Laboratoire MoDyCo / Université Paris Ouest Nanterre
rloth@u-paris10.fr
(doctorant sous la dir. de J.L. Minel et D. Battistelli)

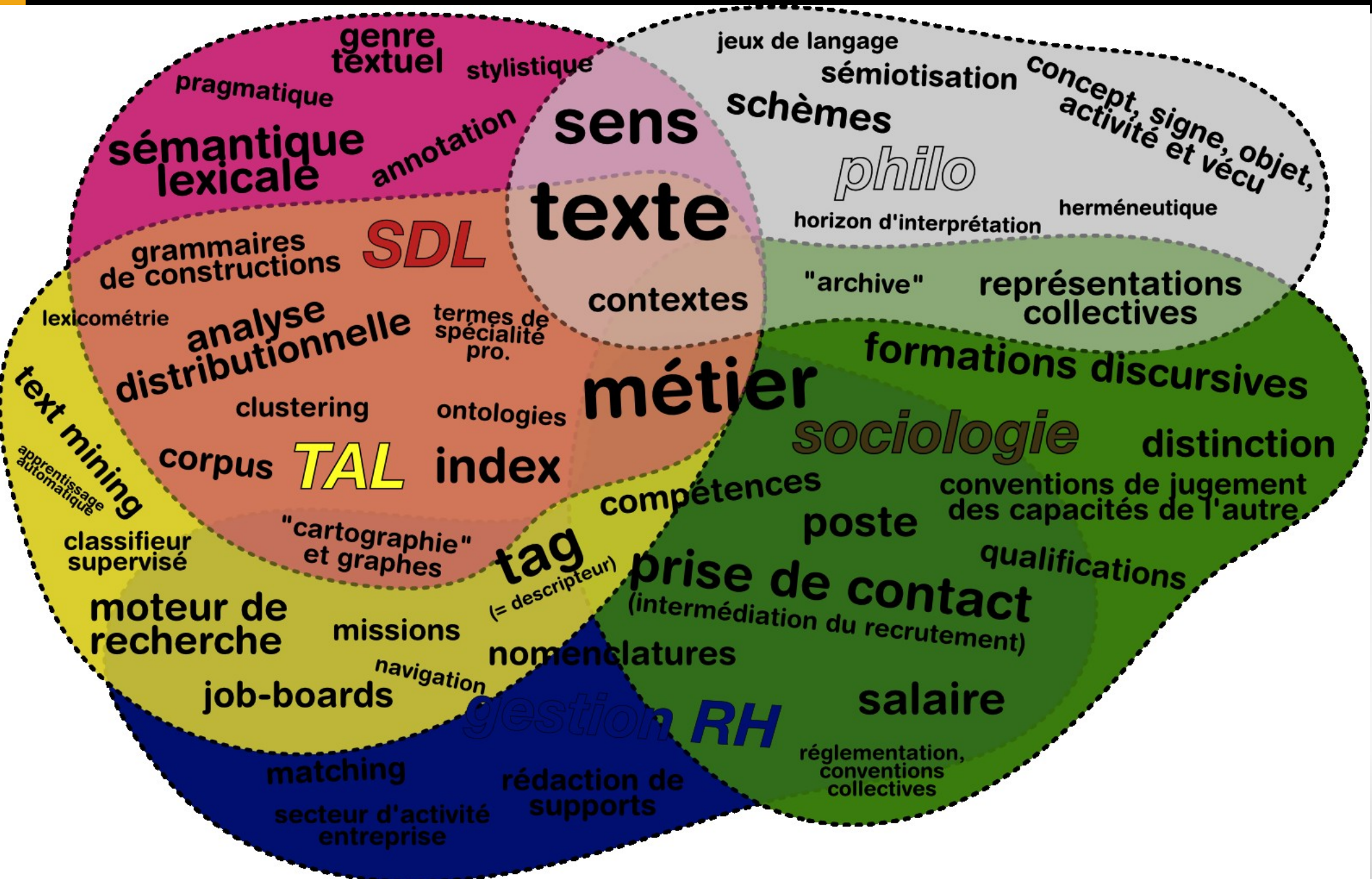
Opérations sur graphes lexicographiques issus de corpus à des fins de « visualisation sémantique »

modyco
Modèles, Dynamiques, Corpus
UMR 7114
Sciences du langage.

PROJET SIRE

l'Europe s'engage en Ile-de-France avec le FEDER

Vocabulaires de métiers : notions concernées



Partie 1

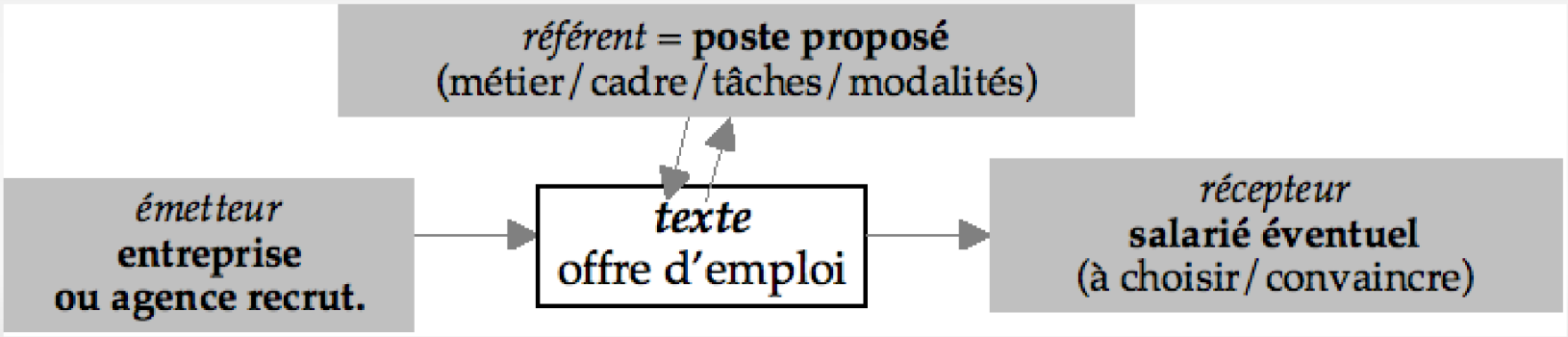
De nos données à la visualisation

Introduction

- Du TAL aux visualisations
  - le « Big data » incite à des visualisations d'ensemble
    - quantités et relations
    - « representing data accurately » (Fry 2008) : sur quels phéno ?
  - Mise en scène d'une description des contenus
    - annotations, moteur de recherche
    - l'info pour classer/naviguer est dans le texte (*« fouille »*)
  - Travaux sur les cooccurrences : un pont vers les graphes
    - famille des réseaux lexicaux
    - particularité : basé sur le corpus (observ. vs modèle ?)

Introduction

- A l'origine, un objectif applicatif sur les offres d'emploi
  - Une genre textuel ordinaire, descriptif, normé par la sphère sociale
  - nombreux besoins recherche d'information
- Matière première de la modélisation
  - lexique des RH («salaire», «équipe»...)
    - phraséologie fonctionnelle structurant le texte
  - lexique des métiers
    - le lexique utilisé suit l'évolution des métiers plus rapidement que les nomenclatures expertes
    - dénominations autorisées pour évoquer les contenus (activités, instruments, lieux, interlocuteurs, etc.)



Contenu du corpus

- Termes simples (N, V, A) et polylexicaux :
  - CDN : « chef de projet », « chiffre d'affaire », « lieu de travail », « structure de loisirs »
  - Adj. qualif. : « expérience significative », « projet associatif », « service gériatrique », « process interne »
  - V-Obj. : « démarrer carrière », « accueillir enfant », « contrôler paramètre », « commercialiser gamme »

EXEMPLE D'OFFRE DE L'APEC

Offre d'emploi Directeur Général H/F
Entreprise :
PME fabricant de luminaires tertiaire et industriel cherche son directeur général (H/F). L'entreprise est en fort développement, structurée avec une vingtaine de personnes dont quatorze en production.

Missions :
Sous la responsabilité du président, il aura en charge l'entier développement de l'entreprise :
* technique (production, BE, maintenance et sécurité),
* administrative/financière (management, juridique, comptabilité/social, informatique/transmission de l'information, relations extérieures),
* marketing (innovation produits, prix, distribution, communication).

La stratégie globale de l'entreprise est définie par le président, les différentes politiques induites sont à mettre en place par le directeur, qui produira un tableau de bord mensuel reflétant la situation de l'entreprise.

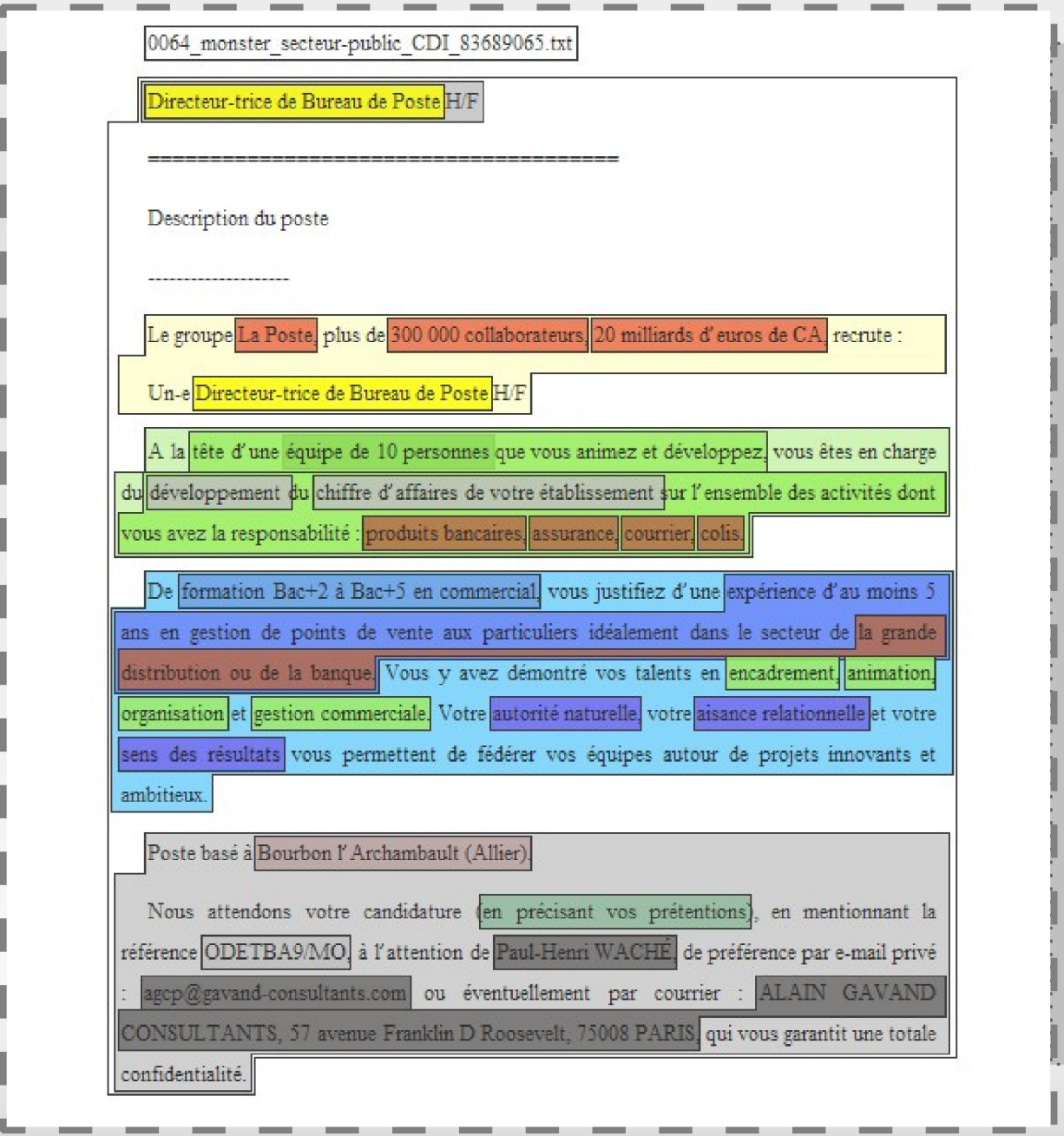
Profil :
Expérience en bureau d'étude tôlerie fine et en management indispensable.
Expérience souhaitée en eléctricité, électronique et/ou optique.
Formation complétée par la holding de l'entreprise. Diplôme ingénieur CNAM souhaité
Rémunération environ 40/45 000€ brut + intéressement suivant diplôme et expérience,
Voiture et téléphone de fonction. Poste à pourvoir rapidement.

EXEMPLE D'OFFRE DE POLE EMPLOI

Métier du ROME H2503 - Pilotage d'unité élémentaire de production mécanique
Chef mécanicien / mécanicienne d'atelier de fabrication
VOUS CONTROLEZ L'ETAT ET LA CONFORMITE DES MOYENS DE PRODUCTION ET DES APPROVISIONNEMENTS,PLANIFIEZ L'ACTIVITE DU PERSONNEL,SUIVEZ ET CONTROLEZ LA QUALITE/QUANTITE DE LA PRODUCTION,IDENTIFIEZ LES DYSFONCTIONNEMENTS. MAITRISE GPAO ET LOGICIELS DE COMMANDES NUMERIQUES.AVOIRAU MOINS 57 ANS

Source	Documents
Pôle Emploi	16974
monster.fr	2817
APEC	1298
fiches ONISEP	540
TOTAL	**21629**

Organisation textuelle (1/2)



0064_monster_secteur-public_CDI_83689065.txt

Directeur-trice de Bureau de Poste H/F

==========

Description du poste

Le groupe La Poste, plus de 300 000 collaborateurs, 20 milliards d'euros de CA, recrute :

Un-e Directeur-trice de Bureau de Poste H/F

A la tête d'une équipe de 10 personnes que vous animez et développez, vous êtes en charge du développement du chiffre d'affaires de votre établissement sur l'ensemble des activités dont vous avez la responsabilité : produits bancaires, assurance, courrier, colis.

De formation Bac+2 à Bac+5 en commercial, vous justifiez d'une expérience d'au moins 5 ans en gestion de points de vente aux particuliers idéalement dans le secteur de la grande distribution ou de la banque. Vous y avez démontré vos talents en encadrement, animation, organisation et gestion commerciale. Votre autorité naturelle, votre aisance relationnelle et votre sens des résultats vous permettent de fédérer vos équipes autour de projets innovants et ambitieux.

Poste basé à Bourbon l'Archambault (Allier).

Nous attendons votre candidature (en précisant vos prétentions), en mentionnant la référence ODETBA9/MO, à l'attention de Paul-Henri WACHÉ, de préférence par e-mail privé : agcp@gavand-consultants.com ou éventuellement par courrier : ALAIN GAVAND CONSULTANTS, 57 avenue Franklin D Roosevelt, 75008 PARIS, qui vous garantit une totale confidentialité.

- Corpus : *annoter/repérer*
- Structure de l'information dans le texte
  - Poste
  - Entreprise, Secteur
  - Missions, activités
  - Expérience
  - Savoir-faire
  - Formation
  - Contact
- Régularités de distribution des occurrences
- Lieux communs descriptifs/argumentatifs du genre
- Affectent aussi les relations lexicales

Organisation textuelle (2/2)

Positions des types d'information observés dans 150 offres APEC

Des données textuelles aux graphes

- Focus sur les proximités sémantiques
  - Problématique : évaluer la proximité paraphrastique
  - Cooccurrences de 2nd ordre à travers ≠ types de contextes
- Exemples de "compétences" en vrac
  - « maîtrise de l'environnement BPL », « pack office », « outils informatiques » (sans plus de précisions)
  - « gestion de projet », « qualités managériales » « meneur(se) d'hommes et de femmes »
  - « dynamique », « goût pour le travail en équipe », «orienté résultats », « sens commercial »
  - « gérontologie », « cryptographie », « radiofréquences (RFID) »

Démarche DSM (1/4)

- Principe de la sémantique distributionnelle
  - La méthode distributionnelle en sémantique,
    - telle qu'envisagée par Harris ou Maurice Gross
    - les co-occurrences d'un mot permettent de le **caractériser** sémantiquement (capturer son sens ?)
  - Exemple Quel est le sens du terme 'bardiwac' ?
    - « Je prendrais un verre de bardiwac. »
    - « Il a renversé le bardiwac… »
    - « Ce bardiwac est fameux ! »
  - Décomptes et analyses dorénavant appliquables sur de grands corpus

Démarche DSM (2/4)

- Les décomptes en contexte saisissent le comportement du terme

Observations dans le corpus

3x réseau aérien
8x réseau électrique
3x centre sanitaire
3x centre municipal
7x transport aérien
4x transport sanitaire
3x entretien municipal
8x système mécanique
6x système électrique
3x système manuel
9x aptitude manuelle
12x aptitude relationnelle
18x poste équivalent
13x poste disponible
14x poste similaire
4x animateur disponible
6x animateur municipal
4x expérience équivalente
6x expérience similaire

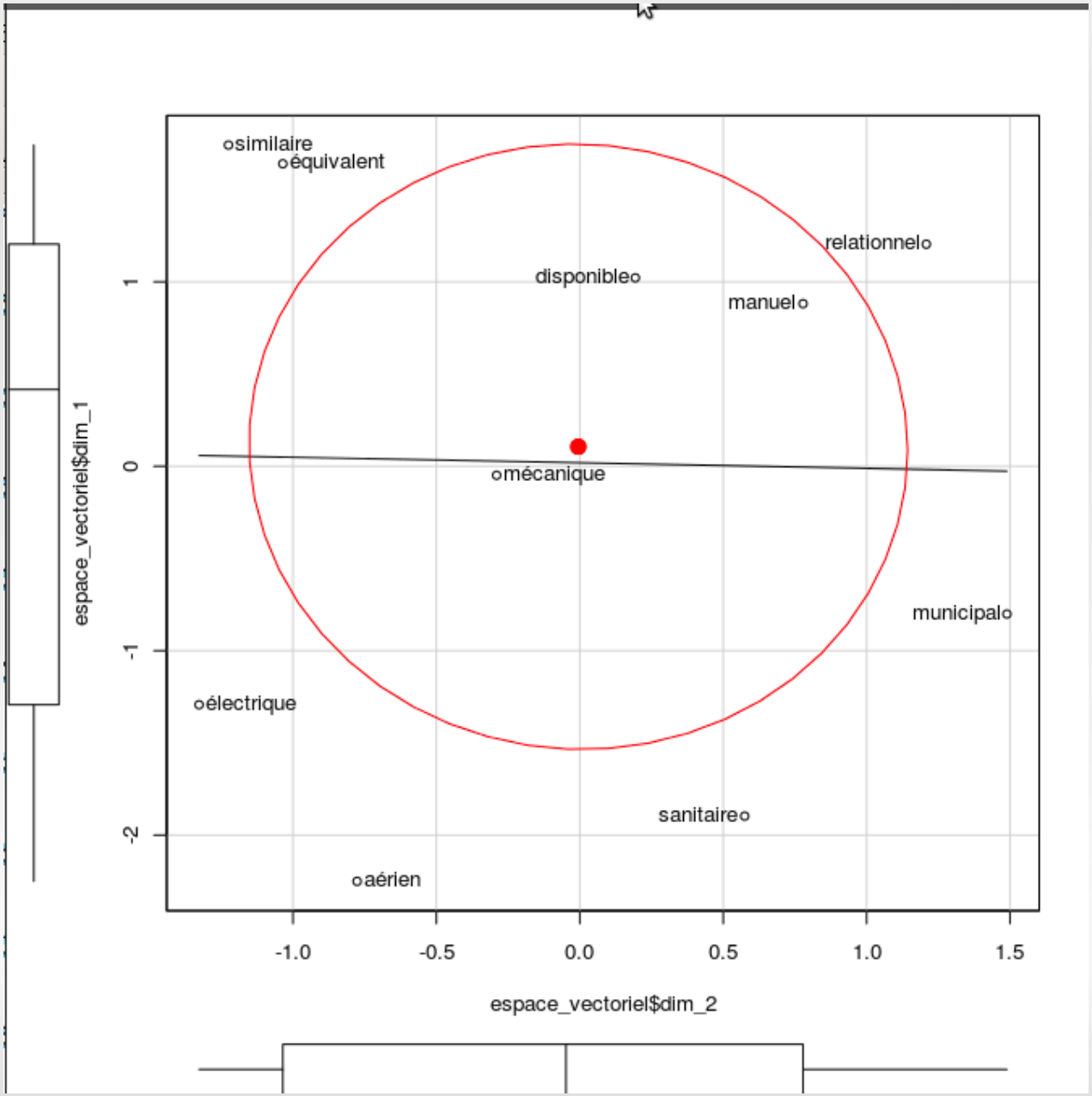
Groupements des adj. par contextes nominaux

$N_1 \leftarrow$ {aérien, électrique}
$N_2 \leftarrow$ {sanitaire, municipal}
$N_3 \leftarrow$ {aérien, sanitaire}
$N_4 \leftarrow$ {électrique, mécanique}
$N_5 \leftarrow$ {manuel, relationnel}
$N_6 \leftarrow$ {similaire, équivalent, disponible}
$N_7 \leftarrow$ {disponible, municipal}
$N_8 \leftarrow$ {équivalent, similaire}

Matrice des adjectifs relevés

	[N_1]	[N_2]	[N_3]	[N_4]	[N_5]	[N_6]	[N_7]	[N_8]
aérien	3	0	7	0	0	0	0	0
municipal	0	3	0	0	0	0	3	0
sanitaire	0	3	4	0	0	0	0	0
électrique	8	0	0	5	0	0	0	0
mécanique	0	0	0	8	0	0	0	0
manuel	0	0	0	3	9	0	0	0
relationnel	0	0	0	0	12	0	0	0
équivalent	0	0	0	0	0	18	0	4
similaire	0	0	0	0	0	14	0	6
disponible	0	0	0	0	0	13	4	0

Démarche DSM (3/4)



- 8 contextes (dims d'observation)
- réduction SVD (≈ ACP)
- ici k=3 dims en sortie
- j'ai choisi les 2 plus intéressantes comme coordonnées
- la dim 1 oppose :

{ similaire, équivalent, disponible, manuel, relationnel }

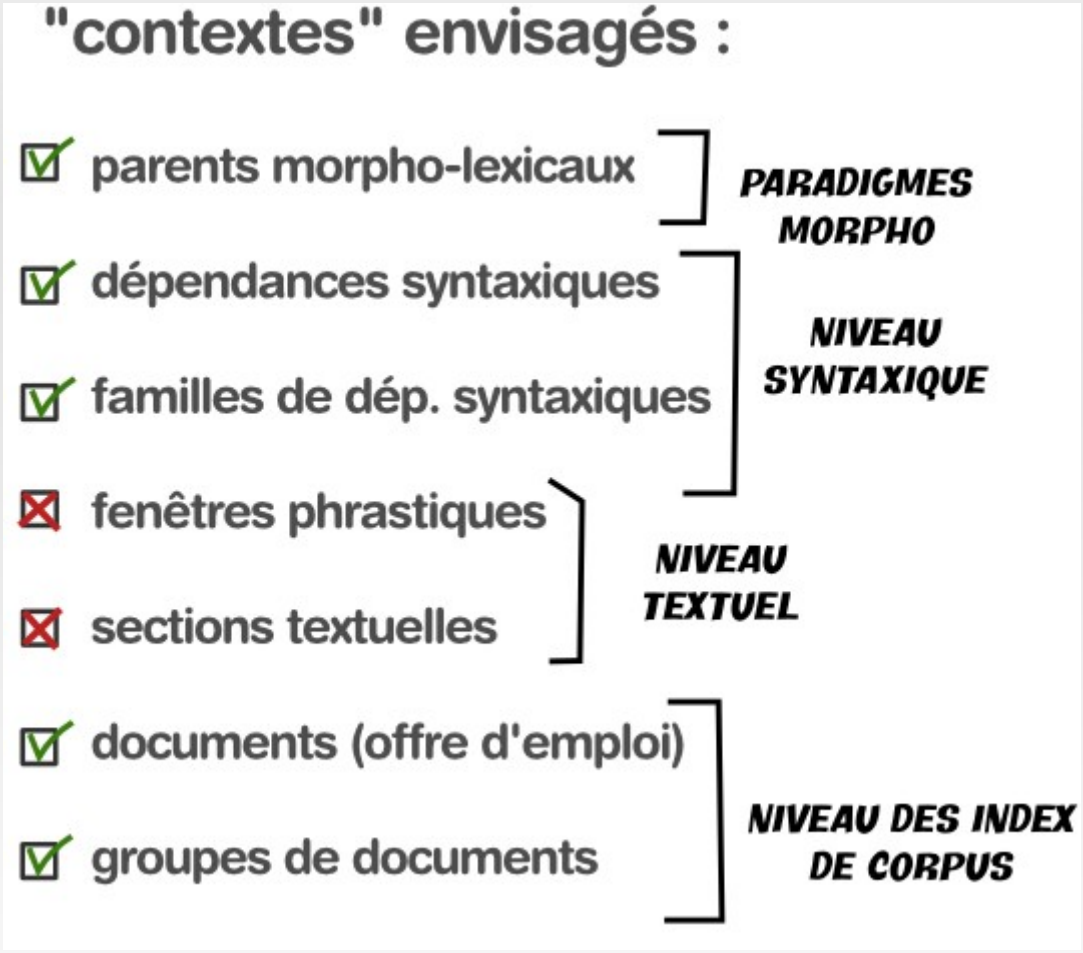
à

{ mécanique, électrique, aérien, sanitaire, municipal }

Démarche DSM (4/4)

- le relevé d'usage d'un terme simule son signifié…
  - le signe comme ensemble d'usages : piste théorique qui rejoint (Sahlgren 2008) et (De Mauro 1969)
- On obtient quelquechose qui ressemble à un espace vectoriel
- Avec une similarité problématique (mesure et interprétation)
  - distance euclidienne ? cosinus ? ou entropie relative KL ?
- L'approximation obtenue du « sens » est :
  - spécifique au corpus (stylistique, discours, thèmes abordés)
  - spécifique aux contextes pris en compte

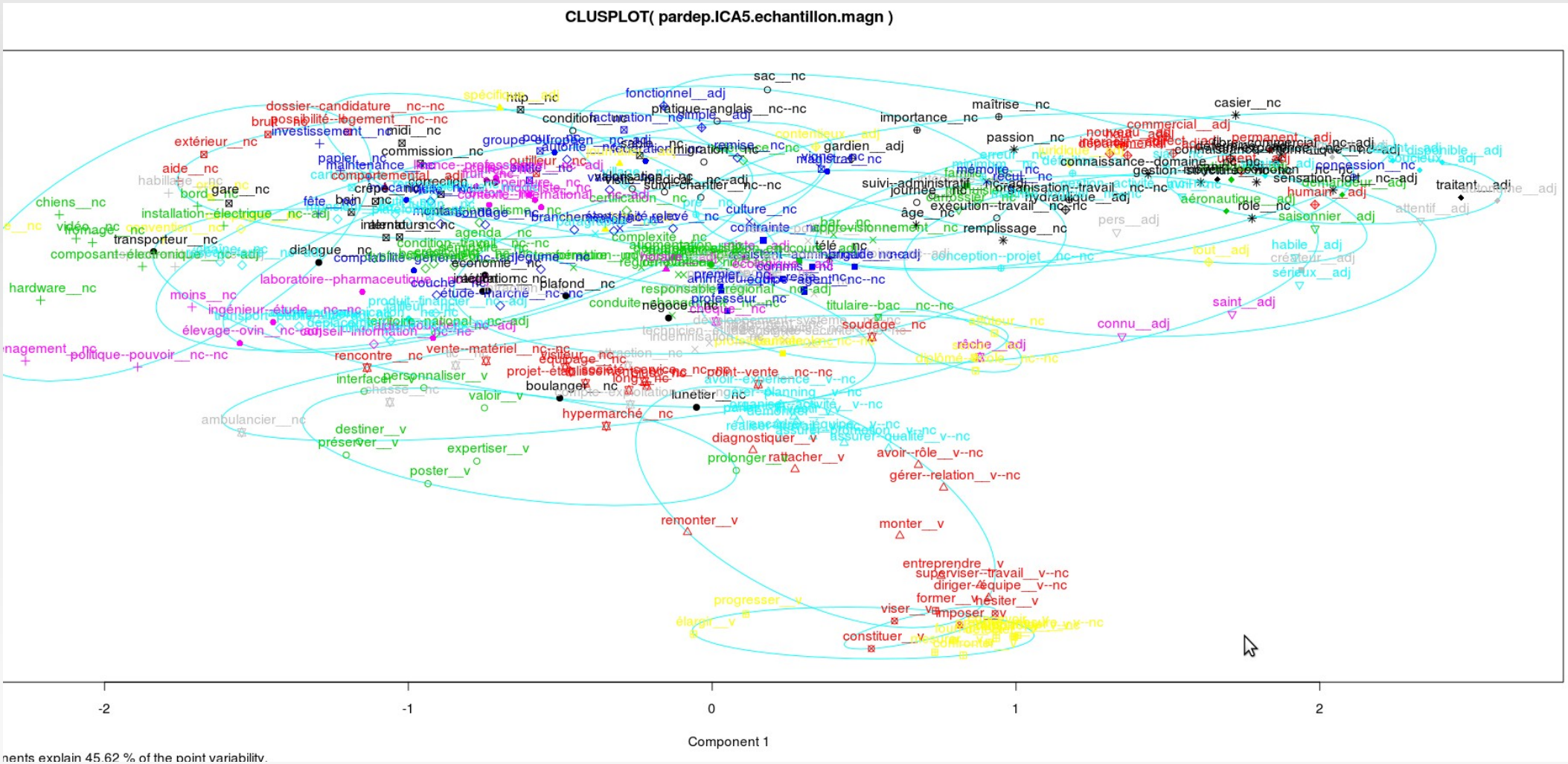
Démarche DSM (4/4)



- ex 1: distribution sur les documents
  - voisinage dans le texte : «*médecin*»; «*chirgurgie*»; «*hospitalier*» ; «*vacation*» ; «*thérapeute*», «*personnel d'accompagnement*» etc.
- ex 2: distribution sur les contextes syntaxiques
  - voisinage selon les contextes de dépendances :
  - contextes comme : «*au sein de H*» ; «*directeur d'H*» «*concours de directeur d'H*» «*H public*» ; «*H de province*» ; (…)
  - voisins comme : *trésor public* ; *collectivité territoriale* ; *siège d'entreprise centre hospitalier* ; *agence bancaire* ; (…)

Hypothèse : les facettes du sens d'un terme se retrouvent graduellement à chaque échelon de son environnement.

Visualisation de DSM ?

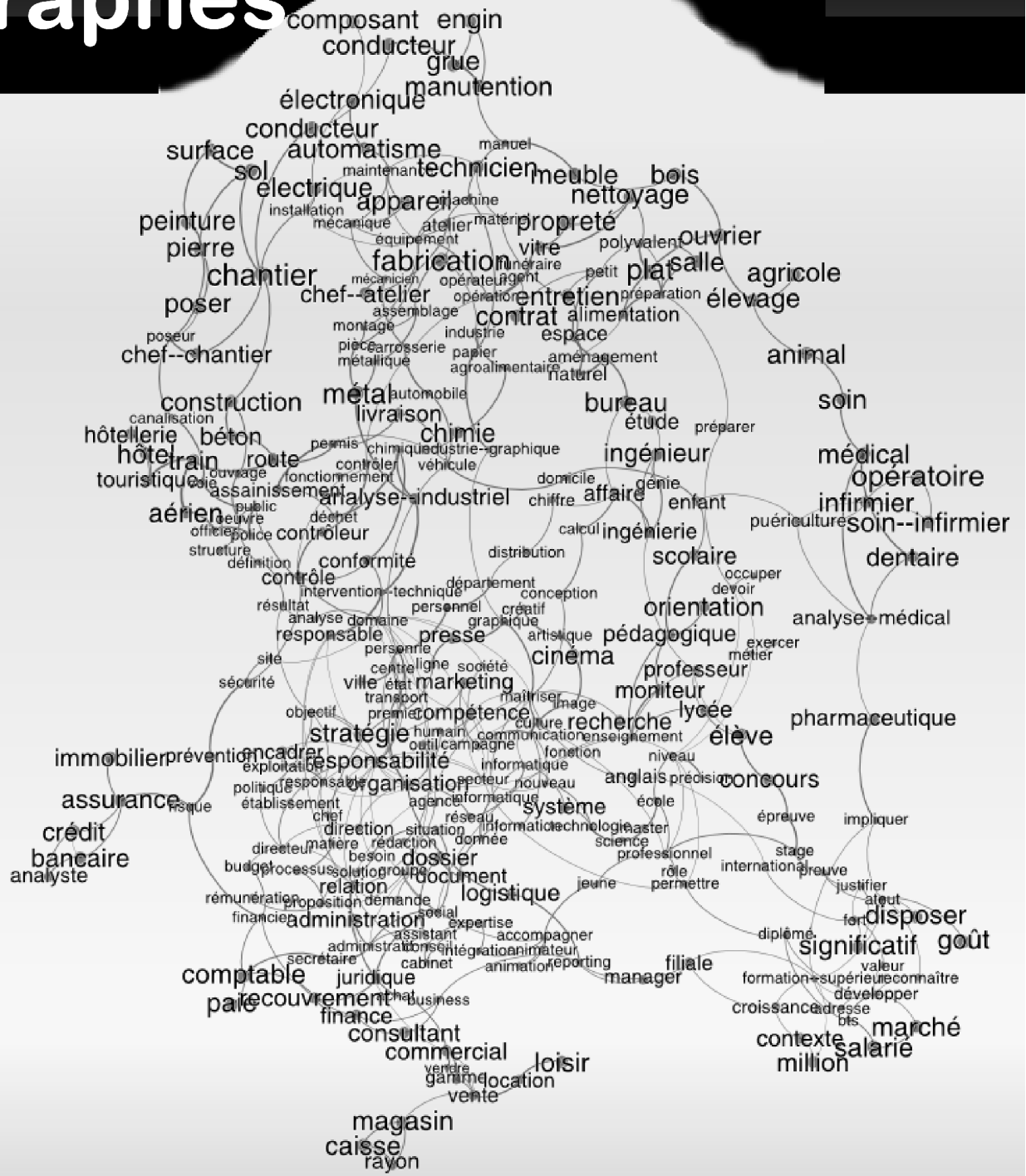


De l'e. v. aux graphes

- mesure de similarité (distance sémantique ?)

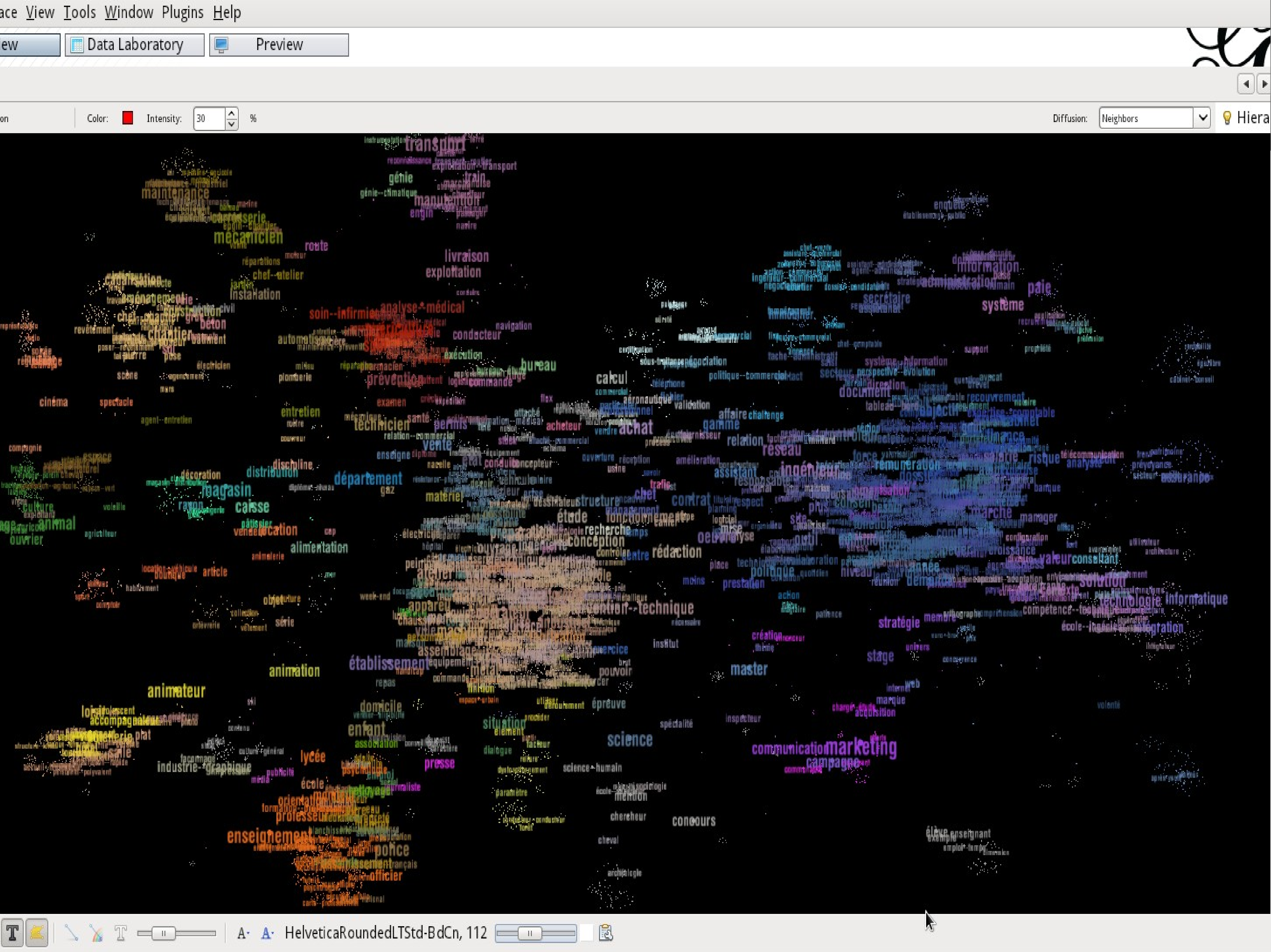
  - - - -

- meilleure similarité donne un arc (knn)
- seuil de similarité donne un arc (fixe ou variable)
- clustering donne un noeud pyramidal avec arcs



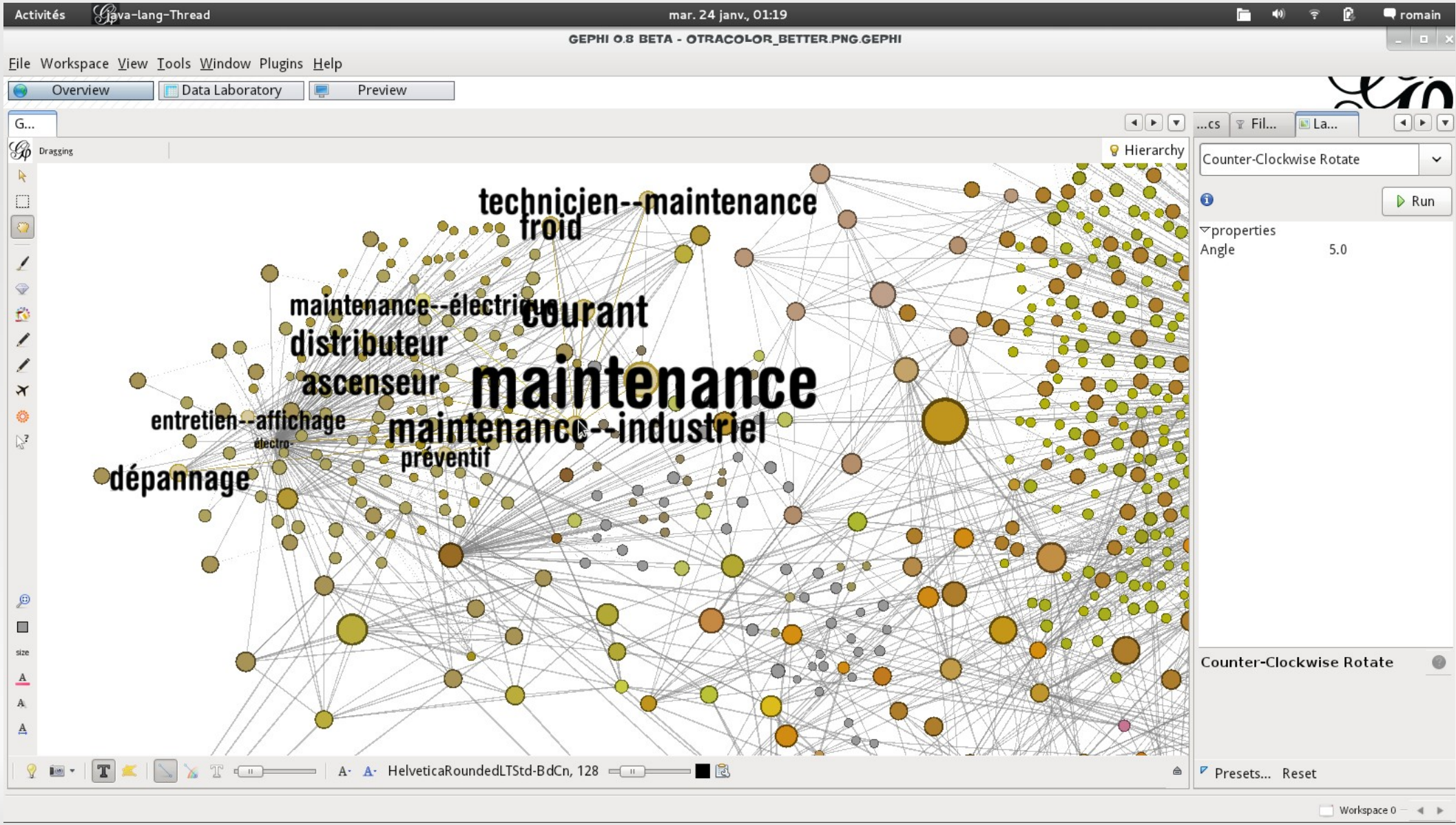
Réseaux comme modèles visuels et/ou comme modèles linguistiques

- Au-delà du réseau dictionnairique, des modèles ML
- Un air du temps multi-disciplinaire :
  - relations sémantiques (relations, polysémie, isotopies) et réseaux lexicaux : (Ploux et Victorri 1998), (Gaume 2004), (Lafourcade 2011)
  - travaux sur les IHM : nuages de tags (`wordle`), interfaces d'ontologies (Cao et al. 2010),
  - modèles textométriques pour le classement thématique et la visualisation `topigraphy` de (Fujimura et al. 2008) et `TreeCloud` de (Gambette et Veronis 2010)
  - travaux sur les espaces vectoriels lexicaux (LSI, DSM, word spaces), induction de lexiques
  - réseaux de terrain : études mathématiques (Mihalcea et Radev 2011), outils de manip/visualisation (`gephi, igraph, csw`)
- Convergence vers une vision topologique du sens

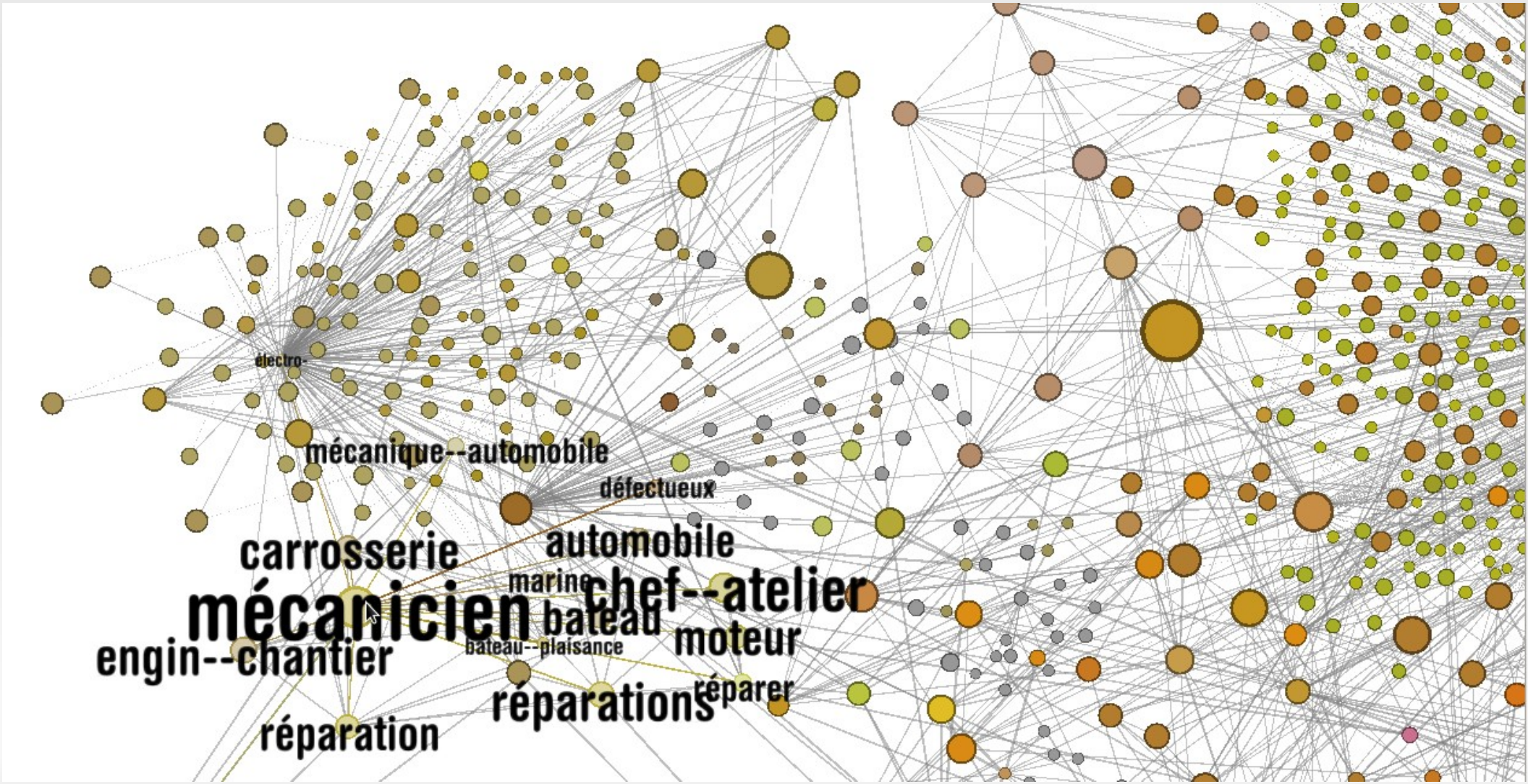


ace View Tools Window Plugins Help
Data Laboratory
Preview
Color:
Intensity:
30
%
Diffusion:
Neighbors
transport
exploitation--transport
génie
train
manutention
engin
maintenance
carrosserie
mécanicien
route
livraison
exploitation
chef--atelier
installation
soin--infirmier
analyse--médical
conducteur
navigation
chantier
béton
bâtiment
exécution
bureau
prévention
commande
cinéma
spectacle
agent--entretien
entretien
technicien
permis
vente
calcul
aéronautique
validation
achat
gamme
relation
réseau
affaire
assistant
ingénieur
information
administration
paie
secrétaire
système
document
rémunération
risque
télécommunication
assurance
discipline
décoration
distribution
magasin
département
caisse
matériel
structure
étude
fonctionnement
recherche
conception
contrat
rédaction
animal
ouvrier
alimentation
article
objet
série
animation
établissement
animateur
domicile
enfant
association
situation
épreuve
science
industrie--graphique
lycée
presse
enseignement
police
officier
stratégie
stage
master
marketing
communication
campagne
solution
technologie
informatique
compétence
valeur
consultant
concours
chercheur
science--humain
HelveticaRoundedLTStd-BdCn, 112

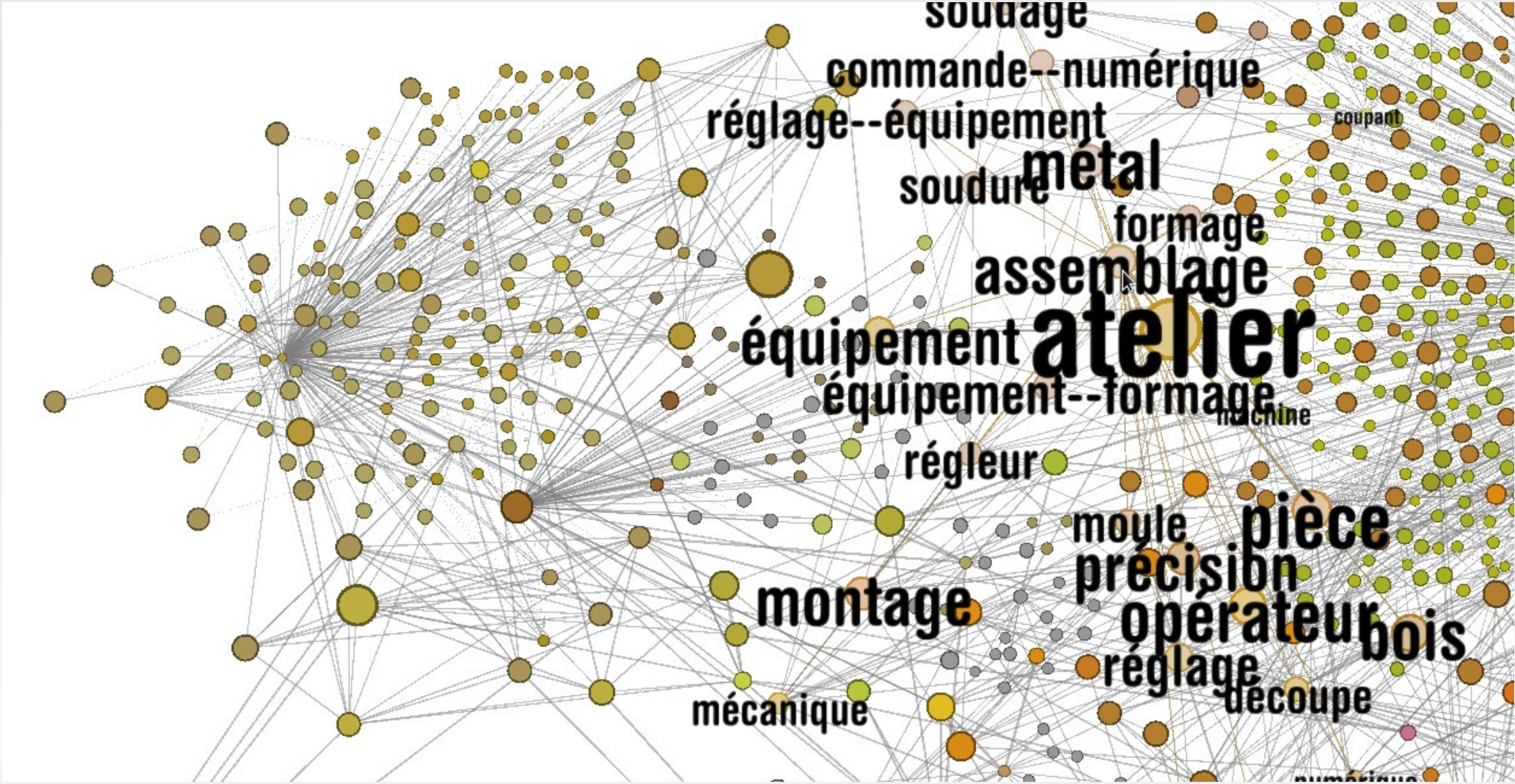
ex 1 : espace des voisinages par docs

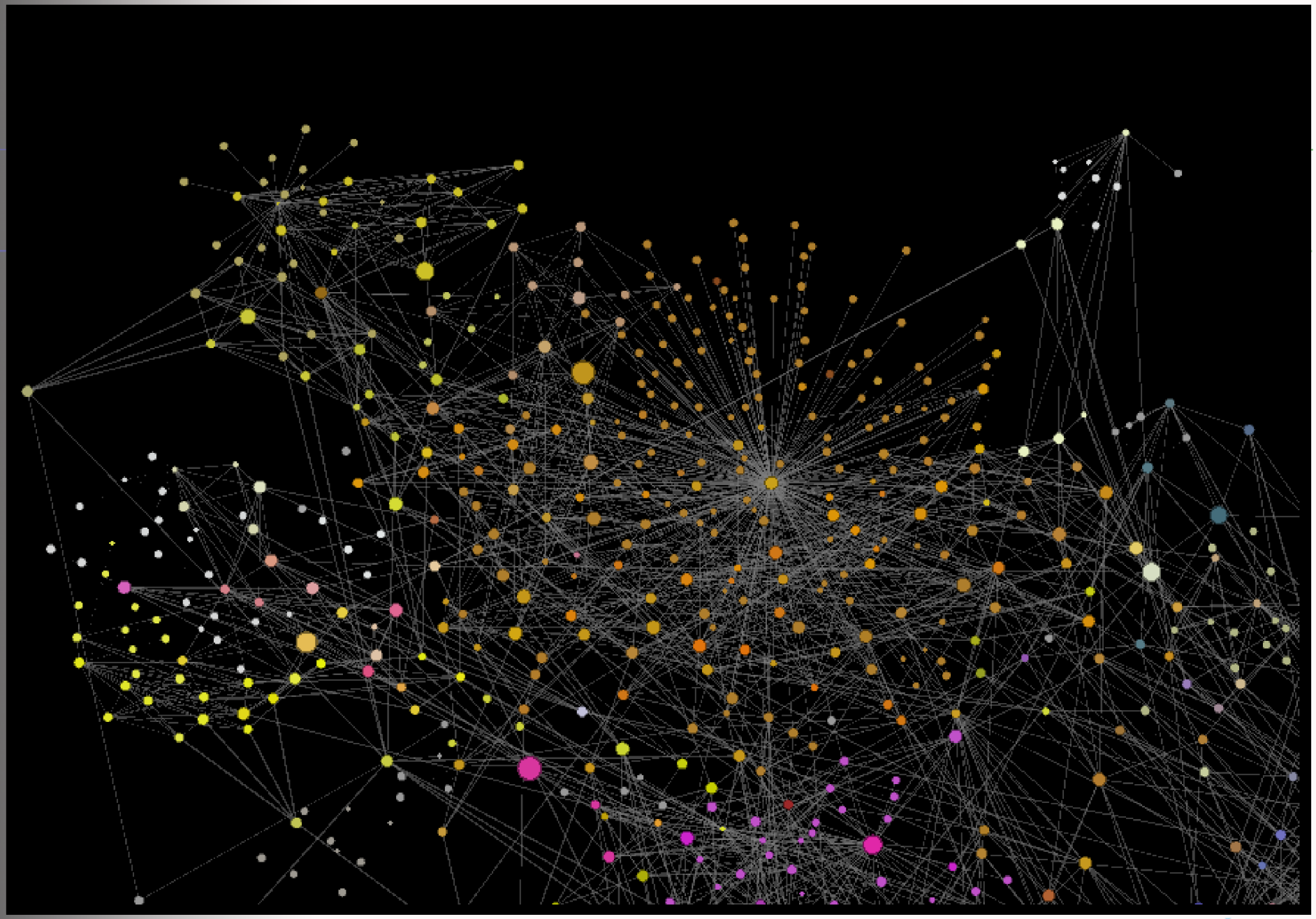


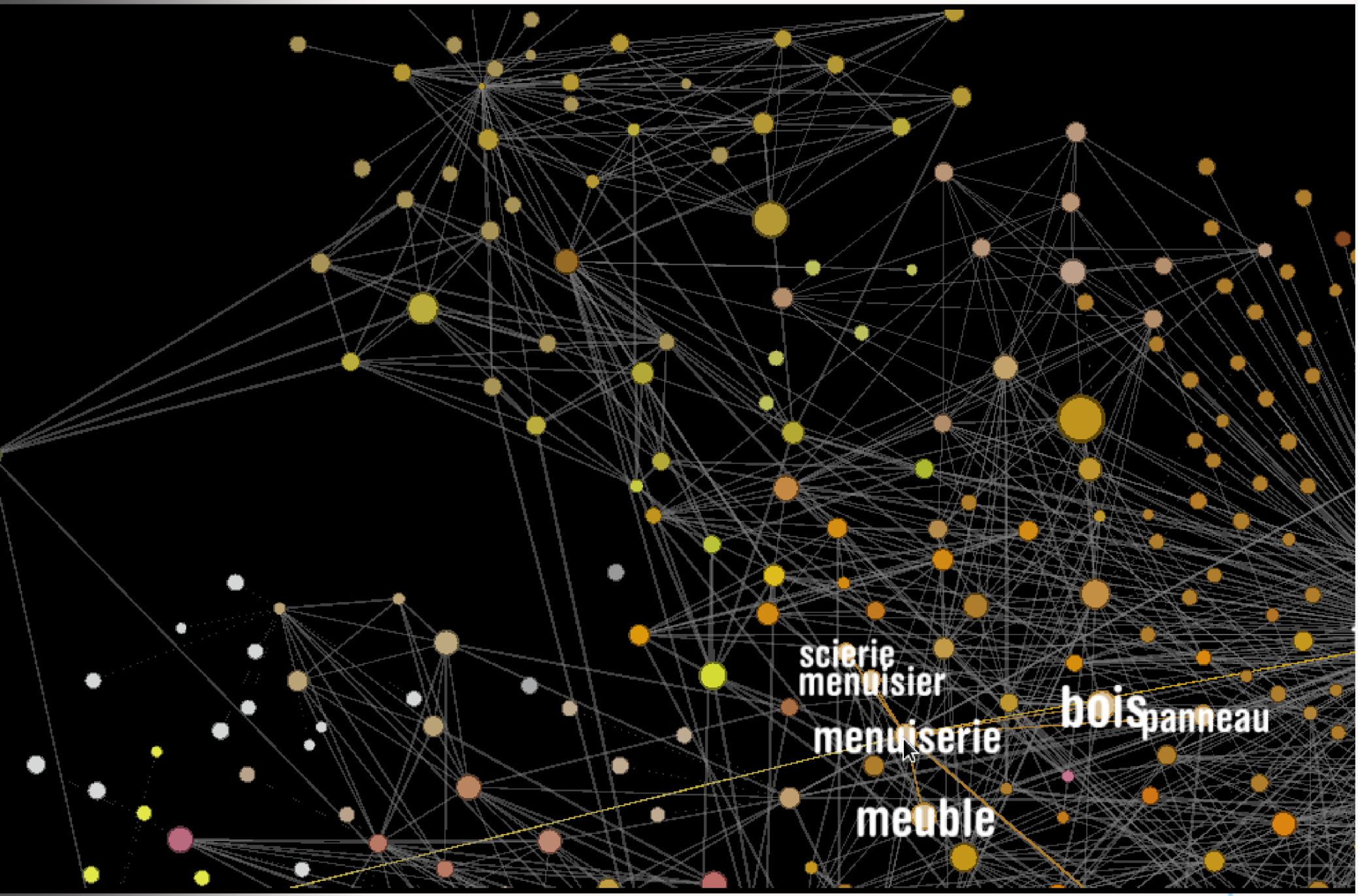
ex 1 : espace des voisinages par docs

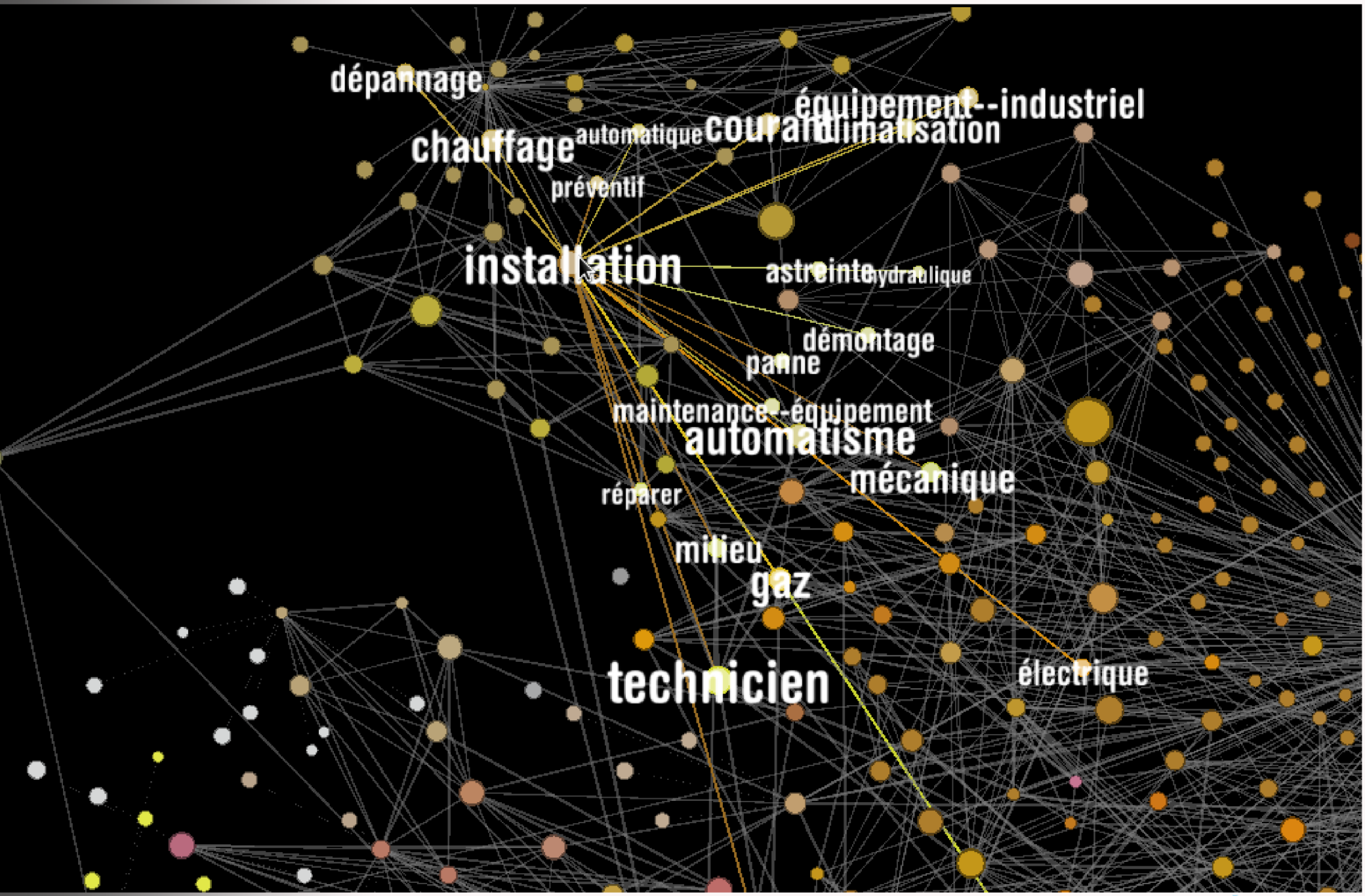


ex 1 : espace des voisinages par docs

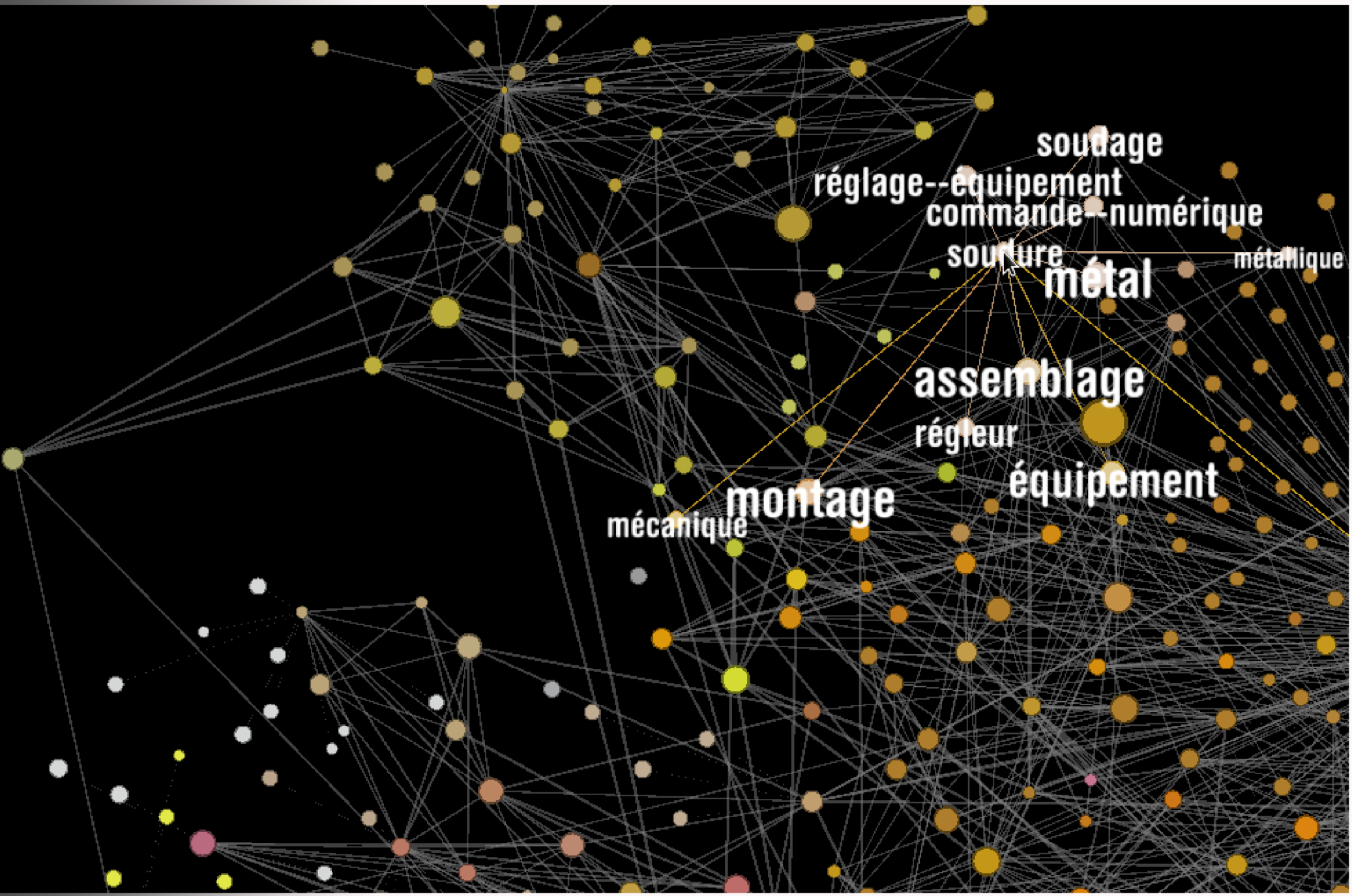




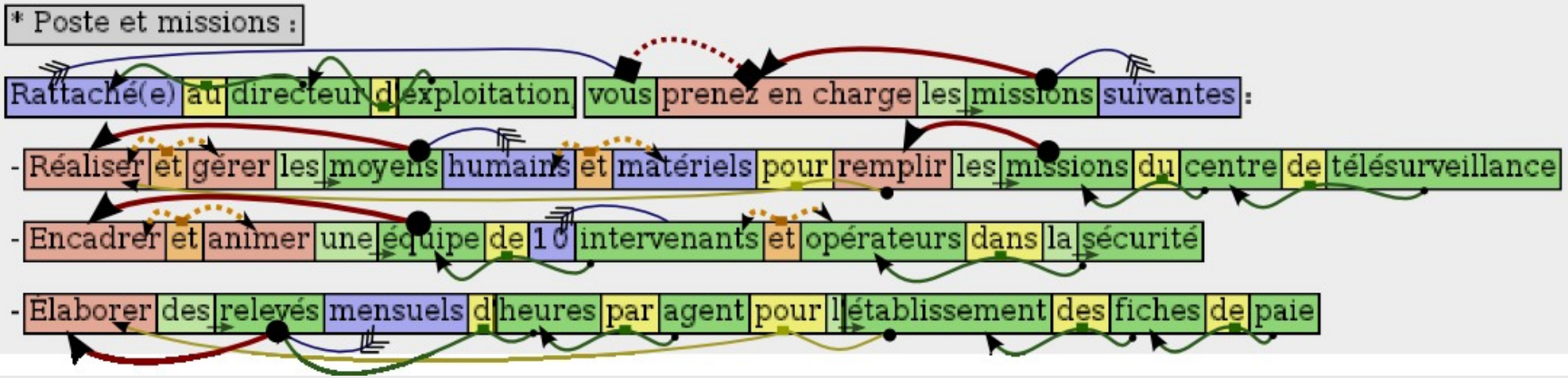

scierie
menuisier
menuiserie
bois
panneau
meuble


dépannage
équipement--industriel
chauffage
automatique
courant
climatisation
préventif
installation
astreinte
hydraulique
démontage
panne
maintenance--équipement
automatisme
mécanique
réparer
milieu
gaz
technicien
électrique


topographique
acoustique
aménagement
intérieur
architecte
construction
extraction


soudage
réglage--équipement
commande--numérique
soudure
métallique
métal
assemblage
régleur
équipement
montage
mécanique

ex 2 : espace des dépendances syn.



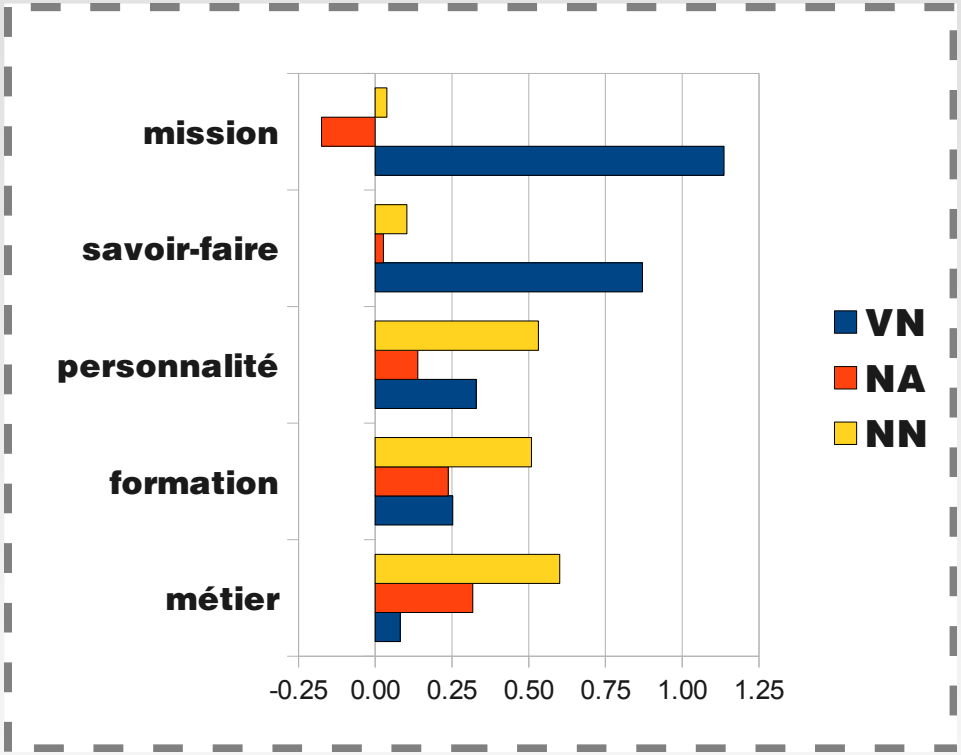
Lexies considérées :

système__nc
ligne__nc
matériel__nc
train__nc
équipement__nc
câblage__nc
nacelle__nc
moteur__nc
bateau__nc
ossature__nc

Extraits des contextes/fréquences relevés dans le DSM

hasAdj.électrique__adj	13
isDeN.conduite__nc	11
hasAdj.industriel__adj	9
isObj.réaliser__v	9
isDeN.installation__nc	8
hasAdj.mécanique__adj	7
hasAdj.électronique__adj	7
isDeN.exploitation__nc	7
isDeN.maintenance__nc	7

Analyse des dépendances



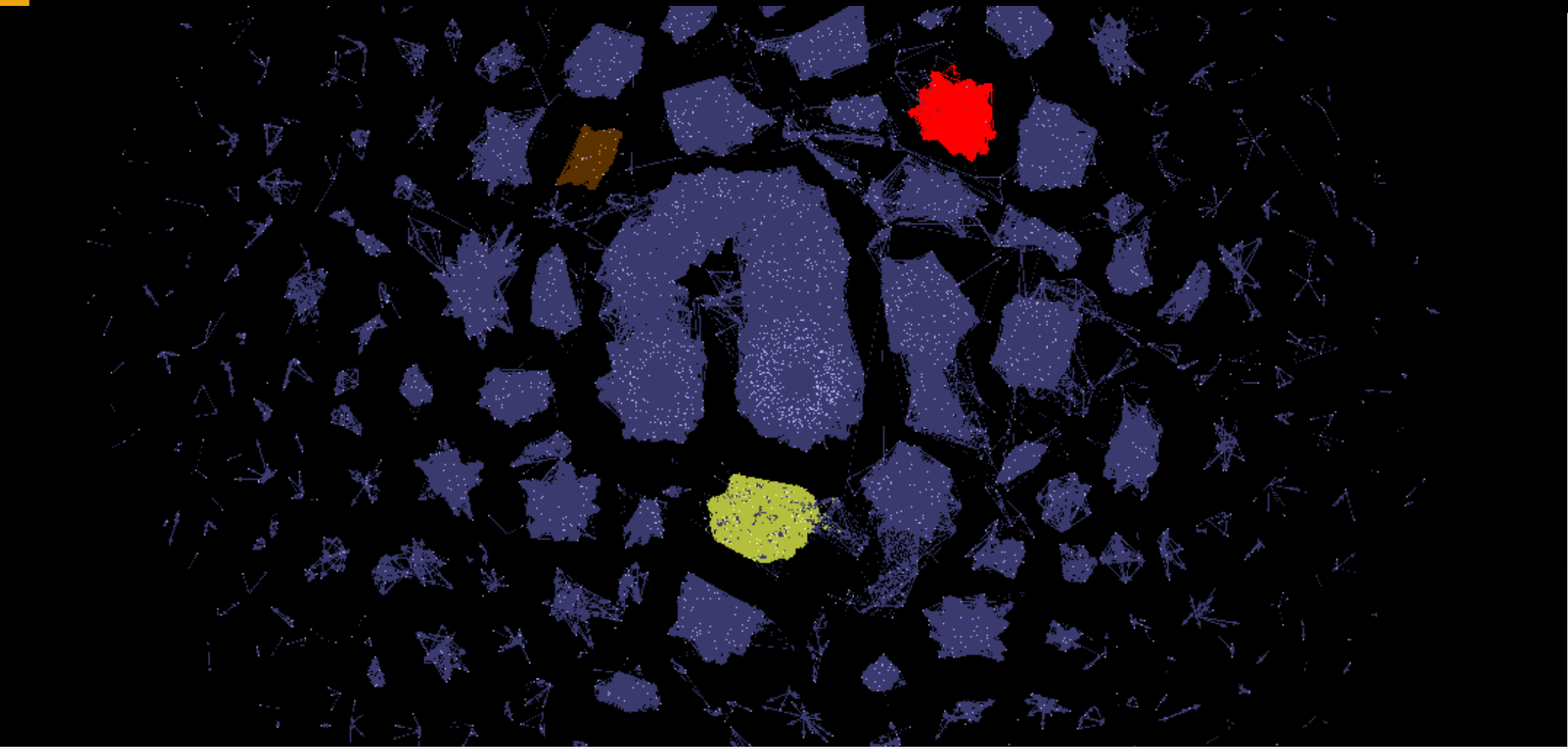
Utilisé aussi pour le figement

1) Relations de dép. entre chaque nom, verbe, adj.
2) Fréq. cooccurences

   => coef. corrélation => «espace»

3) On peut re-tracer le détail des sources et comparer
   - pour métadonnées (eg. *type d'info* selon listes d'unité lexicales)
   - pour les clusters générés nous-même

ex 2 : espace des dépendances syn.



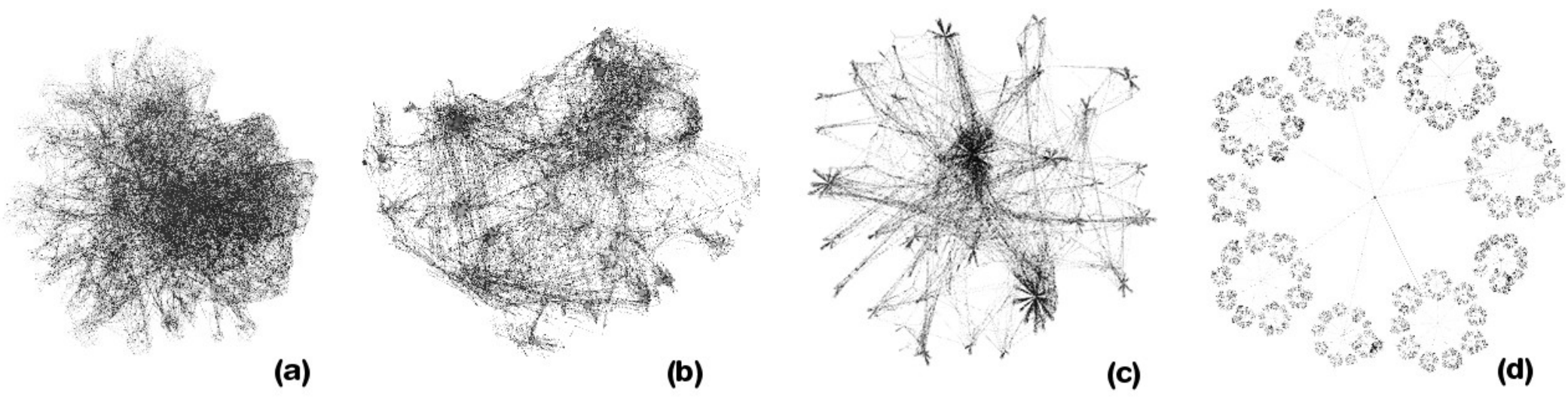
- Les proximités par dépendances communes
  - un espace beaucoup plus fragmenté que par contextes docs

Partie 2

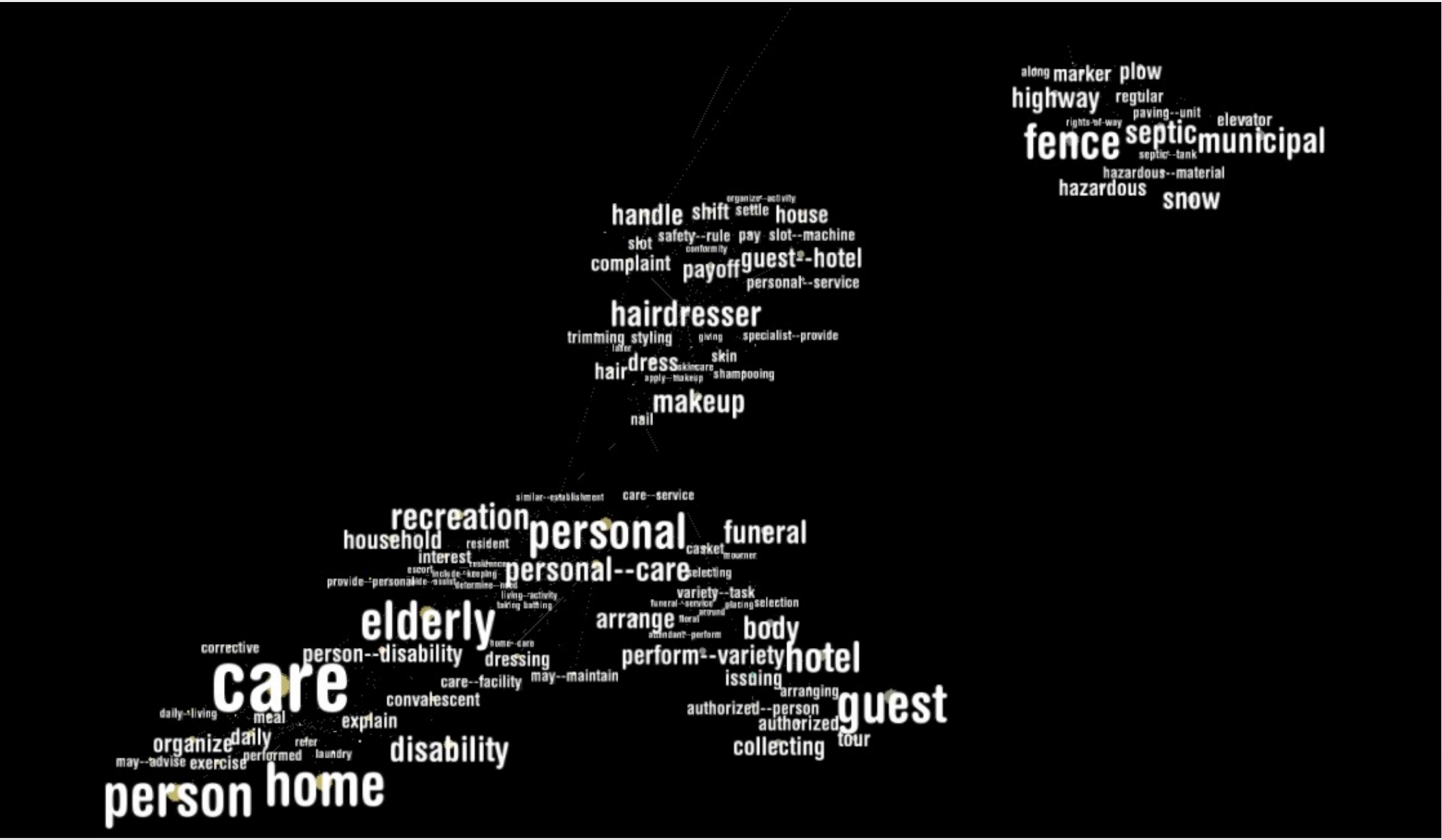
Opérations d'analyse et visualisation-interprétation

Le graphe simplifie tout, ouvre des pistes

- L'obtention d'un espace tangible après un an plutôt axé sur l'e.v.
  - les « petits » points (ici illisibles) sont autant de lexies hyper-spécialisées
  - tendance monosémique des vocabulaires techniques facilite la tâche (positionnement + univoque ou *strong ties*)
  - termes fréquents : souvent plus polysémiques ou « passerelles » (*weak ties*)
    - ex. « soin », « bureau »
- On essaye 4 méthodes de génération des arcs, avec des propriétés différentes
  - (a) nombre d'arcs constant pour chaque noeud, (d) clustering hiérarchique simple
  - (b) et (c) sont des méthodes intermédiaires



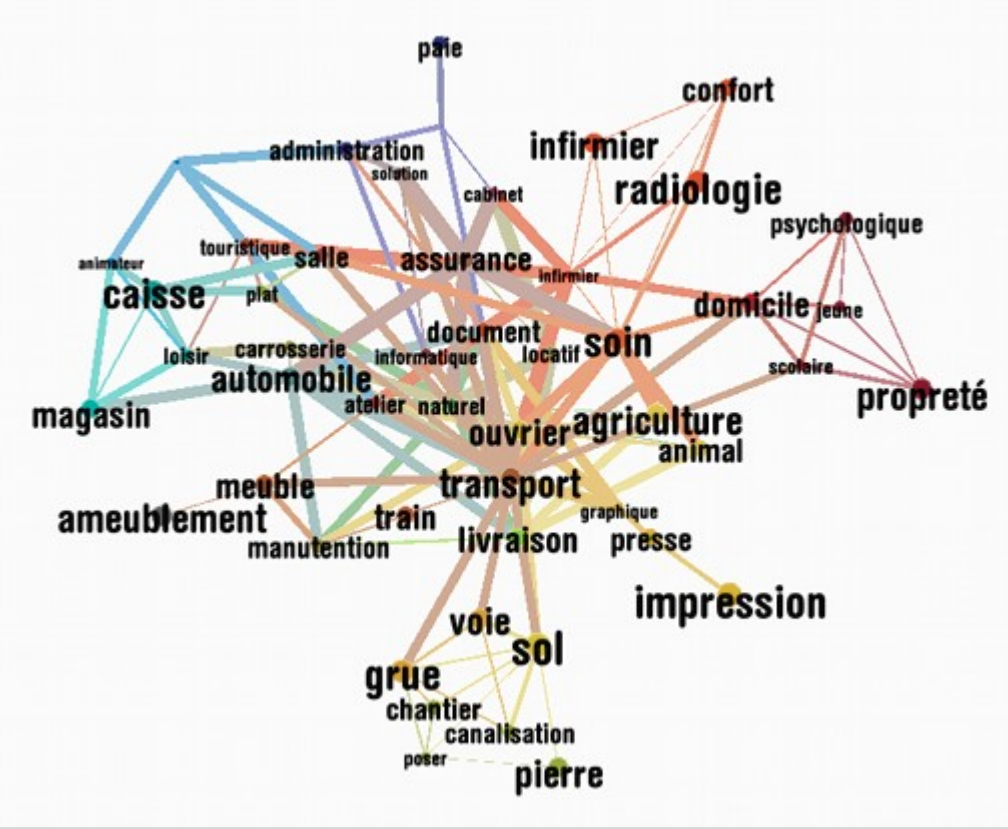
ex 3 : extrait de graphe US-SOC



Le graphe simplifie tout, ouvre des pistes

- Pose la question du modèle théorique sous-jacent ?
  - Reproductibilité ? Suggestivité du visuel ?
- La visualisation est parfois déjà une modélisation théorique
  - *cf.* aspect d'une courbe en analyse, ou arbres syntaxiques

- Dans notre cas : étude de la forme d'ensemble
  - fractalité observée
  - *ie*. des niveaux de précision imbriqués de plus en plus spécifiques autour d'un parangon
- Aussi étude théorique des relations obtenues
  - Qu'est-ce que la « proximité» sémantique ?



Opérations « visuelles » ou « théoriques » ?

- Une fois le graphe lexical obtenu, on peut imaginer une infinité d'opérations dessus
- similarité pré-calculée => forme de donnée plus synthétique que l'espace vectoriel
- indicateurs unaires (ou potentiel, altitude)
  - fréquence, spécificité, etc.
- manips avec la librairie R igraph
  - clustering apcluster : sélection de parangon (prototype)
    - utile en désambigüisation
  - interprétation de la zone qui lui est rattachée ? isotopie ?
- 3 exemples suivent :
  - manipuler des clusters (= groupes émergents de semblables)
  - annoter les arcs
  - identifier les paraphrases d'expressions polylexicales (MWE)

Isoler des clusters et les manipuler

- Clusterings sur l'espace
  - kmeans, kmedoids
  - apcluster
- idem sur le graphe
  - détection de communautés
  - random walks, mesure mincut
- permet immédiatement de distinguer des domaines techniques/thématiques
- permet un prototypage non-supervisé d'ontologies
- permet de propager des métadonnées catégorielles (McLachlan)

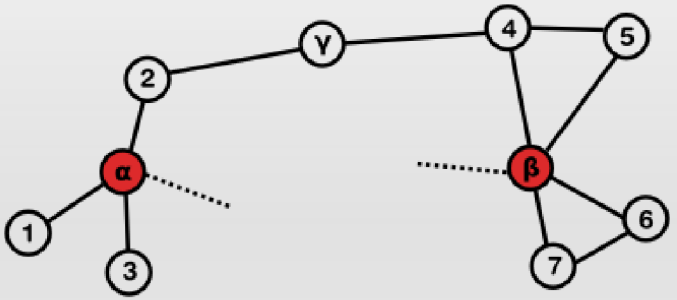
- Applications nombreuses
  - classement thématique
  - nommer les clusters
  - désambigüiser (Victorri)
    - un noeud ayant des voisins réparti dans 2 clusters éloignés est probablement un terme polysémique avec 2 acceptions
  - permet de propager des métadonnées catégorielles (McLachlan, spread activation)

Annoter les arcs

- Baroni & Lenci 2010
  - « distributional memory »
- plusieurs graphes croisés en 1
  - pour les ontologies selon différentes relations conceptuelles
  - pour le lexique différents « paradigmes » contextuels
- intuitivement : opération d'intersection de la zone thématique « médical » avec la zone de dépendances «lieu» => «hôpital»
- autre piste : automate probabiliste
  - annotation en types
  - mesure de cohésion textuelle

Trouver des paraphrases de MWE

- sur sous-espace des figés
  - 1ers résultats moins bons
  - on perd la topologie locale?
  - à re-tenter
- sur graphe d'ens.
  - par substitution d'un seul des deux éléments par un voisin
  - cf. algo simple ci-contre
  - bon résultat moins «logique», plus inattendu
  - marche bien (diapo suivante)
  - à faire évoluer



v(x) voisin de x
a+b concaténation a et b

On cherche les v(a+b)
matchPhrasemes naïf:
$v(a+b) = \{x_i+y_j\}$ tels que :
$x_i \in v(\alpha)$
$y_j \in v(\beta)$
et x_i+y_j figé

==> problème car on oublie les possibilités 2+γ γ+4

Par exemple pour le couple de départ [rédiger__v—document__nc]

=> voisins de « rédiger »

+ voisins de « document »

+ vérification si le nouveau couple est figé

Trouver des paraphrases de MWE

Résultats de l'algorithme page précédente :

précision moyenne, bon rappel

=> très intéressants

[posséder__v--niveau__nc]

[suivre__v--évolution__nc]

[apprécier__v--contact__nc]

[connaître__v--condition__nc]

[connaître__v--évolution__nc]

[recruter__v--professionnel__nc]

[prendre__v--contact__nc]

[organiser__v--temps__nc]

[prendre__v—temps__nc]

[rédiger__v--document__nc]

[étudier__v--plan__nc]

[traiter__v--donnée__nc]

[définir__v--moyens__nc]

[définir__v--procédure__nc]

[définir__v--plan__nc]

[concerner__v--domaine__nc]

[requérir__v—maîtrise__nc]

[réaliser__v--calcul__nc]

[effectuer__v--calcul__nc]

[effectuer__v--test__nc]

[effectuer__v--modification__nc]

[établir__v--chiffrage__nc]

[respecter__v--coût__nc]

[effectuer__v--chiffrage__nc]

[établir__v—coût__nc]

[contrôler__v--fabrication__nc]

[effectuer__v--réglage__nc]

[prendre__v--commande__nc]

[effectuer__v--contrôle__nc]

[effectuer__v--montage__nc]

[réaliser__v--assemblage__nc]

[vérifier__v--élément__nc]

[superviser__v--contrôle__nc]

[effectuer__v--série__nc]

[réaliser__v--réglage__nc]

[établir__v--commande__nc]

[effectuer__v--assemblage__nc]

[gérer__v--commande__nc]

[réaliser__v--élément__nc]

[réaliser__v--montage__nc]

[réaliser__v--pièce__nc]

[prendre__v--contrôle__nc]

[réaliser__v--commande__nc]

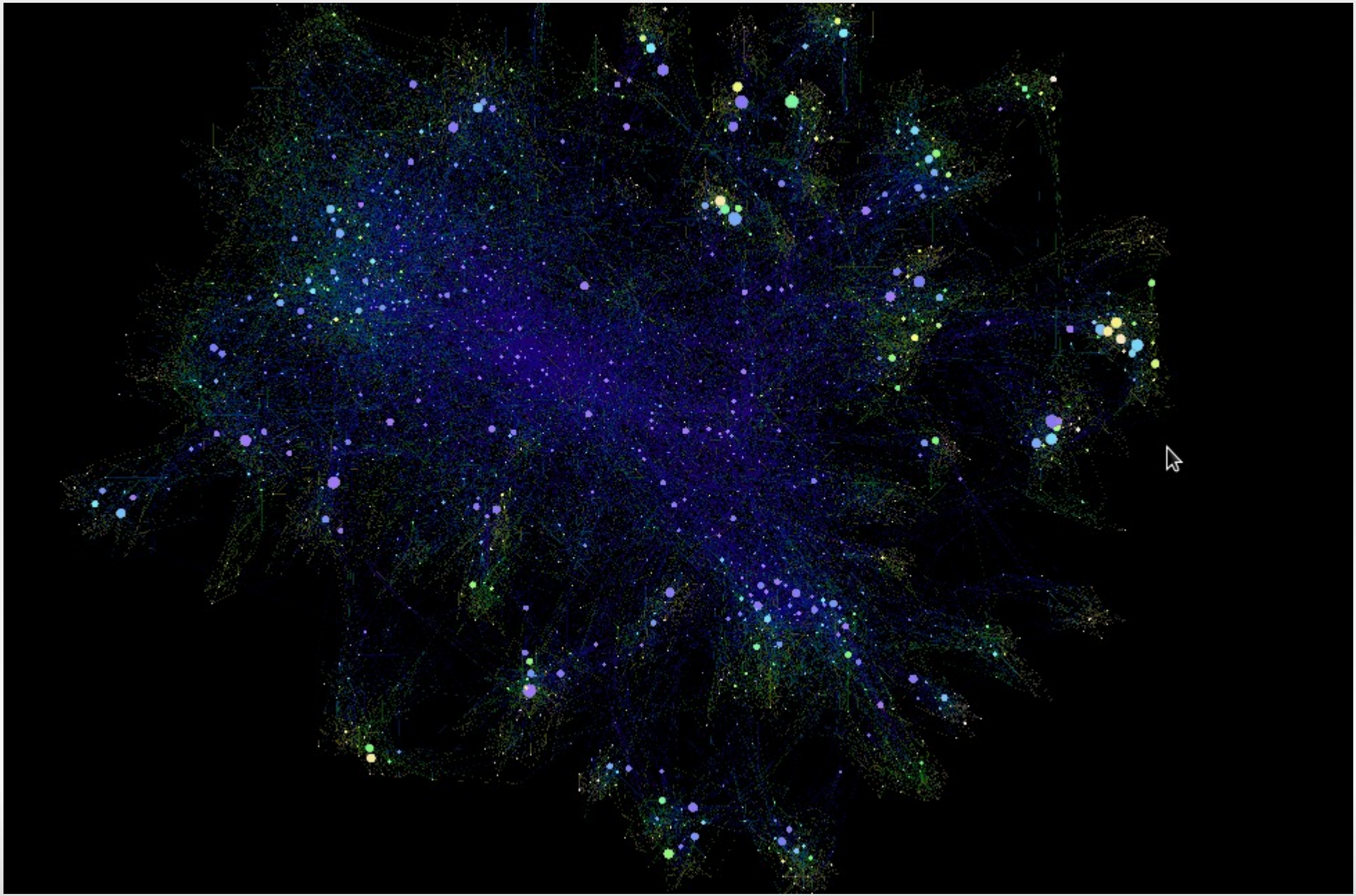
[effectuer__v--commande__nc]

[réaliser__v--équipement__nc]

Croiser un graphe avec un potentiel (1/3)

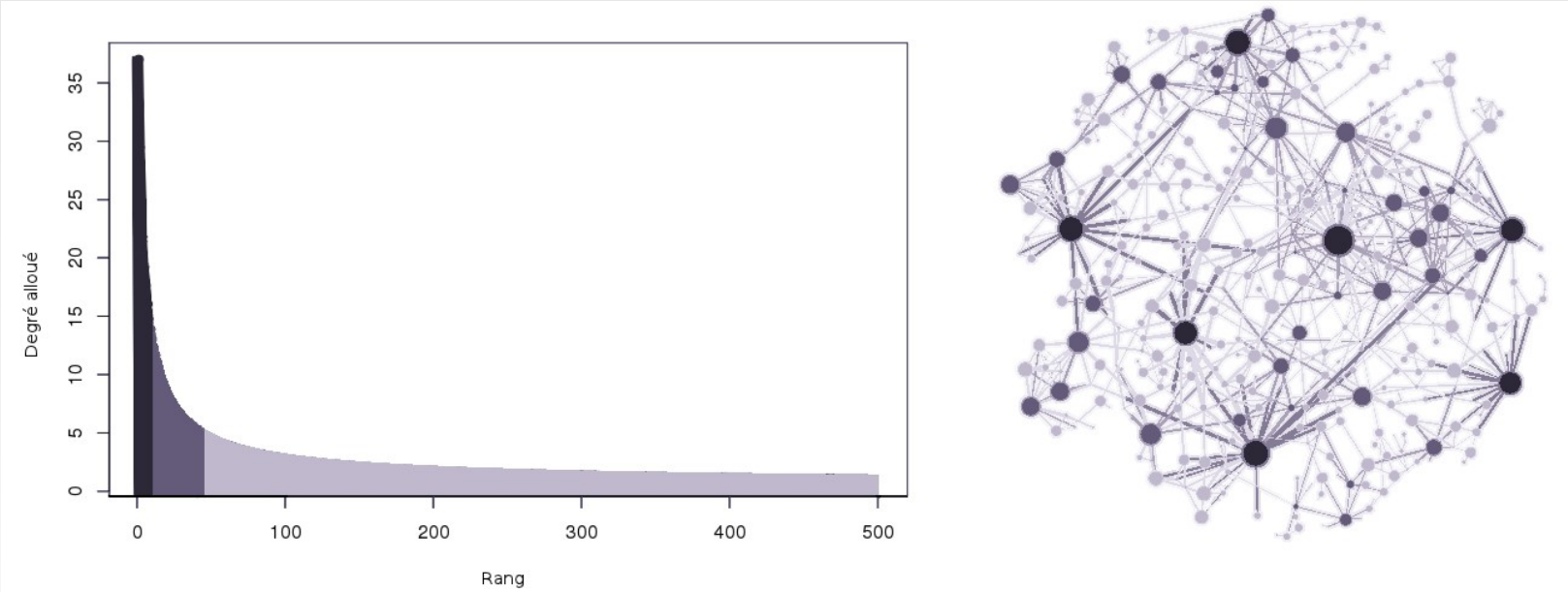
- La phase textométrique donne la fréquence et l'IDF
- L'espace vectoriel donne la 1ère dim de SVD
  - composante factorielle principale u_0
- Le graphe fournit
  - le degré d'un noeud
  - centralité de proximité
  - centralité de « betweenness »
- Heuristiques ad hoc bienvenues ! (ex: prime aux noms de lieux, prime aux verbes, *etc.*)
- Résultat = on peut créer des indicateurs composites de «potentiel», visant à estimer un gradient quelconque sur les noeuds
  - on peut tenter d'approcher toute valeur sémantique graduelle comme l'importance, ou bien la spécificité technique du terme

Croiser un graphe avec un potentiel (2/3)



Croiser un graphe avec un potentiel (3/3)

- Usage 1 : taille des noeuds et étiquettes
- Usage 2 : allocation de degré de voisinage



- autres usages : filtrer un sous-graphe, pondérer des marches aléatoires, *etc.*

Interprétation et visualisation

- Les opérations que l'on peut effectuer sur un format de donnée en graphe nous emmènent au-delà des notions explicatives usuelles en sémantique linguistique classique
- On a par ailleurs toujours deux contraintes très différentes sur les tâches
  - Exigences du modèle d'analyse
  - Exigence de clarté visuelle
- A/R exploratoire toujours possible avec l'espace vectoriel et le corpus : dimensions LSI favorisées par une zone ? types de textes où les mots de la zone apparaissent ?

Conclusions

Visualisation et exploration

- navigation
- découverte
- compréhension

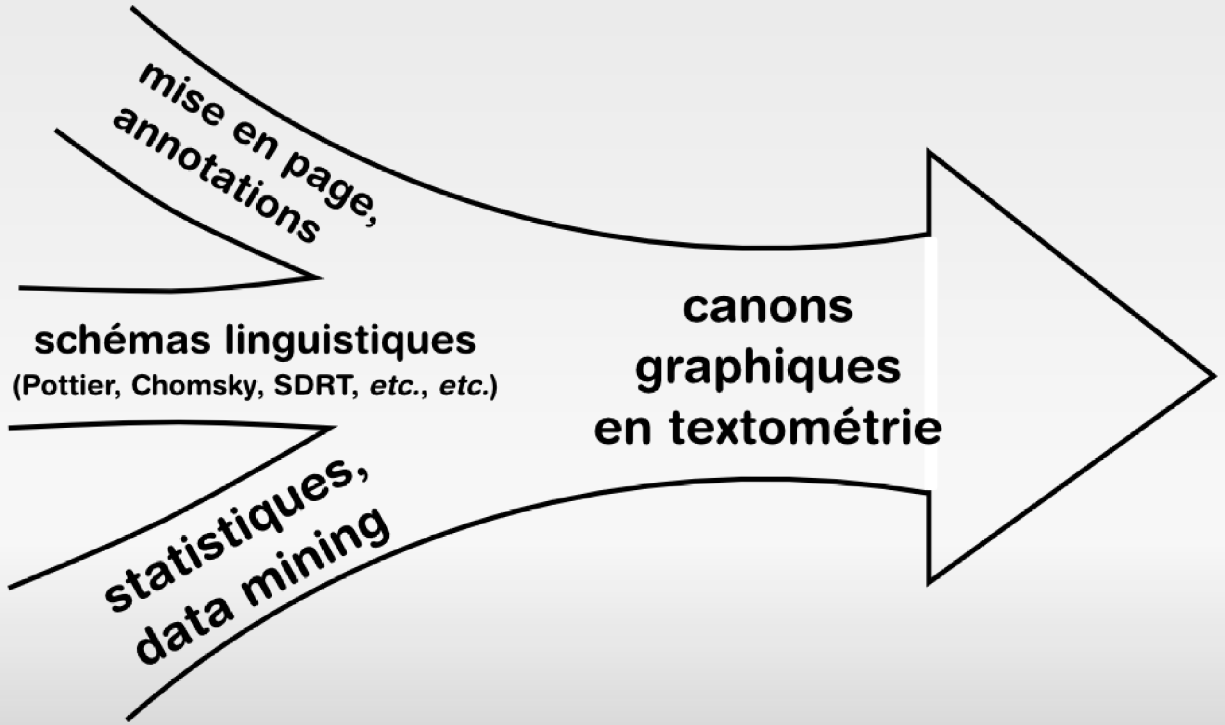
Sérendipité des voisinages sémantiques

- Vocabulaires techniques : la négociation de repères partagés
  - dénomination de concepts, outils, objets, produits visant précision et stabilité
  - une part de sous-entendus (raccourcis elliptiques) refait entrer l'équivoque par la petite porte
    - Ex : technicien réseau => «tu as recruté sur le poste de technicien ?»
    - mais les offres d'emploi sont rarement elliptiques (lectorat large et inconnu)
  - (Mortureux 1995) (Lerat 1995) (Bourigault et Lamé 2002) (Condamines 2006)
  - scénarion de nav. : suggérer des idées de requêtes
- Moteur de recherche
  - un mécanisme de suggestion peut donner corps à la recherche d'emploi tout en ouvrant sur les passerelles inattendues
  - par opposition aux codes comme le ROME qui « mettent les gens dans des cases »

Sérendipité des voisinages sémantiques

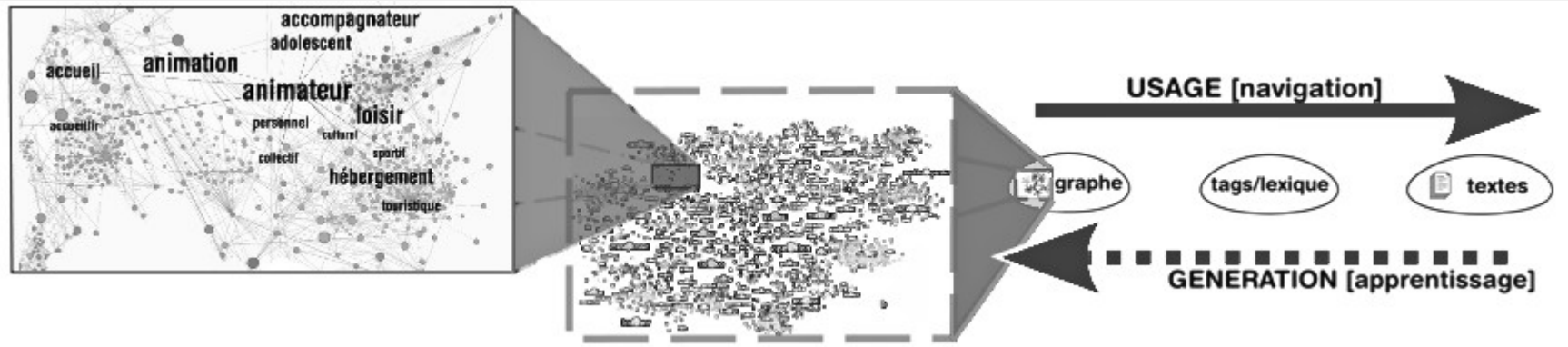
Quelques enseignements méthodo

- texte <=> données <=> visualisation <=> navigation <=> texte
  - « retour au texte »
- Formes de données statistiques ≠ forme des données textuelles
  - faire émerger des quantités et relations
  - visualiser : observer > adopter une perspective sur les données > interpréter
- visualiser : selon des canons graphiques (horizon d'attente)



Visualisation et navigation

- Le visuel comme pré-traitement de la navigation ?
  - une méthode de visualisation correspond à un modèle d'interaction avec les données
  - pré-macher certaines suggestions, relations proposées
    - avec des métadonnées induites
    - donne un aperçu du contenu avant le geste de recherche
  - cartographie des discours (≠concepts)



Pistes pour la linguistique ?

- Un modèle visuel permet d'aller « au concret »
- Symbiose avec l'environnement informatique moderne
  - perspective / geste
  - par exemple
    - les couches d'annotation
    - cf. outil comme Glozz (Widlocher & Mathet 2009)
    - rendent plus concrètes la question des niveaux d'analyse linguistique
- On peut donc avoir une pertinence théorique du visuel

Pistes pour la linguistique ?

- Voisinage => une vision graduelle incluant des contraintes formelles
  - zone + fracture selon contexte = éclairant pour traiter la paraphrase et la polysémie
- Small-world => pseudo-fractalité du lexique
  - pertinent pour les études sur la cognition, théorie du prototype
  - croissance et adaptation du répertoire lexical spécialisé
  - intéressant diagnostic de la dynamique des relations dans un modèle saussurien
  - et utile visuellement !?
- Dimensions SVD et clusters
  - lien avec les traits sémantiques mais en plus « fuzzy »
- Genre textuel et corrélations distributionnelles => attracteurs
  - sens terminologique (postule un concept associé)
  - vs. sens discursif (selon les usages, eg. locution « je vous en prie »)

Comment valider une visualisation ?

- Mais : visuel = intuitif = non scientifique ??
- Problèmes soulevés par (Lebart 2004)
  - caractère « non déterministe » des algorithmes à initialisation aléatoire
  - au pire : des représentations différentes chaque fois
  - souvent : des représentations qui varient à la marge
- indicateurs de confiance ?
  - mincut, variance, effet sur prédictions
  - stabilité de la visualisation (Gambette)

Bibliographie

Baroni et Lenci (2009). *One distributional memory, many semantic spaces*

Bertin (1970). *La graphique*

Fry (2008). *Visualizing Data: Exploring and Explaining Data with the Processing Environment*

Gambette et Véronis (2004). *Visualising a Text with a Tree Cloud.*

Gaume (2004). *Balades aléatoires dans les petits mondes lexicaux*

Grefenstette (1994). *Explorations in automatic thesaurus discovery*

Lafourcade (2011). *Lexique et analyse sémantique de textes : Structures, acquisitions, calculs, et jeux de mots*

Lebart (2004). *Validité des visualisations de données textuelles*

Loth et Rinck (2011). *Les propriétés grammaticales du genre de l'offre d'emploi aux fondements d'une méthode de classement*

Padó et Lapata (2007). *Dependency-based construction of semantic space models*

Ploux et Victorri (1998). *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes*

Rastier (2001). *Arts et sciences du texte*

Sahlgren (2008). *The Word-space Model*

Tutin (2007). *Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques*