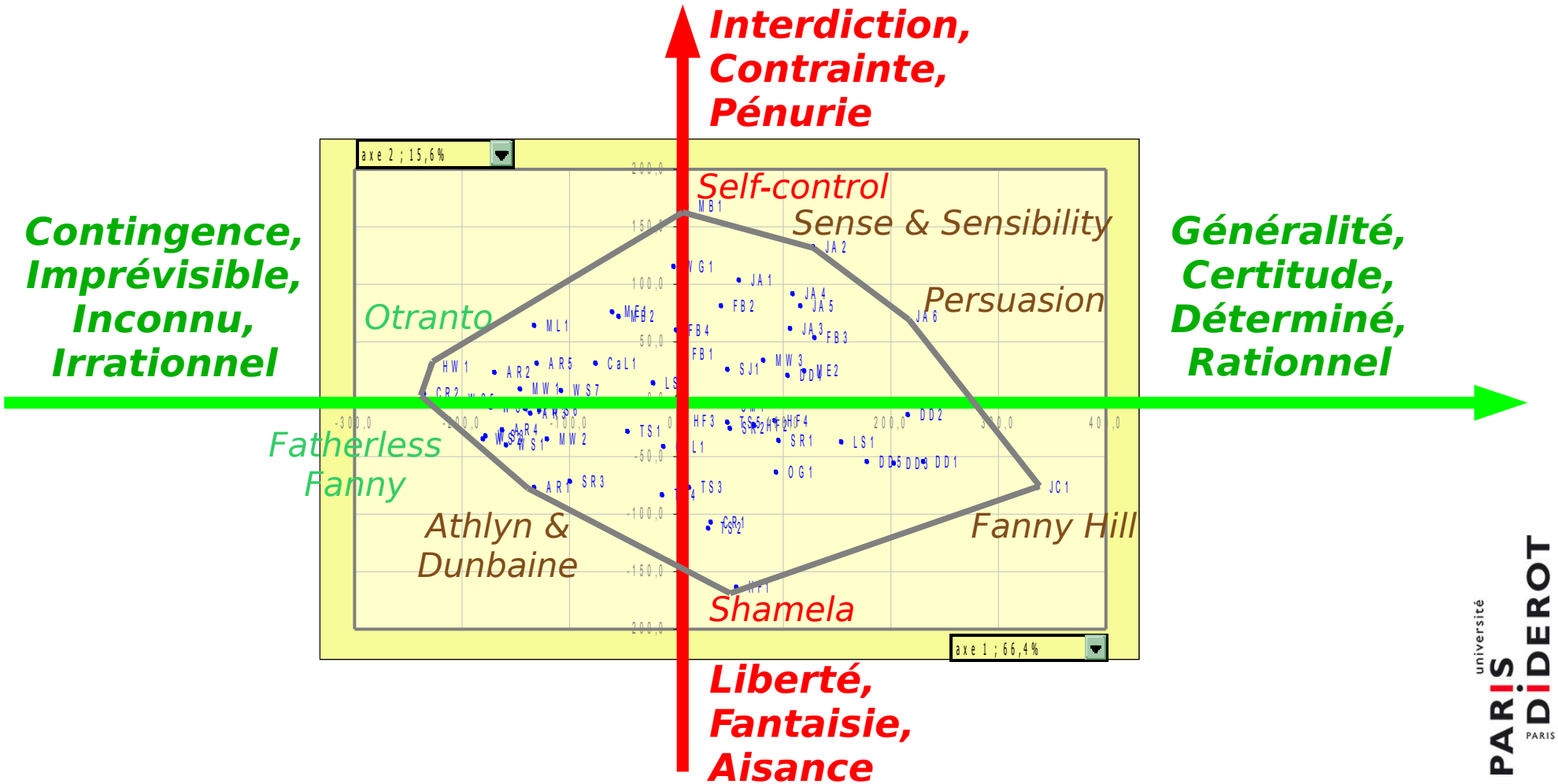
Cartographie des quantifieurs
Interprétations (8/9)



Des auteurs masculins dont la cartographie révèle un versant féminin:
W. Godwin (révolutionnaire, époux de la féministe M. Wollstonecraft)
S. Johnson (épousa une veuve de 20 ans son aînée: scandale et procès)
D. Defoe (Moll Flanders récit au féminin plus empathique que Roxana)

Cartographie des quantifieurs : Interprétations (9/9)

L'usage des quantifieurs structure la vision du monde et caractérise des régimes de discours sur le monde:



Cartographie des quantifieurs: Conclusions (1/2)

→ Les résultats obtenus confirment empiriquement le lien établi par la Théorie Générale des Quantifieurs entre le maniement des quantifieurs et la construction d'ontologies (pris au sens de la spécification des propriétés logiques d'entités liées les unes aux autres au sein d'un univers fini donné.)
→ Toute oeuvre littéraire porte en elle une représentation du monde, une ontologie qui reste souvent implicite, mais qui s'exprime à travers les discours narratifs.
→ La cartographie des quantifieurs produite par l'ACP reflète certains traits saillants de ces représentations du monde (avec un niveau de sensibilité et de cohérence d'ensemble totalement inattendu au départ de cette étude)
Pour moi, le résultat le plus surprenant reste que les 3 quantifieurs QT (le tout) QN (le rien) et QC (le quelque chose) gouvernent l'essentiel de la cartographie !

Cartographie des quantifieurs: Conclusions (2/2)

H1 : la distribution des quantifieurs est un indicateur du sexe de l'auteur
Elle ne l'est qu'indirectement, par le biais du genre littéraire de préférence

H2 : la distribution des quantifieurs est un indicateur de genre littéraire
On peut délimiter des sous-espaces fortement liés à des genres thématiques

H3 : la distribution des quantifieurs est un indicateur de style
Des polygones d'exclusivité existent pour certains auteurs, et dans un genre donné, les différentes oeuvres d'un auteur donné restent proches

H4 : vis-à-vis des quantifieurs, le genre littéraire domine le style
Au sein d'un genre, les auteurs peuvent souvent apparaître en recouvrement mais restent séparables selon certains plans (eg Fielding vs Smollett)

H5 : vis-à-vis des quantifieurs, le style domine le genre littéraire
En cas de changement de genre, la position change (eg M. Shelley)

H6 : la distribution des quantifieurs est un indicateur de nationalité
Il n'y a pas de « domaine réservé » spécifique aux Ecossais ou Irlandais