UNIVERSITÉ GRENOBLE ALPES

LSE LABORATOIRE DES SCIENCES DE L'EDUCATION

espe École supérieure du professorat et de l'éducation Académie de Grenoble

ReaderBench, un système automatique d'aide à Écrire-lire-pour-apprendre ou collaborer

Philippe Dessus

Exposé au LIRIS–Lyon

Le 22 Mai 2015

CC BY NC SA

Plan

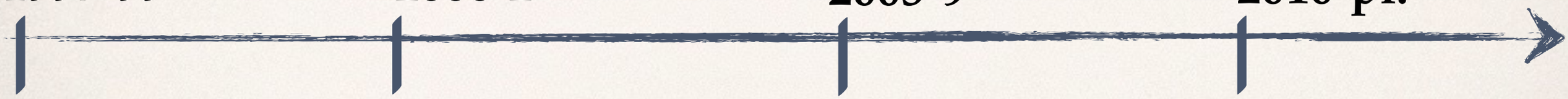
- Introduction

- Présentation et démo de *ReaderBench*, un « atelier » pour analyser plusieurs phénomènes liés à la compréhension en lecture. Identification de
  - mots-clés, réseaux sémantiques, phrases importantes
  - stratégies de lecture
  - complexité des textes lus
  - participation/collaboration à des discussions

➡ **Note :** Les références sont disponibles à **http://www.citeulike.org/user/pdessus/tag/liris** ou **http://bit.ly/liris-pdessus**

Introduction

Quelques étapes de recherche



Premiers essais sur LSA : prédiction de liens hypertextes ; de réponses à QCM (B. Lemaire)

[Dessus 00]

Systèmes d'éval. automatique de résumés de cours : *Apex 1.0* & *Apex 2.0* (A. Vernier, B. Lemaire)

[Lemaire & Dessus 01]

Système d'entraî-nement au résumé : *Résum'Web* (S. Mandin, B. Lemaire, G. Denhière)

[Lemaire et al. 05 ; Mandin 10]

Système d'aide à la lecture : *ReaderBench* (M. Dascalu, S. Trausan-Matu)

[Dascalu et al. 13]

Sciences cognitives appliquées à l'éducation [Gentaz & Dessus 04 ; Dessus & Gentaz 06]

- Disposer de modèles cognitifs de l'activité à étudier (comprendre un texte : McNamara 2004 ; le résumer : Kintsch & van Dijk 1975)

- Réaliser une implémentation informatique qui simule et puisse aider l'activité des élèves, de l'enseignant

- Valider (en amont ou aval) par comparaison avec comportements humains (enseignants, élèves, experts)

- Utiliser et valider l'implémentation dans des contextes de laboratoire ou scolaires, à des fins d'étude de l'enseignement/apprentissage

Ecrire/Lire-pour-apprendre
Lien entre écriture et apprentissage

- Ecriture et apprentissage, 2 activités qui (Bangert-Drowns *et al.*, 2004 ; Emig 1977) :
  - émergent de processus cognitifs symboliques
  - utilisent ou procurent des connexions entre les concepts en jeu
  - sont efficaces si elles sont délibérées et autorégulées
  - sont permises et favorisées par des *feedback* auto-évaluatifs
- Ecrire permet la révision du contenu ; autorise une compréhension plus approfondie ; entraîne un monitorage réflexif => Apprentissage
- Production écrite : Une source et un indicateur du processus d'apprentissage ou de construction de connaissances ; fortement interreliée avec la lecture (de sa production, de celle des autres)

Liens entre les activités

Produire un texte

Caractéristiques du texte écrit (cohésion, mots-clés, complexité, réseaux de mots…)

Lire (sa production, celle des autres)

Mettre en œuvre des stratégies de compréhension

Comprendre, construire des connaissances individuellement

Discuter, collaborer, construire des connaissances collectivement

Légende :
Actions/Paramètres

Questions

- Quelles activités cognitives mises en œuvre dans des activités de lecture/écriture pour comprendre et/ou collaborer ?
  - centration sur la compréhension et la discussion
- Comment les simuler et les analyser ? Ne pas avoir seulement un système qui *fonctionne,* mais qui donne des résultats proches de ceux d'humains

La cohésion comme mesure de nombreux phénomènes

- La cohésion (recouvrement sémantique de 2 propositions) peut être utilisée pour rendre compte des écarts entre mots, propositions, paragraphes, etc.

- Elle rentre donc très souvent dans le calcul de nombreux paramètres des systèmes utilisés : cohérence inter-phrases, inter-paragraphes, mais aussi de 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* calculé : $r = .54$

ReaderBench

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

Histoire de *ReaderBench*

- **2008–11** Début d'une collaboration avec l'univ. de Bucarest (WP commun du FP7 LTfLL) : *Pensum*/France (avec S. Mandin puis M. Loiseau) et *PolyCAFe*/Roumanie (équipe de S. Trausan-Matu)

- **2011-13** Thèse en co-tutelle (UPMF-Univ. Bucarest) de Mihai Dascalu (co-dir avec S. Trausan-Matu), partiellement intégrée au projet DEVCOMP (LSE, porté par M. Bianco, LSE). *ReaderBench*. But : analyse des stratégies en lecture, analyse de la complexité et raffinement de *PolyCAFe*

- **2013-pr.** Ajout/raffinement de modules, avec la participation d'autres étudiants roumains : Luisa Stavarache, Bogdan Oprescu, Ionut Cristian Paraschiv (dir. S. Trausan-Matu)

- **2015**. Séjour Fulbright de M. Dascalu à l'univ. d'Arizona, collaboration avec D. McNamara

4 grands types de fonctionnalités de *ReaderBench*

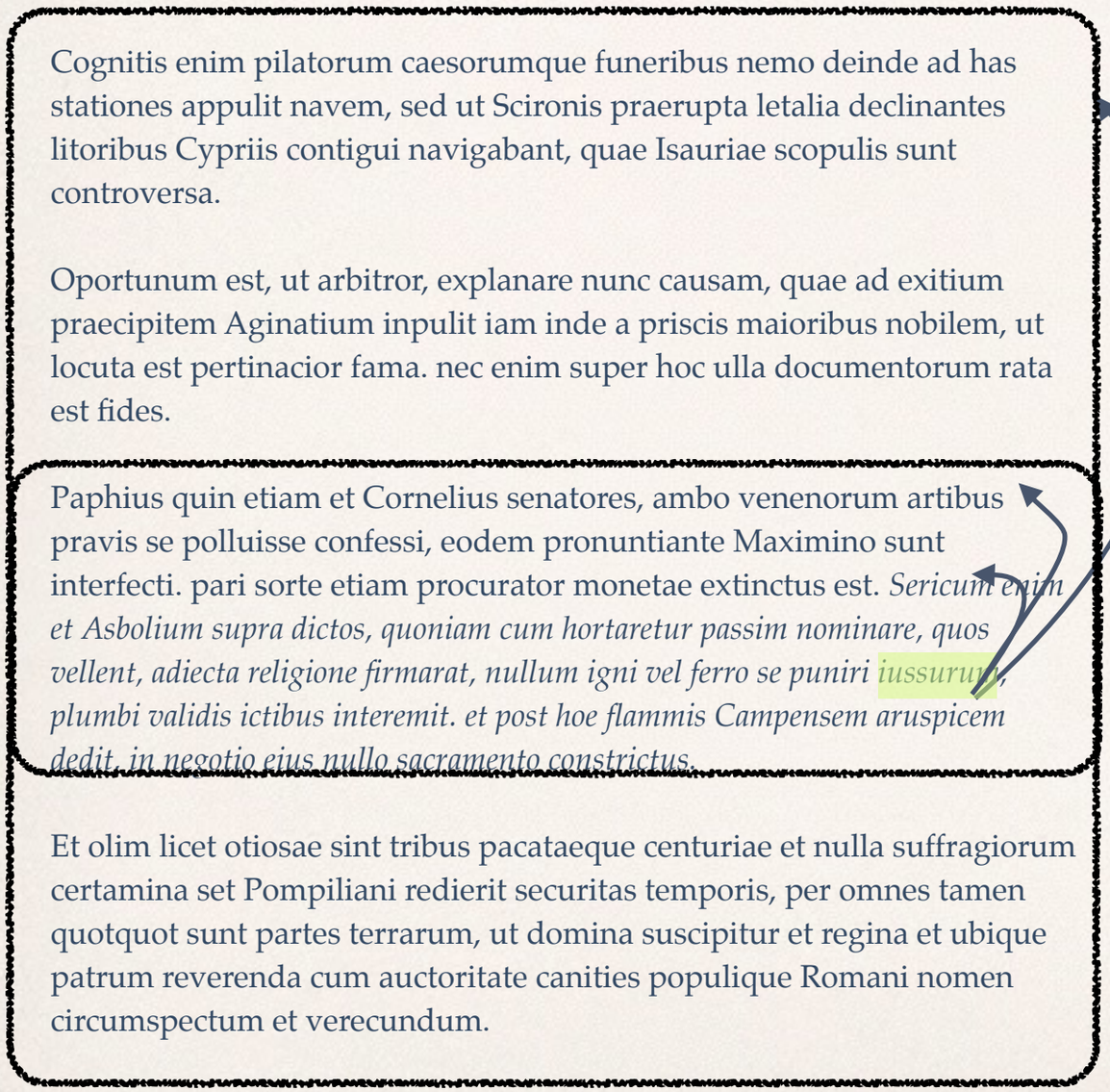
1. Identification de mots-clés, réseau de mots, mots inférés

2. Analyse de la complexité textuelle

3. Identification de stratégies de compréhension en lecture

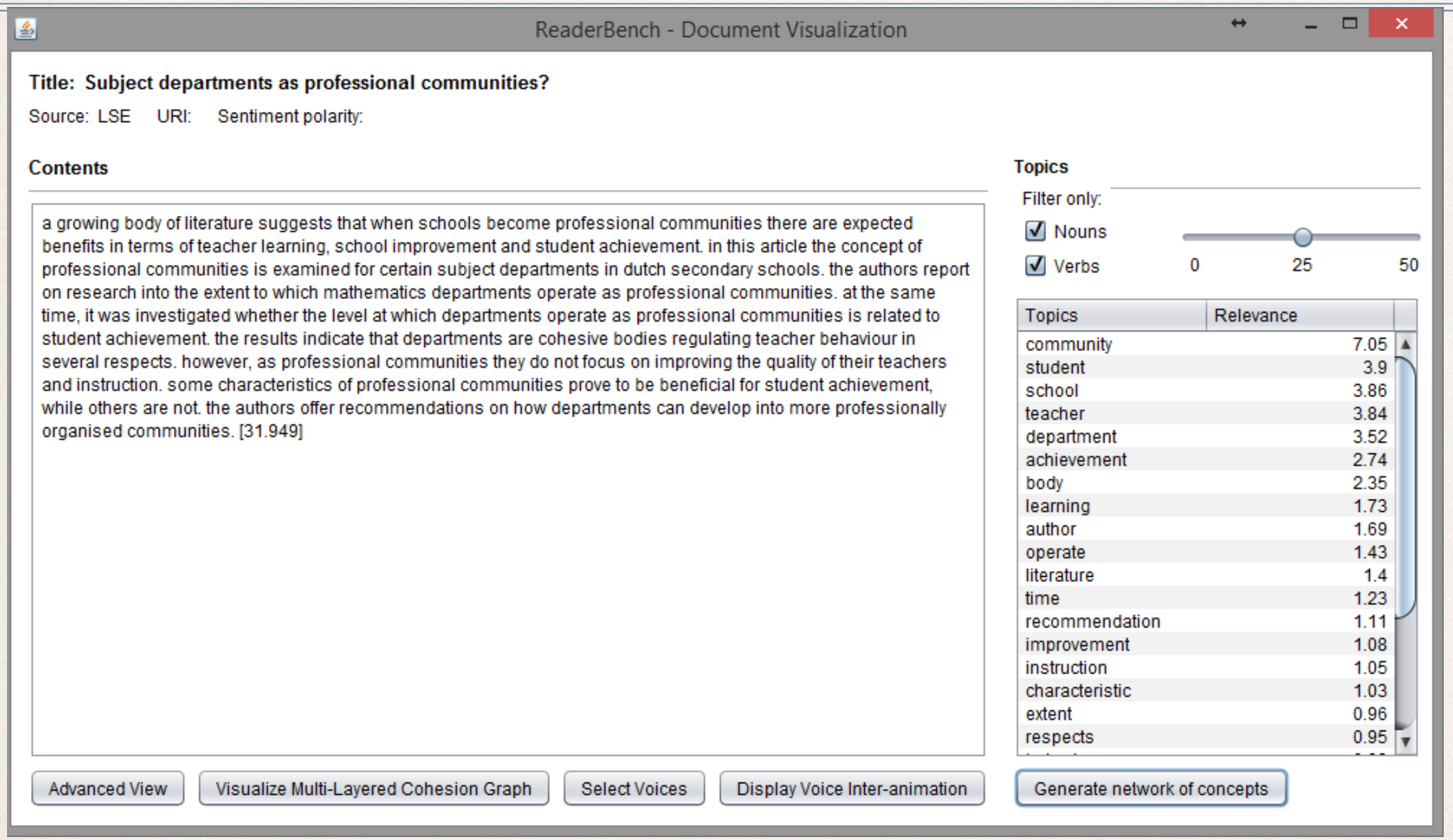
4. Analyse de la participation/collaboration dans les forums

1. Identification de mots-clés (Dascalu 2013)

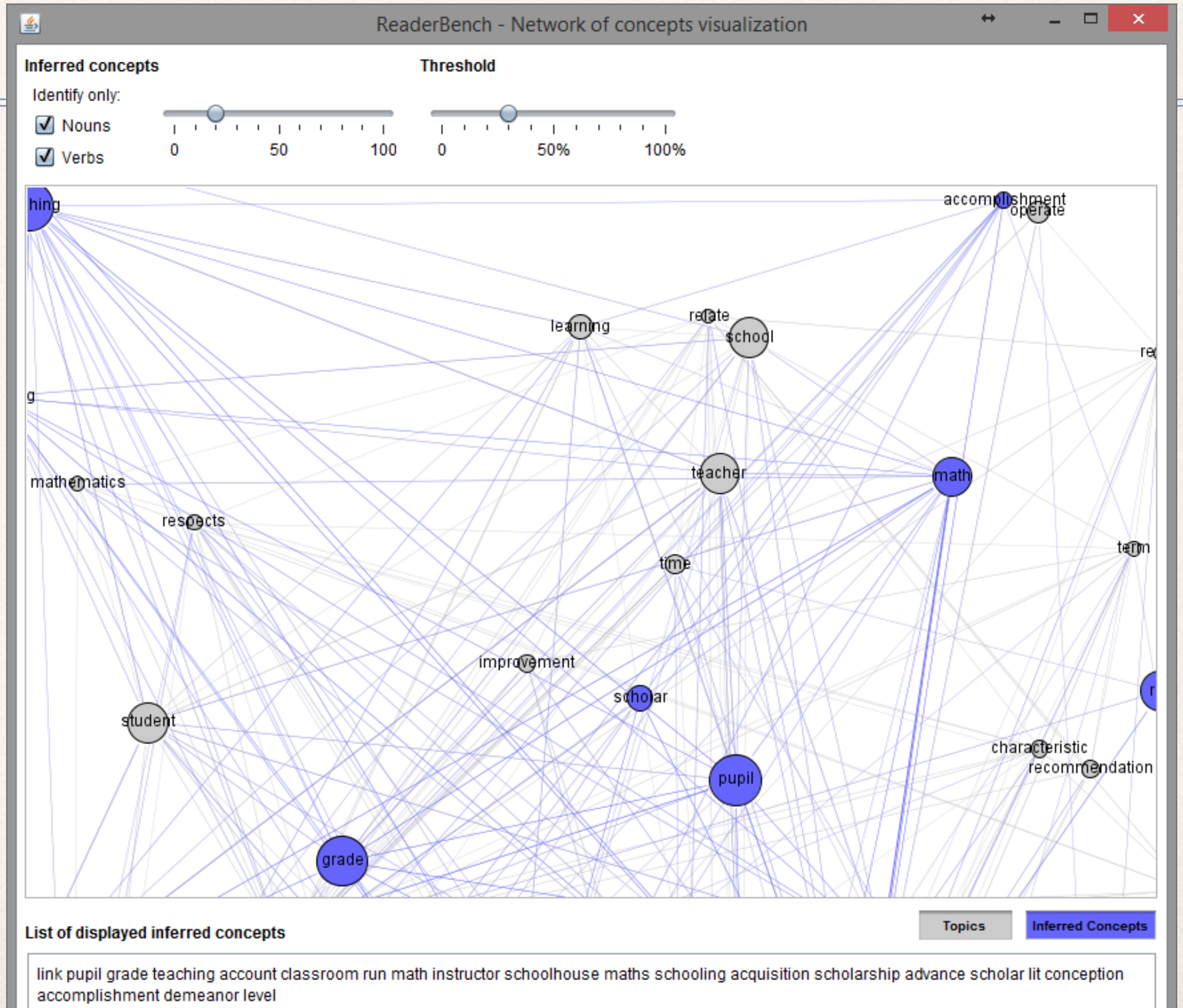


1. Classe gramm. du mot
2. Score d'un mot = *tf*idf* (fréq. locale pondérée par fréq. inv. globale) + similarité mot avec **texte** lu + similarité moyenne mot avec chaque **phrase** + similarité moyenne mot avec chaque **paragraphe**
3. Présenter les mots par score décroissant et par classe (option)

1. Copie d'écran Analyse du texte



1. Copie d'écran 2 - 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.)

- 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. 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 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

2. Lien entre complexité, stratégies et compréhension [Dascalu *et al.* 15]

- Transcription et analyse de 149 résumés de compréhension de 2 textes de Cycle 3. Mesure de la fluence orale des élèves et de leur compréhension effective du texte lu (questions littérales, de vocabulaire, d'inférence, sur le thème de l'histoire)

- Analyse automatique de la complexité de ces résumés, et des stratégies par *ReaderBench*

- SVM pour créer 3 classes de niveau de compréhension, pour essayer de prédire la compréhension finale du texte lu

2. Complexité, stratégies et compréhension : Résultats [Dascalu *et al.* 15*a*]

	Facteurs	C1	C2	C3	**Moy.**
1	Toutes les stratégies	.60	.45	.61	**.55**
2	Toutes les stratégies & Cohésion avec le document lu	.57	.45	.55	**.52**
3	Tous les indices de complexité du résumé	.17	.14	.94	**.42**
4	Facteurs 2 & 3 (stratégies/cohésion/complexité)	.52	.49	.53	**.51**
5.	2 & fluence orale	.67	.53	.78	**.66**
6	5 & complexité morphologique et sémantique	.74	.55	.78	**.68**

3. Auto-explication de stratégies

[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. Exemples de stratégies [Bianco *et al.*, sous presse]

Paraphrases

[P1, CE2] *Ben en fait il y avait des dieux* **et ben ils trouvaient qu'il faisait trop chaud**

Inférences textuelles

[P1, CM2] et y a un serpent qui a Ati qui qui mange enfin qui boit l'eau des nuages **et c'est pour ça qu'il fait aussi chaud**

Inférences de connaissances

[P2, CM1] il veut délivrer un peu l'eau donc du coup **il veut combattre Ati le serpent**

[P1, CM2] **Alors c'est dans le ciel** y a des Dieux et Ati c'est un grand serpent qui va prendre l'eau dans les nuages

Auto-évaluation (contrôle)

[P3, CM2] Bon là déjà **on sait que Indra c'est un garçon pas une fille**

[P5, CE2] **Et j'pense que** après le serpent il était mort…c'est tout.

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]

View document

Verbalization: [redacted] 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 qué **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	

3. Etude sur 19 élèves de Cycle 3

(Bianco *et al.*, sous presse)

- Les verbalisations ont été analysées «à la main» par des juges humains, puis, indépendamment, par *ReaderBench*

- Corrélation par stratégie :
  - Contrôle : .83
  - Causalité : .77
  - Paraphrase : .73
  - Inférence de connaissance : .60

- Une stratégie plus complexe à déterminer automatiquement, le *bridging* (lien entre plusieurs propositions) n'est pas détectée

4. Participation et collaboration dans les forums de discussion

[Dascalu *et al.* 15*b*; Dascalu, subm.]

- 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, interanimation) [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)

4. Participation et collaboration dans les forums de discussion

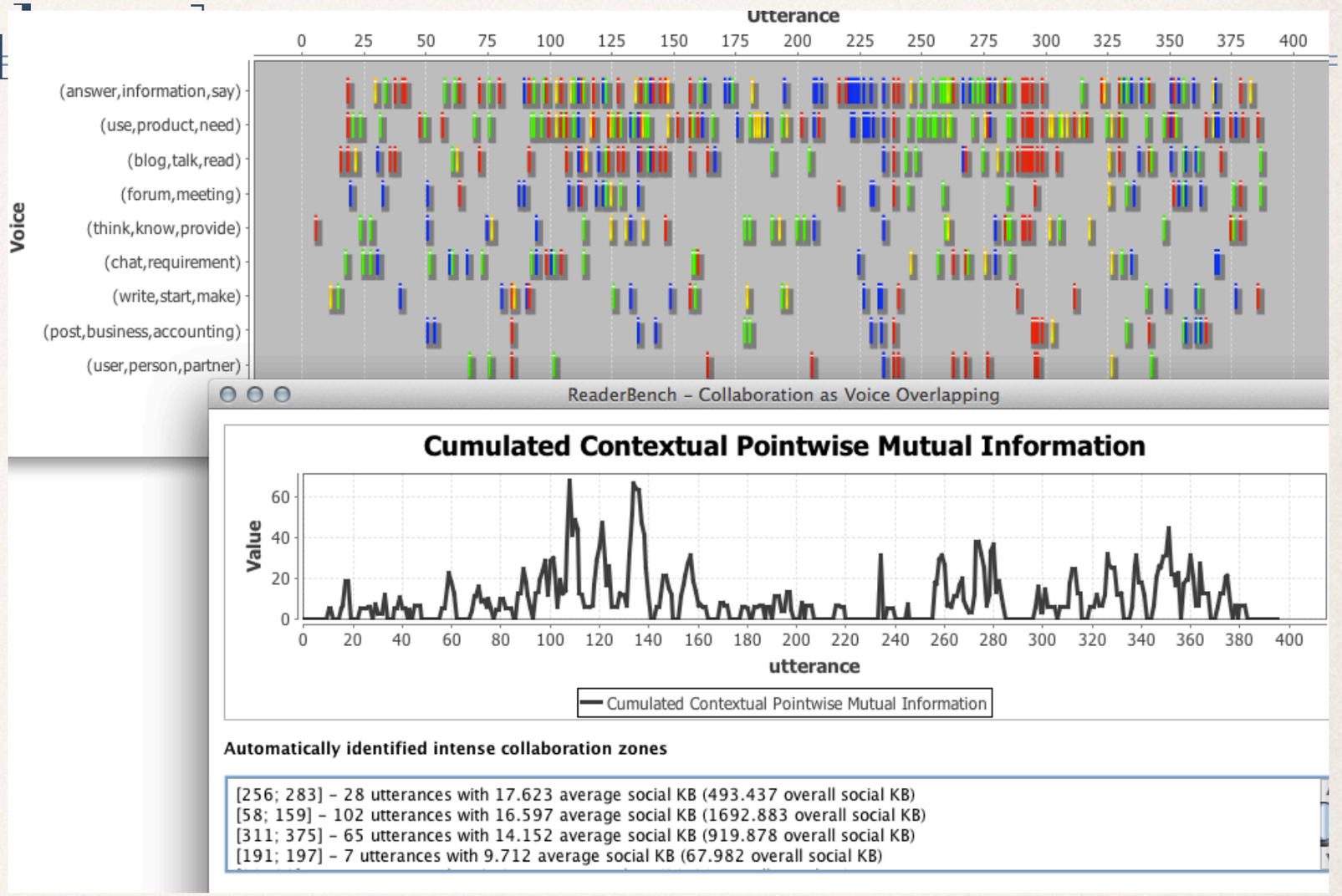
- Définir le score d'un tour de parole en fonction de 3 paramètres liés aux mots le composant : leur fréquence dans le texte, leur similarité sémantique par rapport au tour de parole, à la discussion, la « couverture sémantique » représentée par la chaîne sémantique du mot tout au long de la discussion (extension sémantique de la chaîne lexicale de [Galley & McKeown 03])

- *Pointwise Mutual Information* (mesure du recouvrement de 2 variables) pour déterminer les pics de synergie entre voix (interanimation)

- Analyse de la construction de connaissances
  - individuelle : cohésion des voix d'*un même* participant
  - collaborative : cohésion des voix *interparticipants*

4. Participation et collaboration dans les forums de discussion/*chats*

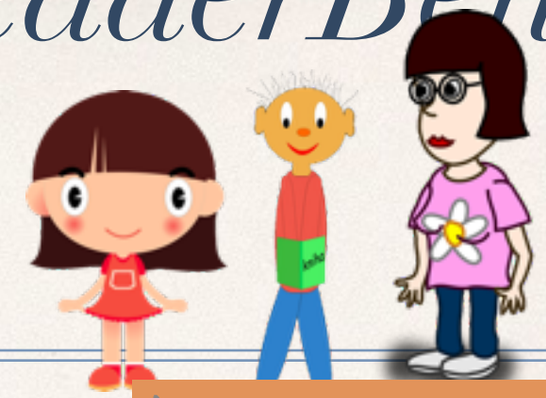
[~~Dascalu *et al.* su~~

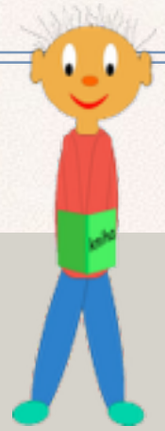


- Validation des évaluations en comparant les évaluations par *ReaderBench vs*. des étudiants de 10 conversations *chat* (110 étudiants codant chacun 3 conversations) : corrélations allant de .59 à .80

Conclusion

Un flux de travail avec *ReaderBench* (d'après Dascalu 2013)





Présélectionne le matériel : **Analyse de la complexité**

Lit le matériel

Fait **évaluer le document** : cohésion mots-clés, phrases importantes

Explique sa compréhension

Analyse des stratégies

Analyse la prod. de l'élève et l'éval. de *ReaderBench*

Discutent à propos du matériel

Analyse de la participation et collaboration

Boucle Lecture

Boucle Evaluation

Boucle Ecriture

Boucle Discussion

Recherches en cours/à venir

En cours

Réseaux

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

Complexité

Lien entre la complexité des explications et d'autres variables socio-linguistiques (liaison facultative, réalisation du *l* dans il(s)). [Nardy *et al*. 15]

Discussion

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

En projet

Réseaux

Analyse de la cohésion de verbalisation comme indice de l'expertise

Complexité

Analyse des tours de parole des enseignants en situation d'enseignement. Complexité comme indice de leur charge cognitive

Validation *in situ*

Scénarisation et test de *ReaderBench* en situation scolaire

Plus tard

Stratégies

Analyser les auto-explications des élèves suite à la lecture de cours ou de MOOC (Ecrire-Pour-Apprendre)

Complexité

Ajout d'un module d'analyse de sentiments, de calcul de complexité évoluées (*word maturity*)

Démo de *ReaderBench*

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. Oprescu, 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

Annexes

Diagnostic de l'auto-explication de stratégies de lecture (McNamara 2004)

- Une phrase verbalisée… (Ph*c* = Phrase cible du document lu)
  - … est la **paraphrase** de Ph*c* si plusieurs mots de la phrase sont identiques (ou synonymes)
  - … utilise une **connaissance inférée**
  - … utilise un ***bridging avec*** Ph*c* : si prox (Ph, Ph*c*)>seuil*b* ET prox (Ph, Ph*k*)>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.

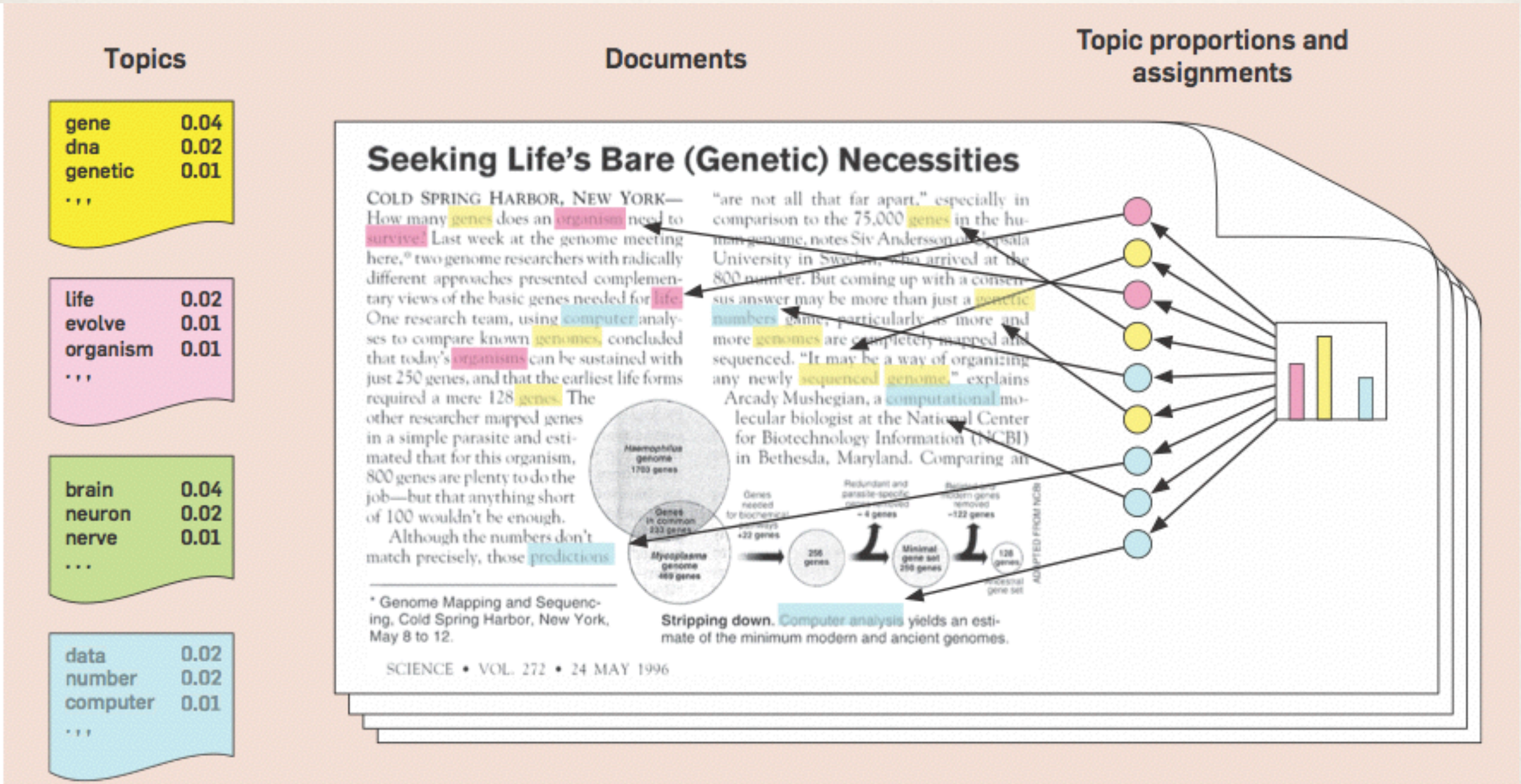
- Un exemple de trois *Topics* :

```
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)



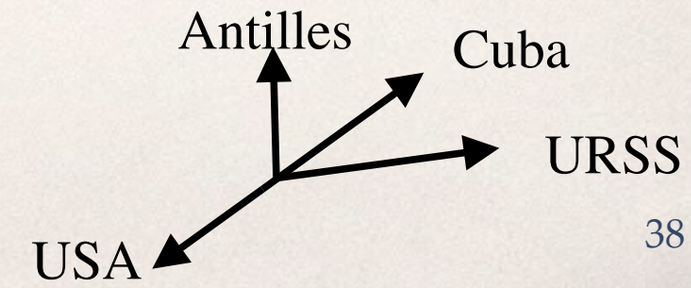
Une parenthèse : Comment marche LSA?

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

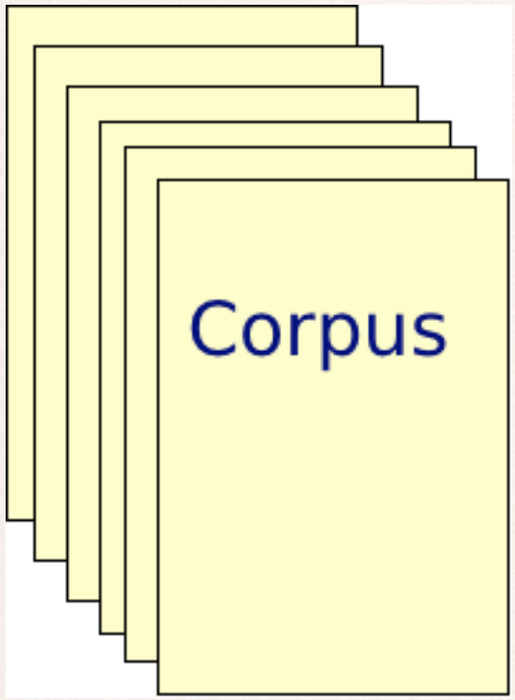


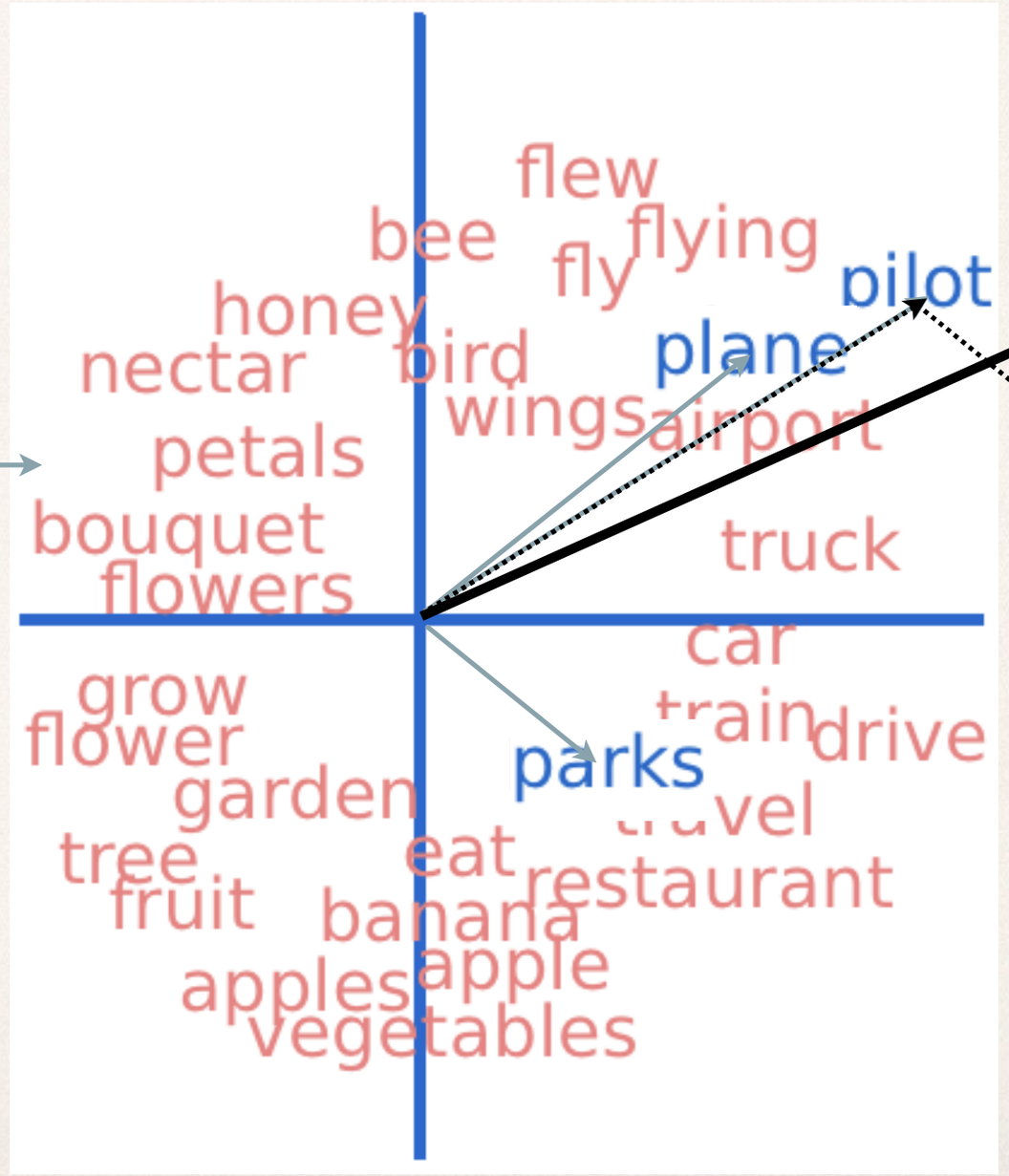
Comment marche LSA (2) ?

- 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)

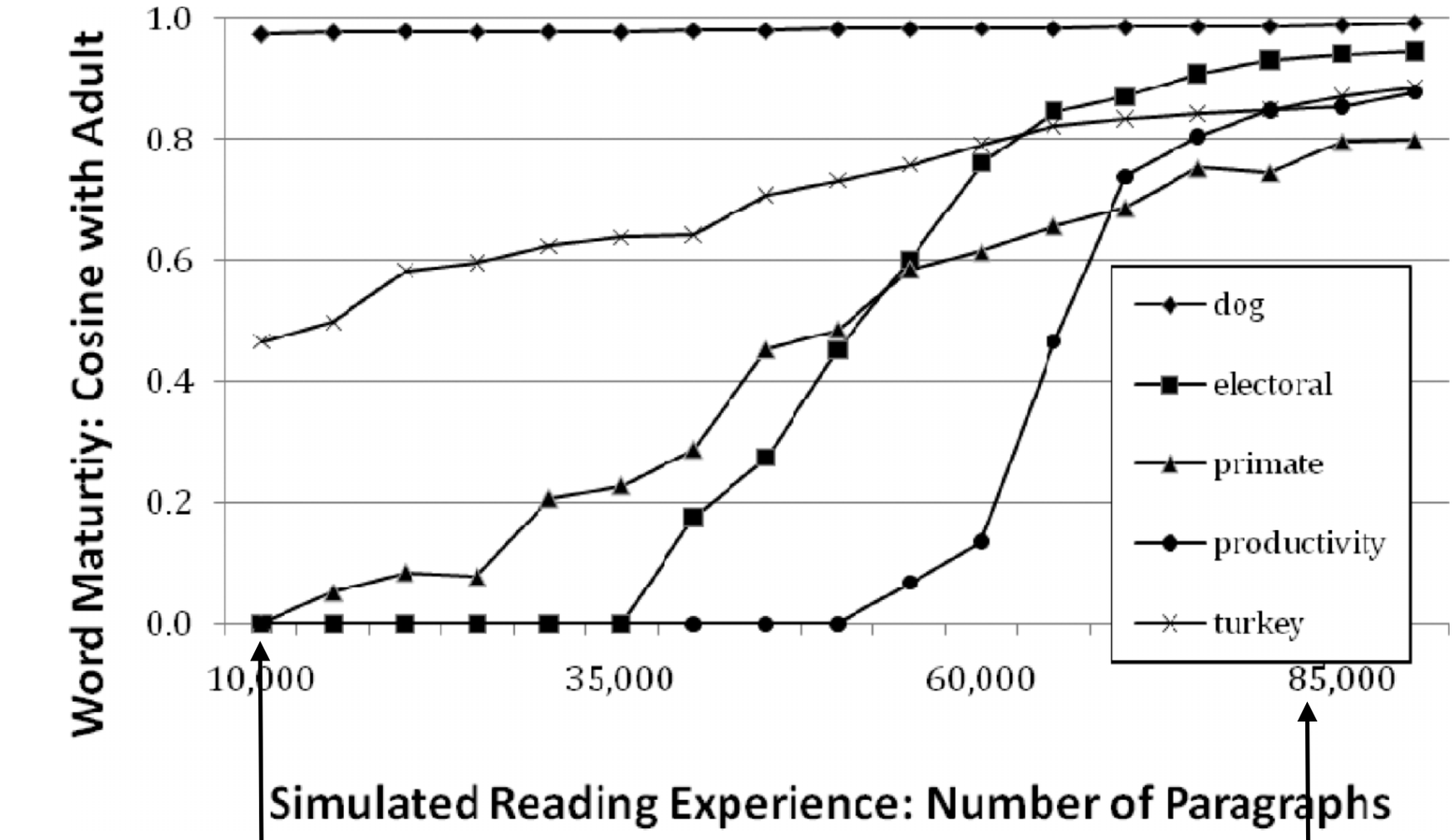
Une phrase est représentée par la somme des vecteurs des mots qui la composent
[Lemaire & Denhière '05]

"The pilot parks the plane"





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



Expos. écrit avant école

Expos. écrit d’un adulte moyen