

# CONSTRUCTION DE CONNAISSANCES PAR EXPOSITION À UN COURS AVEC LSA

Philippe DESSUS

Laboratoire des Sciences de l'Éducation,

Bât. SHM, Université Pierre-Mendès-France, BP 47, F-38040 GRENOBLE Cedex 9, France

Mél : Philippe.Dessus@upmf-grenoble.fr

Toile : <http://www.upmf-grenoble.fr/sciedu/pdessus>

*Et comment t'y prendras-tu, Socrate, pour chercher ce que tu ne connais en aucune manière ? Quel principe prendras-tu, dans ton ignorance, pour te guider dans cette recherche ? Et quand tu viendras à le rencontrer, comment le reconnaitras-tu, ne l'ayant jamais connu ?*

Platon, *Ménon*

## Résumé

Le paradoxe de l'apprentissage, c'est-à-dire la possibilité de dériver de nouvelles connaissances à partir de connaissances existantes, est un problème auquel se sont attelés de nombreux chercheurs. Landauer et Dumais (1997) en proposent une mise en œuvre informatique, l'analyse de la sémantique latente (LSA, pour Latent Semantic Analysis), une méthode d'analyse factorielle multidimensionnelle permettant d'analyser, à partir des co-occurrences de mots, les proximités sémantiques de mots ou paragraphes. Ce modèle a notamment été utilisé pour simuler l'acquisition du vocabulaire humain, et nous répliquons ici ce travail à propos d'une exposition à un cours. Un premier test nous permet d'examiner l'évolution des proximités sémantiques intermots, calculées par LSA, au fur et à mesure de l'exposition à ce cours. Trois types d'évolution de proximités peuvent être dégagées, qui amènent à la considérer comme semblable à celle des humains. Un deuxième test compare les scores d'étudiants à un questionnaire à choix multiple à ceux de LSA, une fois qu'il a traité le cours donné aux étudiants. LSA permet, sans aucun codage humain préalable, de rendre compte de la construction de connaissances par exposition à un cours.

## Abstract

How humans can infer new knowledge from old one is a well known problem named the Learning Paradox. LSA (for Latent Semantic Analysis) is a method implemented by Landauer and Dumais which relies on large corpora of texts to build a semantic high-dimensional space containing all words and texts, by means of a statistical analysis of the words' co-occurrences. Researchers showed that Human vocabulary acquisition, as well as some consequences of the Learning Paradox, can be adequately simulated by LSA. In this paper, two simulations are described. First, we analyse the evolution of semantic similarities between some typical words of an academic course as it advances. Then, we compare students MCQ scores to LSA's, once it has computed the same course given to the students. In these two simulations, LSA performance is close to that of humans. LSA takes into account the construction of knowledge during a lecture without pre-processing.

## 1. Introduction

Deux des présupposés centraux des approches récentes en éducation sont que : — la connaissance est activement construite par l'apprenant, — le contexte dans lequel ce dernier récupère cette connaissance est primordial. Ces présupposés sont issus des travaux de Piaget, et ont même été adoptés par des courants non spécifiquement constructivistes, comme le souligne Bereiter (1985). La construction de connaissances tenant compte du contexte dans lequel elles ont été exposées a fait l'objet de très nombreux travaux, que ce soit dans des situations d'enseignement classiques (Salomon, 1993) ou informatisées (Spiro, 1991).

Cette construction est en butte au paradoxe suivant. Nous avons besoin de connaissances anciennes à partir desquelles nous construisons les nouvelles et, si ces dernières sont vraiment nouvelles, elles ne pourront pas s'incorporer aux anciennes. En revanche, si les nouvelles connaissances ont un degré de similarité

important avec les anciennes, à quoi cela sert-il de les apprendre ? Le paradoxe de l'apprentissage, ou de l'induction, traite de la possibilité de dériver de nouvelles connaissances à partir de connaissances existantes. À ce problème se sont attelés de nombreux philosophes contemporains (Goodman, 1984 ; Popper, 1998 ; Quine, 1977), ou plus anciens (Platon, Hume), des psychologues (Bereiter, 1985 ; Chomsky & Fodor, 1979 ; Smith, 1993), des neuropsychologues (Edelman, 1992 ; Rosenfield, 1994), des chercheurs en éducation et intelligence artificielle (Bereiter & Scardamalia, 1993 ; Heylighen, sous presse ; Roschelle, 1995). Pour certains, comme Chomsky et Fodor, c'est le caractère inné de certaines connaissances qui permet d'en accueillir de nouvelles. D'autres, comme Edelman ou Heylighen, décrivent le mécanisme de *bootstrapping*, qui permet de construire récursivement une syntaxe à partir de l'exposition répétée à des symboles et des phonèmes, sans règles préexistantes. Roschelle (1995) replace ce problème dans un contexte éducatif, celui de

la construction des connaissances scientifiques. Il présente le "paradoxe de la continuité" dans lequel la connaissance initiale, utile pour comprendre des phénomènes physiques, est à la fois nécessaire et problématique.

Ce mécanisme de construction de connaissances a également été étudié dans le domaine de l'acquisition humaine de vocabulaire à partir de l'exposition à des textes. Des recherches, synthétisées par Fukkink et de Glopper (1998), montrent que l'on parvient à acquérir incidemment du vocabulaire simplement par la lecture de textes, sans emploi de dictionnaires ou d'enseignement spécifique. Le lecteur est exposé à une grande quantité de mots à partir de laquelle il peut apprendre le sens de nouveaux mots en se fiant à leur contexte.

Dans cet article, nous utilisons LSA (*Latent Semantic Analysis*, analyse de la sémantique latente, Landauer & Dumais, 1997), un modèle informatisé de la construction de connaissances à partir de larges corpus textuels. Nous avons réalisé deux tests de ce modèle : le premier simule le processus de construction de connaissances à partir de l'exposition à un cours ; le second compare les scores de LSA à ceux d'étudiants dans une tâche de réponse à un questionnaire à choix multiple. Tout d'abord, décrivons le modèle utilisé.

## 2. Présentation de LSA

LSA<sup>6</sup> est un modèle statistique fondé sur un type d'analyse factorielle permettant d'analyser la proximité sémantique intermots ou interparagraphes à l'intérieur d'un grand ensemble d'unités d'informations textuelles. Initialement conçu pour améliorer l'efficacité de l'interrogation de systèmes documentaires informatisés (Dumais, 1991), le modèle de LSA suppose que, étant donné plusieurs "contextes" (unités d'information textuelle : phrases, paragraphes, discours...), il existe une structure latente dans l'usage des mots communs à ces contextes.

### 2.1. Description du fonctionnement de LSA

Une analyse statistique proche d'une analyse en composantes principales permet de mettre en évidence cette structure. La spécificité du modèle de LSA est que les mots et les contextes sont tout d'abord représentés dans un espace vectoriel multidimensionnel, espace qui est ensuite réduit à une centaine de dimensions. C'est cette réduction qui permet de récupérer des associations entre mots plus élaborées qu'un simple relevé de co-occurrences. Par exemple, si un mot X est souvent présent dans les mêmes contextes qu'un mot Y, lui-même souvent relevé dans les mêmes contextes qu'un troisième mot Z, alors X et Z entretiendront pour LSA une grande proximité, même s'ils ne sont pas co-occurents (voir Deerwester *et al.*, 1990, pour une présentation complète du fonctionnement de LSA).

## 2.2. Quelques tests

LSA fait l'objet, depuis une dizaine d'années, de tests visant notamment à comparer ses performances à celles d'humains (voir Landauer *et al.*, 1998a pour une synthèse). Testé tout d'abord à des fins de recherche d'informations, il l'a ensuite été sur des tâches de choix de synonymes, d'évaluation de la cohérence textuelle, et de notation de dissertations. Intéressons-nous de plus près aux tests de choix de synonymes.

Les performances de LSA dans une tâche d'acquisition de la signification des mots à partir d'un contexte correspondent assez fidèlement à des performances humaines. Le modèle de LSA a été testé en simulant l'acquisition de vocabulaire par exposition aux mots entre 2 ans et 20 ans (Landauer & Dumais, 1997). Durant cette période, on estime qu'un humain est exposé à environ 3 500 mots par jour et "apprend" de 7 à 15 mots nouveaux par jour — c'est-à-dire est capable d'évaluer correctement la proximité sémantique entre dix nouvelles paires de mots par jour. Si l'on expose LSA à un nombre équivalent de mots, il apprend 10 mots nouveaux par jour (selon les critères du test du TOEFL, *Test of English as a Foreign Language*). Ce résultat est donc cohérent avec le taux d'acquisition humain du vocabulaire par exposition (Fukkink & de Glopper, 1998).

Ainsi, nous pouvons considérer que la proximité sémantique de deux mots d'un corpus, évaluée par LSA, peut correspondre assez fidèlement à l'acquisition humaine du sens de ces deux mots. Le travail présenté ici consiste en deux tests. Le premier réplique celui de l'acquisition de vocabulaire par exposition à un cours universitaire, c'est-à-dire un corpus beaucoup moins conséquent que celui de Landauer et Dumais (1997). Le second fait passer à LSA un questionnaire à choix multiple d'un examen réel, après lui avoir fait traiter le cours correspondant.

### 3. Premier test : Exposition cumulative de LSA à un cours

Comment évolue le sens de mots lors de l'exposition de LSA à un cours ? Si l'on observe le voisinage sémantique de quelques mots (les mots dont ils sont sémantiquement les plus proches, selon LSA), constate-t-on des variations au fur et à mesure que le cours est exposé ? Nous tentons de répondre à cette question en posant que le sens d'un mot est largement déterminé par son contexte (c'est-à-dire les mots qui lui sont co-occurents), que ce sens pourra donc être modifié si la quantité de mots co-occurents varie au cours de l'exposition du cours. Ce sens pourra ainsi se stabiliser si le contexte traité est suffisant, ou, au contraire, se dégrader si les co-occurences se raréfient au fur et à mesure de l'exposition (voir Rapaport & Ehrlich, à paraître, pour une autre mise en œuvre de cette procédure, avec des moyens informatisés différents). Comme nous ne pouvons examiner le voisinage sémantique de tous les mots du corpus, nous avons sélectionné une dizaine de mots parmi les plus typiques. Détaillons maintenant la procédure utilisée.

---

<sup>6</sup> LSA est écrit en langage C et fonctionne sur une station de travail Unix, il est déposé en 1990 par *Bell Communications Research Inc.* Voir <http://lsa.colorado.edu> pour des informations sur ce logiciel.

### 3.1. Procédure

Le corpus utilisé est un cours de sociologie de l'éducation de niveau licence (230 ko, environ 28 000 mots). Nous avons traité le cours entier avec LSA et récupéré les dix mots les plus centraux, hors mots-outils — un mot est d'autant plus central, qu'il est typique, c'est-à-dire sémantiquement proche d'un plus grand nombre de mots.

Ensuite, nous avons récupéré, pour chacun de ces mots, les mots calculés par LSA comme sémantiquement les plus proches de ces derniers (toujours en supprimant les mots-outils), en filtrant arbitrairement les mots de proximités supérieures à 0,62, ce afin d'en conserver une dizaine pour chaque mot-cible<sup>7</sup>. Nous avons ensuite observé l'évolution de la proximité sémantique de ces paires de mots, en introduisant cumulativement 10 % du cours dans le corpus traité par LSA. Cela nous donne, pour chaque mot, une courbe d'évolution de ses proximités, en fonction de la taille du cours traitée par LSA. Nous avons réalisé des régressions simples sur chaque courbe afin d'en récupérer la pente, et ainsi avoir un indicateur de l'évolution de la proximité sémantique des paires de mots.

### 3.2. Résultats

Cette exposition de LSA à un contenu de cours fait varier les proximités sémantiques des mots qui le composent et nous observons que cette variation n'est pas identique pour toutes les paires de mots. Trois types d'évolution de proximités peuvent être ainsi dégagés (voir tableau 1 et figure 1, page suivante).

*Une proximité croissante* au fur et à mesure que le cours est exposé (pente supérieure ou égale à 0,008, valeur arbitraire). Ces couples de mots dont la proximité sémantique relative évolue de manière croissante tout au long de l'exposition au cours sont majoritaires.

*Une proximité décroissante* (pente inférieure ou égale à - 0,008, valeur arbitraire). Ces couples de mots, peu nombreux, ont une proximité sémantique relative qui évolue de manière décroissante tout au long du cours.

Nous avons ici, dans ces deux catégories, des couples de mots dont le voisinage sémantique évolue suivant l'exposition au cours. Même si certaines notions centrales du cours apparaissent dans ces deux catégories (par exemple "réussite scolaire", "performance [des] élèves", "origine sociale"), elles sont moins susceptibles de s'y trouver que dans la catégorie suivante.

*Une proximité constante* (comprise strictement entre les deux valeurs précédentes). Les couples de mots, peu nombreux, dont la proximité sémantique relative n'évolue pas ou peu tout au long de l'exposition au cours appartiennent donc au même voisinage sémantique tout au long du cours. On retrouve dans cette catégorie des notions comme "effet-écoles",

"efficacité [des] écoles", "performances [des] écoles". Ces couples correspondent pour la plupart à des notions centrales du cours, puisque leur voisinage sémantique ne varie pas au long de l'exposition.

### 3.3. Interprétation

Proposons maintenant une interprétation de ces phénomènes observés. Bereiter et Scardamalia (1993) proposent une distinction intéressante entre deux types de construction des connaissances : l'assimilation directe d'une information nouvelle (*direct assimilation*) et les schémas (*knowledge-building schemas*).

Dans le premier type, le sujet, ne disposant pas des connaissances appropriées pour appréhender la connaissance nouvelle, la transforme immédiatement sous une forme assimilable, bien qu'inexacte, par exemple par le biais de synonymes. Dans une étude de Carey (1978, citée par Bereiter & Scardamalia, 1993), des enfants, lorsqu'on leur montre pour la première fois une nouvelle couleur, olive, s'y réfèrent en tant que couleur verte.

Dans le second type, la construction de schéma, le sujet raffine progressivement la connaissance qu'il a d'un nouveau mot, jusqu'à ce qu'elle soit appropriée, sans tenter immédiatement des associations avec des connaissances anciennes. Dans l'étude de Carey, des sujets utilisent cette couleur et ce nom inappropriés "odd color, odd name", le temps qu'ils comprennent qu'il s'agit d'une nouvelle couleur.

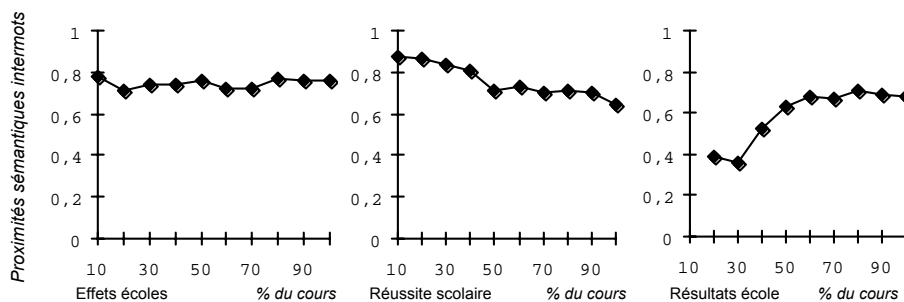
Ces deux types de construction de la connaissance peuvent correspondre respectivement, dans notre test, aux courbes décroissantes et croissantes. Dans le premier cas, les proximités intermots sont de moins en moins élevées, au fur et à mesure de l'exposition au cours, comme si LSA raffinait progressivement le voisinage sémantique du mot-cible, en corrigeant des valeurs initiales trop élevées. Dans le deuxième cas, le sens du mot-cible converge, au fur et à mesure de l'exposition au cours, vers une proximité maximale stable avec d'autres mots.

La seule exposition à un corpus relativement restreint permet à LSA de simuler une construction de connaissances dans le domaine de ce corpus. Cette évolution de proximités sémantiques intermots est un indicateur de la manière dont le vocabulaire est acquis : au fur et à mesure de l'exposition à un contexte, LSA est sensible à la richesse sémantique du vocabulaire, qui est fonction du nombre de liens entre mots (Baker, Simmons & Kameenui, 1995). Toutefois, il convient maintenant de comparer directement les performances de LSA à celles d'humains. C'est l'objet du deuxième test que nous présentons maintenant.

<sup>7</sup> Les dix mots les plus centraux sont : "scolaire", "école", "élèves", "effets", "travaux", "écoles", "réussite", "scolaires", "enseignement", "sociale". "Travaux" n'apparaît pas dans la suite de ce test, car aucun mot du corpus n'entretient avec lui une proximité sémantique supérieure au seuil défini.

**Tableau 1** — Quelques exemples d'évolutions de proximités intermots et la valeur de leurs pentes, en italiques, les paires de mots représentées dans la figure 1 (les évolutions croissantes ne sont pas toutes représentées ici). Le premier mot de chaque paire est un des mots les plus typiques du corpus

Pente nulle		Pente négative (= - 0,008)		Pente positive (= 0,008) (extraits)	
Paires de mots	Pente	Paires de mots	Pente	Paires de mots	Pente
sociale individus	-0,007	scolaire enseignement	-0,027	scolaire classe	0,008
sociale scolaire	-0,006	<i>réussite scolaire</i>	<i>-0,025</i>	sociale origine	0,009
réussite enseignement	-0,006	sociale société	-0,021	élèves différences	0,013
élèves apparaissent	-0,005	élèves variables	-0,017	école exemple	0,017
écoles efficacité	-0,003	écoles acquisitions	-0,017	écoles recherches	0,017
écoles performances	-0,002	réussite social	-0,016	[...]	[...]
<i>effets écoles</i>	<i>0,001</i>	élèves performances	-0,015	effets études	0,033
école classe	0,002	scolaires effet	-0,014	réussite résultats	0,034
école différences	0,003	sociale position	-0,014	effets élèves	0,035
effets résultats	0,003	sociale sociales	-0,013	élèves écoles	0,037
école réussite	0,003	scolaires acquisitions	-0,010	<i>résultats école</i>	<i>0,042</i>



**Figure 1** – Évolution des proximités sémantiques selon le taux d'exposition de LSA au cours, respectivement nulle (effets écoles), négative (réussite scolaire) et positive (résultats école)

#### 4. Second test : Construction de connaissances évaluées par QCM

Nous voulons ici comparer les résultats d'étudiants à un examen à ceux de LSA, sous la forme de réponses à un questionnaire à choix multiple. Nous avons utilisé ici un cours de didactique des sciences et des mathématiques de niveau licence (119 ko, environ 17 000 mots). Nous avons de plus récupéré les copies d'examen des trente-cinq étudiants inscrits à ce cours, examen qui consistait en des réponses à une trentaine de questions à choix multiple.

Si LSA est un modèle adéquat de la construction de connaissances à partir de textes, nous devrions obtenir, d'une part un score au questionnaire proche de la moyenne des scores humains. D'autre part, les choix des items par LSA devraient être proches de ceux des étudiants. Ce test reprend la méthode d'un travail de Landauer *et al.* (1998b, cités par Landauer *et al.*, 1998a ; et Landauer *et al.*, 1998c), où ces derniers font sélectionner à LSA des réponses à un questionnaire à choix multiples après lui avoir fait traiter un manuel

d'introduction à la psychologie. LSA obtient 60 % de bonnes réponses, soit une valeur nettement supérieure à celle du hasard (25 %), mais toutefois inférieure à celle nécessaire pour réussir l'examen.

##### 4.1. Procédure

Nous avons fait traiter le cours par LSA, puis chacune des vingt-sept questions<sup>8</sup> de l'examen a été tour à tour traitée, afin de calculer sa proximité sémantique avec chacune des cinq réponses possibles. La réponse ayant la proximité maximale a été relevée comme étant la réponse à la question " choisie " par LSA, suivant en cela la procédure décrite par Landauer et Dumais (1997) à propos du test du TOEFL.

Voici un exemple de question et les valeurs de proximité sémantique, calculées par LSA, entre chaque question candidate et l'énoncé de la question. Dans cet exemple,

<sup>8</sup> Nous avons supprimé trois questions énonçant des problèmes arithmétiques, donc non traitables par LSA.

la réponse choisie est la réponse E, puisqu'elle est la plus proche de l'énoncé. C'est également la réponse exacte.

Énoncé : Les principes de la causalité naturelle sont...

- A. Des principes causaux qui existent réellement dans la nature. (0,37),
- B. Des principes que tous les scientifiques utilisent. (0,32),
- C. Des principes des scientifiques modernes. (0,27),
- D. Des principes qui ont fait leur preuve mais qu'on n'utilise plus beaucoup. (0,37),
- E. Des principes inexacts qu'on a naturellement tendance à appliquer. (0,47).

Deux types de résultats sont mis au jour, tout d'abord la note globale de LSA au questionnaire, obtenue en sommant les bonnes réponses. Ensuite, comme LSA permet de calculer la proximité sémantique d'une question avec chacune des cinq réponses possibles, il a été calculé la corrélation entre ces valeurs et la distribution des choix des étudiants.

#### 4.2. Résultats

La note qu'obtient LSA au test est de 12/27 (soit 44,4 % de bonnes réponses) ce qui correspond à une note nettement inférieure à la note moyenne des étudiants (18,8/27 ; écart-type 3,7). Cette note est d'une part nettement supérieure à celle du hasard (5,4/27 ou 20 % de bonnes réponses). D'autre part, elle est comparable à celle obtenue par Landauer *et al.* (1998b). En d'autres termes, seuls trois étudiants ont une note inférieure ou égale à celle de LSA.

La corrélation entre les scores de proximité questions-items et les effectifs des réponses des étudiants est faible :  $r = .30$  ;  $p < .0001$  (corrélation des rangs). Afin de mieux comprendre le fonctionnement de LSA, nous avons ensuite distingué les questions réussies par LSA de celles qui le mettaient en échec. Les valeurs de corrélations entre les valeurs de proximité de LSA et des distributions des réponses des étudiants concernant les bonnes et les mauvaises réponses sont respectivement de  $.50$  ( $p < .0001$ ) et de  $.09$  (n.s.). Cela signifie que les erreurs de LSA ne sont pas liées aux erreurs des étudiants, alors que ses réussites, elles, sont fortement liées à celles des étudiants.

#### 4.3. Commentaires

LSA obtient de moins bons résultats à ce test qu'à celui du test de synonymie du TOEFL. Toutefois, le contexte n'est pas le même : la base de connaissances pour répondre au test de synonymie est un large corpus encyclopédique de 4,5 millions de mots, alors que le nôtre est un cours de faible taille. L'exercice n'est pas non plus le même : le test du TOEFL est conçu pour une évaluation de connaissance du vocabulaire alors que le nôtre est une évaluation de connaissances de notions de cours. Enfin, le matériau même du test diffère : les items du test de synonymie du TOEFL comportent un court texte d'une dizaine de lignes, ce qui permet au lecteur de tenir compte d'un contexte plus riche que celui

de notre test, où seule la question d'une dizaine de mots est comparée tour à tour à chaque réponse possible.

Il ne fait aucun doute que les étudiants, lorsqu'ils répondent à un tel questionnaire, font appel à bien d'autres connaissances que celles du cours. Ils en suivent d'autres sur des domaines connexes et lisent d'autres ouvrages que le cours. Ils bénéficient, en outre, des réponses aux éventuelles questions des étudiants pendant le cours, qui, elles, ne sont pas intégrées à ce dernier. Ainsi, il est normal que les performances de LSA soient quelque peu en deçà de celles des humains.

### 5. Discussion

Discutons ici les principaux résultats de nos tests en les reliant aux discussions de la littérature. Nous commençons par reprendre les deux critiques les plus souvent rencontrées dans la littérature à propos de LSA : celle de la nature du corpus initial et celle de sa non-prise en compte des informations syntaxiques, nous continuerons par la discussion des résultats du deuxième test.

#### 5.1. Le rôle de la nature du corpus

Le corpus textuel traité par LSA joue bien évidemment un rôle primordial dans la qualité de l'analyse sémantique. À notre connaissance, son influence n'a pas été encore rigoureusement testée, et son choix dépend plutôt des capacités de traitement de l'ordinateur utilisé. Les résultats à ce sujet sont donc très empiriques. À la question "quelle taille doit avoir le corpus à traiter ?", on est tenté de répondre, comme Perfetti (1998) : "la plus large possible". Cette réponse est insatisfaisante, car non réaliste si l'on travaille dans le domaine de l'éducation où les corpus à traiter — notamment les cours — ont une taille relativement faible. Dans nos tests, seul le contenu du cours a été traité par LSA. Aucune connaissance de la langue, sous la forme d'autres textes, n'a été ajoutée, afin de ne pas allonger la durée de traitement, déjà conséquente. Cette absence rend la comparaison avec des performances humaines moins pertinente. Mais la question se pose alors de l'éventuel corpus à ajouter au cours, afin d'ajouter de la "connaissance" à LSA. Cette question aussi soulève plus de problèmes qu'elle n'en résout : si l'on ajoute de nombreux textes hors du domaine, on risque de diluer les connaissances du cours proprement dit ; si l'on ajoute des textes du domaine, on risque de modifier plus ou moins profondément les connaissances de la base.

Autre problème inhérent à notre premier test, au fur et à mesure que l'on ajoute 10 % du cours, ces 10 % nouvellement introduits ont un poids de moins en moins important par rapport au cours déjà traité. Or, on a déjà noté par ailleurs (Dessus, 1999) l'influence de la taille du corpus traité sur l'évaluation des proximités intermots. Toutefois, pondérer cet apport de 10 % aurait rendu la simulation moins proche de l'acquisition humaine de connaissances.

Enfin, il nous faut souligner que ces tests ont été réalisés avec un corpus de textes en français, alors que la grande majorité des tests existants utilisent des corpus en anglais. Le fait que les performances relevées ici soient voisines des tests anglais montrent que le modèle de LSA fonctionne de manière relativement indépendante de la langue du corpus traité.

## 5.2. La non-prise en compte de la syntaxe

Comme LSA transforme les textes qu'il traite en paquets de mots, les informations sur la syntaxe sont perdues. De nombreuses critiques, souvent justifiées, ont été faites à ce sujet. Perfetti (1998) signale par exemple que LSA ne pourrait pas identifier un texte d'un auteur original d'un texte pastiché. C'est exact, mais ce qui fait l'intérêt du pastiche, c'est justement qu'un humain a également de la peine à faire cette identification. De manière plus générale, les informations syntaxiques pourraient ne pas être indispensables pour établir une information sémantique, même si elles la déterminent en partie. Nous renvoyons à ce sujet à la revue de Redington et Chater (1998), qui montrent l'intérêt théorique et pratique de méthodes connexionnistes et statistiques pour l'étude de l'acquisition humaine du langage.

## 5.3. Évaluer les connaissances par QCM

Bien évidemment, l'évaluation de connaissances par QCM est insuffisante à montrer toute la richesse de ces dernières. L'utilisation de connaissances dans des situations plus ouvertes et complexes est nécessaire. Des travaux, qu'il serait trop long de décrire ici, ont montré que les notes de dissertations données par des juges humains corrôlaient fortement avec celles données par LSA (Foltz *et al.*, 1999 ; Lemaire & Dessus, à paraître).

## 5.4. Les erreurs de LSA sont-elles différentes de celles des humains ?

Vérifier si LSA traite la connaissance de la même manière que les humains passe aussi par la vérification que ses erreurs sont humaines. Sur ce point, notre deuxième test produit un résultat négatif : les valeurs de LSA corrôlent assez fortement avec les valeurs choisies par les étudiants, mais seulement pour les questions où LSA répond correctement. En revanche, là où LSA répond à côté, les étudiants ont une distribution de choix très différente.

## 6. Conclusion : LSA et les sciences cognitives

Le mécanisme utilisé par LSA (réduction de vecteurs au sein d'un espace multidimensionnel) est connu depuis longtemps comme un modèle efficace de processus psychologiques (par exemple, chez Osgood et son différenciateur sémantique). Perfetti (1998) a écrit, pour le critiquer, que LSA faisait tout assez bien. Il nous semble que cela est plutôt une qualité. En effet, le mécanisme de LSA est relativement simple et autorise une utilisation dans de nombreux domaines des sciences cognitives.

Les performances de LSA dans le domaine de la recherche documentaire ont été évaluées depuis longtemps. Plus récemment, leurs auteurs (Landauer, 1999 ; Landauer & Dumais, 1997) le promeuvent en tant que théorie — et non plus seulement outil — du langage et de l'esprit, s'appuyant pour cela sur le nombre croissant d'études empiriques, souvent prometteuses (Landauer *et al.*, 1998a). Toutefois, certaines études récentes sont plus mitigées. Desai (1998) montre notamment que LSA ne peut distinguer les synonymes des antonymes d'un mot-cible. Robertson et Glenberg, (s. d.) montrent que LSA ne parvient pas à distinguer, lorsqu'on lui présente une description de situation, sa suite vraisemblable de celle qui présente un non-sens. De plus, LSA est

incapable de distinguer la paraphrase d'un proverbe d'un énoncé de sens contraire.

Nous avons montré ici que LSA est un moyen utile pour déterminer automatiquement l'évolution de l'utilisation, au sein d'un document didactique, des différents mots qui le composent. Il pourrait être utilisé pour repérer les notions principales traitées dans un cours, ainsi que pour assister la correction de questionnaires ouverts.

Le fait que des chercheurs de champs disciplinaires divers — recherche documentaire, psycholinguistique, intelligence artificielle et éducation — s'intéressent à LSA montre qu'il peut avoir sa place dans le champ des sciences cognitives.

## Remerciements

Nous remercions Benoît Lemaire, Jacques Baillé, Sylvain Dionnet et Catherine Pellenq pour leurs commentaires d'une précédente version de cet article, ainsi que Pascal Bressoux et Christian Dépret pour nous avoir communiqué leurs cours.

## Références bibliographiques

- [Baker *et al.*, 1995] Baker, S. K., Simmons, D., Kameenui, E. J. (1995). Vocabulary acquisition: synthesis of the research. Eugene : Université de l'Oregon, *rapport technique n° 13*.
- [Bereiter, 1985] Bereiter, C. (1985). Toward a solution of the learning paradox. *Review of Educational Research*. 55(2). 201-226.
- [Bereiter et Scardamalia, 1993] Bereiter, C., Scardamalia, M. (1993). *Surpassing ourselves, an inquiry into the nature and implications of expertise*. Open Court : Chicago.
- [Carey, 1978] Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, G. A. Miller (Eds), *Linguistic Theory and Psychological Reality*. MIT Press : Cambridge. 265-293.
- [Chomsky et Fodor, 1979] Chomsky, N., Fodor, J. (1979). Exposé du paradoxe. In M. Piattelli-Palmarini (Ed.), *Théories du langage, théories de l'apprentissage*. Seuil : Paris. 379-382.
- [Deerwester *et al.*, 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*. 41(6). 391-407.
- [Desai, 1998] Desai, R. (1998). *A review of two statistical models of knowledge representation, acquisition, and memory*. Bloomington : Université de l'Indiana, rapport de recherche non publié.
- [Dessus, 1999] Dessus, P. (1999). Vérification sémantique de liens hypertextes avec LSA. In J.-P. Balpe, A. Lelu, S. Natkin, I. Saleh (Eds), *Hypertextes, hypermédiâs et internet (H2PTM'99)*. Hermès : Paris. 119-129.
- [Dumais, 1991] Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*. 23(2), 229-236.
- [Edelman, 1992] Edelman, G. M. (1992). *Biologie de la conscience*. Paris : Seuil, coll. Points.
- [Foltz *et al.*, 1999] Foltz, P. W., Laham, D., Landauer, T. K. (1999). Automated essay scoring : applications to Educational Technology. *Proc. ED-MEDIA'99*. Seattle.
- [Fukkink et de Glopper, 1998] Fukkink, R. G., de Glopper, K. (1998). Effects of Instruction in deriving word meaning from context : a meta-analysis. *Review of Educational Research*. 68(4). 450-469.

[**Goodman, 1984**] Goodman, N. (1984). *Faits, fictions et prédictions*. Paris : Minuit.

[**Heylighen, sous presse**] Heylighen, F. (sous presse). Bootstrapping knowledge representations: from entailment meshes via semantic nets to learning webs. *International Journal of Human-Computer Studies*.

[**Landauer, 1999**] Landauer, T. K. (1999). Latent semantic analysis : a theory of the psychology of language and mind. *Discourse Processes*. 27(3). 303-310.

[**Landauer et Dumais, 1997**] Landauer, T. K., Dumais, S. T. (1997). A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*. 104. 211-240.

[**Landauer et al., 1998a**] Landauer, T. K., Foltz, P. W., Laham, D. (1998a). An Introduction to Latent Semantic Analysis. *Discourse Processes*. 25. 259-284.

[**Landauer et al., 1998b**] Landauer, T. K., Foltz, P. W., Laham, D. (1998b). *Latent Semantic Analysis passes the test: Knowledge representation and multiple-choice testing*. Manuscrit non publié.

[**Landauer et al., 1998c**] Landauer, T. K., Laham, D., Foltz, P. W. (1998c). Learning human-like knowledge by singular value decomposition: a progress report. In M. I. Jordan, M. J. Kearns, S. A. Solla (Eds), *Advances in Neural Information Processing Systems*. 10. 45-51.

[**Lemaire et Dessus, à paraître**] Lemaire, B., Dessus, P. (à paraître). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*.

[**Perfetti, 1998**] Perfetti, C. A. (1998). The limits of co-occurrence : tools and theories in language research. *Discourse Processes*. 25(2/3). 363-377.

[**Popper, 1998**] Popper, K. R. (1998). *La connaissance objective*. Flammarion, coll. Champs : Paris.

[**Quine, 1977**] Quine, W. v. O. (1977). *Le mot et la chose*. Flammarion : Paris .

[**Rapaport et Ehrlich, à paraître**] Rapaport, W. J., Ehrlich, K. (à paraître). A computational theory of vocabulary acquisition. In S. C Iwanska, S. Shapiro (Eds), *Natural language processing and knowledge representation : language for knowledge and knowledge for language*. MIT Press : Cambridge.

[**Redington et Chater, 1998**] Redington, M., Chater, N. (1998). Connectionist and statistical approaches to language acquisition : A distributional perspective. *Language and Cognitive Processes*. 13(2/3). 29-191.

[**Robertson et Glenberg, s. d.**] Robertson, D. A., Glenberg, A. M. (s. d.). *Grounding symbols and computing meaning: A supplement to Glenberg & Robertson*. Madison : Université du Wisconsin, rapport non publié disponible sur la toile à l'URL : [http://psych.wisc.edu/glenberg/supplement\\_jml\\_g&r.html](http://psych.wisc.edu/glenberg/supplement_jml_g&r.html)

[**Roschelle, 1995**] Roschelle, J. (1995). Learning in interactive environments: prior knowledge and new experience. In J. Falk, L. Dierking (Eds). *Public institutions for personal learning: establishing a research agenda*. The American Association of Museums.

[**Rosenfield, 1994**] Rosenfield, I. (1994). *L'invention de la mémoire*. Flammarion, coll. Champs : Paris.

[**Salomon, 1993**] Salomon, G. (Ed.) (1993). *Distributed Cognitions*. Cambridge University Press : Cambridge.

[**Smith, 1993**] Smith, L. (1993). *Necessary knowledge, piagetian perspectives on constructivism*. Erlbaum : Hove.

[**Spiro, 1991**] Spiro, R. (1991). Constructivism, old and new : cognitive flexibility theory and the promotion of advanced knowledge acquisition. *Educational Technology*. 11(5). 24-33.

## L'auteur



Philippe Dessus est maître de conférences en sciences de l'éducation à l'IUFM de Grenoble et chercheur au laboratoire de sciences de l'éducation de la même ville. Au sein de ce laboratoire, il a soutenu, en 1994, une thèse sur la planification de séquences d'enseignement assistée par ordinateur. Depuis, il consacre la majeure partie de ses recherches sur diverses expérimentations d'applications éducatives d'un modèle informatisé d'acquisition des connaissances, Latent Semantic Analysis.

