

# ***Présentation de***

**READER  
BENCH**



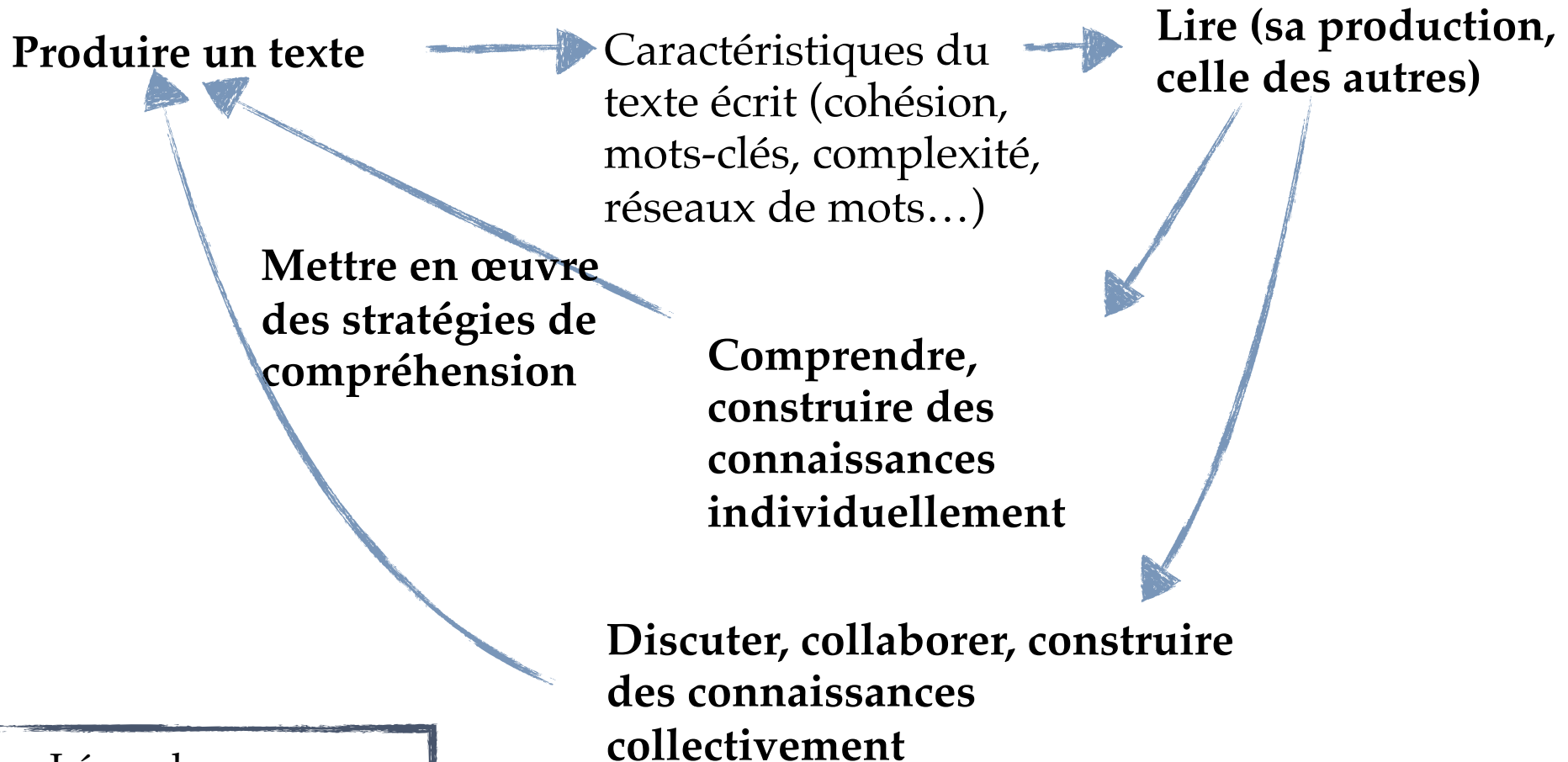
**Séminaire du LSE  
19 novembre 2015**

**Philippe Dessus**

# INTRODUCTION

# Liens entre activités liées à la lecture

► pour comprendre => stratégies de compréhension



Légende :  
**Actions/Paramètres**

## 4 La cohésion comme élément de mesure de nombreux phénomènes

- La cohésion (recouvrement sémantique de 2 propositions) peut être utilisée pour rendre compte des écarts entre mots, propositions, paragraphes, etc.
- Entre souvent dans le calcul de nombreux paramètres des systèmes utilisés : cohérence inter-phrases, inter-paragraphes, complexité
- De nombreux travaux montrent que des techniques d'analyse sémantique comme LSA capturent de manière assez satisfaisante la cohésion (Bellissens *et al.* 2010)
- Validation : 10 histoires dont liens interparag. évalués par étudiants ; lien jugement humain vs. LSA/LDA/WordNet :  $r = .54$



**Un atelier multilingue pour évaluer diverses informations à propos de lecture et compréhension (en *input*, en *output*), individuellement et collaborativement**

## 4 grands types de fonctionnalités de *ReaderBench*

1. Identification de mots-clés, réseau de mots, mots inférés
2. Analyse de la complexité textuelle
3. Identification de stratégies de compréhension en lecture
4. Analyse de la participation/collaboration dans les forums

# 1. Identification de mots-clés (Dascalu 2013)

1. Classe gramm. du mot
2. Score d'un mot =  $tf^*idf$  (fréq. locale pondérée par fréq. inv. globale) + similarité mot avec texte lu + similarité moyenne mot avec chaque phrase + similarité moyenne mot avec chaque paragraphe
3. Présenter les mots par score décroissant et par classe grammaticale (en option)

Cognitis enim pilatorum caesorumque funeribus nemo deinde ad has stationes appulit navem, sed ut Scironis praerupta letalia declinantes litoribus Cypriis contigui navigabant, quae Isauriae scopulis sunt controversa.

Oportunum est, ut arbitror, explanare nunc causam, quae ad exitium praecipitem Aginatum inpulit iam inde a priscis maioribus nobilem, ut locuta est pertinacior fama. nec enim super hoc ulla documentorum rata est fides.

Paphius quin etiam et Cornelius senatores, ambo venenorum artibus pravis se polluisse confessi, eodem pronuntiante Maximino sunt interfecti. pari sorte etiam procurator monetae extinctus est. *Sericum enim et Asbolium supra dictos, quoniam cum hortaretur passim nominare, quos vellent, adiecta religione firmarat, nullum igni vel ferro se puniri iussurum, plumbi validis ictibus interemit. et post hoc flammis Campensem aruspice dedit, in negotio eius nullo sacramento constrictus.*

Et olim licet otiosae sint tribus pacataeque centuriae et nulla suffragiorum certamina set Pompiliani redierit securitas temporis, per omnes tamen quotquot sunt partes terrarum, ut domina suscipitur et regina et ubique patrum reverenda cum auctoritate canities populi que Romani nomen circumspectum et verecundum.

# 1. Copie d'écran Analyse du texte

ReaderBench - Document Visualization

**Title:** Subject departments as professional communities?

Source: LSE    URI:    Sentiment polarity:

**Contents**

a growing body of literature suggests that when schools become professional communities there are expected benefits in terms of teacher learning, school improvement and student achievement. in this article the concept of professional communities is examined for certain subject departments in dutch secondary schools. the authors report on research into the extent to which mathematics departments operate as professional communities. at the same time, it was investigated whether the level at which departments operate as professional communities is related to student achievement. the results indicate that departments are cohesive bodies regulating teacher behaviour in several respects. however, as professional communities they do not focus on improving the quality of their teachers and instruction. some characteristics of professional communities prove to be beneficial for student achievement, while others are not. the authors offer recommendations on how departments can develop into more professionally organised communities. [31.949]

**Topics**

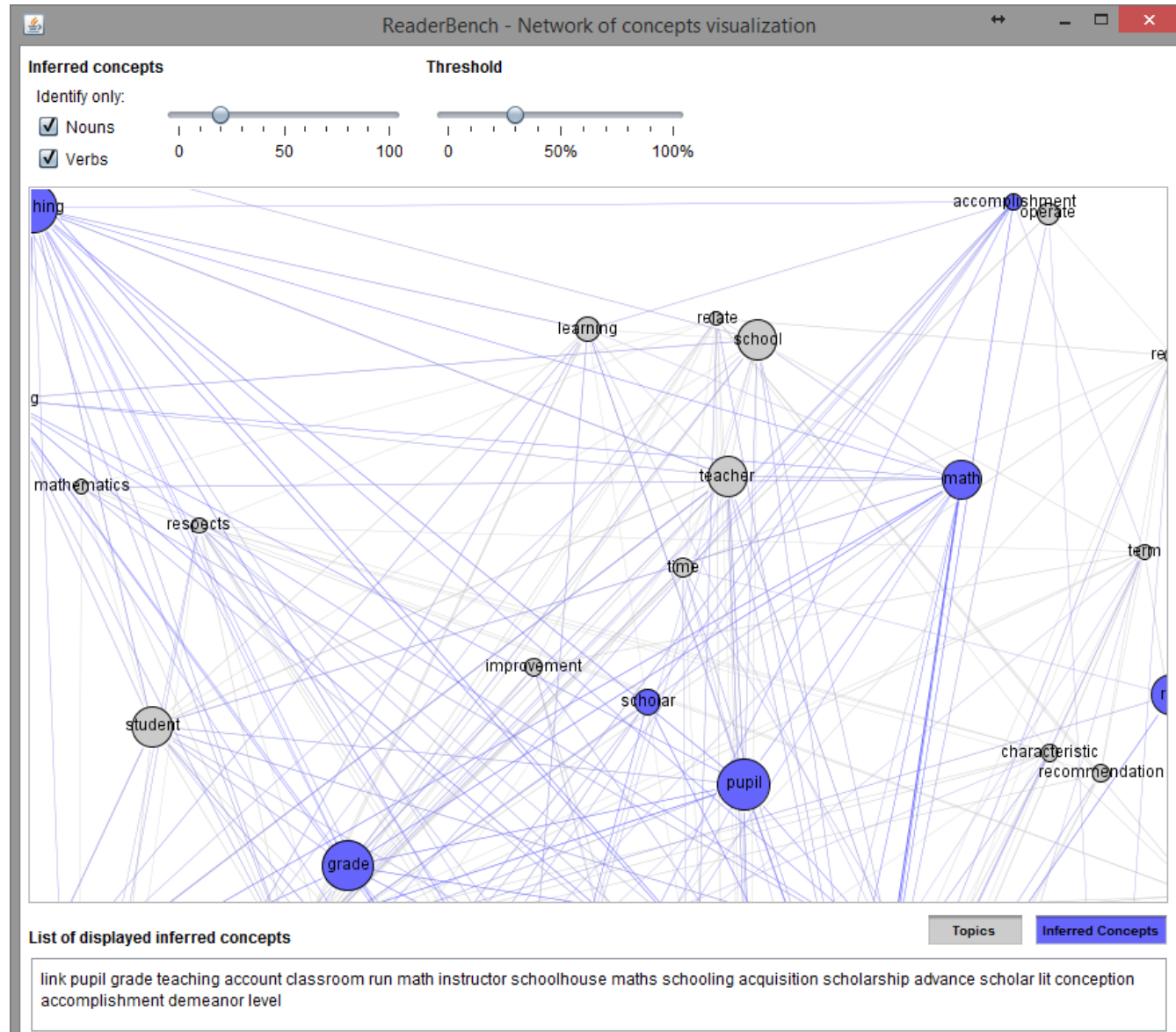
Filter only:

☒ Nouns    ☒ Verbs    0    25    50

| Topics         | Relevance |
|----------------|-----------|
| community      | 7.05      |
| student        | 3.9       |
| school         | 3.86      |
| teacher        | 3.84      |
| department     | 3.52      |
| achievement    | 2.74      |
| body           | 2.35      |
| learning       | 1.73      |
| author         | 1.69      |
| operate        | 1.43      |
| literature     | 1.4       |
| time           | 1.23      |
| recommendation | 1.11      |
| improvement    | 1.08      |
| instruction    | 1.05      |
| characteristic | 1.03      |
| extent         | 0.96      |
| respects       | 0.95      |

Advanced View    Visualize Multi-Layered Cohesion Graph    Select Voices    Display Voice Inter-animation    **Generate network of concepts**

## 2. Copie d'écran - Réseau de mots



## 2. Complexité textuelle

- Partiellement en français pour l'instant (école élémentaire, nécessité de disposer d'une grande base de textes pré-calibrés pour entraîner *ReaderBench*)
- But : Essayer de quantifier, sur un nombre de facteurs important, des indices de complexité de textes donnés à lire. Pouvoir aisément ajouter de nouveaux facteurs ; multilingue (fr, eng., es, it, bientôt ro)
- Nécessité de combiner plusieurs niveaux d'évaluation (lexical, morphosyntaxique, sémantique), car chaque niveau en lui-même n'est pas suffisamment relié à la complexité [Bailin & Grafstein 01]

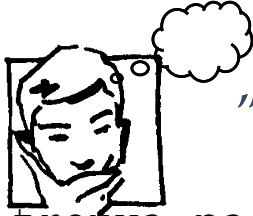
## 2.<sup>11</sup> Quelques niveaux de complexité (Dascalu *et al.* 2012; Dascalu *et al.* 2015)

- De surface
  - ▶ Mesure de lisibilité
  - ▶ Entropie (diversité)
  - ▶ Complexité des mots (longueur moyenne des mots, etc.)
- Morphologie et syntaxe
  - ▶ Mesure intégrée de complexité lexicale et syntaxique (Schultze *et al.* 2010)
  - ▶ Profondeur et taille de l'arbre d'analyse (*parsing tree*)
- Sémantique
  - ▶ Cohésion (proximité sémantique interphrases) avec LSA ou LDA
  - ▶ Chaînes lexicales (mots sémantiquement reliés), montrent l'importance d'un concept
- La technique de classification *Support Vector Machine* permet de classifier des documents après «entraînement» avec un grand nombre de doc. pré-classifiés

### 3. Auto-explication de stratégies [Projet DEVCOMP]

[Dascalu *et al.* 13 ; Bianco *et al.*, 13]

- ✧ Le soleil brillait de toutes ses forces et les dieux qui vivaient sur la Terre trouvaient qu'il faisait trop chaud. Les dieux racontèrent à Indra qu'ils avaient vu le grand serpent Ati étirer sa tête jusque dans le ciel pour avaler les nuages qui passaient. Ati avait si soif qu'il buvait toute l'eau des nuages : voilà pourquoi il ne tombait plus une goutte de pluie.



"Qu'as-tu compris de ce passage ?"

- ✧ Indra ne trouva pas cela drôle du tout. À cause de ce glouton, la Terre entière mourait de soif !

Etc...

### 3. Heuristiques d'identification des stratégies

- **Heuristiques d'identification des stratégies**
  - ▶ **Causalité** : mise en évidence de mots-clés («parce que», «pour», «donc», «à cause de», etc.)
  - ▶ **Contrôle** : idem («je me souviens», «je crois», «ils racontent», etc.)
  - ▶ **Paraphrase de mots** : Similarité sémantique + similarité des lemmes via *WordNet*
  - ▶ **Inférence de connaissances** : Mot non paraphrasé mais : proche d'un mot du texte, ces deux étant proches du texte lu.
  - ▶ **Reprise (*bridging*)** : Segment de phrase du texte évoqué de manière non littérale (établir une relation avec un passage précédent, qui a été repris)

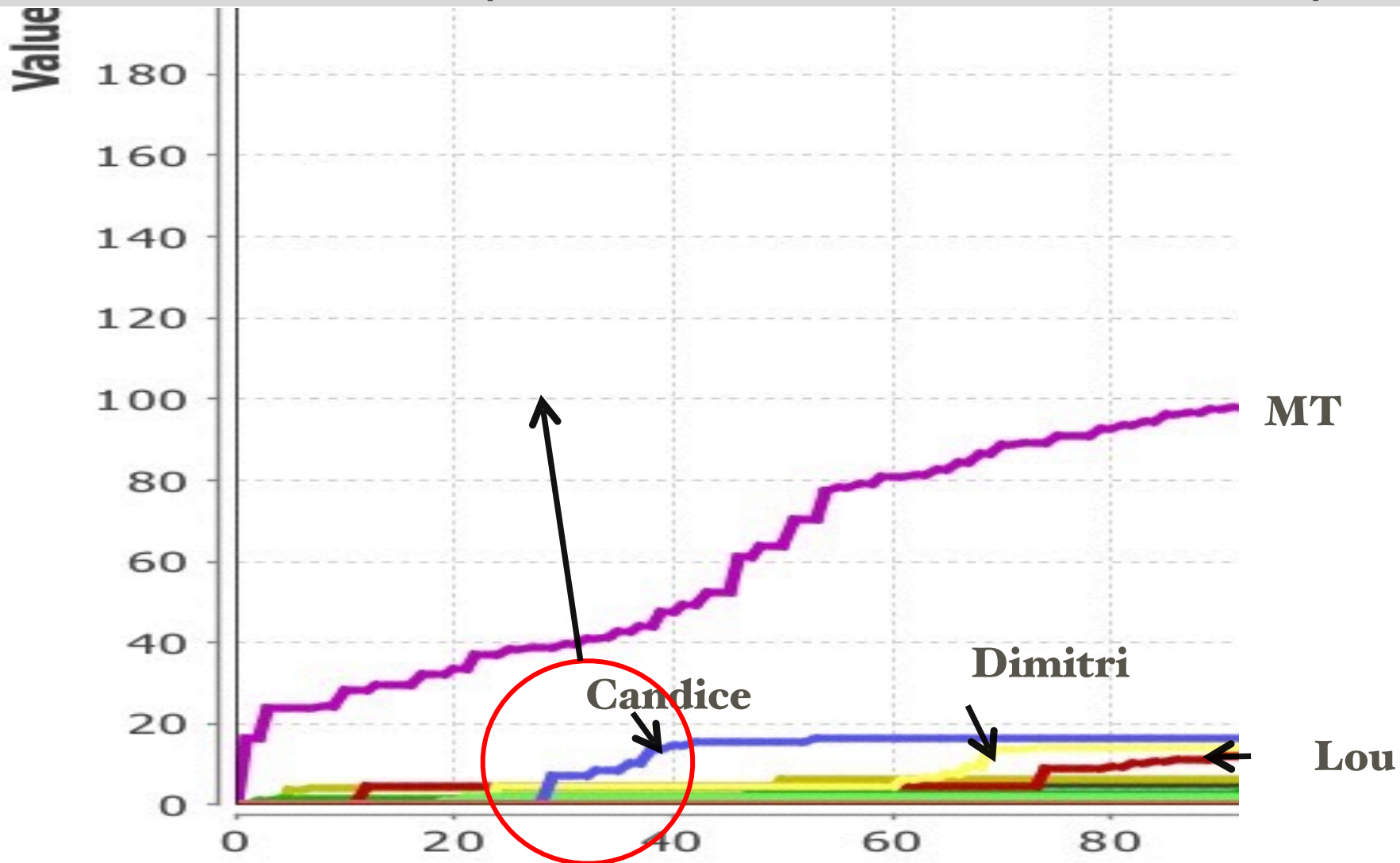
### 3. Copie d'écran - Analyse de stratégies

| Document title: Matilda [config/LSA/lemonde_fr, config/LDA/lemonde_fr]                                                                                                                                                                                                                                                                                                                                                                              |           |         |            |           |          |          | <a href="#">View document</a> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|---------|------------|-----------|----------|----------|-------------------------------|
| Verbalization: <span style="background-color: red; color: black;">[REDACTED]</span> Alice.xml)                                                                                                                                                                                                                                                                                                                                                      |           |         |            |           |          |          |                               |
| Contents                                                                                                                                                                                                                                                                                                                                                                                                                                            |           |         |            |           |          |          |                               |
| Text                                                                                                                                                                                                                                                                                                                                                                                                                                                | Causality | Control | Paraphr... | Knowle... | Bridging | Cohesion |                               |
| la mère[8] devint toute blanche . elle dit[5] à son mari il y a quelqu' un dans la maison[2] . ils arrêtèrent[9] tous de manger[10] . ils étaient tous sur le qui - vive . la voix[7] reprit[11] salut[6] , salut[6] , salut[6] . le frère[12] se mit à crier ça recommence[13] ! matilda se leva et alla éteindre la télévision[3] .                                                                                                               |           |         |            |           |          | 0.315    |                               |
| Je ai compris[4] que c' est une famille[2] la famille[2] dans laquelle il ? suis qui dinent[1] devant la télé[3] . et qui . tout de un coup il z entendent[4] une voix[7] qui leur dit[5] salut[6] . et du coup ils ont peur donc parce que la mère[8] de matilda ? donc c' est que je pense que ils ont peur . alors ils arrêtent[9] de manger[10] . puis le frère[12] commence à comprendre quelque cho quelque chose en disant ça recommence[13] | 5         | 1       | 13         | 0         | 1        |          |                               |
| la mère . paniquée . dit à son mari : henri . des voleurs[15] . ils sont dans le salon , tu devrais[14] y aller . le père . raide sur sa chaise ne bougea pas . il n' avait pas envie de jouer au héros . sa femme lui dit : alors , tu te décides ? ils doivent[14] être en train de faucher l' argenterie[16] !                                                                                                                                   |           |         |            |           |          | 0.294    |                               |
| alors je pense que c' est une famille[*] peut - être assez riche parce que il y a de l' argenterie[16] . et qui pensent que ceux qui doit[14] être riche ou que y a beaucoup de voleurs[15] dans notre dans leur maison donc                                                                                                                                                                                                                        | 2         | 1       | 3          | 1         | 1        |          |                               |
| monsieur verdebois s' essuya nerveusement les lèvres avec sa serviette et proposa d' aller[17] voir[18] tous ensemble . la mère attrapa un tisonnier au coin de la cheminée . le père[19] s' arma d' une canne de golf posée dans un coin . le frère attrapa un tabouret . matilda prit[9] le couteau avec lequel elle mangeait . puis ils se dirigèrent tous les quatre vers la porte du salon en marchant sur la pointe des pieds .               |           |         |            |           |          | 0.399    |                               |
| à ce moment - là , ils entendirent à nouveau la voix . matilda fit alors irruption dans la pièce en brandissant son couteau et cria haut[20] les mains[21] , vous êtes pris[9] ! les autres la suivirent en agitant leurs armes .                                                                                                                                                                                                                   |           |         |            |           |          | 0.189    |                               |
| donc la c' est on sait déjà comment s' appelle la famille . et puis ils racontent que là vu que le père[19] veut pas y aller[17] tout seul . il est accompagné de toute sa famille pour aller[17] voir s' y a un voleur . et y a la le la parole[*] ça le bruit aussi ? qui recommence . et du coup elle , la petite fille[*] qui s' appelle matilda commence à avoir peur . donc elle lui dit haut[20] les mains[21] vous êtes pris[9]             | 4         | 2       | 5          | 2         | 1        |          |                               |

## 4.<sup>15</sup> Participation et collaboration dans les forums de discussion [Dascalu *et al.* 15b; Simon *et al.* 15]

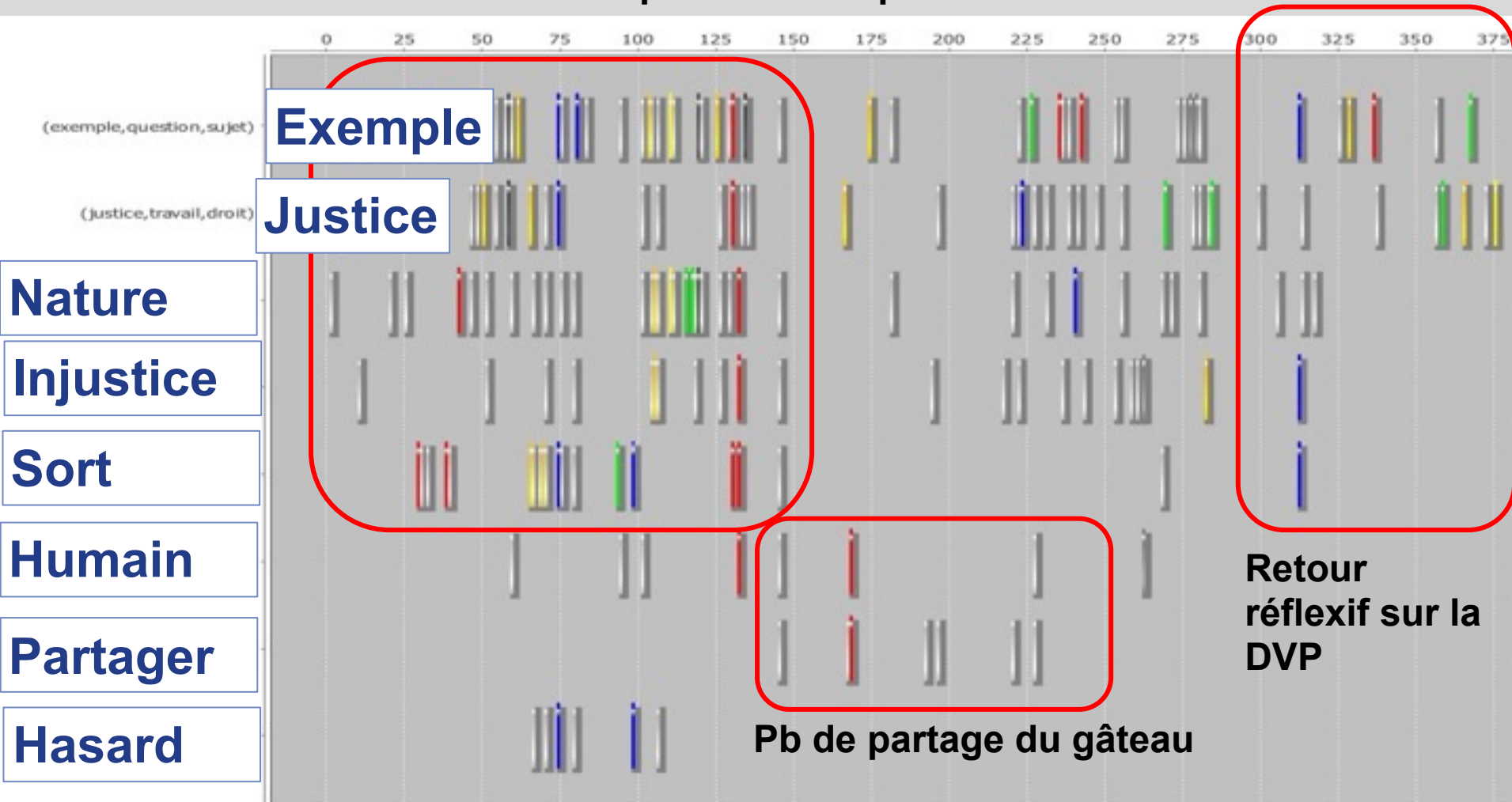
- Des participants à un débat (*via* forum, chat, ou *de visu*...)
- Analyse de la multivocalité (de multiples voix, points de vue sont exprimés) et de la polyphonie (chaque voix interagit avec les autres, inter-animation) [Bakhtine 81]...
- ... par la cohésion des tours de parole entre eux (liens lexicaux, syntaxiques, sémantiques)
- Identification de la *participation* (nombre de tours de parole d'un participant) et de la *collaboration* (tours de parole en réponse à d'autres)
- Analyse de Discussions à Visée Philosophique (Simon *et al.*, 2015) à partir d'une discussion sur « Ce qui est juste » en CM2

# Visualisation du « poids » des contributions sur la discussion totale (début de la discussion affichée)



# Les thèmes abordés

## Brainstorming à partir d'exemples



# CONCLUSION

# Recherches en cours/à venir

## Réseaux

En cours

Analyse sémantique des résumés d'articles en éducation pour déterminer leur proximité et des tendances scientométriques. Collaboration avec ENS-Lyon (EducMap)

Analyse de la cohésion de verbalisation comme indice de l'expertise à partir d'études de cas

## Complexité

Ajout d'un module d'analyse de sentiments, de calcul de complexité évoluées (*word maturity: Age of exposure* : Dascalu, Crossley, McNamara, Trausan-Matu, en prép.)

## Discussion

Analyse des contributions des élèves en séance de philosophie (cycle 3). Coll. avec J.-P. Simon, LIDILEM

En projet

## Validation *in situ*

Scénarisation et test de *ReaderBench* en situation scolaire

<sup>20</sup>  
**Merci de votre attention !**

## Des questions ?

- La plupart des références citées (n'hésitez pas à me demander les autres) sont à <http://bit.ly/liris-pdessus>
- Cette présentation sera tout bientôt disponible à <http://webcom.upmf-grenoble.fr/sciedu/pdessus/>
- Ces travaux ont été réalisés avec la collaboration de M. Bianco, M. Dascalu, D. Dupré, G. Denhière, B. Lemaire, D. Lupan, M. Loiseau, S. Mandin, D. McNamara, A. Nardy, B. Opreescu, C. Paraschiv, S. Trausan-Matu, E. Villiot-Leclercq, V. Zampa ...
- ... et avec des financements des projets EC-FP7 LTfLL & ERRIC, et de l'ANR DEVCOMP



# <sup>22</sup>Diagnostic de l'auto-explication de stratégies de lecture (McNamara 2004)

- Une phrase verbalisée... (Phc = Phrase cible du document lu)
  - ▶ ... est la paraphrase de Phc si plusieurs mots de la phrase sont identiques (ou synonymes)
  - ▶ ... utilise une connaissance inférée
  - ▶ ... utilise un *bridging* avec Phc : si  $\text{prox}(\text{Ph}, \text{Phc}) > \text{seuil}_b$  ET  $\text{prox}(\text{Ph}, \text{Phk}) > \text{seuil}_b$
  - ▶ utilise un contrôle : si des *patterns* comme «j'ai compris que», «je pense que...» apparaissent dans la phrase.

# LDA (1/2) : 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 2/3 (Blei 2012 p. 78)

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

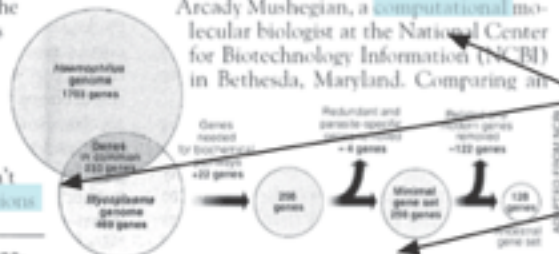
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

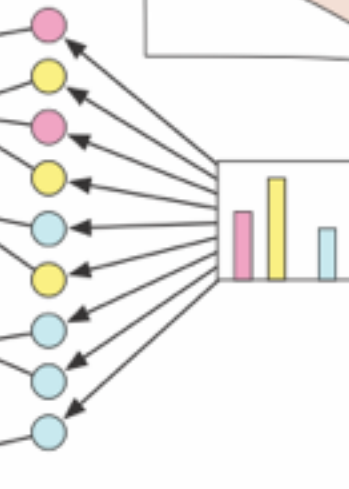


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments

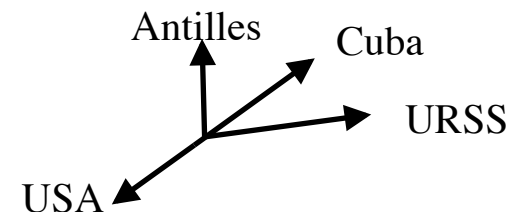


# Une parenthèse : Comment marche LSA?



# 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 (paragraphe, phrases) ont un sens similaire (contiennent des informations similaires) s'ils contiennent des mots de sens similaire

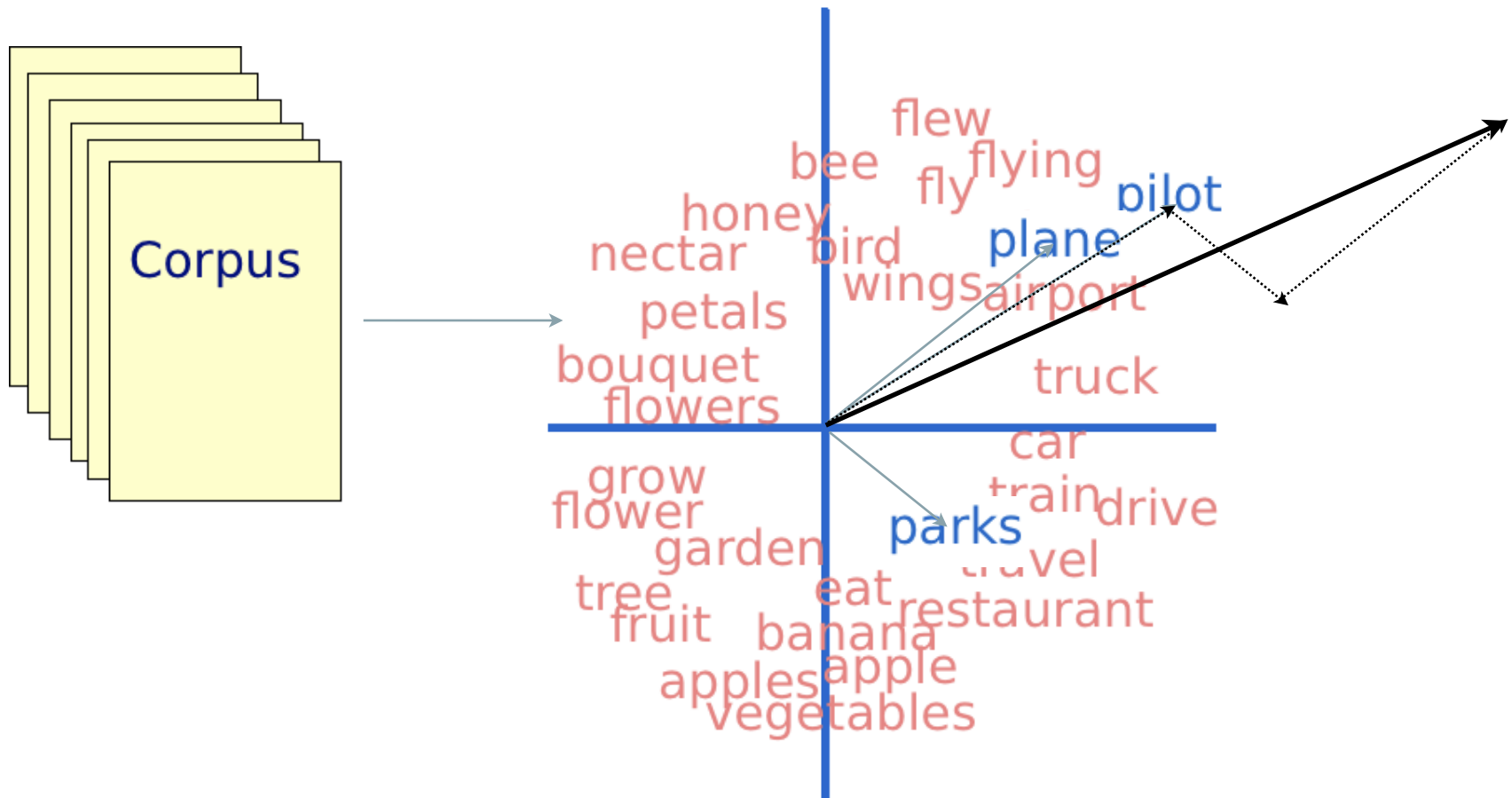


## Comment marche LSA (2) ?

- 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-occurents 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)

28  
Une phrase est représentée par la somme des vecteurs des mots qui la composent  
[Lemaire & Denhière '05]

“The pilot parks the plane”



# La maturité des mots (Landauer et al. 2011)

