



EVOLLEX : analyses statistiques et exploration des embeddings

Thématiques Actuelles de la Recherche en TAL

18 décembre 2017





QUELQUES ANALYSES STATISTIQUES

■ Questions de recherche

- Y a-t-il un lien entre les mesures calculées sur les paires et les autres caractéristiques des réponses ?

- *Annotations :*

- Certaines méthodes de mesure de la similarité privilégient-elles des relations sémantiques particulières ?

- *Distribution des réponses*

- Peut-on identifier les réponses les plus attendues

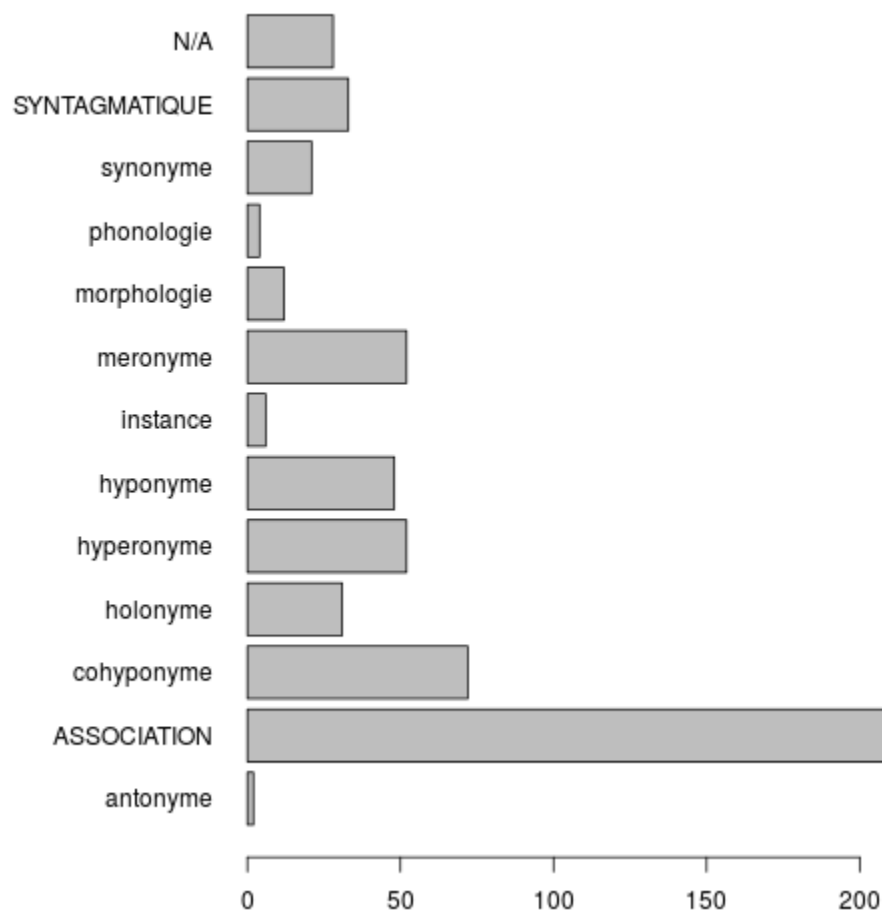
- Globalement, quels sont les comportements globaux observables sur ces données

- *En croisant l'ensemble des caractéristiques*

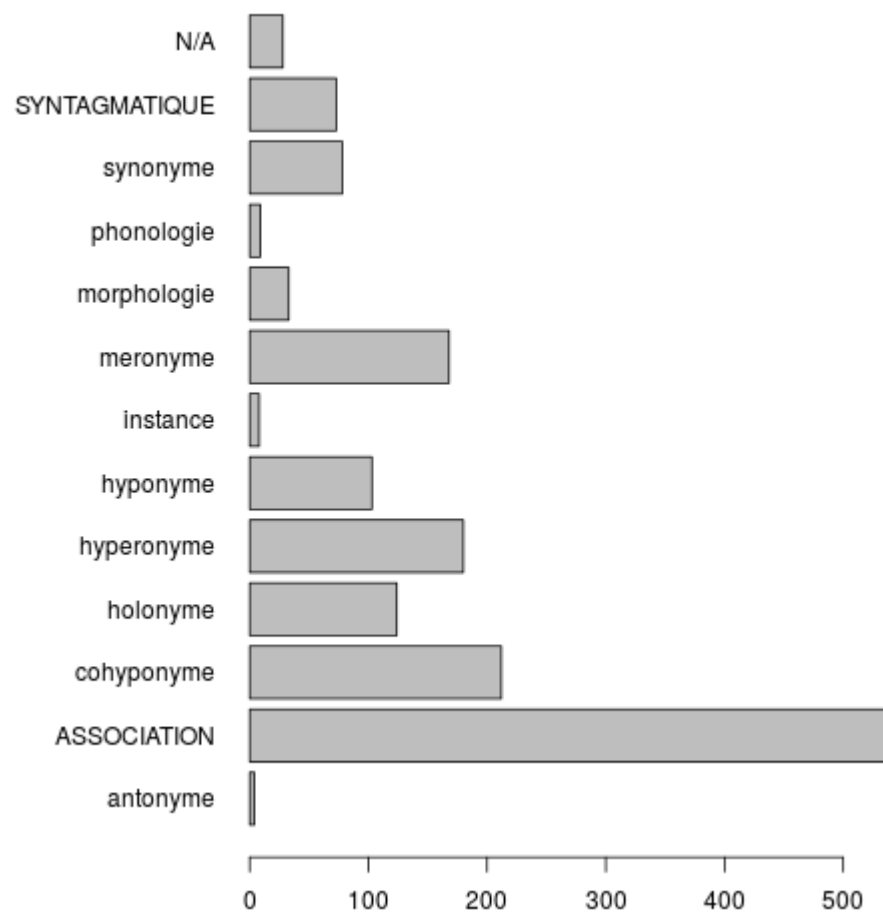


Annotation des relations et fréquence

Nb de paires



Nb de réponses



■ Regroupement des annotations

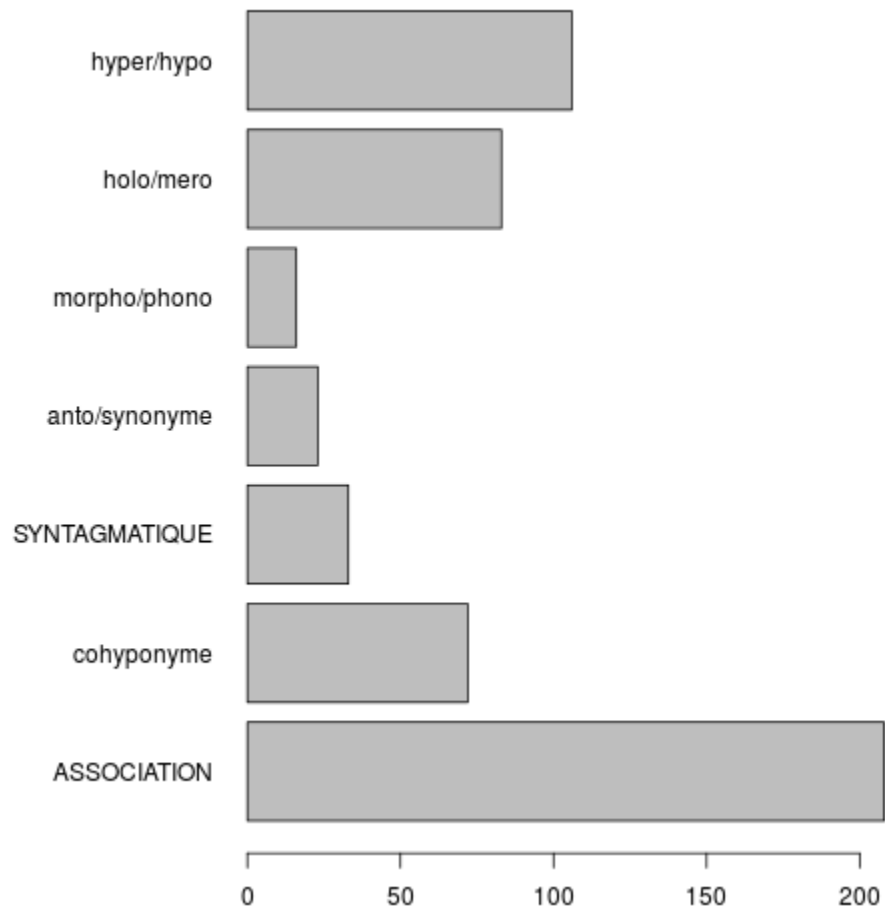
□ Niveau 1 :

- *Hyper+hypo+instance*
- *Antonyme+synonyme*
- *Morphologie+phonologie*
- *Holonymie+meronymie*

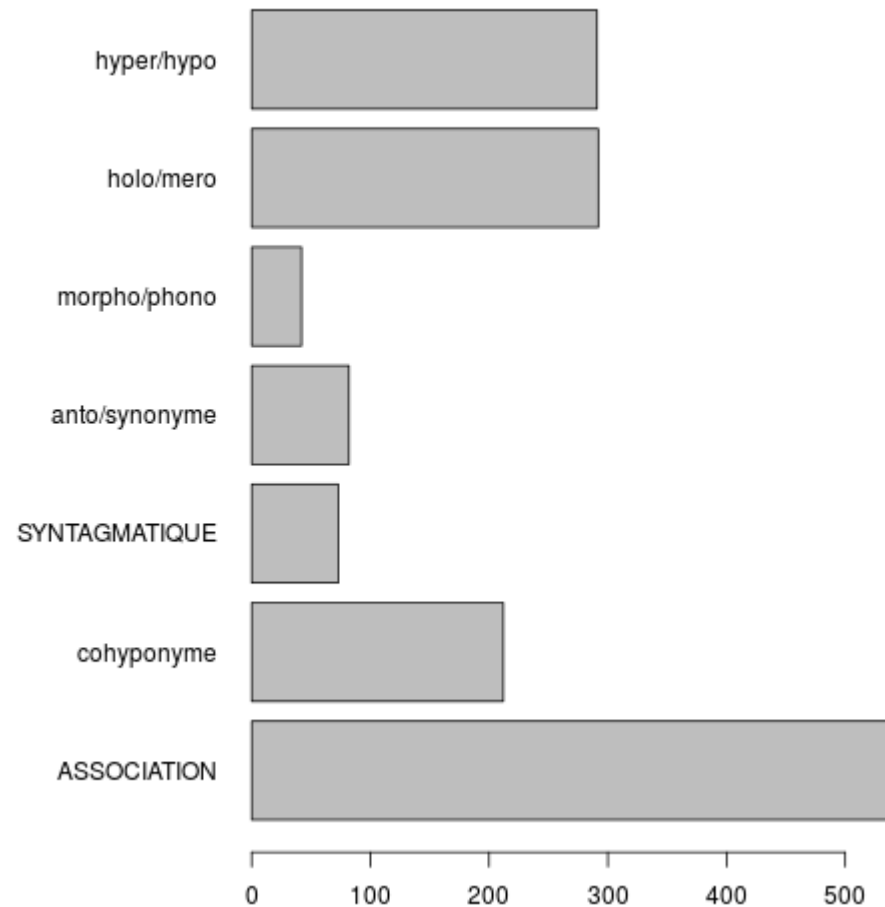
□ Niveau 2 :

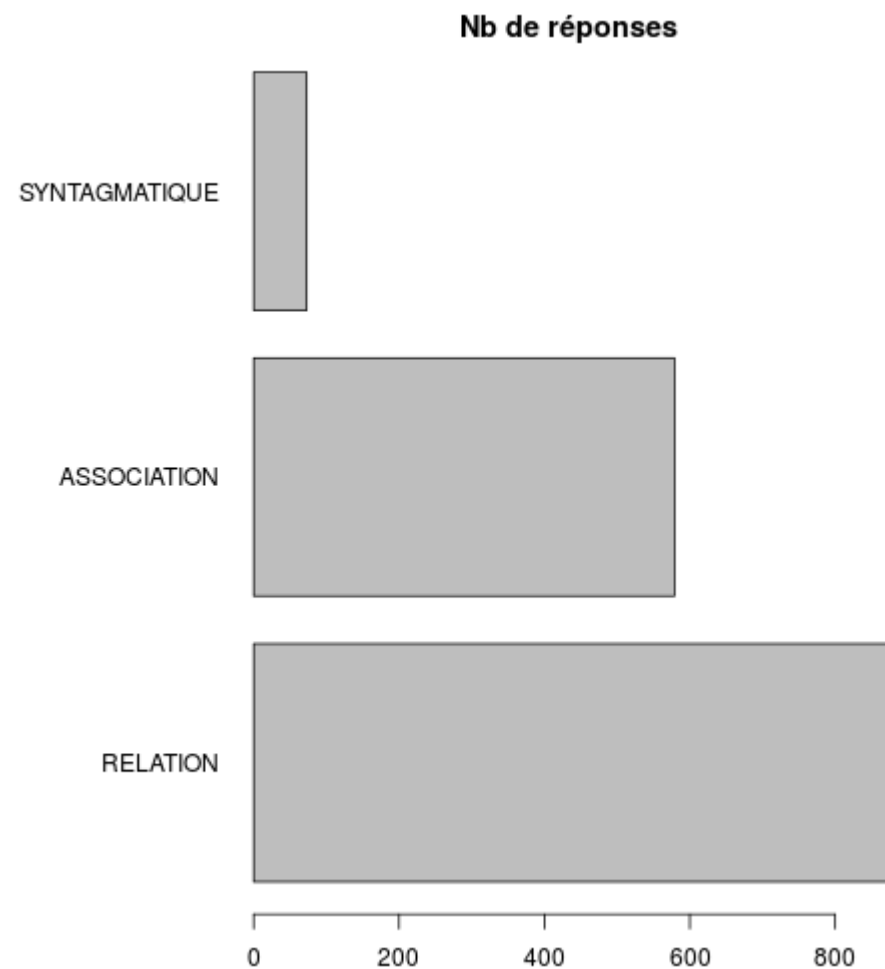
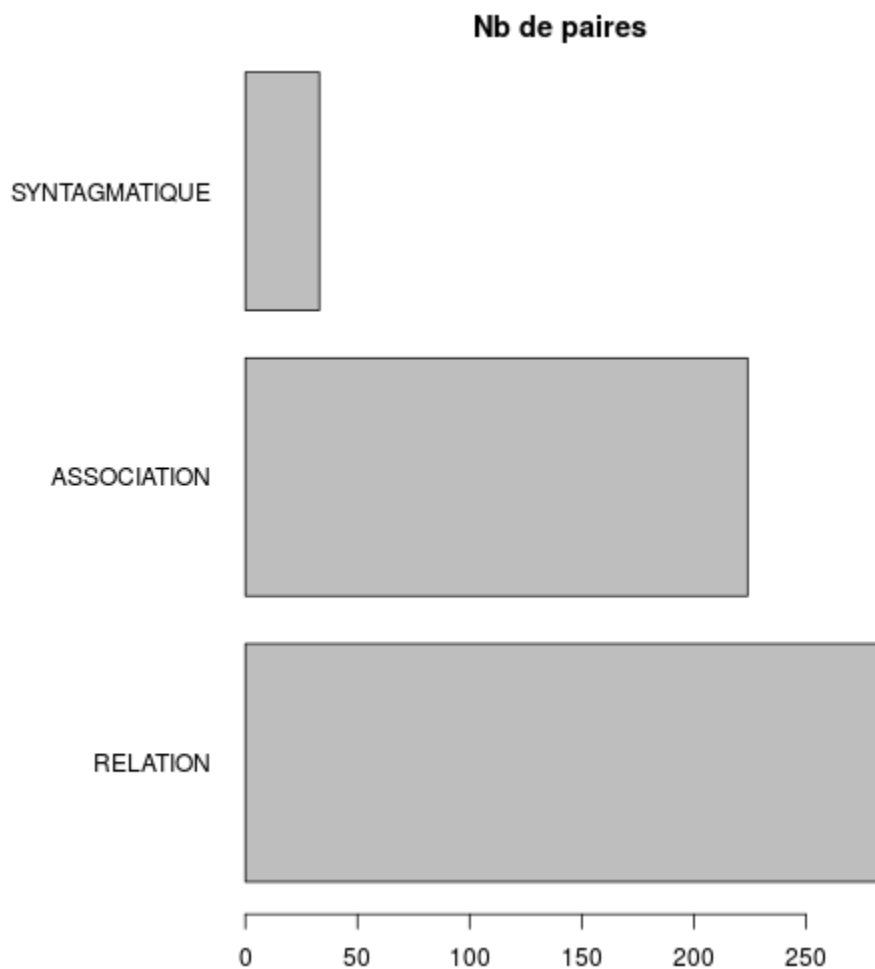
- *Syntagmatique*
- *Association (ou morpho/phono)*
- *Relation sémantique (quelconque)*

Nb de paires



Nb de réponses







Fréquence (en corpus) des réponses

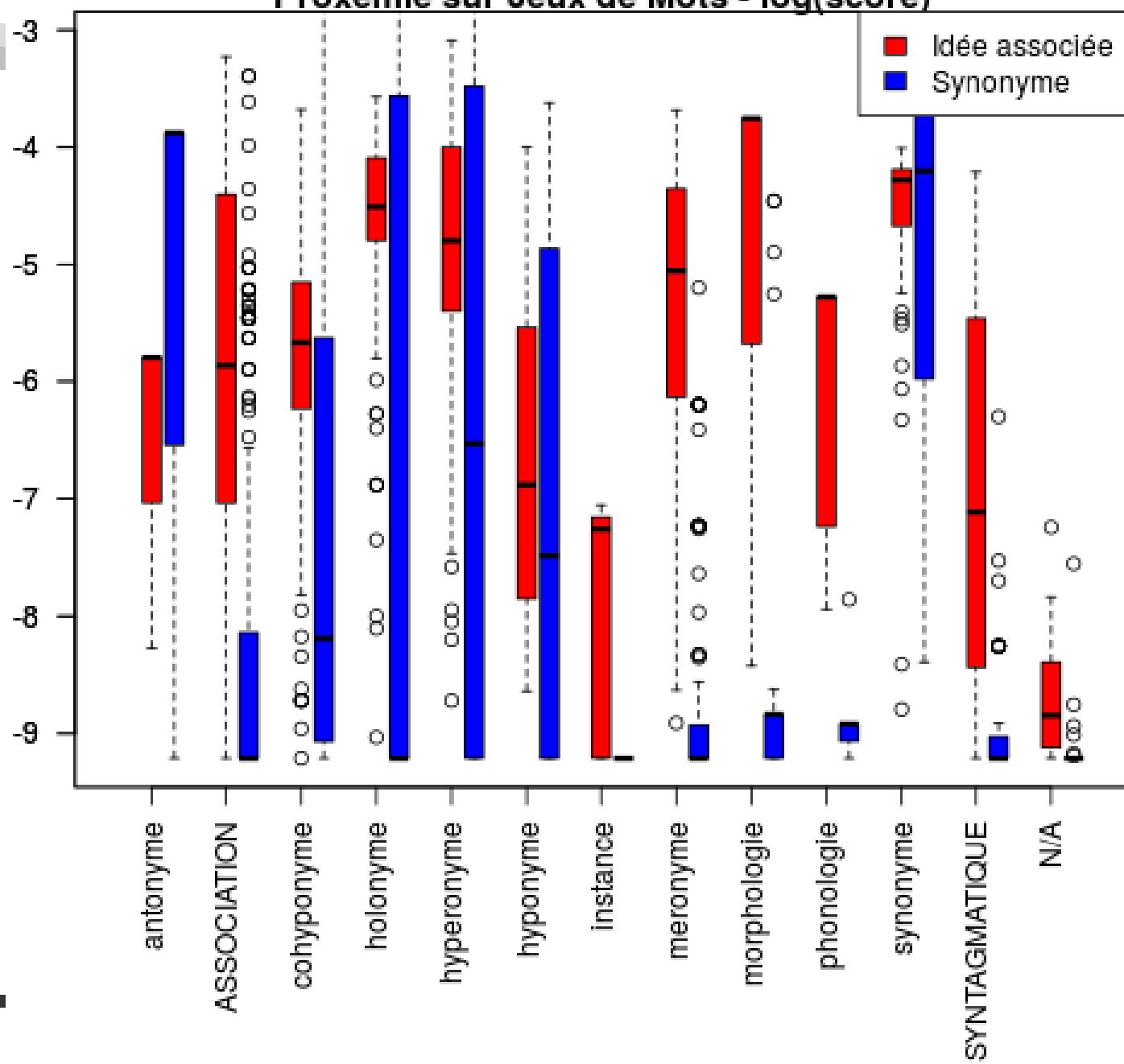
- Calcul de la fréquence de la réponse en corpus (FRWAC, log)
- Corrélation avec le nombre de sujets l'ayant proposée :
 - +0.21 ($p < 0.01$)
- Plus un mot est fréquent en corpus plus il est proposé (indépendamment du stimulus)
- Pas de différence significative entre les différentes relations sémantiques



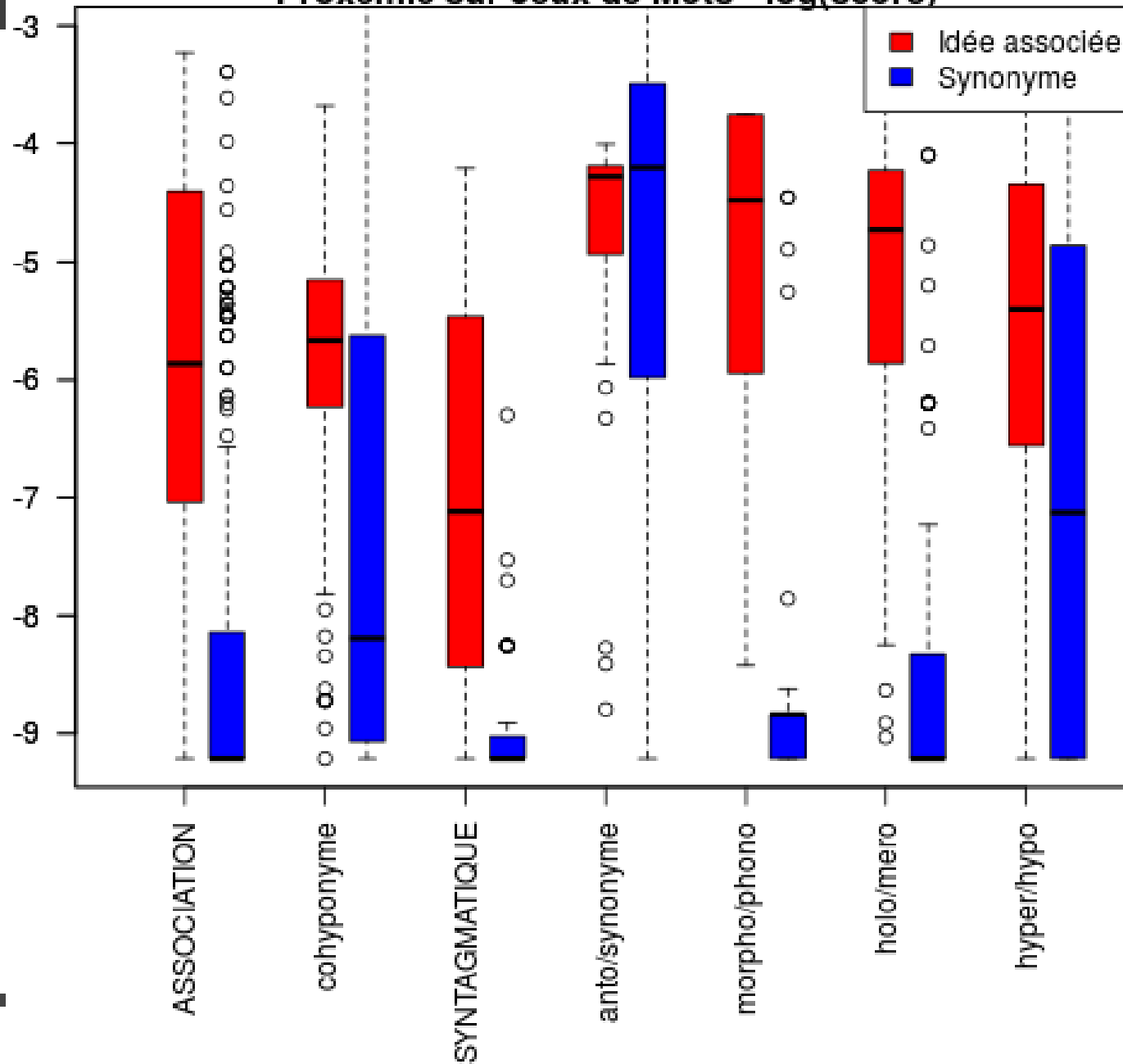
Proxémie sur la base « Jeux de mots »

- Corrélation avec le nombre de sujets l'ayant proposée :
 - Idée associée :
 - $+0.39$ ($p < 0.01$)
 - Synonymie
 - $+0.13$ ($p < 0.05$)

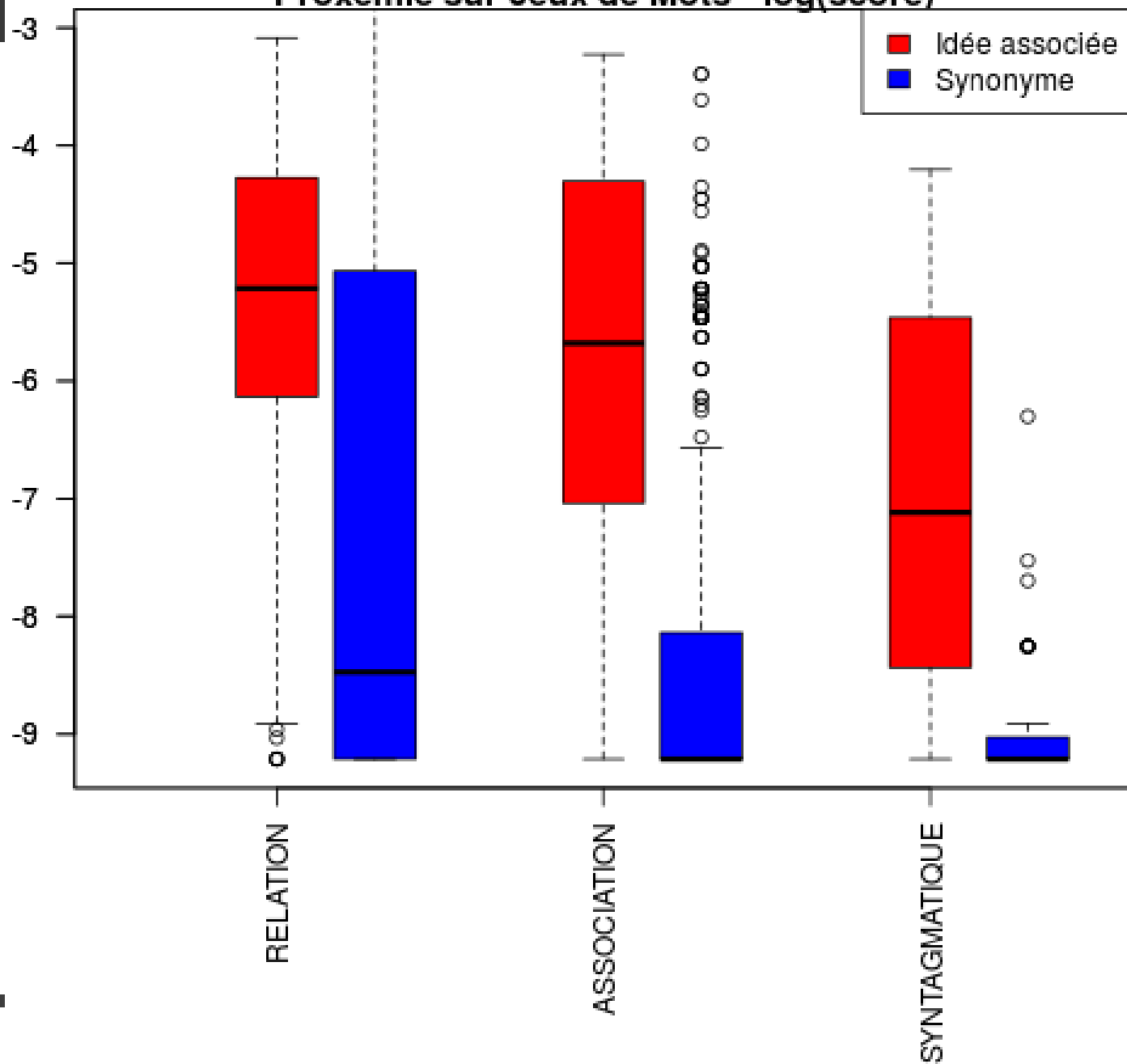
Proximité sur Jeux de Mots - log(score)



Proxémie sur Jeux de Mots - log(score)



Proxémie sur Jeux de Mots - log(score)

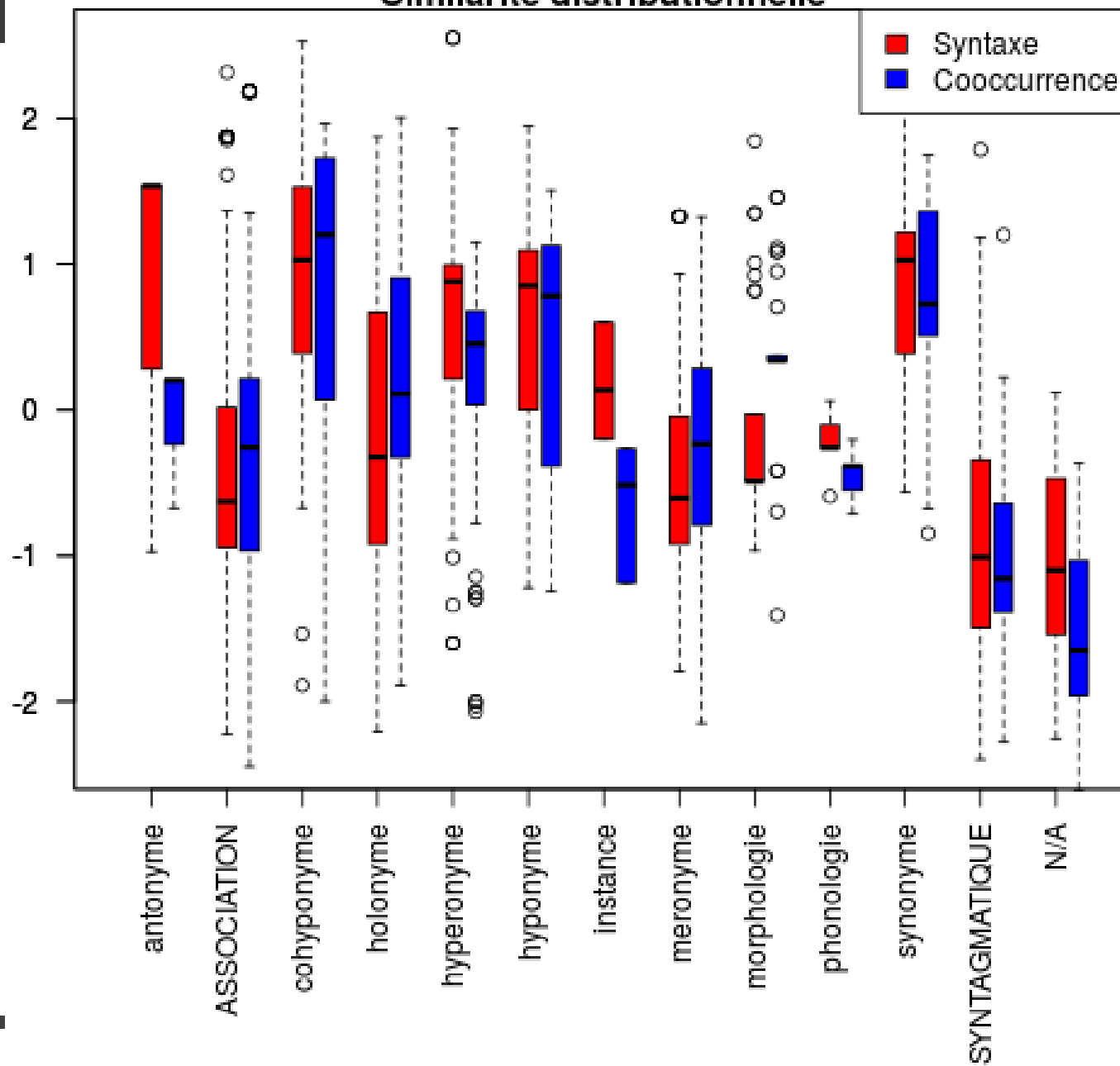




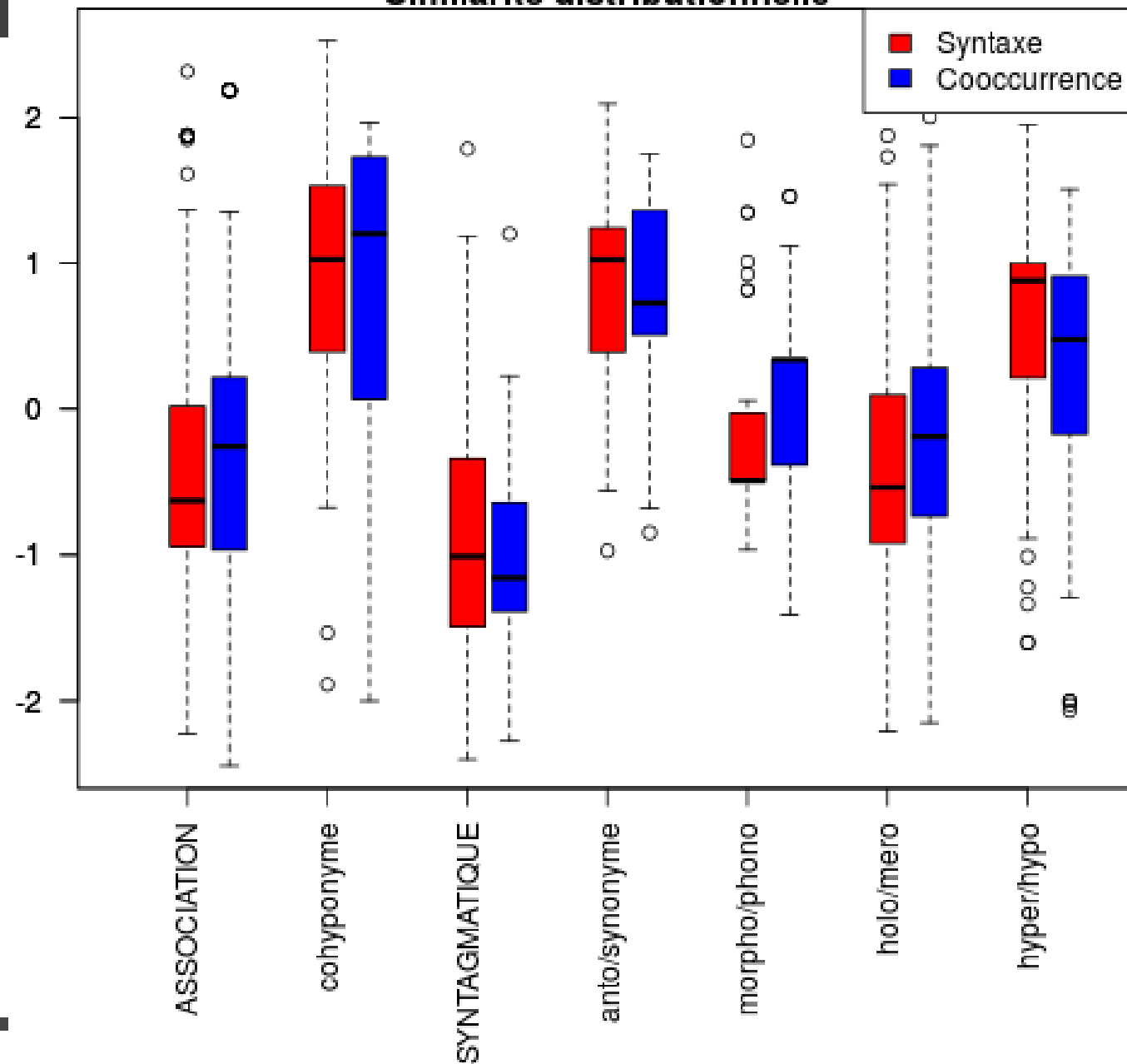
Sémantique distributionnelle

- Corrélation avec le nombre de sujets ayant proposé la réponse :
 - Méthode Word2vec Skipgram classique, cosinus brut
 - $+0.23$ ($p < 0.01$)
 - Méthode Word2vec syntaxique, cosinus brut
 - $+0.09$ ($p < 0.05$)
 - Méthode Word2vec Skipgram classique, rang parmi les voisins du stimulus
 - -0.24 ($p < 0.01$)
 - Méthode Word2vec syntaxique, rang parmi les voisins du stimulus
 - -0.05 (NS)

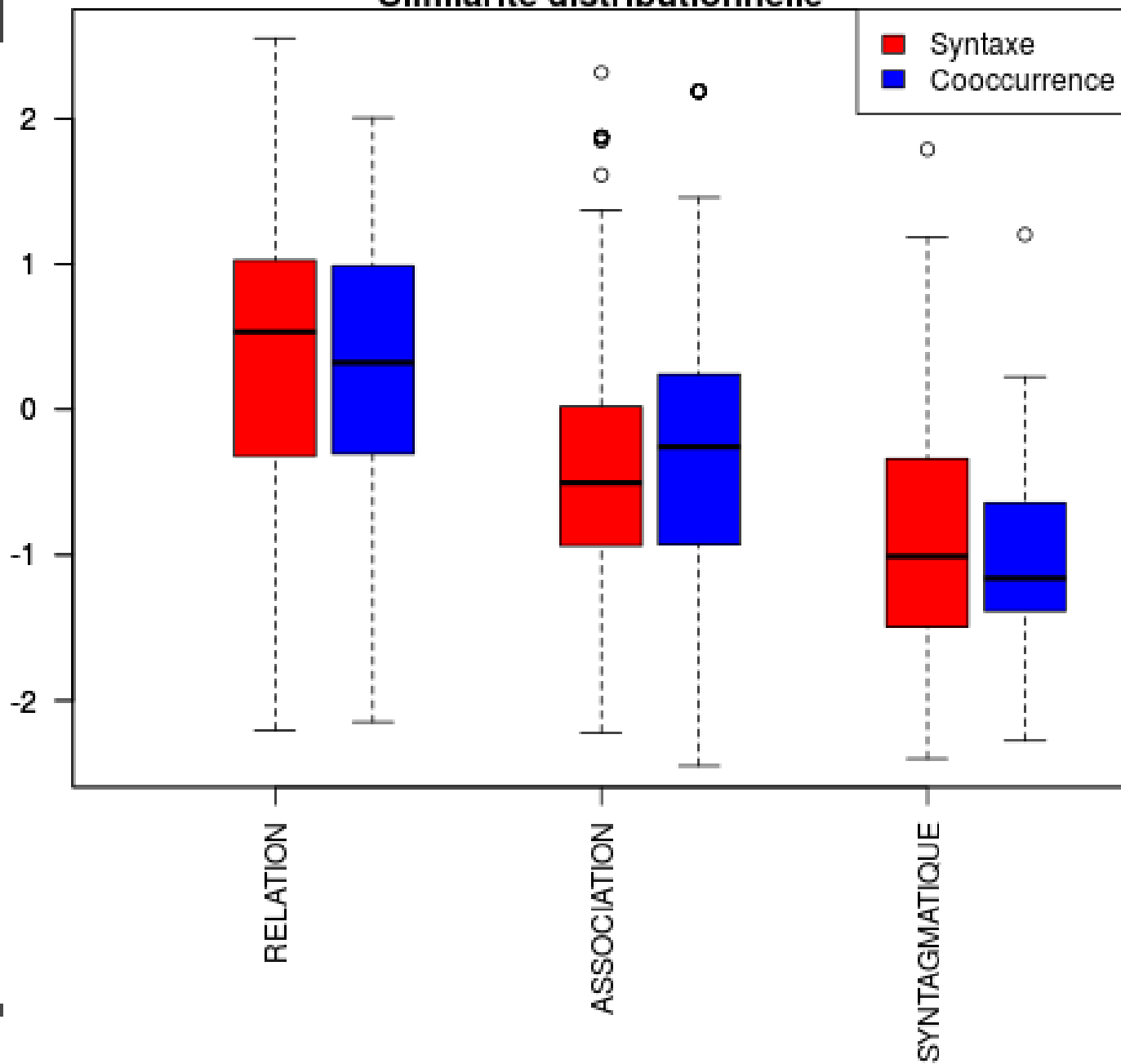
Similarité distributionnelle



Similarité distributionnelle



Similarité distributionnelle

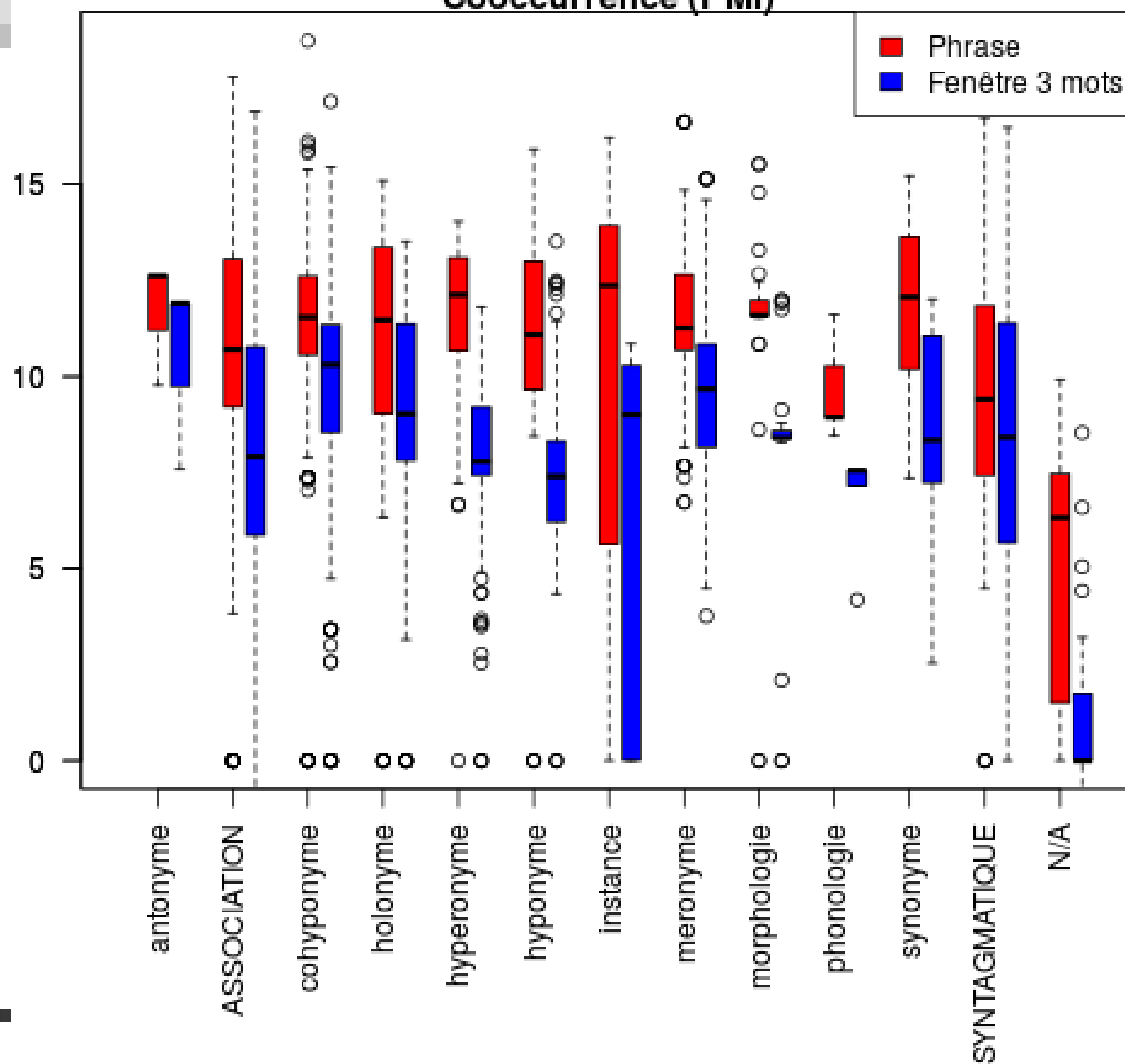




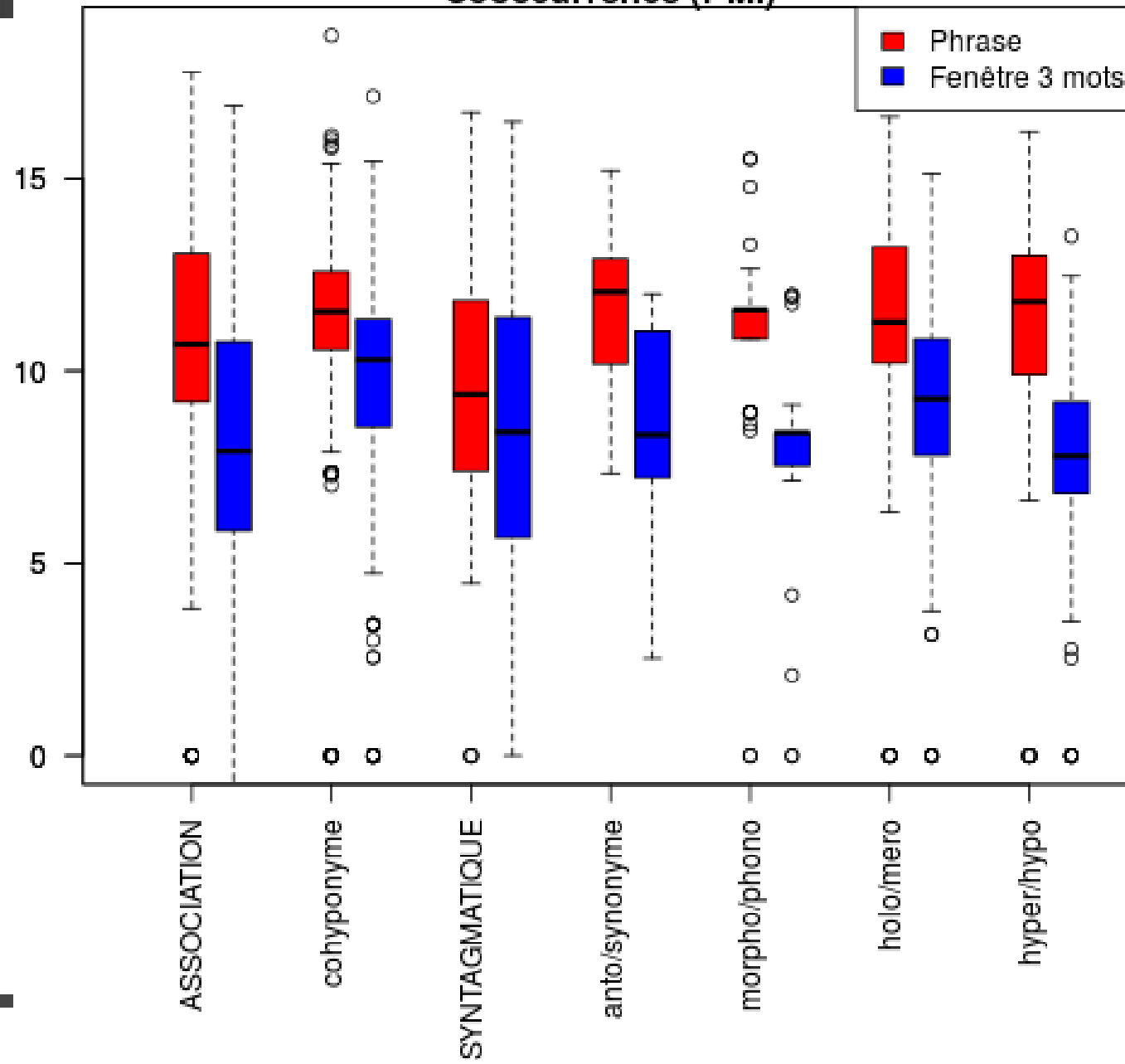
Cooccurrence

- Corrélation avec le nombre de sujets ayant proposé la réponse :
 - InfoMutuelle cooccurrence dans la phrase :
 - $+0.18$ ($p < 0.001$)
 - InfoMutuelle cooccurrence dans une fenêtre de 3 mots :
 - $+0.25$ ($p < 0.001$)

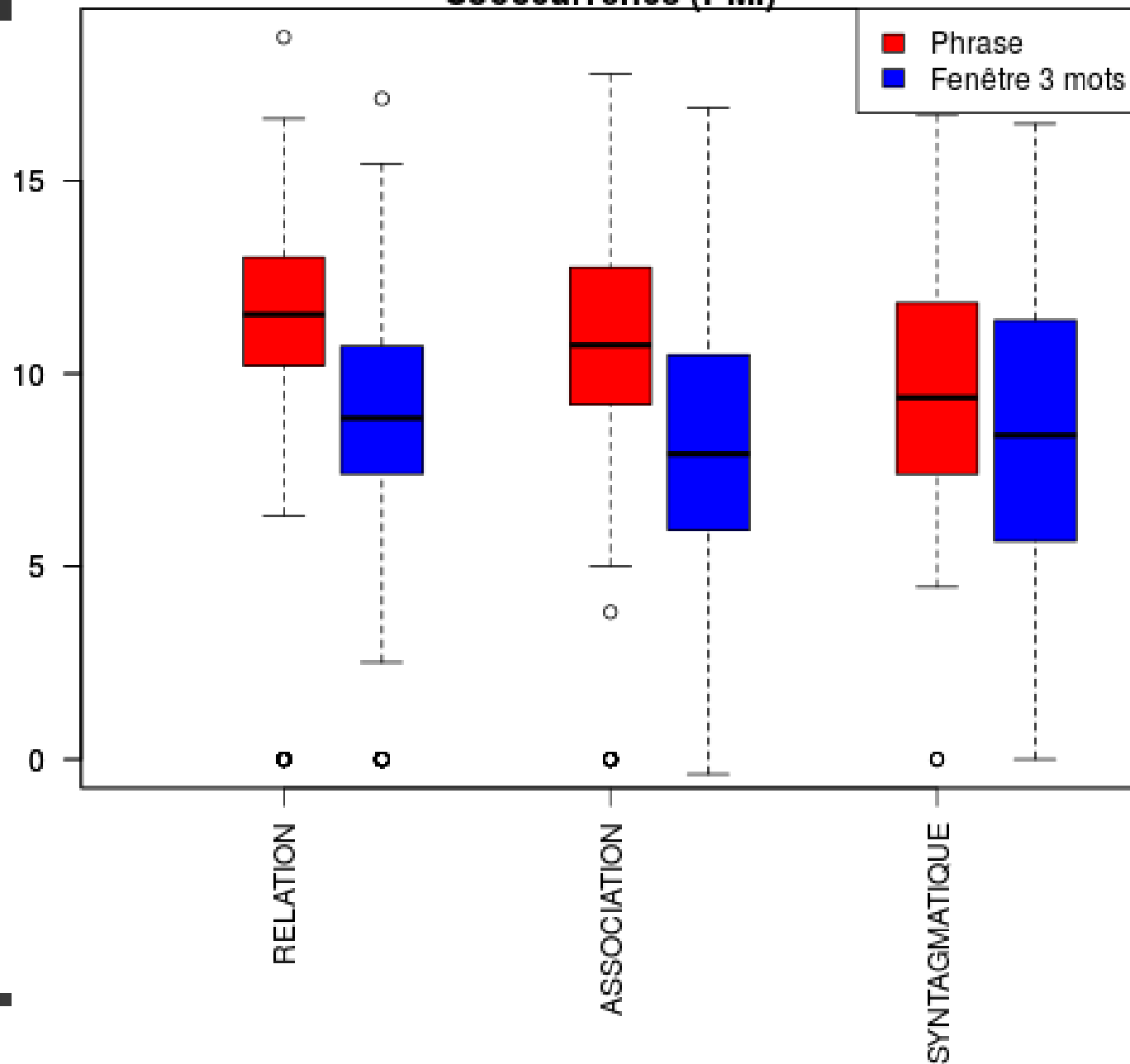
Cooccurrence (PMI)



Cooccurrence (PMI)

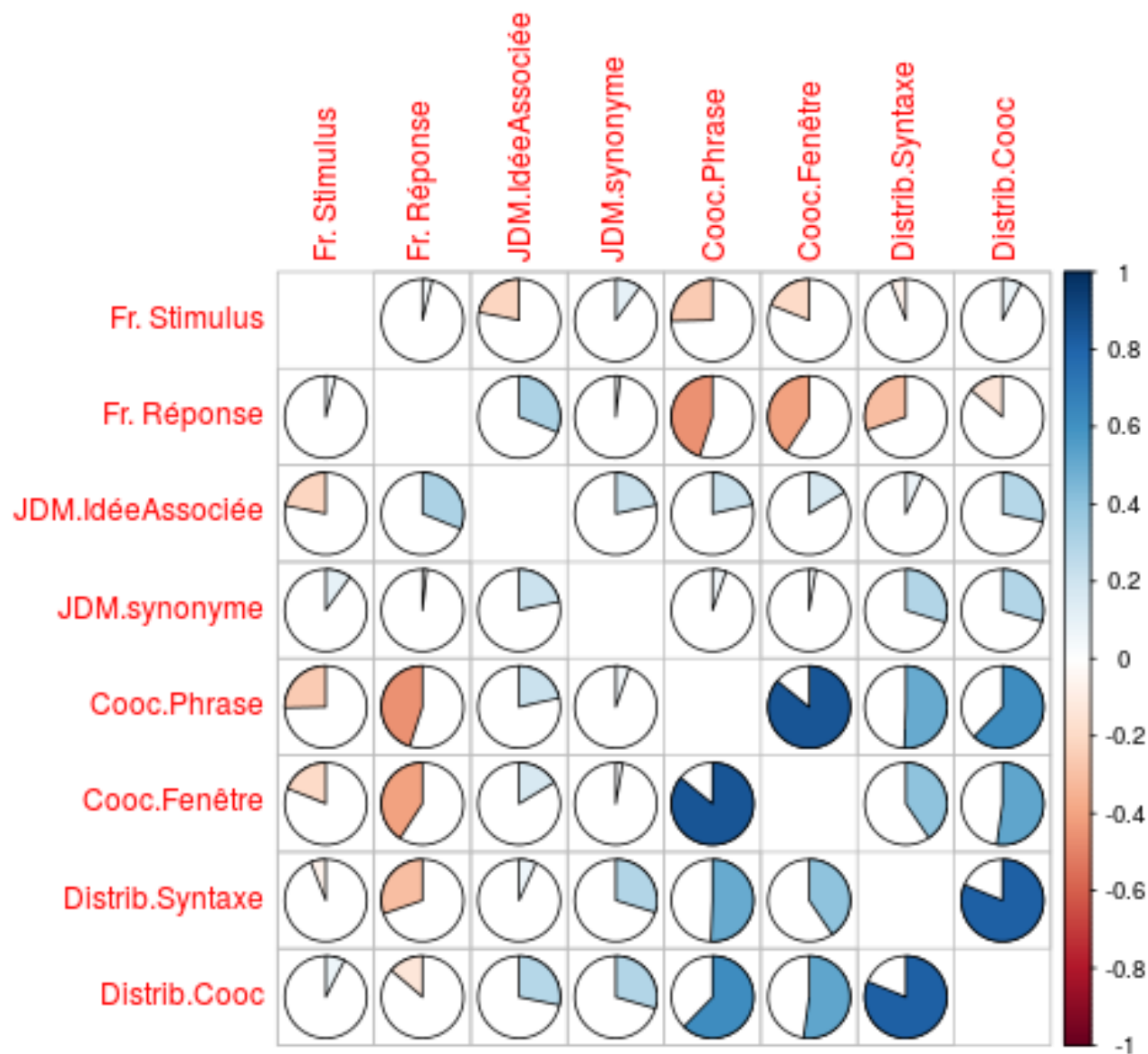


Cooccurrence (PMI)





Corrélations entre les annotations





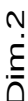
JOUONS AVEC LES EMBEDDINGS...

■ Questions :

- ☐ Que peut-on apprendre sur ces données en observant les embeddings ?
- ☐ Et sur les embeddings en regardant ces données ?
- ☐ En observant les embeddings
 - *des réponses globalement ;*
 - *des réponses par rapport aux stimuli ;*
 - *des réponses d'un individu particulier...*
 - *des réponses à un stimulus donné ;*

■ Méthodes :

- ☐ Word2vec classique, skipgram fenêtre 5, 500 dimensions sur FRWaC taggé et lemmatisé
- ☐ Calcul des vecteurs des S et des R
- ☐ Réduction des dimensions par analyse en composantes principales





Clusters obtenus après ACP (ici N=8)

abricot	ampoule	aile	aliment	Inuit	appareil	aide	Alice
apéritif	armoire	anneau	bactérie	abeille	apéro	aiguillage	ami
artichaut	bague	anse	brûlure	abolement	argent	alliance	amie
boeuf	baguette	arbre	cancer	alligator	arnaque	ampère	amusement
bolognaise	baaignoire	baïonnette	carie	amanite	astuce	aquarelle	artiste
brocoli	balançoire	bec	démangeaison	animal	auto	arcane	ballade
cake	bandoulière	bois	grippe	antilope	bagage	assise	capitaine
caramel	biberon	boule	infection	ara	bicyclette	attraction	cirque
carbonara	bidon	caillou	insecte	arlequin	bidule	aube	clown
carotte	bocal	catapulte	maladie	bestiole	bonbon	aurore	compagnie
champignon	boite	chapeau	microbe	brebis	bouffe	Australie	crêpuscule
chou	bougie	chapiteau	moisissure	bélier	bébé	automne	dame
chèvre	bouton	chêne	morsure	bête	cartable	avion	demoiselle
concombre	cabas	cire	médicament	chat	chose	bateau	enfance
confiture	cage	clapier	peau	chenille	course	Bruxelles	fée
courgette	canapé	clou	piqûre	chien	couverture	campagne	femme
cèpe	cercueil	cocon	plante	chienne	dodo	chaîne	Figaro
dessert	chaise	corde	pollen	cobra	enfant	château	fil
entrecôte	chandelle	cornette	rhume	coq	flipper	chemin	frère
figue	chatière	couleur	rougeole	crinière	friandise	chute	funéraille
fruit	chaussure	creux	vaccin	crocodile	gourmandise	communion	guignol
féculent	cheminée	cristal	varicelle	cygne	hygiène	cour	homme
girolle	couette	dent	venin	esquimau	jeu	danger	honneur
grillade	coussin	doigt	virus	fauve	jouet	densité	jumeau
gâteau	crayon	enclume	éruption	fouine	joujou	dessin	magicien
haricot	dos	faucille		félin	lessive	deux-roues	magicienne
huile	édredon	feu		hirondelle	machin	Dundee	magie
huître	évier	feuille		lapin	magasin	eau	manège
kiwi	fauteuil	filament		libellule	maison	échec	mariage
lait	genou	flamme		lion	mal	électricité	marionnette
légume	guidon	fleur		lionceau	moto	emploi	marionnettiste
menthe	laisse	fourchette		loup	mère	éphémère	matin
miel	lampe	glace		léopard	médecin	essence	Merlin
nouille	lavabo	gorge		lézard	nourriture	faubourg	mort
œuf	lit	crayon		marion	parapluie	fil	pass



Corrélations entre les annotations

■ Etude des vecteurs R-S

□ Réponse – Stimulus

- *Supposé représenter le mouvement sémantique de l'un à l'autre (ajout/suppression d'éléments de sens)*

□ Regroupement des vecteurs par similarité

■ *Classification Hiérarchique Ascendante*

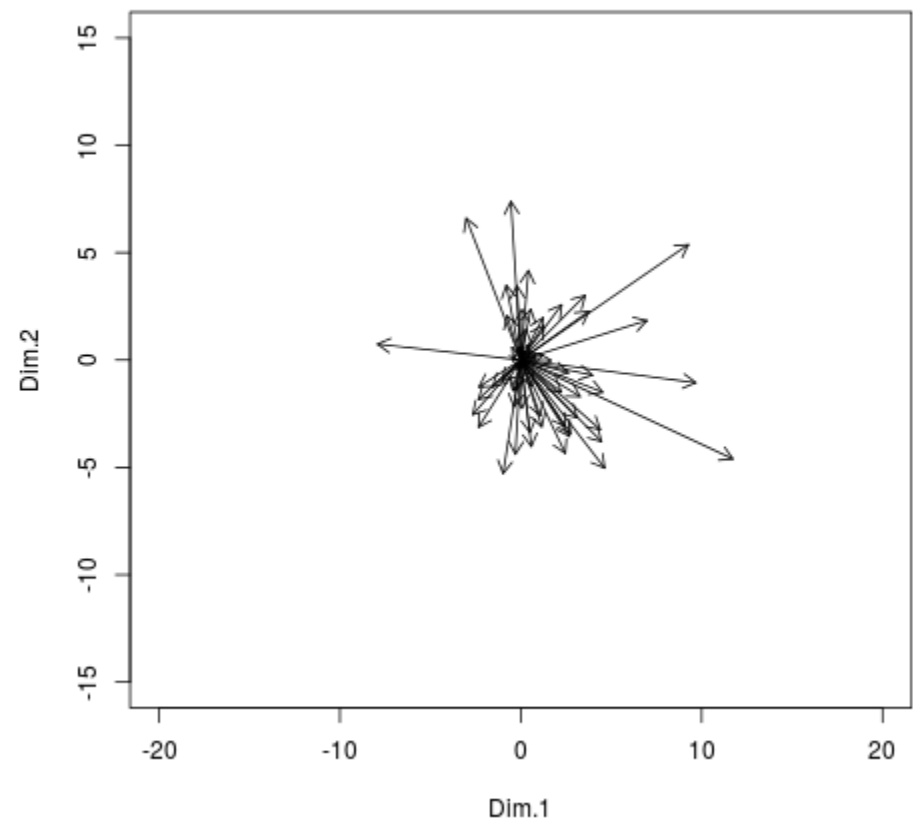
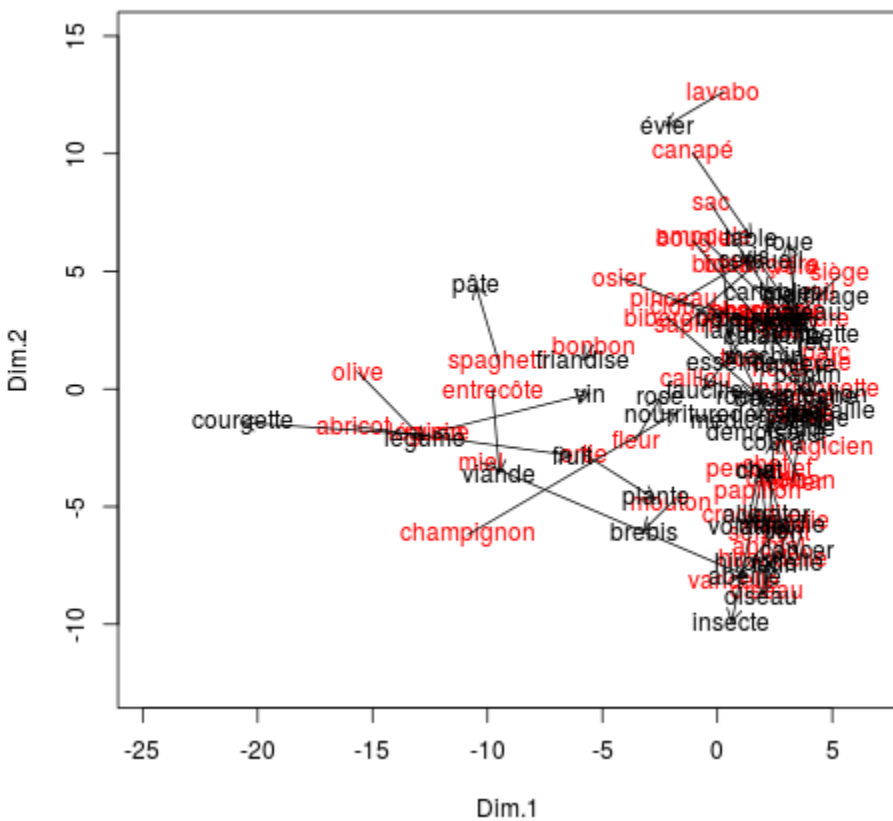
- Sur la base des 500 dimensions des embeddings
- Cosinus, critère de Ward
- *Première constatation : les paires regroupées sont celles qui ont un mot en commun*
 - Essentiellement S puisque c'est lui qui se répète
- *Recherche de cas où deux paires sont regroupées sans avoir de mot commun*

- Un joli cas d'analogie à la Mikolov :
 - *chien-aboiement // lion-rugissement*
 - Mais un seul...
- Des intrus dans un cluster normal :
 - *sac-cabas // sac-chaussure // sac-plastique // sac-sachet // truc-astuce*
- Des cas qui laissent songeur
 - *légume-fruit // mouton-Panurge // olive-huile*



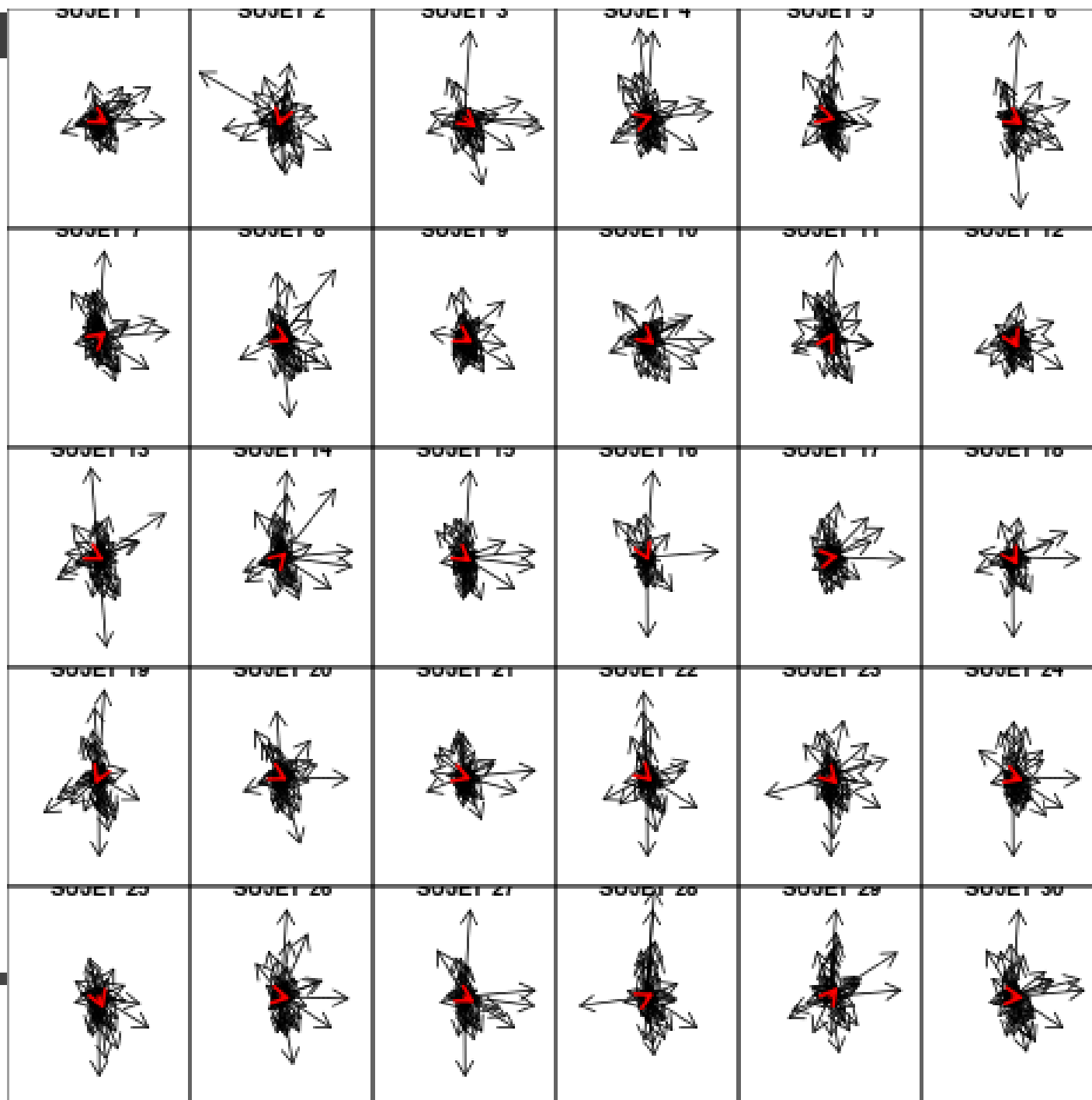
Profil d'un individu

- Projection des réponses d'un individu dans l'espace sémantique
 - Tracé des vecteurs allant de l'embedding du stimulus à celui de la réponse
 - *En restant dans l'espace réduit par l'ACP sur l'ensemble des données*









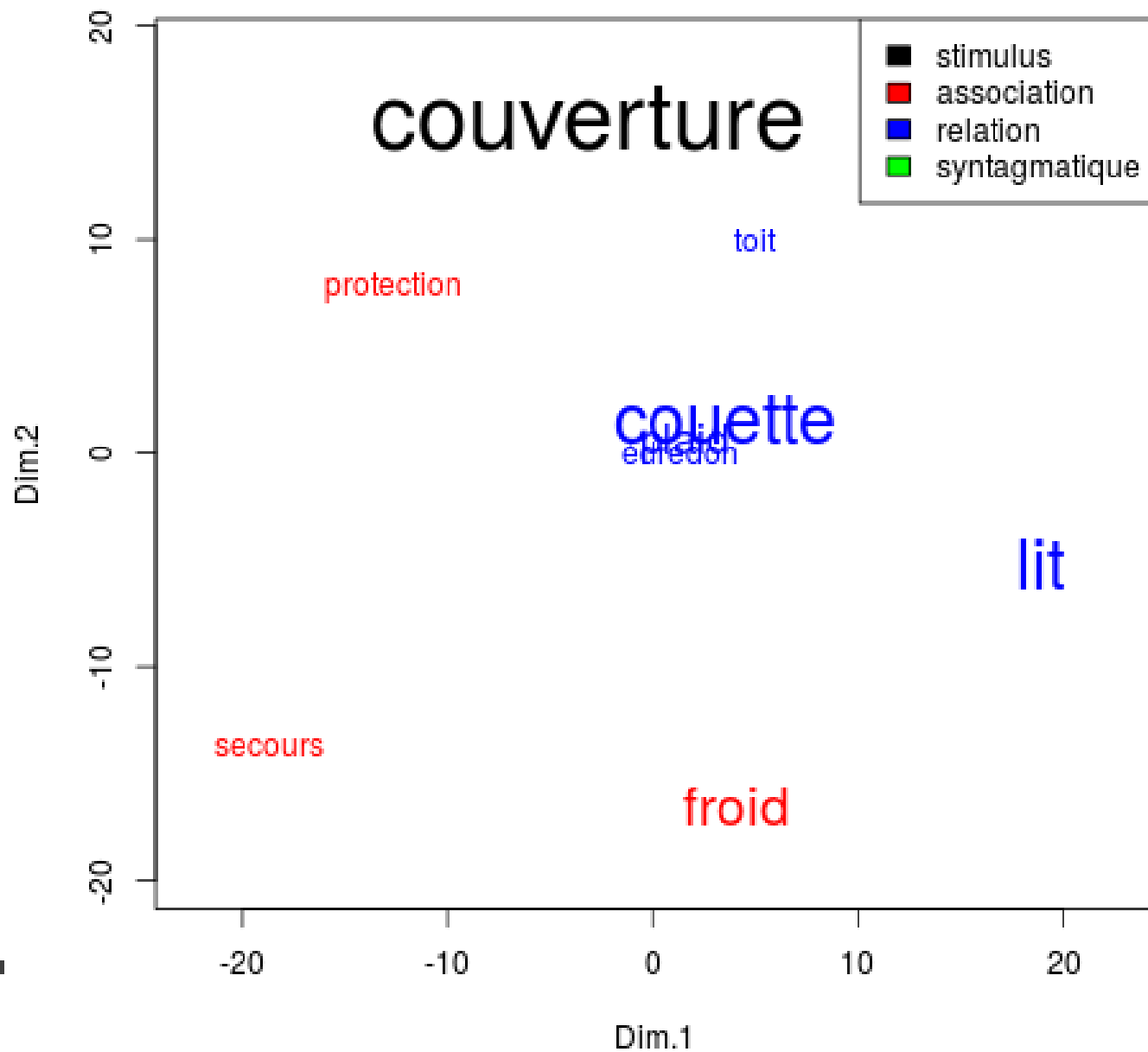


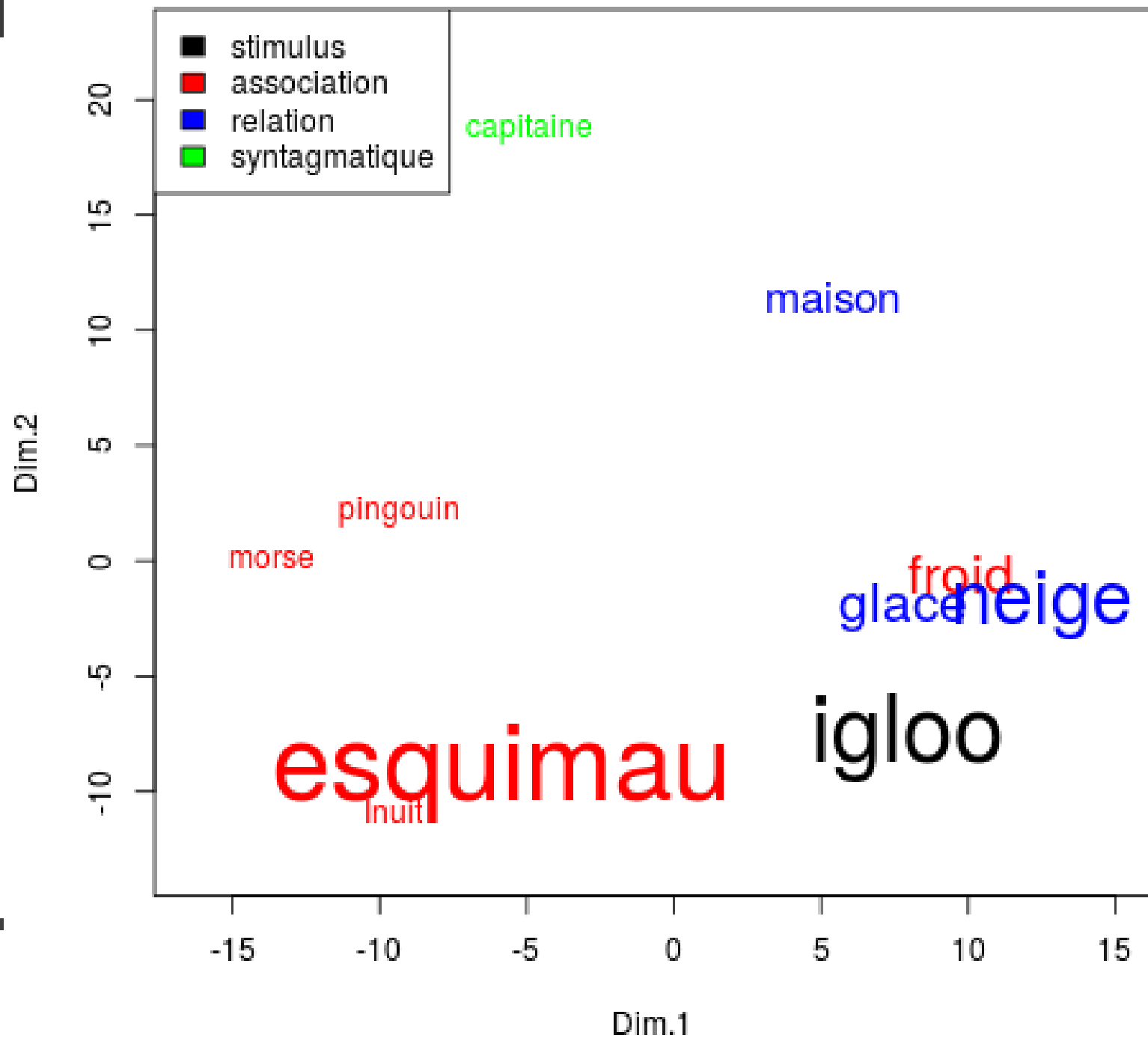
Profil d'un stimulus

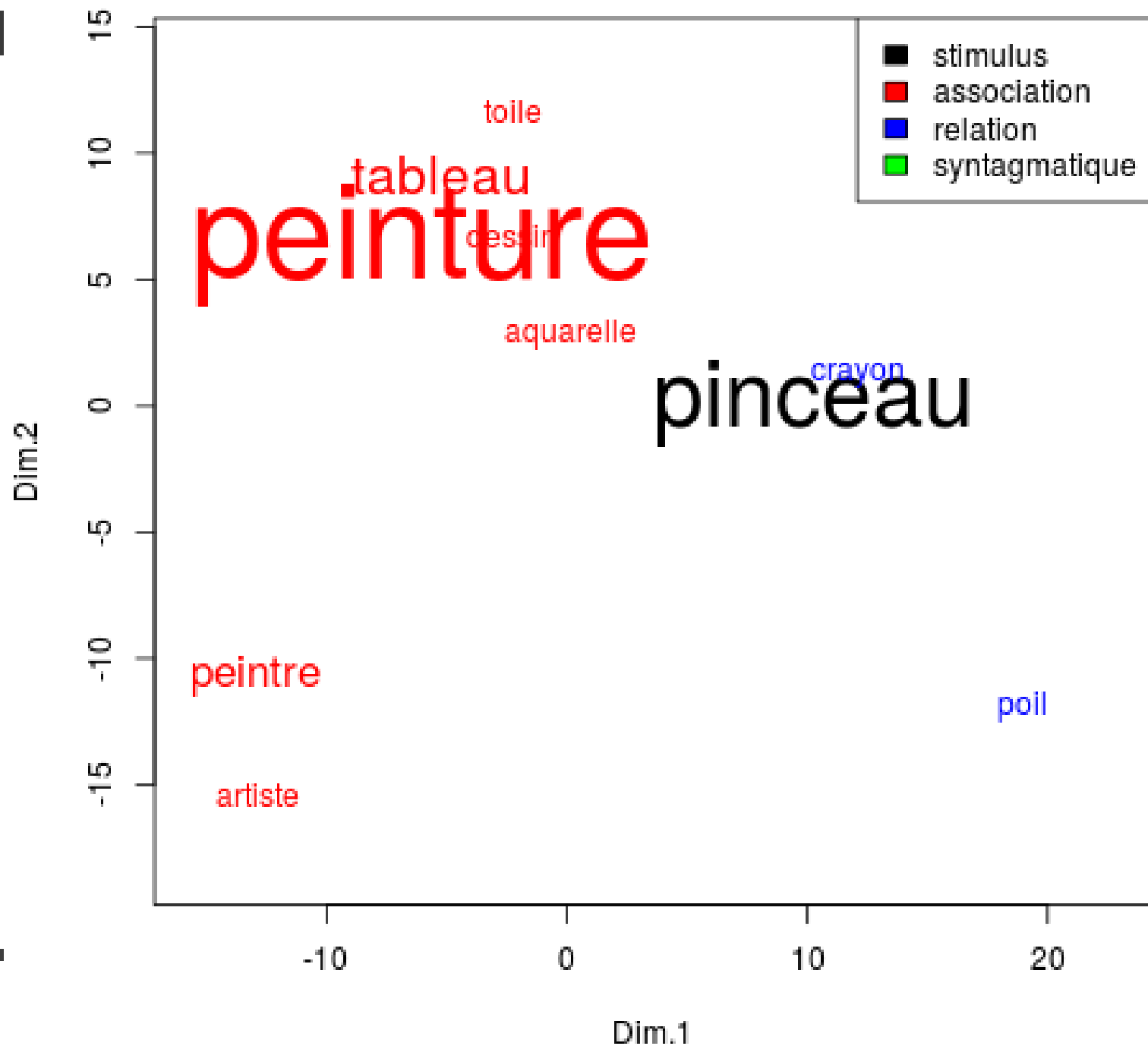


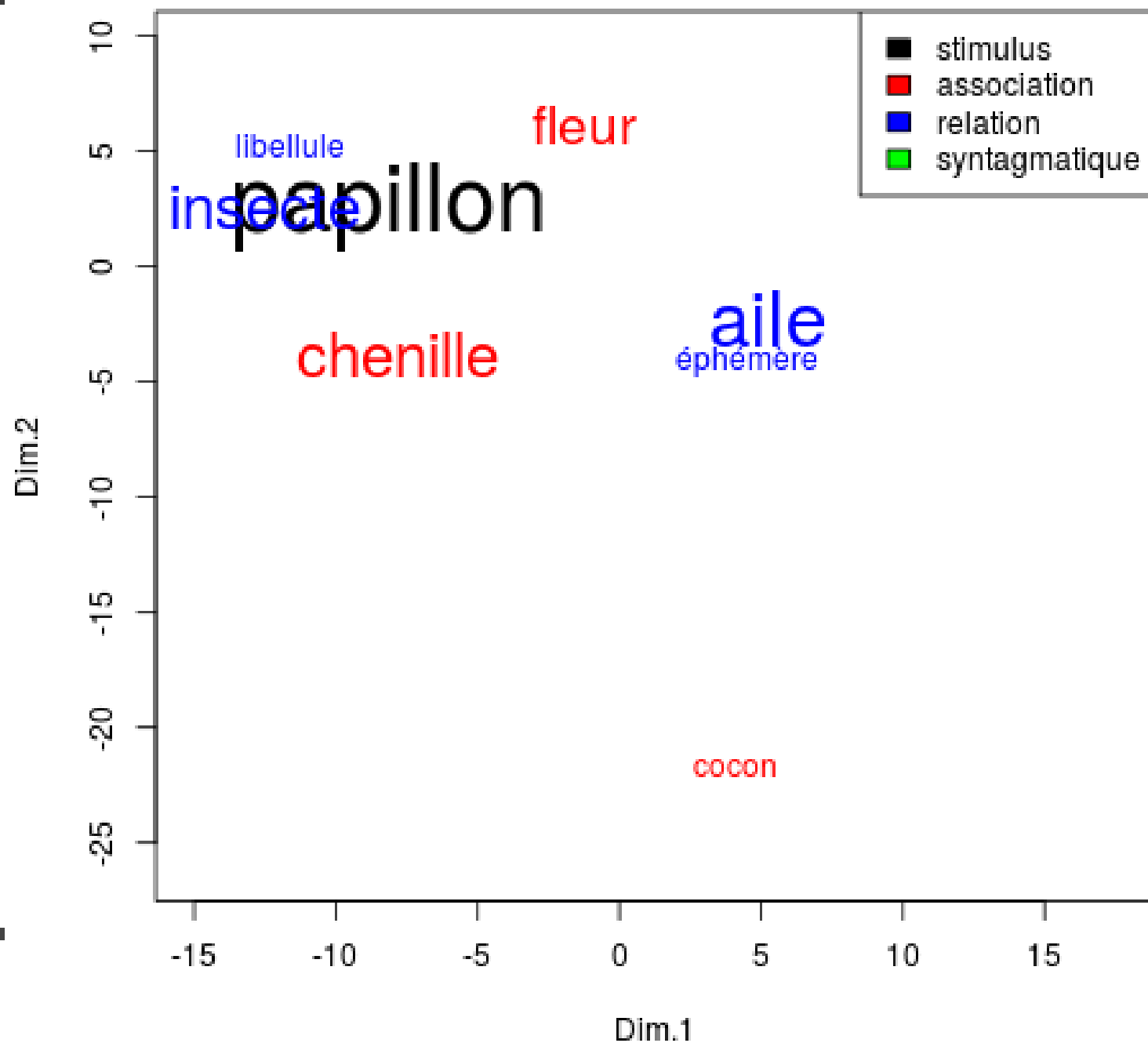
■ Pour un stimulus donné :

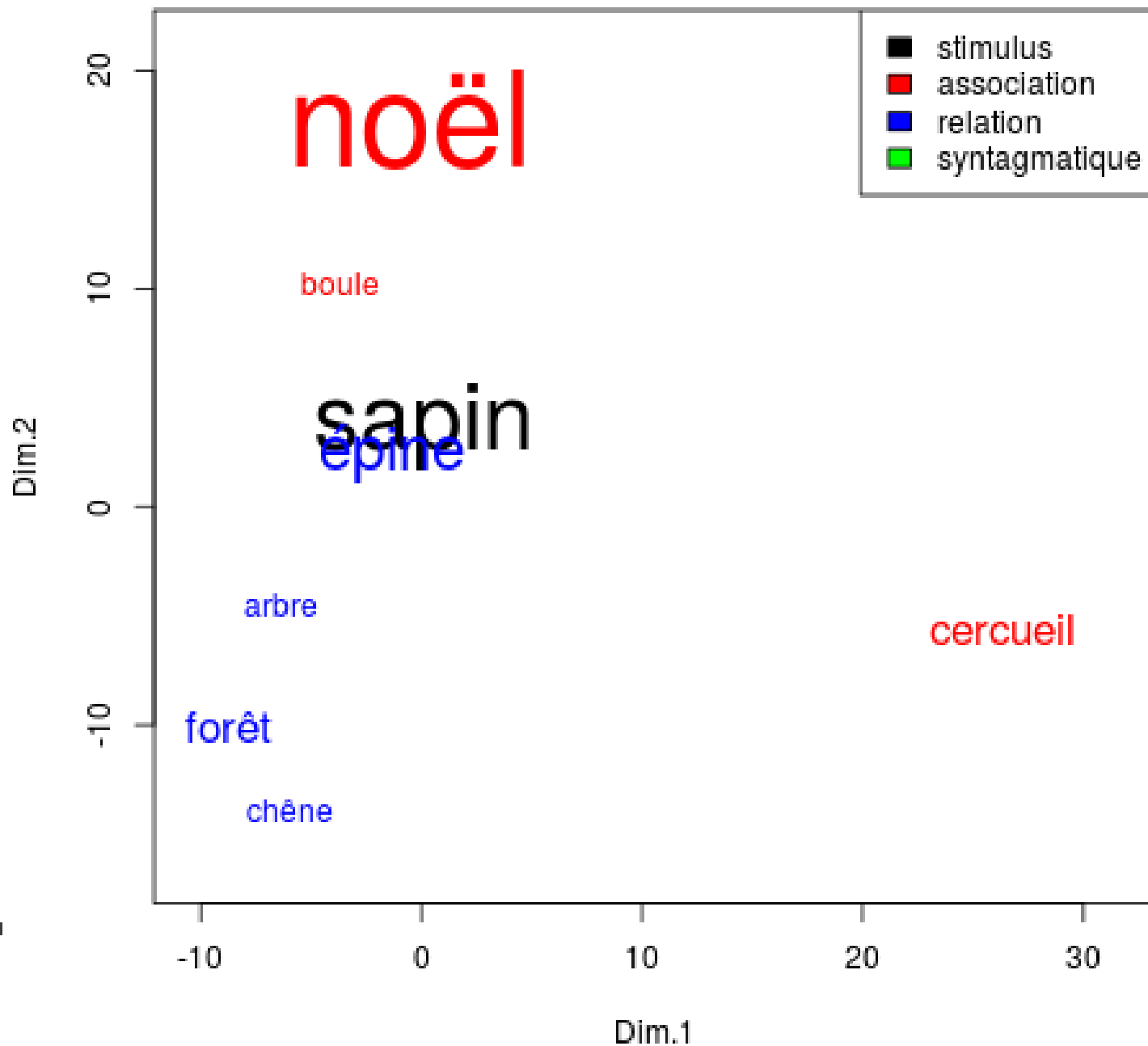
- Regroupement de toutes les réponses proposées
- ACP sur les embeddings du S et des R
- Positionnement des mots en marquant :
 - *La nature (Stimulus = noir)*
 - *La fréquence des réponses (taille du mot)*
 - *La relation annotée entre S et R*













QUELQUES RÉFLEXIONS ET PERSPECTIVES

- Possibilité de prédire la popularité d'une réponse
 - Mais méthodes peu complémentaires au final (expliquent 20% de la variance en les combinant)
- Les relations sémantiques annotées ne décrivent qu'une partie des phénomènes
- Les embeddings :
 - Reproduisent dans leur organisation globale l'espace thématique imposé par les stimuli (nourriture, animaux, objets du quotidien, etc.)
 - Les différences relatives dans l'espace vectoriel entre S et R ne semblent a priori pas pertinentes
 - Par contre, ils permettent localement de bien séparer les réponses à un même stimulus, mais pas forcément en lien avec la relation sémantique (*nymie)
 - *Piste du côté des relations plus fines : fonctions lexicales de Melcuk ou structure de qualia de Pustejovsky ?*

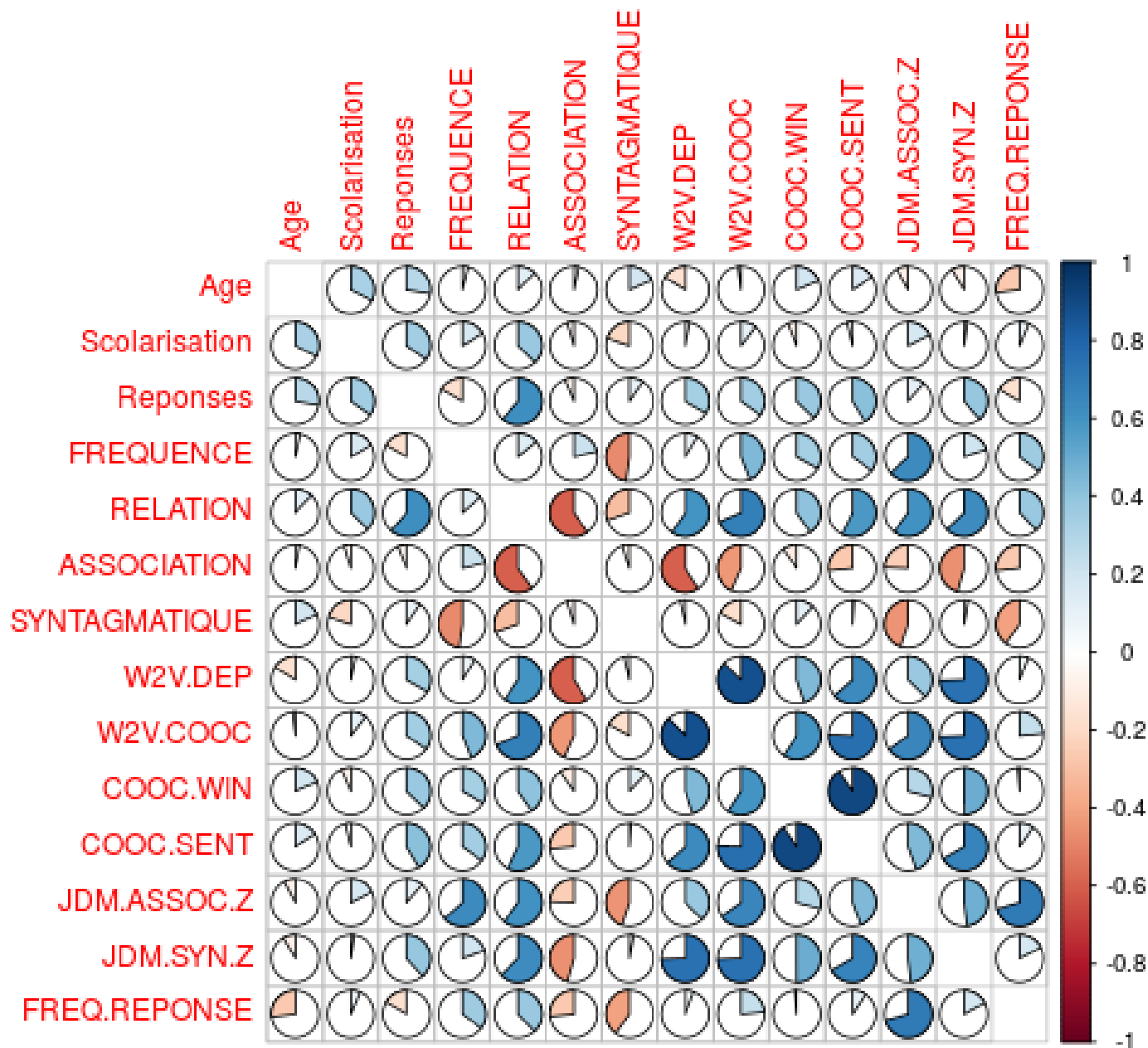


CARACTÉRISTIQUES DES SUJETS ET DES RÉPONSES

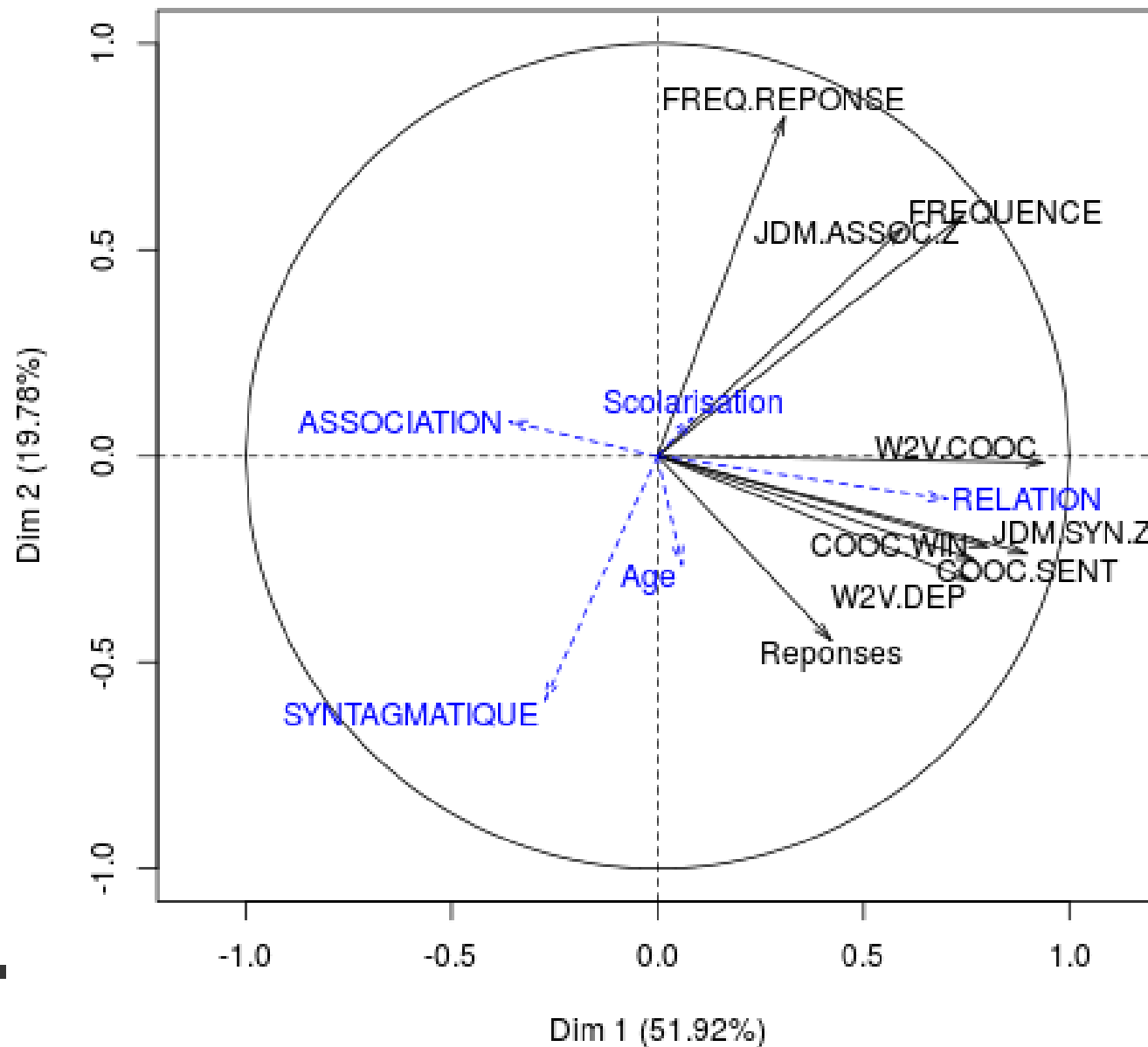


■ Bilan par sujet :

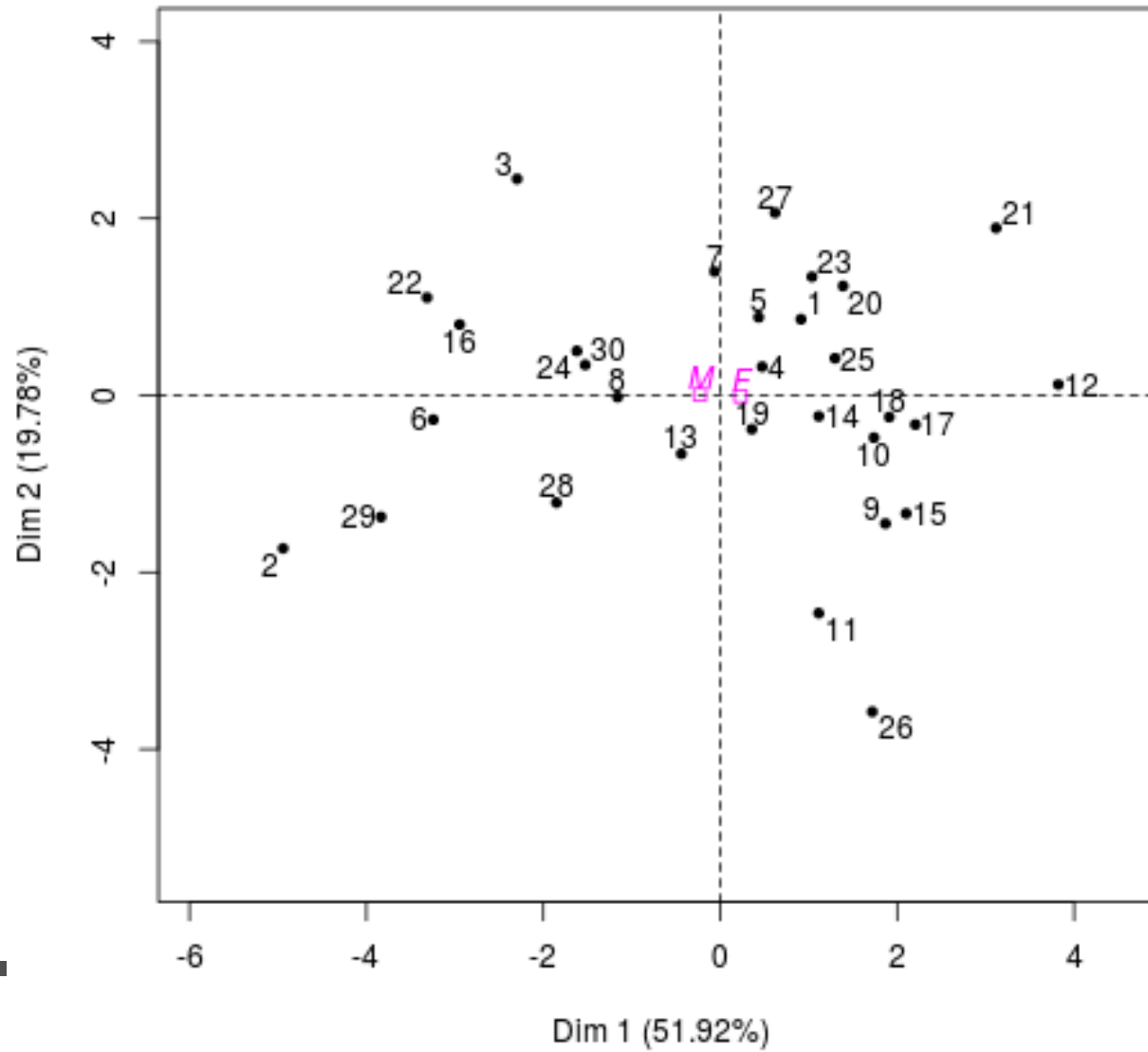
- ☐ Nb de réponses valides (/60)
- ☐ Fréquence moyenne des réponses
- ☐ Fréquence moyenne en corpus des réponses
- ☐ Nb de réponses par type de relation
- ☐ Scores moyens
 - *Proxémie (2 versions)*
 - *Cooccurrence (2 versions)*
 - *Similarité distributionnelle (2 versions)*



Variables factor map (PCA)



Individuals factor map (PCA)



- 
- Calcul des types de réponses pour chaque sujet
 - Schéma à 3 catégories (relation, association, syntagmatique)
 - Résidus de Pearson normalisés

SUJ30
SUJ29
SUJ28
SUJ27
SUJ26
SUJ25
SUJ24
SUJ23
SUJ22
SUJ21
SUJ20
SUJ19
SUJ18
SUJ17
SUJ16
SUJ15
SUJ14
SUJ13
SUJ12
SUJ11
SUJ10
SUJ9
SUJ8
SUJ7
SUJ6
SUJ5
SUJ4
SUJ3
SUJ2
SUJ1

■ SYNTAGMATIQUE
■ ASSOCIATION
■ RELATION

-2 0 2 4