Analyser les pratiques scientifiques et éditoriales avec des outils scientométriques

❊

Philippe Dessus
LSE (EA 602), Univ. Grenoble Alpes

❊

Séminaire EducMap
26 juin 2014, Ifé-ENS, Lyon

Préambule

- Je ne suis pas spécialiste de scientométrie
- Mais ai un intérêt pour l'appliquer *via*
  - les sciences de l'éducation
  - les aspects évaluatifs dans la recherche
  - le traitement automatique du langage

Propos et plan

- Des techniques scientométriques peuvent être utilisées pour mettre au jour des réseaux de collaborations en sciences humaines

- Quelles questions et problèmes ? Quelles solutions ?

A. Vue d'ensemble des protagonistes et du flux de l'information

B. 5 questions posées + premières réponses

C. Démo d'un logiciel lié à certaines de ces réponses

- Les références citées sont disponibles à http://www.citeulike.org/user/pdessus/tag/educmap

A. Une vue d'ensemble

La scientométrie, une nouvelle religion?

- La scientométrie devient courante pour remplir de nombreux buts

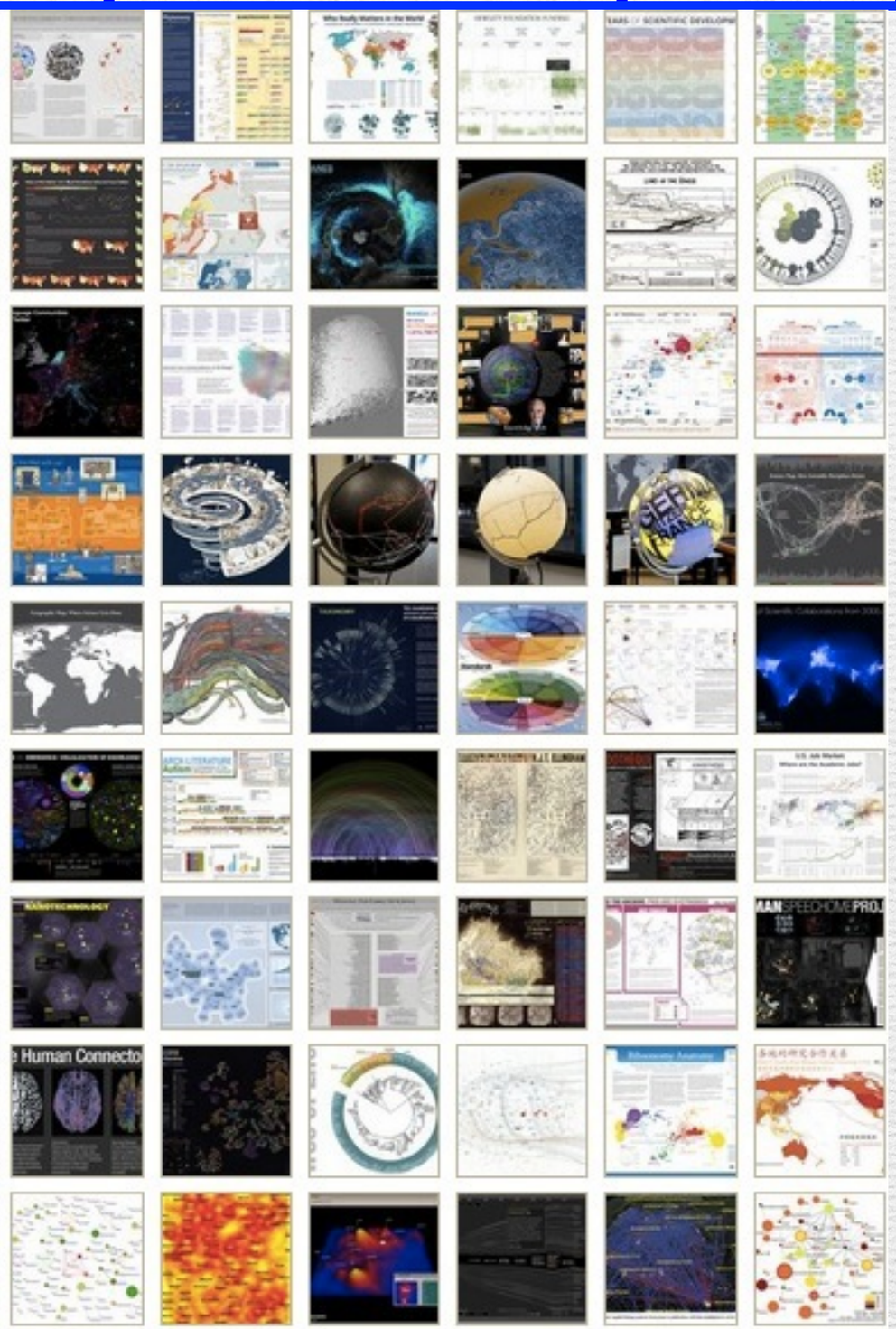
  - **politiques** (évaluation externe, prospective), **académiques** (évaluation interne), **de recherche** (analyse de domaines)

- Méthodes de plus en plus sophistiquées (méthodes avancées de TAL et fouilles de données) et utilisées

- Utilité pour des recherches **internationales** et **inter-disciplines** (Noyons 04)

Champ florissant

(http://scimaps.org/)



Comment ça marche ?

Pratiques

Communauté académique

Langage

Pratiques

Editeurs

Quand?

Où ?

Quoi ?

Avec qui ?

Classements

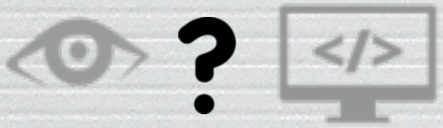
Réseaux

Cartes

D'après Noyons (2004, p. 240) ; Börner & Polley (2014)

Mais…

- Quels problèmes ces méthodes résolvent-elles ?
- Les informations produites sont-elles valides ?
- Ne produit-on pas du *chartjunk* (Tufte, 2007) ?
- Les informations produites sont-elles utiles ? À qui ? (éditeurs, communautés, politiques)



B. 5 Q/R

Q1. Les cartes sont-elles valides?

- La scientométrie (et la fouille de données) permet d'apporter peut-être trop souvent des réponses précises à des questions vagues ; cela peut être assez souvent un problème (Marcus & Davis, 2014)

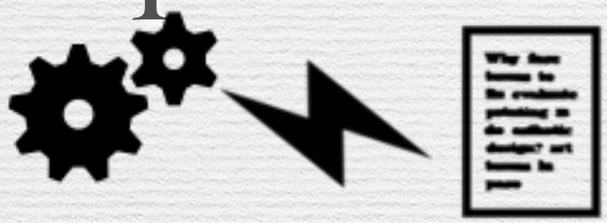
- Le projet Pantheon (MIT Media Lab) classe très précisément des hommes et femmes célèbres (mais Nostradamus est le 20e écrivain)… [http://pantheon.media.mit.edu/]

R1. Utiliser les jugements d'experts

- Bon rappel, mais faible précision. Il serait donc intéressant de confronter les résultats à des jugements d'experts du domaine cartographié

- Les différences entre avis d'experts et cartes permettrait de questionner à la fois le traitement et les représentations, et mettre en valeur ce qui a du sens pour les experts

Q2. Comment améliorer les processus de traitement du langage?



- En amont du processus de construction de la carte se réalisent des processus de traitement statistique de texte qui sont importants et conditionnent la validité de cette construction

- Beaucoup de traitements se fondent sur des analyses de mots-clés ou de résumés, qui peuvent être insuffisants pour capturer toute la finesse de productions scientifiques

R2 - Méthodes d'analyse factorielle textuelle

- Il existe notamment des méthodes d'analyse factorielle, comme *Latent Semantic Analysis* (Landauer & Dumais, 1990) capturant des relations sémantique d'ordre 2 entre les mots-clés, donc arrivent à déterminer que 2 mots-clés sont similaires même s'ils n'apparaissent pas dans les mêmes contextes

- LDA (*Latent Dirichlet Allocation,* Blei, 2003, 2012), regroupe de plus les mots analysés en thèmes (Günneman et al., 2013)

Comparer les approches

Occ. mots

12	enseignant
12	élèves
11	interactions
11	qualité
9	observation
7	méthodes
6	classe
5	**analyse**
5	**question**
4	**article**
4	**CLASS**
4	compétences
4	**dimensions**
4	**développement**
4	enseignants

ReaderBench (LSA+LDA+Wordnet)

élève	8,9
enseignant	8,9
classe	6,5
interaction	5,8
qualité	5,8
méthode	4,8
relation	3,7
compétence	3,3
question	3,2
observation	3,1
outil	2,9
lier	2,6
apprentissage	2,6
utiliser	2,5
approche	2,4

Q3*a* - Peut-on tenir compte des différences de pratiques académiques…

- Les pratiques d'écriture et les stratégies de co-autorat des différentes communautés académiques sont très diverses (et varient même intra-communautés) (voir Sword, 2012). Ne pas en tenir compte est sans doute une erreur

- Sans compter les multiples à-côté, souvent difficiles à détecter par les outils scientométriques, et même parfois par les humains :

  - (auto-)plagiat, tricherie, canulars
  - faible réaction à la rétractation d'articles (Davis, 2012)

Q3*b* - … ou éditoriales ?

- Les éditeurs ont, de leur côté, des pratiques de publication qui interagissent nécessairement avec l'analyse, car la scientométrie est devenue une source de procédés d'évaluation, amenant des pratiques

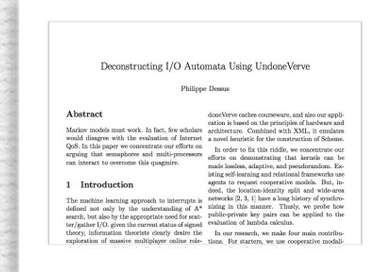
- soit positives (déterminer le niveau des journaux)

- soit négatives (le détournement de journaux, ou des journaux à compte d'auteur), voir http://scholarlyoa.com/tag/journal-hijacking/ ; http://scholarlyoa.com/2014/06/12/serbian-journal-accepts-paper-in-24-hours-with-no-peer-review-demands-eur-1785/

R3a - La textométrie pour aider—et troubler—le système

- Des outils d'analyse textométrique sont donc utiles pour analyser les caractéristiques des textes produits (p. ex., complexité syntaxique, nombre de mots abstraits, pronoms personnels, etc.), et aussi les textes à problèmes

- Voir, par exemple :
  - Déjà Vu (auto-plagiat) : http://dejavu.vbi.vt.edu/dejavu/
  - SciGen (canulars) : http://pdos.csail.mit.edu/scigen/


Deconstructing I/O Automata Using UndoneVerve

Philippe Dussa

Abstract

Markov models must work. In fact, few scholars would disagree with the evaluation of Internet QoS. In this paper we concentrate our efforts on arguing that semaphores and multi-processors can interact to overcome this quagmire.

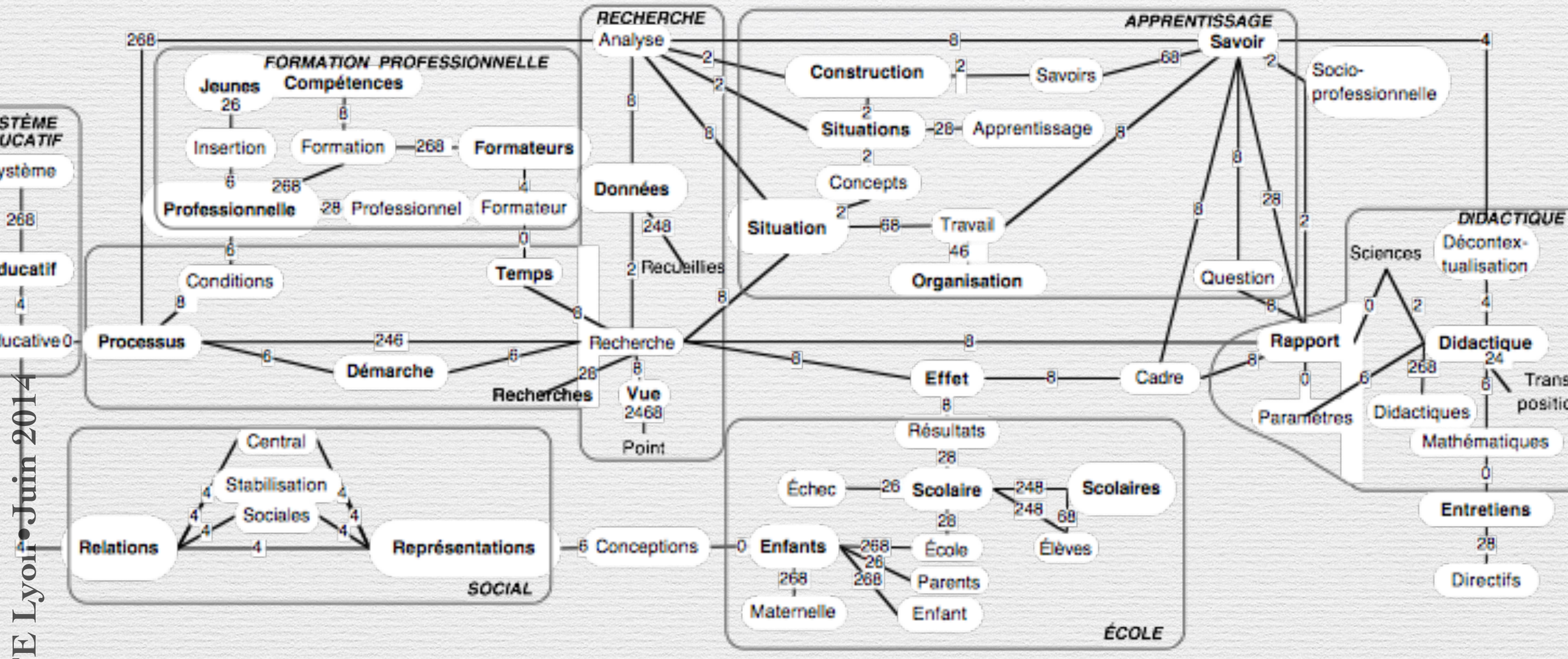
1 Introduction

The machine learning approach to interrupts is defined not only by the understanding of A* search, but also by the appropriate need for scatter/gather I/O. given the current status of signed theory, information theorists clearly desire the exploration of massive multiplayer online role-doneVerve caches courseware, and also our application is based on the principles of hardware and architecture. Combined with XML, it emulates a novel heuristic for the construction of Scheme.

In order to fix this riddle, we concentrate our efforts on demonstrating that kernels can be made lossless, adaptive, and pseudorandom. Existing self-learning and relational frameworks use agents to request cooperative models. But, indeed, the location-identity split and wide-area networks [2, 3, 1] have a long history of synchronizing in this manner. Thusly, we probe how public-private key pairs can be applied to the evaluation of lambda calculus.

In our research, we make four main contributions. For starters, we use cooperative modali-

R3b - Observer les dynamiques des proximités entre éditions d'un colloque (Dessus 1999)



Editions 1992 à 1998 des Biennales de l'éducation et de la formation

Q4 - Comment les visualisations sont-elles interprétées?

- Il est souvent tenu pour acquis que les cartes réalisées sont lisibles. Mais le public lisant de telles cartes peut en avoir des lectures variables : les politiques et les chercheurs ont des lectures tout à fait différentes des cartes (Noyons, 2001)

- Les buts de conception de telles cartes sont multiples et il convient de les préciser avant leur création : politique prospective, bilan, scientométrie

R4. Recherche sur la *graphicacy*

- Les processus cognitifs engagés dans la lecture de telles cartes devraient être étudiés de plus près

- S'il est aisé de construire des cartes multicolores et avec de nombreuses informations, il n'en reste pas moins qu'il faut qu'elles soient comprises de leur public, ce qui est, la plupart du temps, assumé plus que vérifié

- Quels classement et opinion sur des cartes produites (Ashraf, 2014; de Vries & Ashraf, sous presse) ?

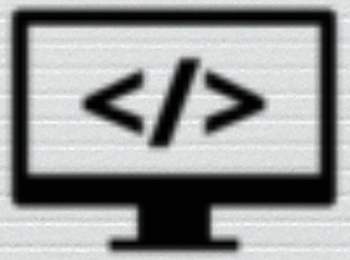
Q5. Cartes descriptives vs. normatives

- On oppose classiquement les analyses *prescriptives* (calcul de facteur d'impact, classements) aux études *descriptives* (souvent fondées sur des analyses en réseaux sociaux), en montrant que ces dernières sont moins évaluatives, et donc plus descriptives (Pansu, Dubois & Beauvois, 2013)

R5. Les cartes, ça sert, d'abord, à faire la guerre (Lacoste 1976)

- Mais même les cartes dites « collaboratives » peuvent avoir un aspect évaluatif marqué, p. ex., montrer les chercheurs isolés, les thèmes de recherche non « tendance », etc.

- Donc, que veut-on *vraiment* mettre au jour ? Quelles conséquences (politiques, individuelles, sociales) cela aura-t-il ?

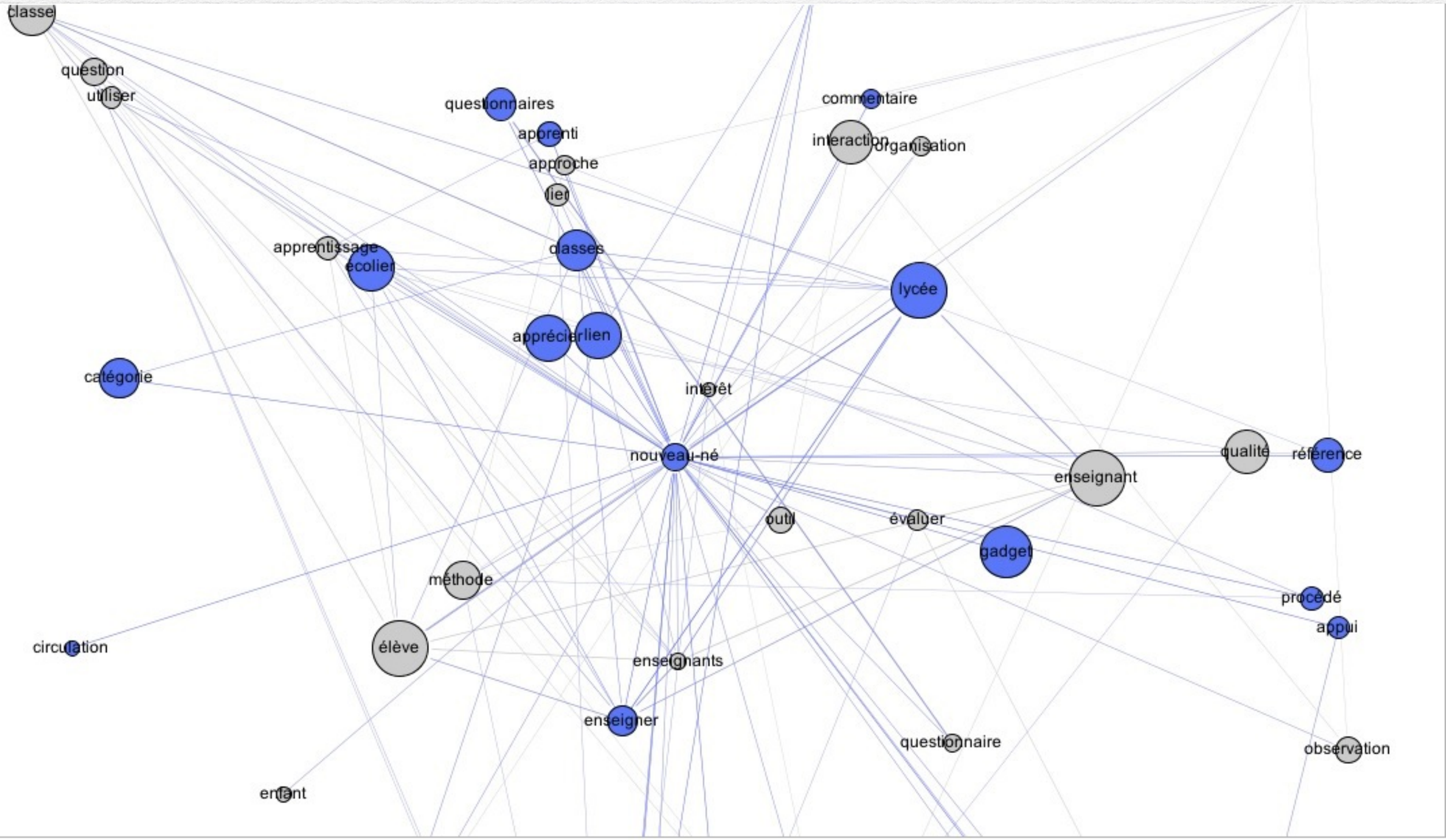


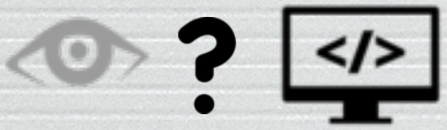
C. Une démo

conçu en collaboration avec l'univ. "Polytehnica", Bucarest, Roumanie

Une démo de ReaderBench (Dascalu *et al.*, 2013)

- Un outil pour déterminer
  - les mots-clés de documents (réels et inférés)
  - leur niveau de complexité textuelle selon de nombreux paramètres (nécessite un étalonnage)
  - les contributions relatives de documents par rapport au thème général, mais aussi les termes « échos » (posts de forums, articles de *workshops*)


classe
question
utiliser
questionnaires
apprenti
approche
lier
commentaire
interaction
organisation
apprentissage
écolier
classes
lycée
apprécier
lien
catégorie
intérêt
nouveau-né
qualité
référence
enseignant
outil
évaluer
gadget
méthode
procédé
appui
circulation
élève
enseignants
enseigner
questionnaire
observation
enfant
Topics
Inferred Concepts



Pour résumer

Pour résumer, une approche interdisciplinaire

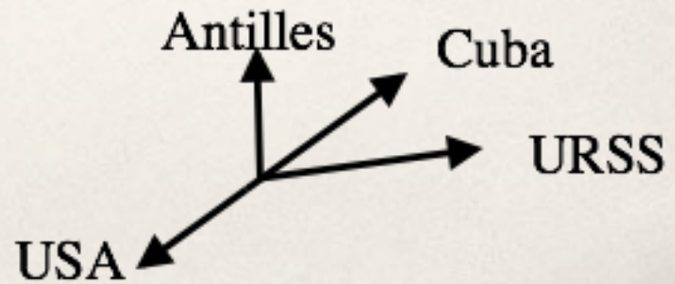
- un réseau d'experts pour évaluer la validité des cartes
- des *benchmarks* pour évaluer les outils textométriques (informatique)
- mieux caractériser les pratiques académiques et éditoriales (sociologie des sciences)
- mieux comprendre comment les cartes sont lues (cognition)
- prédire l'appropriation politique de ces outils

Merci de votre attention !

- philippe.dessus@upmf-grenoble.fr
- Présentation disponible à http://webcom.upmf-grenoble.fr/sciedu/pdessus/
- Les références citées sont disponibles à http://www.citeulike.org/user/pdessus/tag/educmap
- Merci à Pascal Pansu pour ses commentaires d'une version précédente de cette présentation

Comment marche l'Analyse sémantique latente ? (1/2)

- LSA méthode d'analyse statistique de grands corpus textuels (type d'analyse factorielle). Part du principe que
  - deux mots ont un sens similaire s'ils apparaissent dans des contextes similaires
  - deux contextes (paragraphes, phrases) ont un sens similaire (contiennent des informations similaires) s'ils contiennent des mots de sens similaire

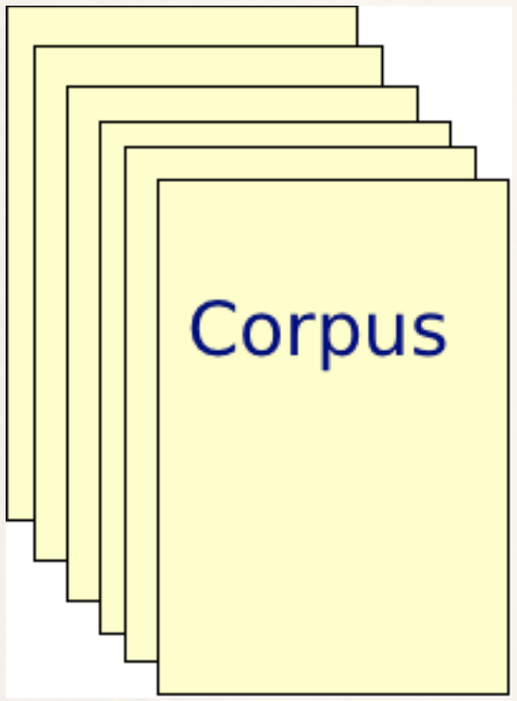


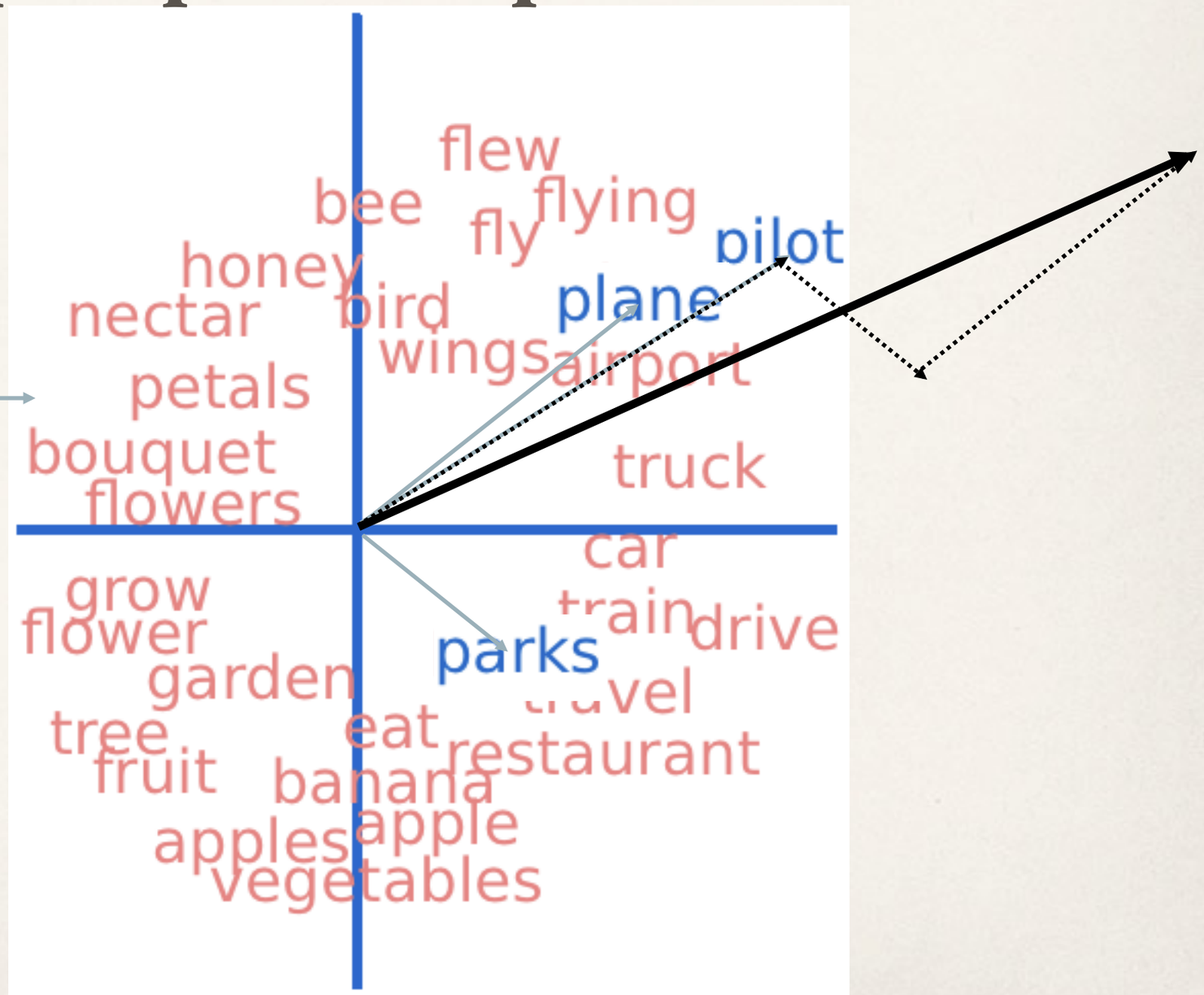
Comment marche LSA (2) ?

- Rend compte de la structure *latente* du sens des mots
- Importance de la composition du corpus pour «entraîner» LSA
- Pas seulement fondée sur la co-occurrence de mots : 2 mots jamais co-occurrents peuvent avoir une grande proximité
- Approche «paquets de mots», pas de prise en compte de la ponctuation, de la syntaxe, du style, etc.
- Pas de prise en compte de «mots outils» (déterminants, prépositions)

Une phrase est représentée par la somme des vecteurs des mots qui la composent [Lemaire & Denhière '05]

"The pilot parks the plane"





LDA : Allocation de Dirichlet Latente (Blei 2012)

- Package *Mallet* (http://mallet.cs.umass.edu/)

- Méthode probabiliste, analyse en arrière plan de grands corpus (*TASA*/*Le Monde*) de plusieurs millions de mots. Chaque document est un « mix » de thèmes (*Topics*) de probabilité décroissante. Chaque mot a une probabilité d'occurrence par *Topic* décroissante.

- Un exemple de trois *Topics* :

```
père(1750.0) famille(1267.0) mère(1221.0) fils(1139.0) enfant(1088.0)
  jeune(771.0) grand(644.0) parent(589.0) ...

guerre(1074.0) armée(381.0) soldat(294.0) résistance(227.0) combat(214.0)
  général(211.0) officier(187.0) ...

hôpital(975.0) médecin(774.0) médical(549.0) service(407.0) santé(405.0)
  malade(363.0) docteur(306.0) ...
```

LDA (Blei 2012 p. 78)