La cooccurrence à l'heure du web : graphes bipartites de catégories linguistiques et de sites web

Eglantine Schmitt

Doctorante en épistémologie, UTC, EA Costech

Enjeux

- Techniques : « univers sémantiques »
- SHS
- Marketing

Analyse linguistique « assistée par ordinateur »

- Premiers outils de CAQDAS dans les années 70-80
- Textométrie vs linguistique de corpus (Pincemin 2012)
- Différents modes d'approche (Jenny 1997) :
  - lexicométrie, socio-sémantique, réseaux de mots associés, analyse propositionnelle et prédicative, ingénierie textuelle, complément du traitement d'enquêtes (post-codification), systèmes experts
- « objectivation partielle de l'analyse » (Brugidou 2001)
- « les logiciels ne sont pas de simples instruments d'administration de la preuve, mais constituent des ressources mobilisables - parmi d'autres - pour tester des lectures interprétatives, enrichir des théorisations provisoires » (Demazière 2005)
- Culture de l'interprétation du texte – approches ascendantes

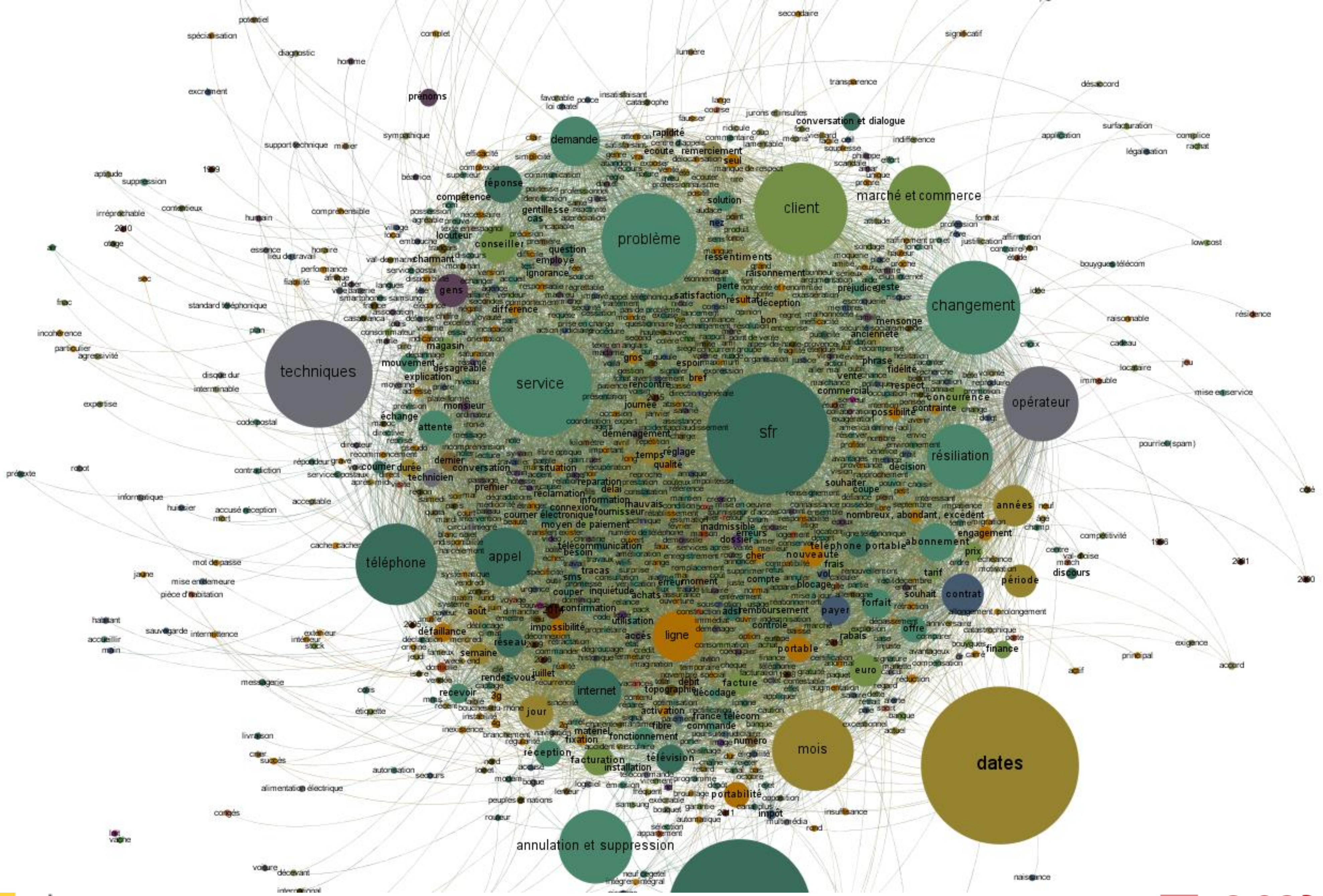
Ubiq (Proxem)

- Détection d'entités nommées
  - groupes lexicaux + activateurs ou inhibiteurs
  - taxonomie
  - score

Cooccurrence linguistique

- « Le passage de l'analyse lexicale à l'analyse thématique conduit de signes non interprétés à des unités sémantiques qui résultent d'un parcours interprétatif. Il se concrétise par le passage des cooccurrents aux corrélats. » (Rastier 2001)
- « Ce que produit le calcul textométrique, ce sont donc des cooccurrents, au plan des signifiants ; et ce qui est visé, c'est l'obtention de corrélats, au plan des signifiés. On passe des premiers aux seconds par une interprétation qui reconnaît la présence d'un trait sémantique commun entre le ou les mots servant d'amorce à la recherche, et le cooccurrent alors qualifiable comme corrélat. » (Pincemin 2012)

- Question de la distance
- Question de l'unité (mot, ngram, terme, collocation)


techniques
service
problème
client
sfr
changement
opérateur
résiliation
marché et commerce
demande
réponse
téléphone
appel
ligne
internet
mois
dates
annulation et suppression
années
contrat
période
payer
forfait
abonnement
attente
gens
prénoms
conseiller
conversation et dialogue
jour

Analyse de réseaux sociaux (SNA)

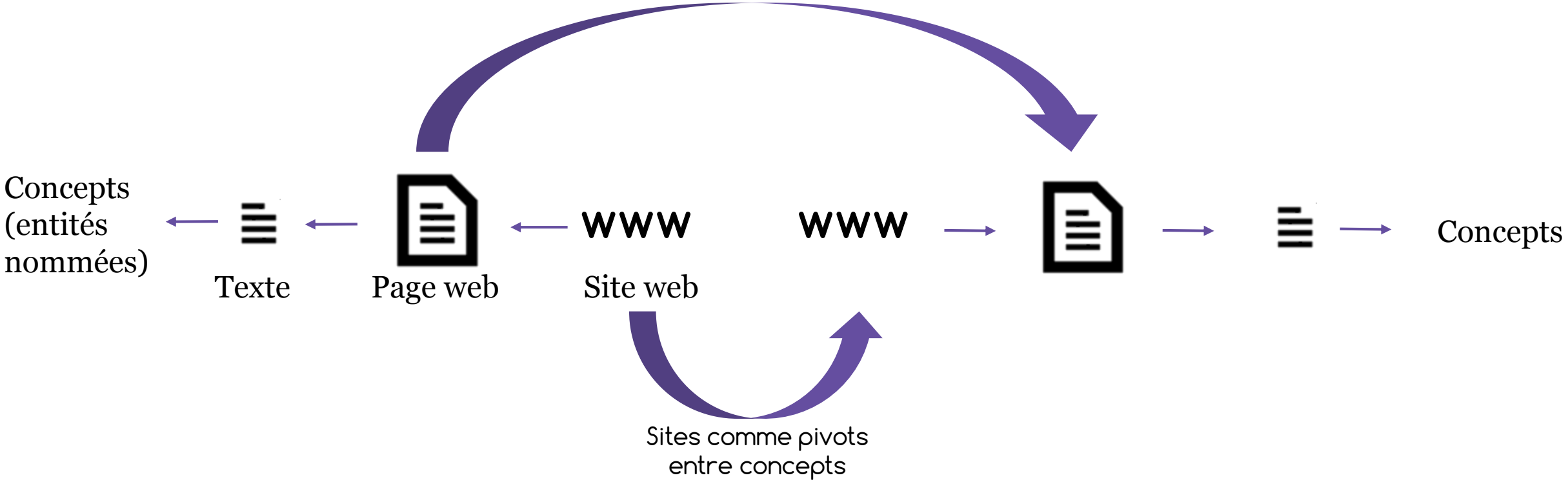
- Théorie des graphes
- Nouvelle science des réseaux (Granovetter, Watts)
- Réseaux invariants d'échelle (Barabási et Albert), degré, attachement préférentiel, intermédiarité, centralité, hubs, trous structuraux, clique
- « ensemble de méthodes sans théorie explicite » (Mercklé 2011)
- Web comme réseau (Castells, Page) physique et social
- Gephi
  « Les logiciels de graphes sont très puissants pour interpréter des structures qui autrement nous resteraient cachées, mais ils ne valent que par les hypothèses qu'ils permettent de formuler et qu'il faut éprouver par ailleurs. » (Jacomy et al. 2007)
- « Il est de la responsabilité des sciences sociales de ne pas laisser aux département de R&D des grands groupes ou aux entreprises de marketing le soin exclusif de produire une connaissance de plus en plus systématique sur les interdépendances et les processus sociaux. » (Lazega 2007)
- Graphes généralement unipartites

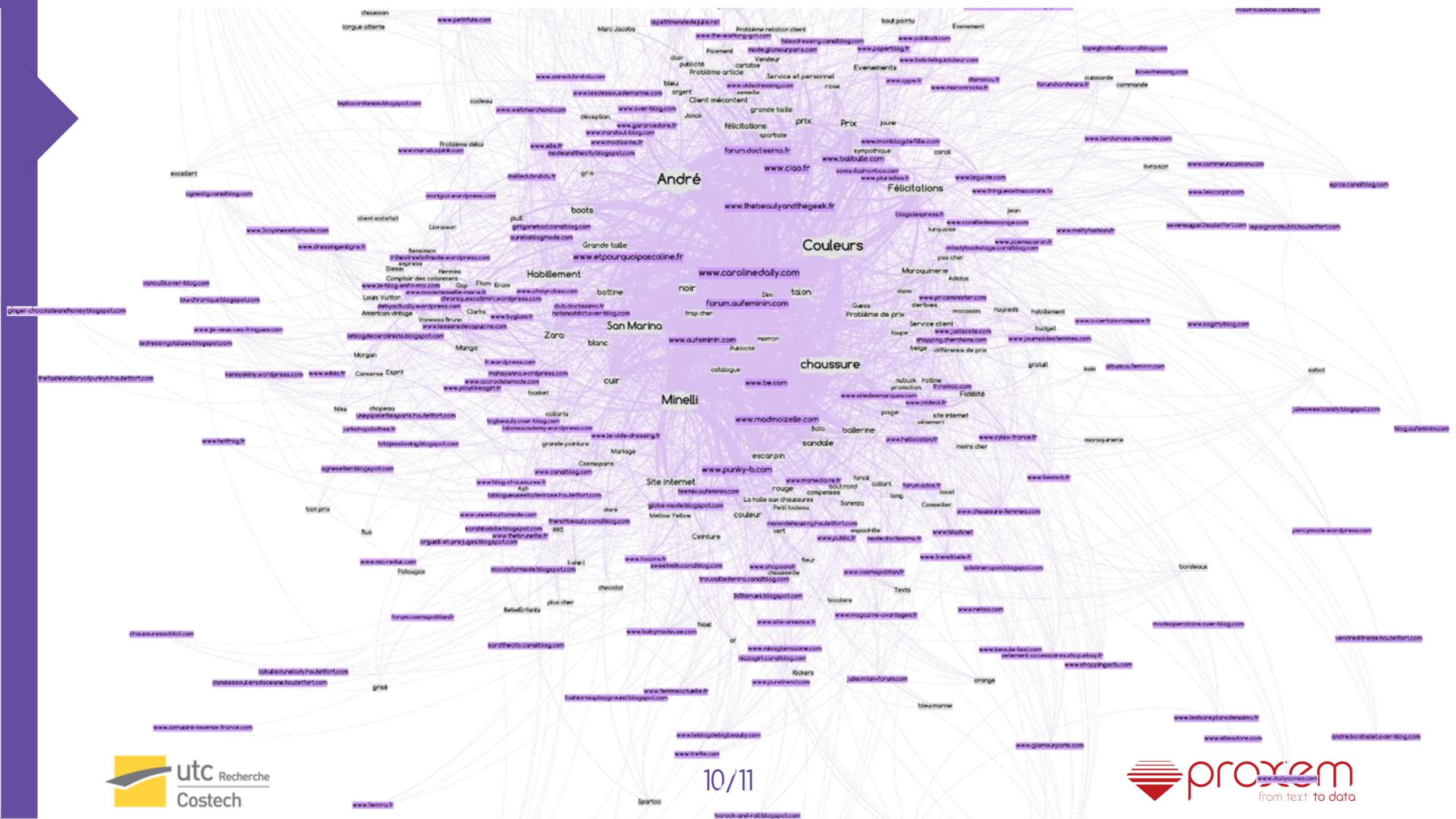
Collecte web

- Aubaine pour la linguistique de corpus (Tanguy 2013) :
  - Corpus de textes largement disponibles
  - Phénomènes linguistiques rares
  - *Web as corpus* vs *web for corpus*
- « it is not at all clear that data derived from social media communications provide social researchers with an alternative to ethnographic immersion » (Edwards et al. 2013)

- Proxem Discovery
  - Requêtes sur un moteur de recherche
  - Combinaison de grappes de ressources linguistiques
  - Crawling, scraping

Constitution du graphe
Cooccurrence indirecte





utc Recherche Costech

proxem from text to data

Pistes et interrogations

- Arbitraires
  - Délimitation du corpus
  - Choix des concepts
  - Algorithme de spatialisation
  - Simplification du graphe
  - Interprétation
- Pistes
  - Partir d'un corpus moins homogène dans ses thématiques
  - Documents plus courts, plus propres
  - Sites mono-thématiques
  - Preuve par le corpus

Merci.

Eglantine Schmitt

eglantine.schmitt@utc.fr

Twitter : @egschmitt

http://bigdata.hypotheses.org