Comportements et sentiments.
De l’ambiguïté dans les émotions ?

Pierre Molette
LERASS PsyCom
Toulouse mai 2014

www.tropes.fr
www.owledge.org
www.lerass.com

Différentes approches pour l'analyse de textes

Plusieurs méthodes utilisant des logiciels :

- Post codification (« Stabilo »). On va annoter manuellement les mots des textes.
- Search « minimaliste ». On cherche une chaine de caractères et on lit les textes (par exemple, Google).
- Lexicométrique. On fait des statistiques poussées, sans chercher à résoudre les polysémies.
- Sémantique, supervisée. Le logiciel propose une classification, qui sera rectifiée par l'utilisateur.
- Sémantique, purement automatique. Il faut « faire confiance » au logiciel.

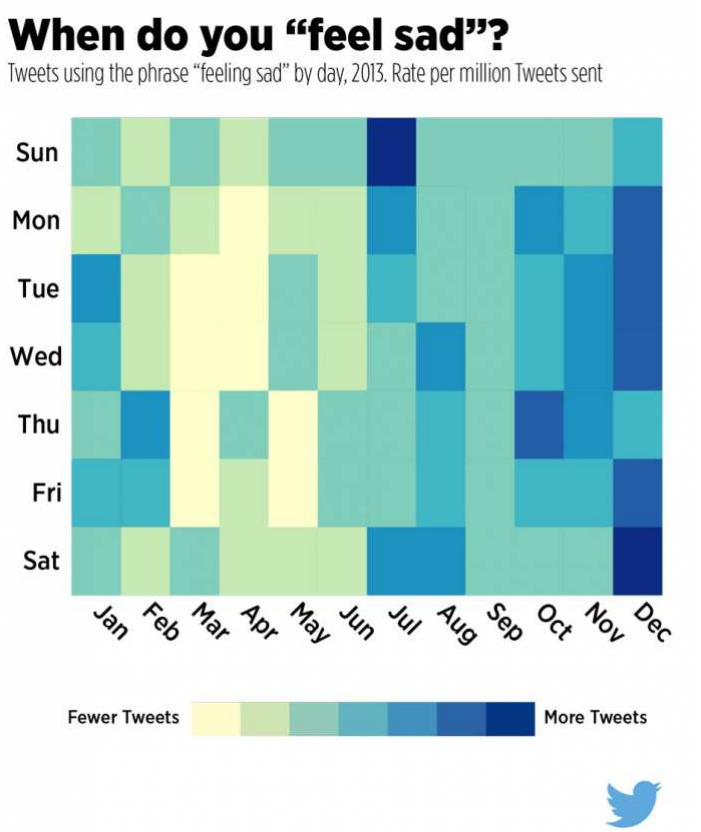
Chacune de ces approches présente des avantages et des inconvénients.

Pour un panorama des méthodes d'analyse de textes, lire Marchand (1998).

Plus la quantité de données est grande, plus il est difficile d'analyser les résultats.

Problème sémantiques sur les gros corpus

C’est pour cette raison que des analyses se limitent souvent à compter le nombre de réponses pour une simple expression. Par exemple, cette étude Twitter se limite à une recherche (« feeling sad ») :



Cela fonctionne sur un corpus énorme. Mais on ne peut pas considérer que c’est suffisant.

Les logiciels de classification sémantiques

Classer, c'est prendre le risque de se tromper. Mais c'est nécessaire pour comprendre.

A contrario, les moteurs de recherche internet affichent - sans les classer - d'interminables listes de résultats correspondant à ce que « tout le monde devrait chercher » (avec une forte orientation marketing). Cette approche est généralisée. Bien que scientifiquement peu satisfaisante quand on veut analyser objectivement.

Une classification sémantique automatique n'est réellement exploitable que si le taux d'erreur reste acceptable pour l'utilisateur. C'est possible sur de gros volumes de textes.

Comme dans tous les logiciels traitant le langage naturel, de nombreux termes polysémiques perturbent les résultats d'analyse. Comment améliorer les résultats ?

Partant du principe qu'il est préférable d'utiliser des termes spécifiques pour réduire le champ sémantique de certains mots, il convient de retirer (ou de spécialiser) certains concepts pour améliorer la qualité de l'analyse. En empruntant aux tests médicaux, disons qu'il faut étudier le rapport bénéfice/risque des mots ambigus pour isoler ceux qui ont trop d'effets indésirables (c'est-à-dire qui sont pratiquement toujours utilisés à contresens).

Les scénarios « Concepts » de Tropes ont été conçus pour réduire fortement les ambiguïtés. Ils s'appuient sur un réseau sémantique et une logique d'intelligence artificielle, qui évoluent depuis une vingtaine d'années.

Les logiciels Tropes et OwlEdge

Tropes (1994-2014) :

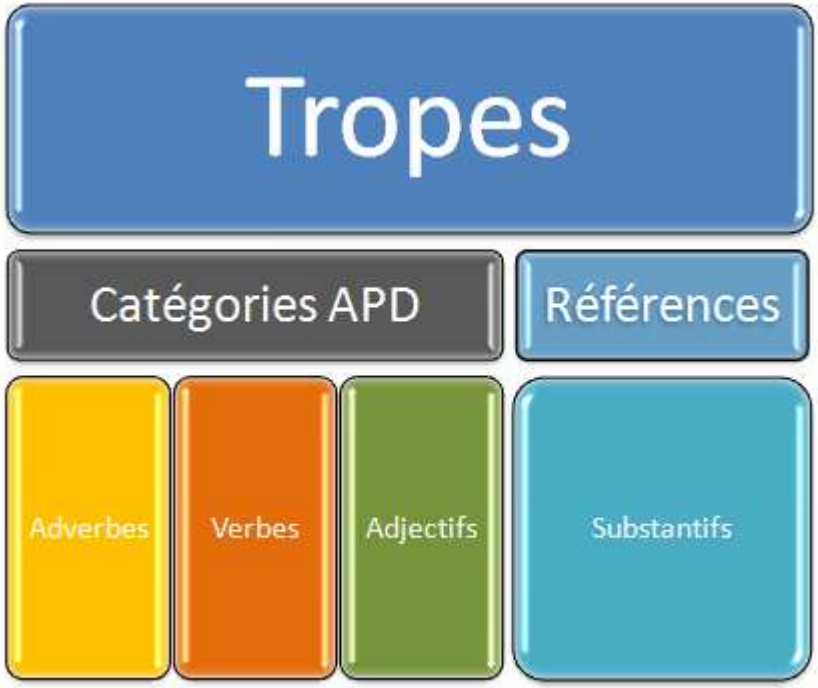
- Logiciel d'analyse sémantique de textes
- Développé par Pierre Molette et Agnès Landré, sur la base des travaux de Rodolphe Ghiglione
- Fondé sur l'Analyse propositionnelle du discours et l'Analyse cognitivo discursive (entre autres)
- www.tropes.fr

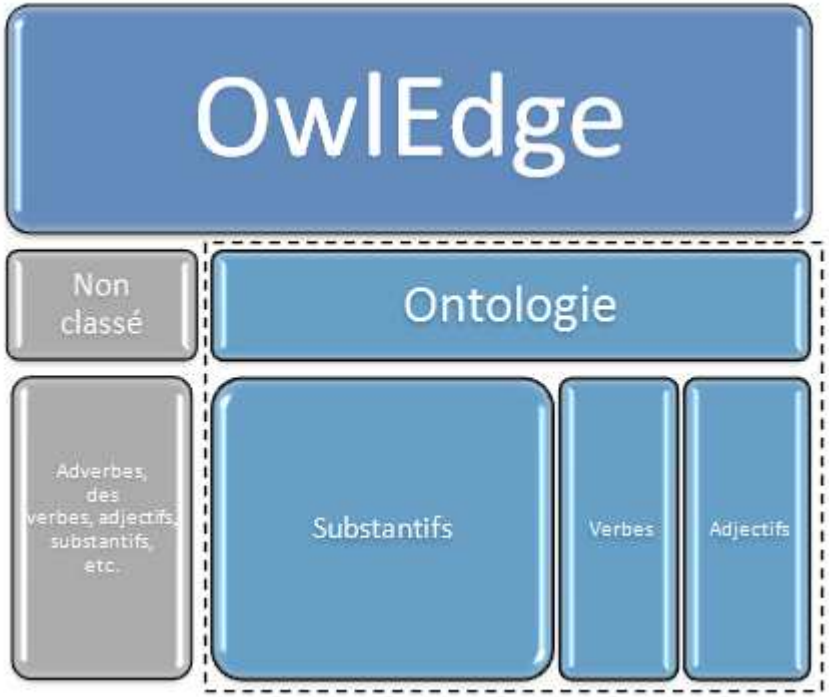
OwlEdge (2011-2014) :

- Moteur de recherche et d'analyse par concepts
- Développé par Pierre Molette
- Exploite des ontologies sémantiques, dont certaines sont compatibles avec Tropes
- www.owledge.org

Pour une explication du fonctionnement de Tropes, lire Ghiglione et al. (1998) ou Molette (2009).

TROPES VS OWLEDGE





Tropes, qui se fonde sur la théorie de l'APD, catégorise les mots à partir de la grammaire.

OwlEdge exploite une classification unifiée sans cloisonnement grammatical.

Ontologies et Scénarios

Les *scénarios* de Tropes sont des classifications sémantiques personnalisables (ontologies).

Il existe plusieurs scénarios de référence. Par exemple :

- Concepts Fr V8 (recouvrant 100.000 termes français) - Molette, Landré (1998-2014)
- Concepts US V8 & Natural Sciences US V8 (environ 100.000 termes anglais) - Molette et al. (200 ?)
- EMOTAIX - Lexique des émotions (4.600 termes) - Piolat, Bannour (2009)
- MeSH (Medical Subject Heading), thésaurus médical américain (NLM, MEDLINE), obsolète (2002).
- AGROVOC - Vocabulaire de l'agronomie multilingue (FAO, INRA, Cirad), 2006.
- etc.

Certains scénarios sont livrés avec le logiciel Tropes (*Concepts V8*…), d'autres sont disponibles sur demande auprès de leurs auteurs (par exemple EMOTAIX).

➔ Tous les utilisateurs du logiciel Tropes peuvent (et doivent) définir leurs propres classifications.

Comparaison de deux corpus contenant des sentiments

Etude de deux corpus, a priori peu comparables :

	Date de recueil	Nombre de documents	Format	Occurrences de mots	Taille totale du texte
Assemblée Nationale (Fr)	Janvier 2014	61 159	HTML	389 millions	5 Go
Classiques littéraires (ABU)	XXème siècle	100	Texte	5 millions	32 Mo

Le premier est une copie du site web de l'Assemblée nationale, collecté début janvier 2014. Ce corpus comprend des rapports, les comptes-rendus des débats et de multiples informations parlementaires. Ces textes ont été essentiellement rédigés durant les trente dernières années.

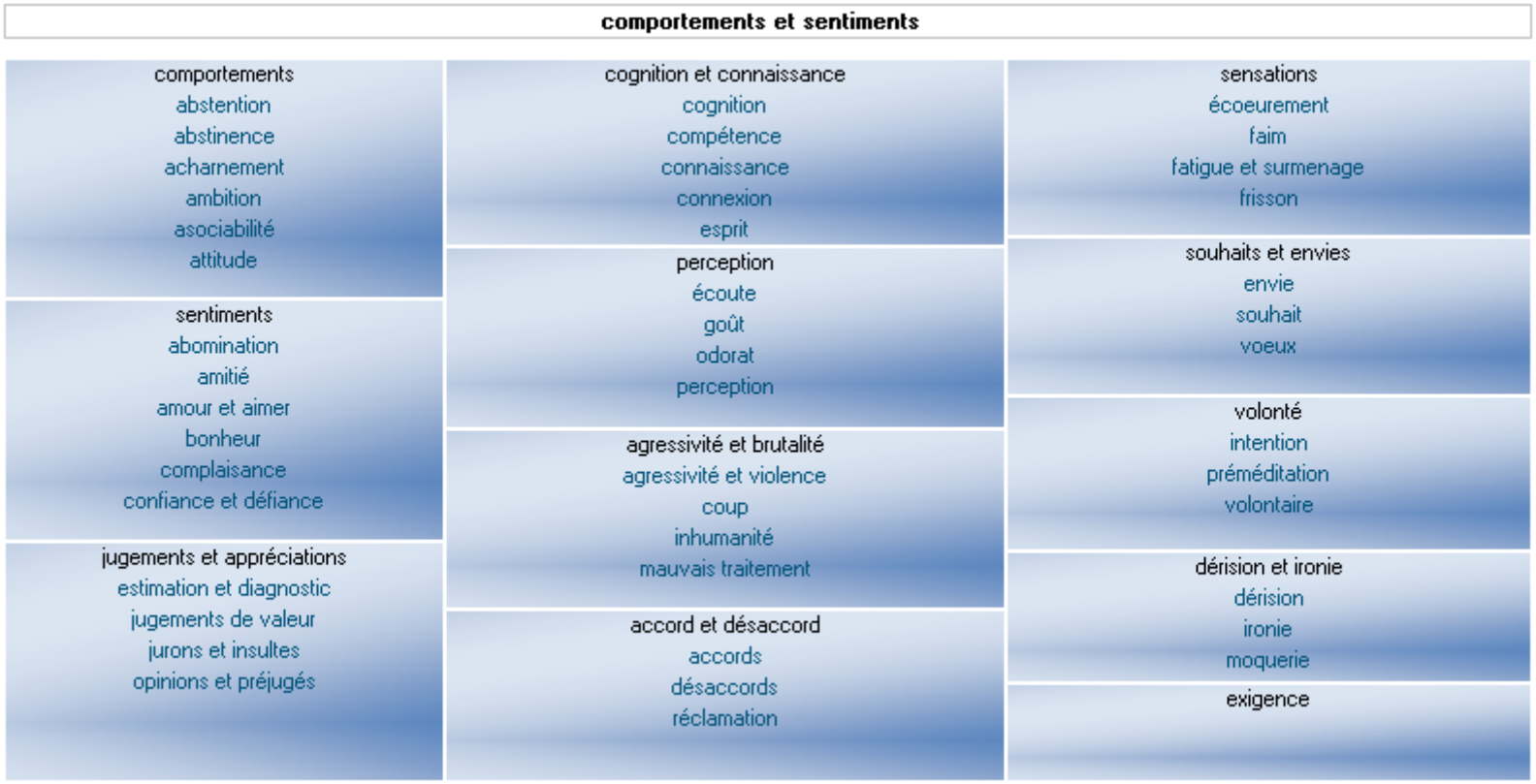
L'autre corpus contient cent classiques littéraires, qui ont été recueillis sur le site de l'ABU (Association de Bibliophiles Universels), http://cedric.cnam.fr/ABU, 1999

➔ Ces deux corpus ont été analysés avec l'ontologie fournie par défaut avec Tropes (le scénario Concepts Fr V8), sans personnalisation des classifications. On y a toutefois ajouté une liste de 5000 noms propres. Le traitement est donc purement automatique.

➔ Une personnalisation des classifications permettrait d'augmenter la précision des résultats.

Contenu de la branche « Comportement et sentiments »

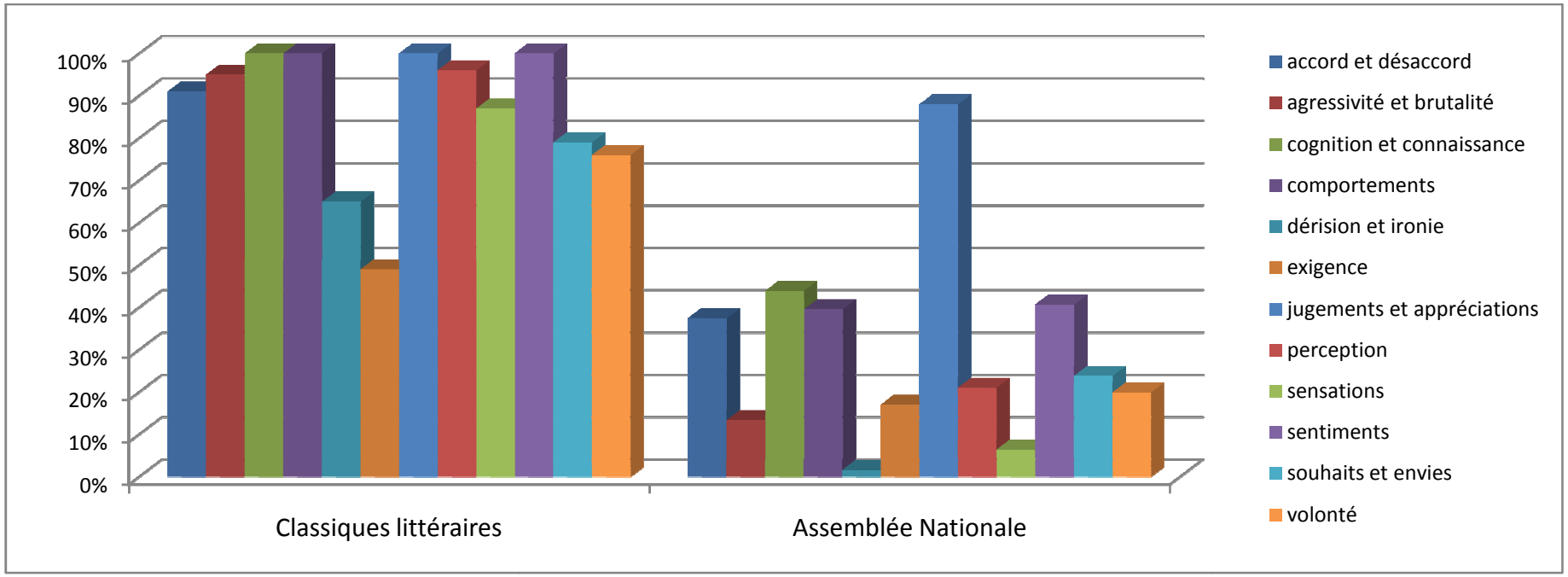
Treemap d'OwlEdge – Scénario Concepts Fr V8 de Tropes



Nombre de documents. Données affichées pour les Classiques littéraires.

Comparaison de corpus asymétriques (suite)

Une approche simple consiste à comparer le nombre de documents classés, en pondérant les valeurs. Par exemple voici les résultats d'OwlEdge (sur « comportement et sentiments ») :

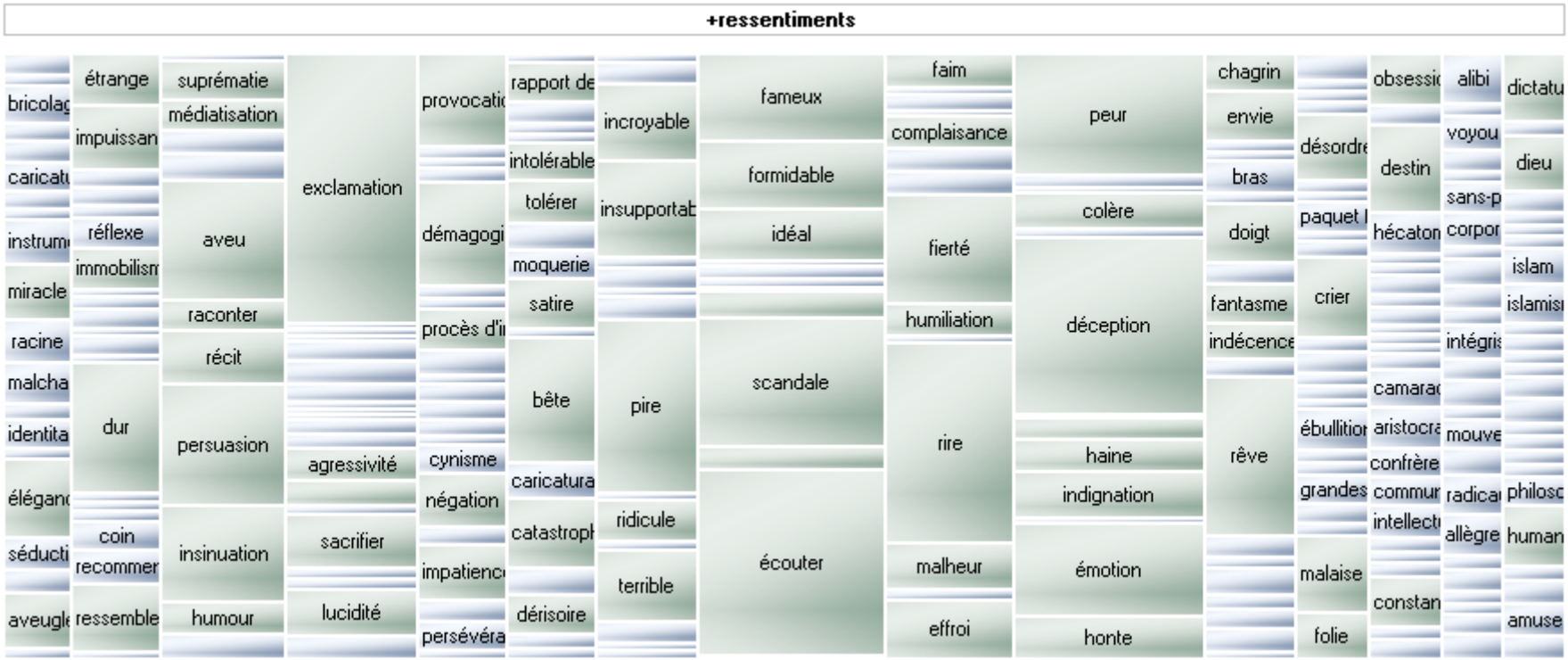


Pour les classiques littéraires, pratiquement toutes les branches de cette classification sont activées.

Ce n'est pas le cas pour l'Assemblée nationale, excepté pour les « jugements et appréciations », présents dans 88% des documents. Il ne faut pas en déduire qu'il y a peu d'émotions à l'Assemblée nationale…

Clustering d'OwlEdge sur les ressentiments, documents

Résultat d'un clustering sur le concept de « ressentiment », corpus Assemblée nationale :

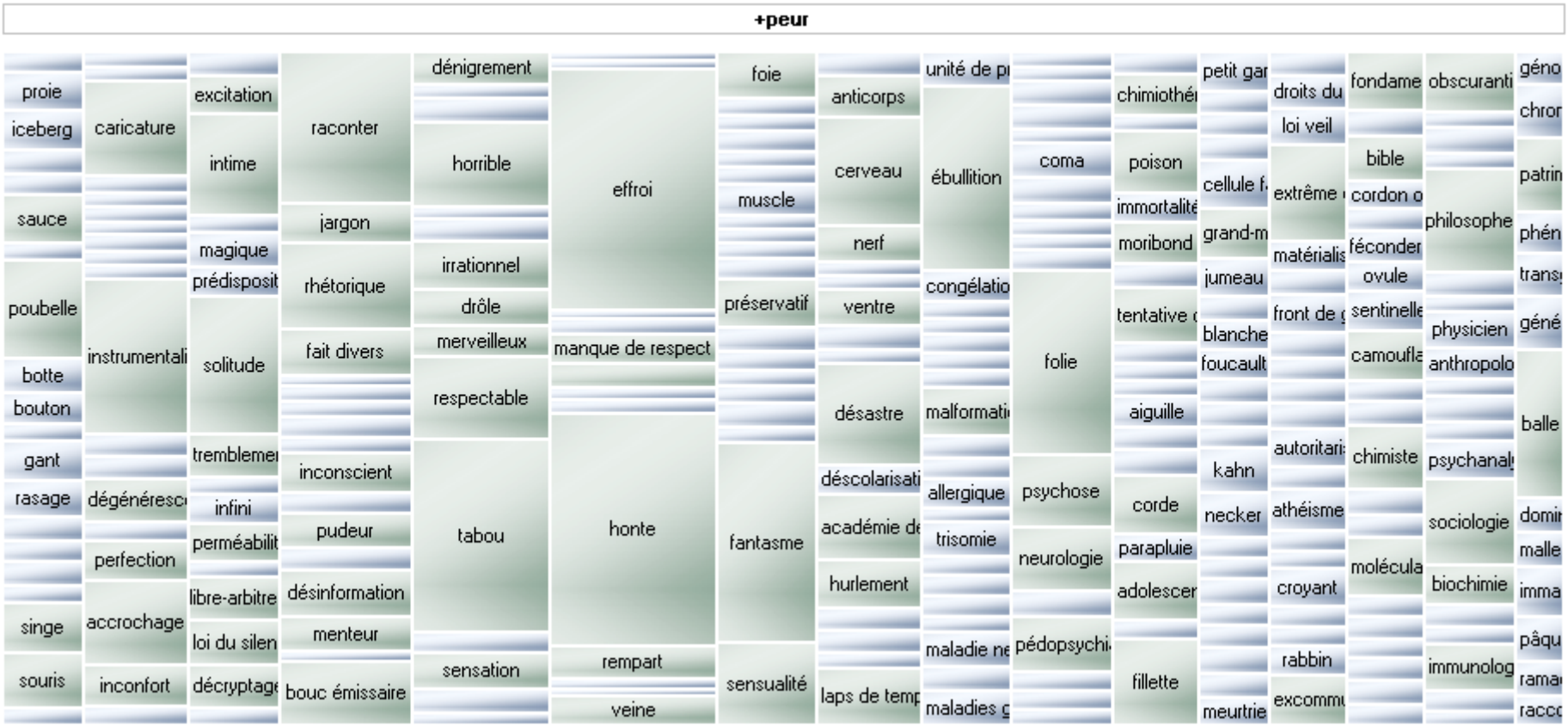


La recherche intègre les sens associés au concept de ressentiment (déception, colère, etc.)

Les résultats (proportionnels au nombre de documents trouvés), résultent du clustering d'OwlEdge (concepts significatifs pour une requête avec expansion sémantique). On voit très peu de noms propres ici...

Clustering d'OwlEdge sur la peur, documents

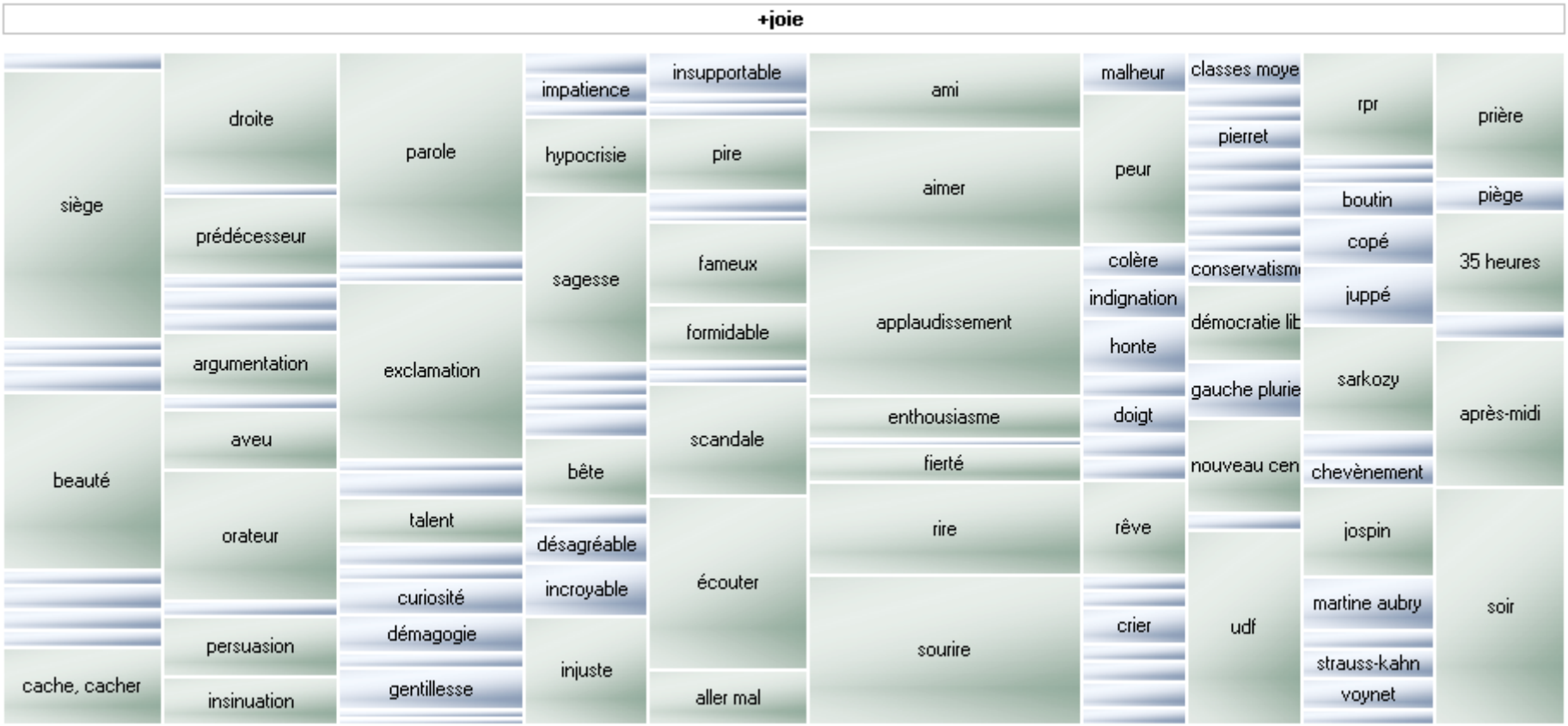
Résultat d'un clustering sur le concept de « peur », corpus Assemblée nationale :



La recherche intègre les sens associés à la peur (incluant l'angoisse et la crainte). On parle de troubles mentaux et de folie. Mais aussi d'autre chose... Est-ce que la science fait peur ?

Clustering d'OwlEdge sur la joie, documents

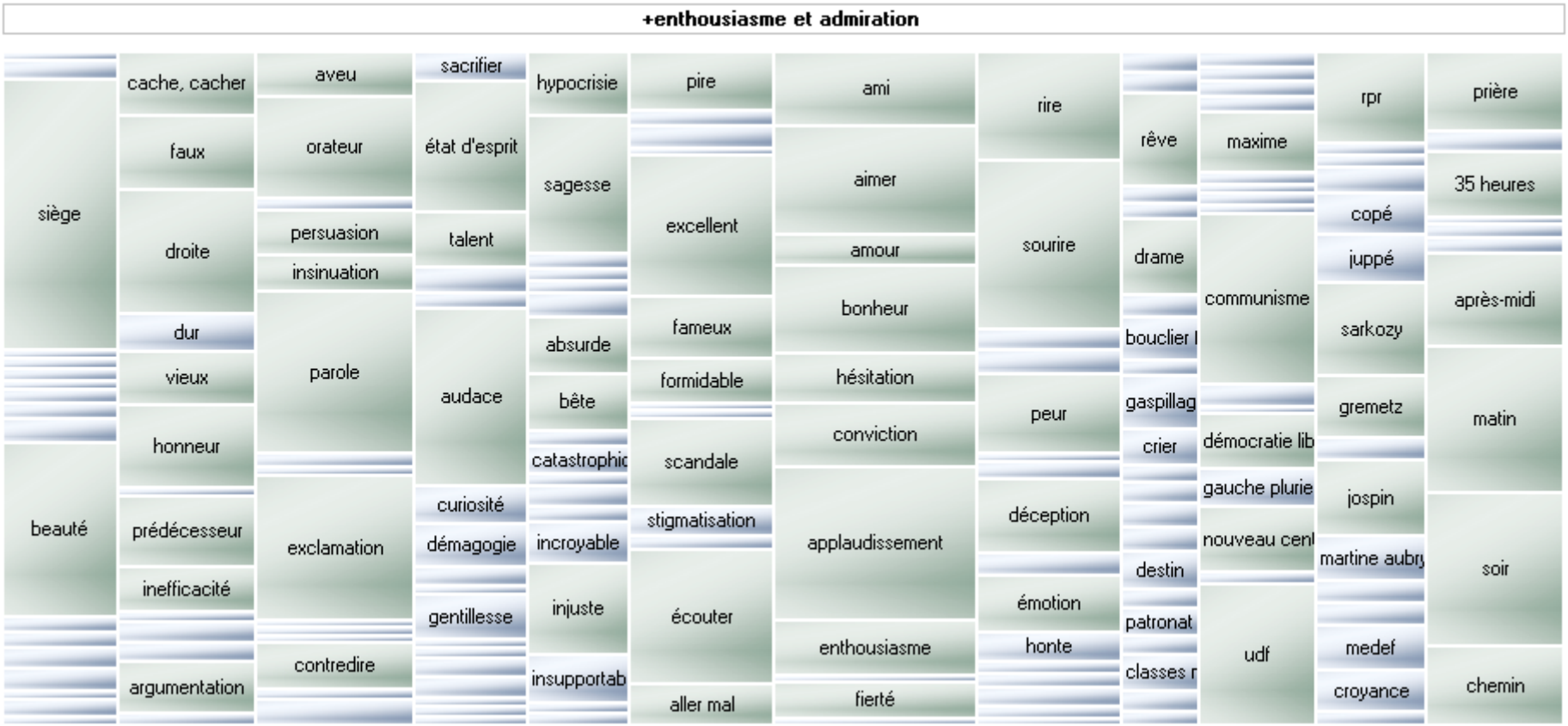
Résultat d'un clustering sur le concept de « joie », corpus Assemblée nationale :



La recherche intègre les sens associés à la joie (incluant rire et sourire). On voit apparaître des partis politiques et des noms propres, sur la partie droite du graphe.

Clustering d'OwlEdge sur l'enthousiasme, documents

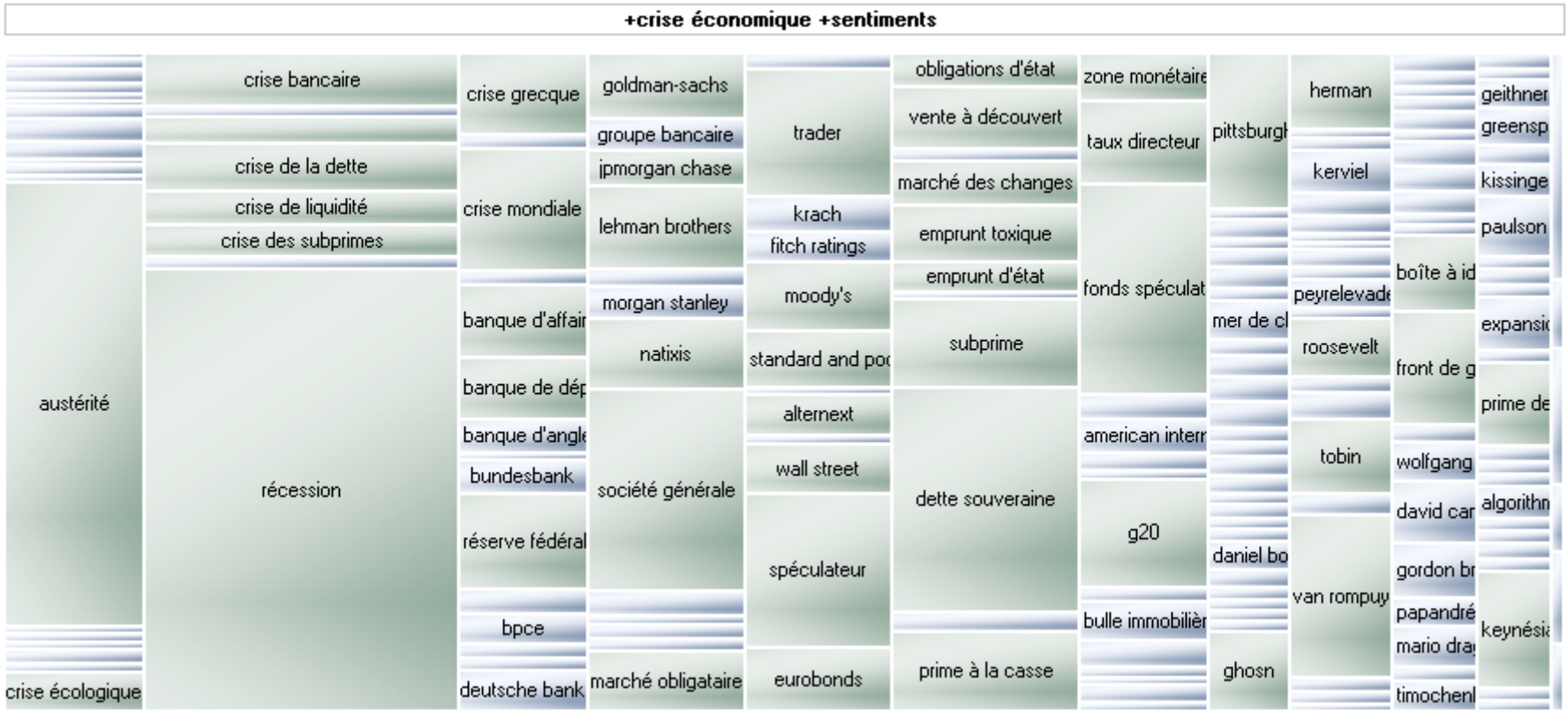
Résultat d'un clustering sur le concept « d'enthousiasme et admiration », corpus Assemblée nationale :



La recherche intègre les sens associés (incluant les applaudissements). Comme dans le cas précédent, on voit apparaître des partis politiques et des noms propres, sur la partie droite du graphe.

Clustering d'OwlEdge sur la crise économique, documents

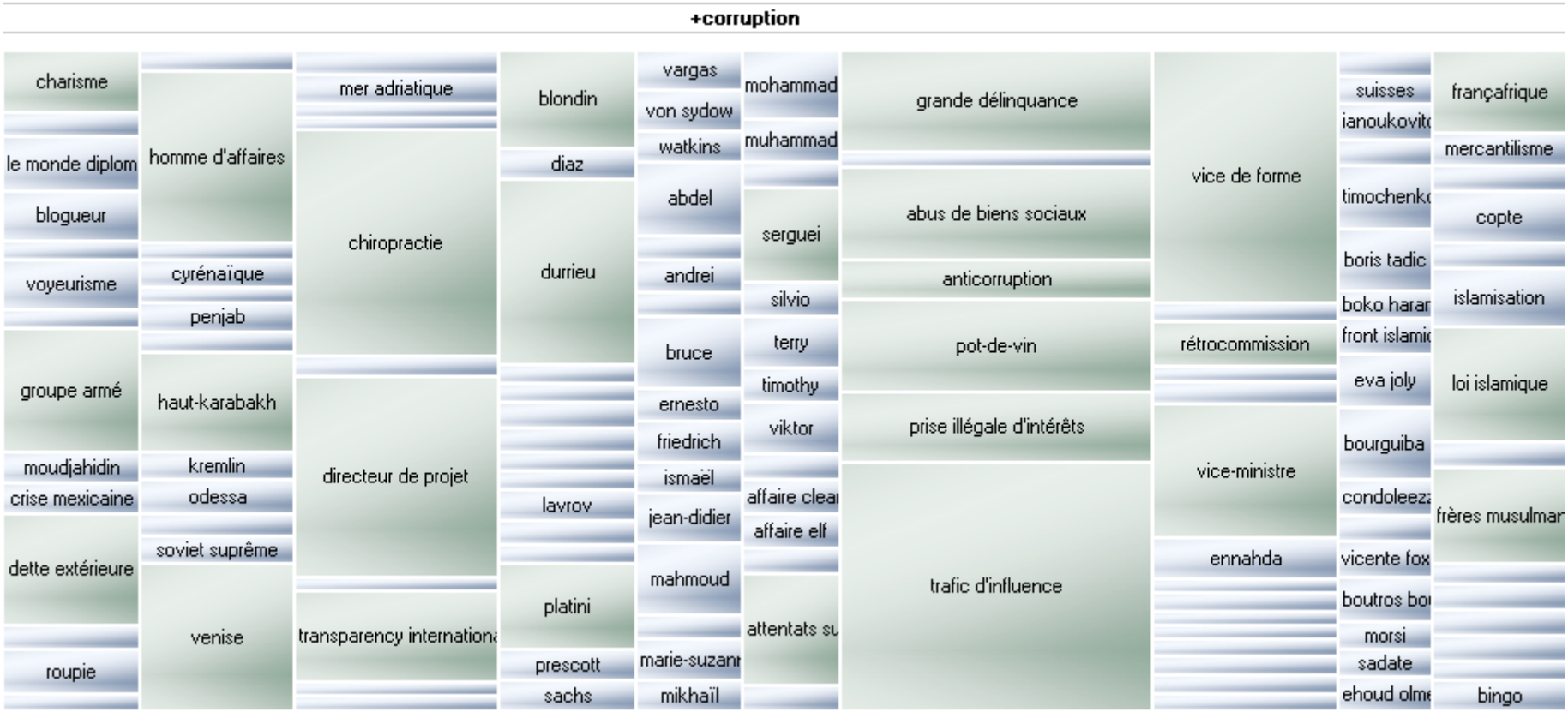
Résultat d'un clustering sur « crise économique » et « sentiments », corpus Assemblée nationale :



Les acteurs économiques et de nombreuses personnalités politiques étrangères sont présents sur ce graphe. Les comportements et sentiments de Tropes ne sont pas significatifs. Ils sont implicites ! Notez que les politiques français ne sont pas sur ce graphe, excepté le Front de gauche.

Clustering d'OwlEdge sur la corruption, documents

Résultat d'un clustering sur le concept de « corruption », corpus Assemblée nationale :



Certaines confessions religieuses, des corps de métiers et personnalités étrangères sont bien représentés sur ce graphe. Les personnalités françaises sont atypiques (ce ne sont pas celles qu'on trouve en analysant un corpus de presse avec la même requête). Que faut-il en déduire ?

BIBLIOGRAPHIE

Rodolphe GHIGLIONE et Al. *L'analyse automatique des contenus.* Paris, Dunod, 1998.

John LYONS. *Semantics.* Cambridge University Press. 1977.

Pascal MARCHAND. *L'Analyse du Discours Assistée par Ordinateur.* Paris, Armand Colin, 1998.

Pierre MOLETTE. *De l'APD à Tropes : comment un outil d'analyse de contenu peut évoluer en logiciel de classification sémantique généraliste.* Colloque Psychologie Sociale et Communication. Tarbes, 2009.

Pierre MOLETTE. *OwlEdge. Un moteur d'analyse de corpus par concepts.* Communication au séminaire du Labex SMS. Toulouse, mars 2014.

Annie PIOLAT, Rachid BANNOUR. *EMOTAIX : un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif.* L'Année Psychologique, 109, 657-700.

TRANSPARENCY INTERNATIONAL. *Transparence de la vie publique et maintenant ?* Rapport 2013.

ANNEXES

Problèmes d'identification sémantique

L'Assemblée nationale regorge d'expressions difficiles à décoder, par exemple :

La Commission est saisie de l'amendement CL 16 du rapporteur.

M. le rapporteur. Je propose la création, s'agissant du blanchiment, de la corruption et du trafic d'influence, d'un statut de « repenti », à l'instar de ce que prévoit déjà le code pénal dans d'autres domaines. Une personne qui empêcherait la commission de l'infraction pourrait ainsi être exemptée de peine, et celle qui dénoncerait une infraction commise, bénéficier d'une réduction de peine.

De telles dispositions seraient de nature à donner à la justice un avantage important, notamment pour ce qui concerne les affaires les plus complexes.

M. Patrick Devedjian. Pourquoi ne pas également faire nôtre l'institution de la *bocca di leone*, qui encourageait les citoyens vénitiens à pratiquer la délation anonyme au nom de la protection de la république ?

M. le président Jean-Jacques Urvoas. Cela mérite un voyage d'étude ! (*Sourires.*)

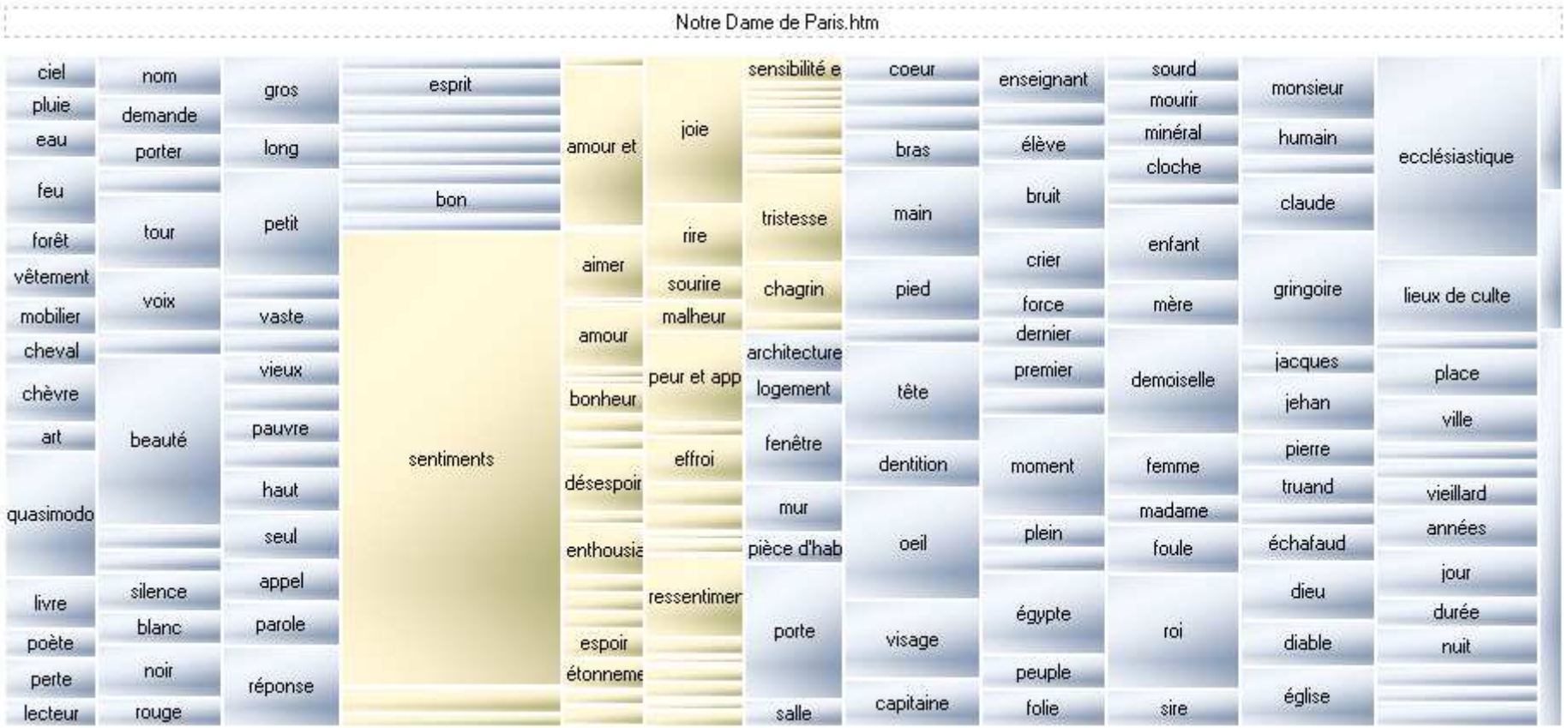
La Commission ***adopte*** *l'amendement.*

Commentaires sur le rapport de Yann GALUT. Nos 1130 et 1131. Enregistrement du 12 juin 2013.

Le traitement de ces informations nécessite le décodage d'implicites, complexes…

Les sentiments, dans un classique littéraire

Analyse d'un unique texte avec OwlEdge :

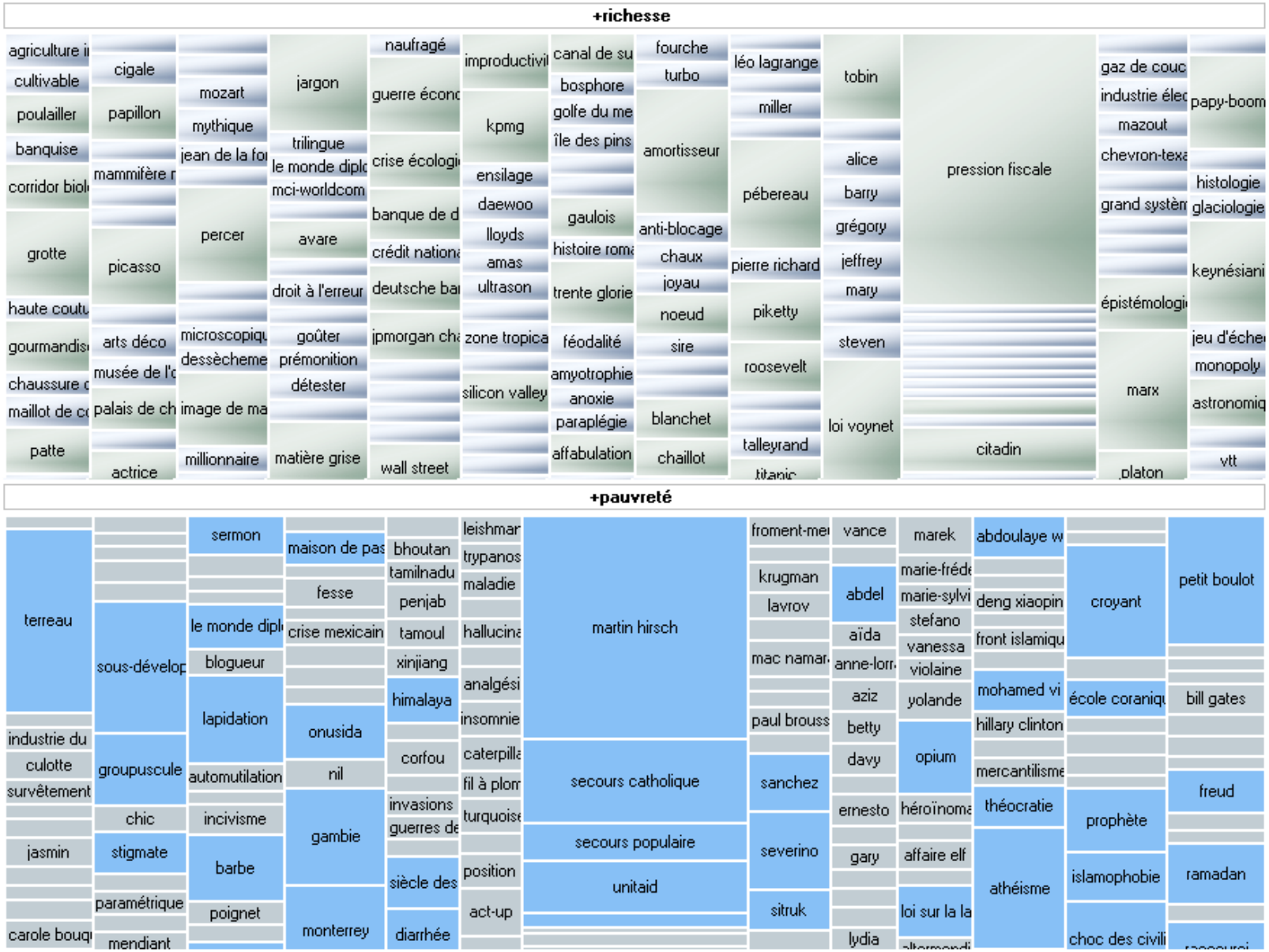


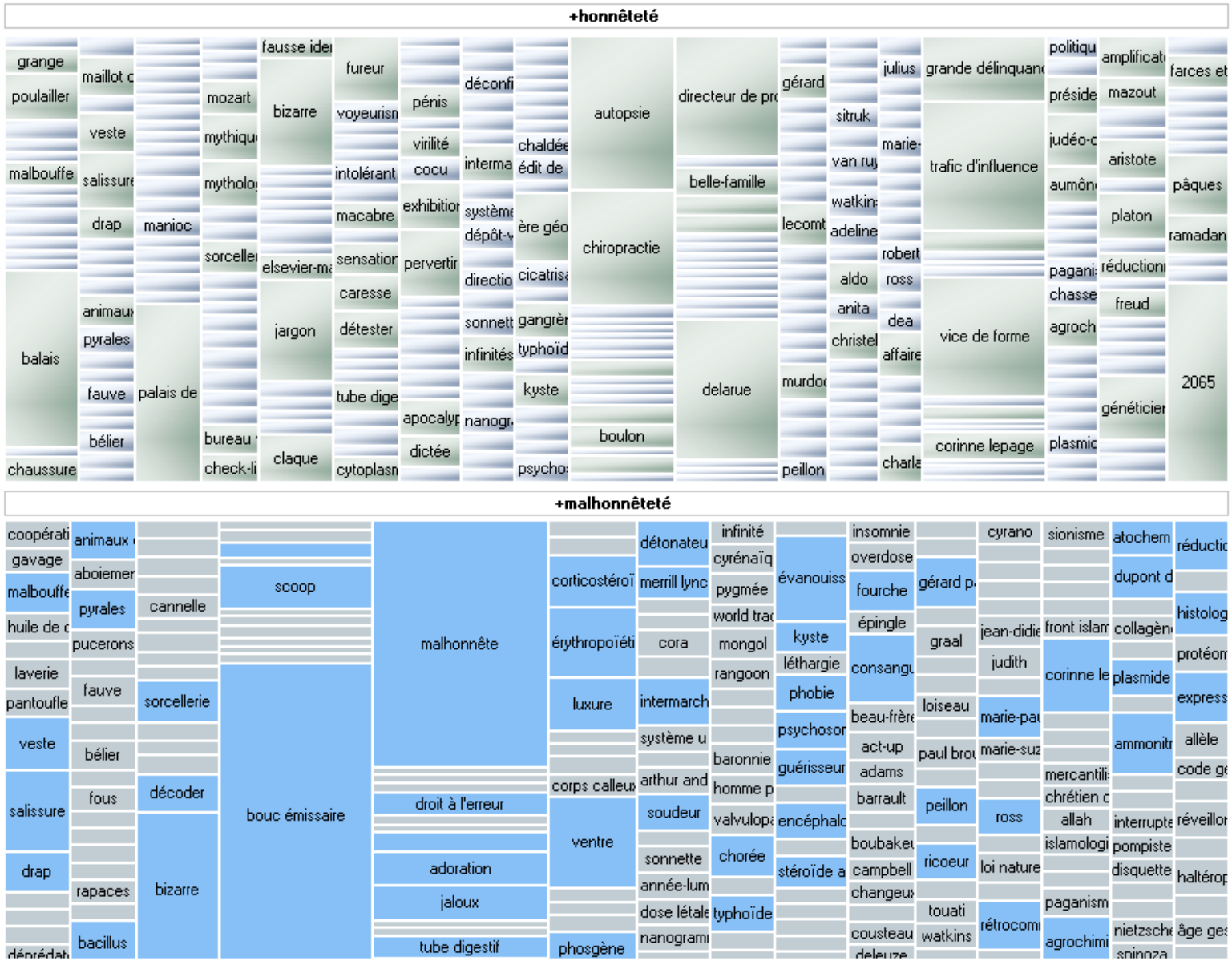
Nombre d'occurrences de mots. Notre Dame de Paris (Victor Hugo). Les sentiments sont fréquents.

Paires de graphes, Assemblée nationale

Suivent différents résultats de clustering des documents d'OwlEdge, sur des grappes de concepts antonymes, sur le corpus Assemblée nationale.

Tout ceci mériterait d'être complété par une étude plus longue.


+richesse
agriculture i
cultivable
poulailler
banquise
corridor biol
grotte
haute coutu
gourmandis
chaussure c
maillot de c
patte
cigale
papillon
mammifère r
picasso
arts déco
musée de l'c
palais de ch
actrice
mozart
mythique
jean de la fo
percer
microscopiqu
dessècheme
image de ma
millionnaire
jargon
trilingue
le monde diplo
mci-worldcom
avare
droit à l'erreur
goûter
prémonition
détester
matière grise
naufragé
guerre écono
crise écologi
banque de d
crédit nationa
deutsche ba
jpmorgan cha
wall street
improductivi
kpmg
ensilage
daewoo
lloyds
amas
ultrason
zone tropica
silicon valley
canal de su
bosphore
golfe du me
île des pins
gaulois
histoire roma
trente glorie
féodalité
amyotrophie
anoxie
paraplégie
affabulation
fourche
turbo
amortisseur
anti-blocage
chaux
joyau
noeud
sire
blanchet
chaillot
léo lagrange
miller
pébereau
pierre richard
piketty
roosevelt
talleyrand
titanic
tobin
alice
barry
grégory
jeffrey
mary
steven
loi voynet
pression fiscale
citadin
gaz de couc
industrie élec
mazout
chevron-texa
grand systèm
épistémologi
marx
platon
papy-boom
histologie
glaciologie
keynésiani
jeu d'éche
monopoly
astronomiq
vtt
+pauvreté
terreau
industrie du
culotte
survêtement
jasmin
carole bouq
sous-dévelop
groupuscule
chic
stigmate
paramétrique
mendiant
sermon
le monde dipl
blogueur
lapidation
automutilation
incivisme
barbe
poignet
maison de pas
fesse
crise mexicain
onusida
nil
gambie
monterrey
bhoutan
tamilnadu
penjab
tamoul
xinjiang
himalaya
corfou
invasions
guerres de
siècle des
diarrhée
leishmar
trypanos
maladie
hallucina
analgési
insomnie
caterpilla
fil à plom
turquoise
position
act-up
martin hirsch
secours catholique
secours populaire
unitaid
froment-me
krugman
lavrov
mac namar
paul brouss
sanchez
severino
sitruk
vance
abdel
aïda
anne-lorr
aziz
betty
davy
ernesto
gary
lydia
marek
marie-frédé
marie-sylvi
stefano
vanessa
violaine
yolande
opium
héroïnoma
affaire elf
loi sur la la
abdoulaye w
deng xiaopin
front islamiqu
mohamed vi
hillary clinton
mercantilisme
théocratie
athéisme
croyant
école coraniqu
prophète
islamophobie
choc des civili
petit boulot
bill gates
freud
ramadan


+honnêteté
grange
poulailler
malbouffe
balais
chaussure
maillot c
veste
salissure
drap
animaux
pyrales
fauve
bélier
manioc
palais de
mozart
mythiqu
mytholo
sorcelle
bureau
check-li
fausse ide
bizarre
elsevier-m
jargon
claque
fureur
voyeurisn
intolérant
macabre
sensation
caresse
détester
tube dige
cytoplasn
pénis
virilité
cocu
exhibitio
pervertir
apocalyp
dictée
déconfi
interma
système
dépôt-v
directio
sonnett
infinités
nanogr
chaldée
édit de
ère géo
cicatris
gangrèr
typhoïd
kyste
psycho
autopsie
chiropractie
boulon
directeur de pr
belle-famille
delarue
gérard
lecomt
murdo
peillon
sitruk
van ru
watkin
adeline
aldo
anita
christel
julius
marie-
robert
ross
dea
affaire
charle
grande délinquanc
trafic d'influence
vice de forme
corinne lepage
politiqu
préside
judéo-c
aumôn
pagani
chasse
agroch
plasmic
amplificat
mazout
aristote
platon
réduction
freud
généticie
farces et
pâques
ramadan
2065
+malhonnêteté
coopérati
gavage
malbouffe
huile de c
laverie
pantoufle
veste
salissure
drap
animaux
aboiemer
pyrales
pucerons
fauve
bélier
fous
rapaces
bacillus
cannelle
sorcellerie
décoder
bizarre
scoop
bouc émissaire
malhonnête
droit à l'erreur
adoration
jaloux
tube digestif
corticostéroï
érythropoïéti
luxure
corps calleux
ventre
phosgène
détonateu
merrill lync
cora
intermarch
système u
arthur and
soudeur
sonnette
année-lum
dose létale
nanogram
infinité
cyrénaïq
pygmée
world trac
mongol
rangoon
baronnie
homme p
valvulop
chorée
typhoïde
évanouiss
kyste
léthargie
phobie
psychosor
guérisseur
encéphalc
stéroïde a
insomnie
overdose
fourche
épingle
consangu
beau-frèr
act-up
adams
barrault
boubake
campbell
changeu
cousteau
deleuze
gérard p
graal
loiseau
paul brou
peillon
ricoeur
touati
watkins
cyrano
jean-didi
judith
marie-pa
marie-suz
ross
loi nature
rétrocom
sionisme
front islam
corinne le
mercantili
chrétien c
allah
islamologi
paganism
agrochim
atochem
dupont d
collagèn
plasmide
ammonit
interrupt
pompiste
disquette
nietzsch
spinoza
réductio
histolog
protéom
express
allèle
code g
réveillo
haltérop
âge ge