

Titre	Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition d'une langue étrangère de spécialité		
Sous-titre	<i>Sous-titre de l'article</i>		
Auteur	Virginie Zampa	Labo / Université	LSE Université Pierre-Mendès-France Grenoble
	Mél Virginie.Zampa@upmf-grenoble.fr	Adresse	
Mots-clés	LSA, acquisition, langue étrangère de spécialité		
Abstract	<p>Cet article présente l'utilisation de LSA (Latent Semantic Analysis) dans un prototype d'acquisition de langue étrangère de spécialité. Ce prototype nommé RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) a pour finalité d'optimiser l'acquisition d'une langue étrangère en fournissant des lectures à l'apprenant. Les textes fournis à l'apprenant dépendent de ses connaissances ainsi que des connaissances du domaine de spécialité. L'article est divisé en trois parties, la première présente LSA, son fonctionnement et ses utilisations. La seconde partie présente RAFALES, son architecture et son fonctionnement. Enfin la dernière partie présente les résultats de l'expérimentation du prototype auprès d'apprenants ainsi que les raisons qui nous ont poussé à utiliser LSA dans RAFALES.</p>		

1. Introduction

Pour acquérir des connaissances dans une langue étrangère, ou une langue étrangère de spécialité, un apprenant a le plus souvent recours à des livres ou manuels. Mais il est rare qu'ils répondent pleinement à ses attentes : ils ne traitent que peu de la langue de spécialité, ils ne sont pas adaptés à son niveau de connaissances du domaine de spécialité ou de la langue étrangère.

Dans cet article, nous allons présenter notre prototype nommé RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) qui est fondé sur l'utilisation de LSA (Latent Semantic Analysis).

Nous allons dans un premier temps exposer les principes de LSA puis, dans un second temps nous présenterons RAFALES, c'est-à-dire son architecture, son fonctionnement ainsi qu'une expérimentation. Enfin nous justifierons notre choix d'utilisation d'une analyse automatique de la sémantique dans notre prototype d'acquisition de langue étrangère.

2. Analyse de la sémantique latente (LSA)

LSA a été développé par les laboratoires Bellcore en 1989, comme outil de recherche documentaire (Deerwester et al, 90). Mais très vite, grâce à ses performances, son utilisation s'est étendue à d'autres domaines comme nous allons le voir.

2.1. La méthode

LSA s'appuie sur une représentation multidimensionnelle de la signification des mots dans la langue. Grâce à une analyse statistique, le sens de chaque mot est caractérisé par un vecteur dans un espace de grande dimension, avec la propriété que la proximité entre deux vecteurs (leur cosinus) correspond à la proximité de sens des mots qu'ils représentent. Le modèle d'apprentissage prend donc en entrée un ensemble de textes et prédit les proximités qui vont résulter de la lecture de ces textes.

LSA analyse l'ensemble des textes source pour en représenter les mots dans un espace sémantique multidimensionnel. Cette analyse statistique (présentée plus loin) permet de faire ressortir les relations sémantiques entre mots ou entre textes. Deux mots peuvent être considérés sémantiquement proches s'ils sont utilisés dans des contextes similaires. Le contexte d'un mot est ici défini comme l'ensemble des mots qui apparaissent conjointement à lui. Ainsi, les mots « vélo » et « bicyclette » sont considérés comme sémantiquement proche puisqu'ils apparaissent tout les deux avec des mots tels que « guidon », « pédaler », etc. et ils n'apparaissent que rarement avec des mots comme « ordinateur », « bouilloire », etc. Cette notion de cooccurrence est statistique : la méthode fonctionne si un nombre suffisant de textes est utilisé. Mais il ne s'agit pas simplement de comptage, il faut aussi disposer d'une procédure pour établir les liaisons sémantiques. Cette procédure est la réduction de la matrice.

Le principe est le suivant. LSA construit dans un premier temps la matrice d'occurrences. Il s'agit d'une matrice dont les lignes représentent les unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes les mots (plus précisément les graphies). L'élément (i,j) de la matrice correspond ainsi au nombre d'occurrences du mot j dans le paragraphe i . L'étape suivante consiste à réduire ces dimensions à environ 200. Ce nombre est important car une réduction à un espace trop grand ne fait pas suffisamment émerger les liaisons

sémantiques entre les mots, et un espace trop petit conduit à une trop grande perte d'informations. Ce nombre de dimensions est issu de tests empiriques (Deerwester et al., 90). Cette réduction est réalisée par le biais d'une décomposition aux valeurs singulières. La réduction à n dimensions va consister à ne conserver que les n premières de ces valeurs pour reconstituer une matrice approchée, de dimension n . Chaque mot et chaque paragraphe, traité de la même façon dans cette procédure, est ainsi représenté par un vecteur à n dimensions.

L'espace sémantique construit, il faut choisir une mesure appropriée afin de mesurer la proximité entre deux éléments. Les tests empiriques ont privilégié la méthode du cosinus. La proximité entre deux vecteurs est le cosinus de leur angle. La proximité sémantique entre deux mots, entre deux paragraphes ou entre un mot et un paragraphe est donc une valeur entre -1 et 1 où 1 indique une très forte proximité sémantique.

2.2. Quelques applications et validations

Au départ, LSA a été développé comme outil de recherche d'information (Dumais, 94 ; Dumais, 97). Avec les problèmes de choix de mots-clés liés à la polysémie, aux inflexions et à la synonymie, il est aisé de postuler que la recherche devrait se faire sur le sens des mots et non sur leur « forme ». D'ailleurs des expérimentations avec LSA ont mis en évidence un gain allant de 16 à 30%.

Un second domaine d'application est l'apprentissage. Ce modèle a été testé par Landauer et Dumais (1997). Ils ont simulé l'acquisition entre 2 et 20 ans, ce qui correspond à une exposition de 3500 mots par jour et un apprentissage de 7 à 15 mots nouveaux par jour. Avec une même exposition, qui correspond à un corpus de 4,6 millions de mots tirés d'une encyclopédie, LSA apprend 10 mots par jour. LSA passe les tests de synonymie du TOEFL, en choisissant, parmi les quatre mots proposés, le plus proche du mot initial dans l'espace sémantique. Il obtient un résultat (64,4% de bonnes réponses) comparable à la moyenne des sujets non anglophones admis dans les universités américaines (64,5%). Le comportement de LSA est donc un modèle intéressant de l'apprentissage du vocabulaire chez les humains.

Un troisième domaine d'application concerne l'acquisition de connaissances. Ces acquisitions peuvent concerner les langues (Redington et Chater, 98) ou un domaine particulier comme celui traité dans un cours (Dessus, 00). Elles peuvent concerner un langage et non une langue naturelle. C'est le cas par exemple avec l'apprentissage des jeux tels que le tic-tac-toe (Lemaire, 98) ou kalah (Lemaire, 99).

LSA est aussi utilisé de diverses manières dans des EIAH (environnement informatique d'apprentissage humain). Des travaux ont ainsi porté sur l'évaluation de copies, c'est le cas du système APEX (Aide à la Préparation aux EXamens) (Dessus et Lemaire, 99 ; Dessus et al. 00). D'autres travaux ont porté sur la notation de copies (Foltz et al, 99) dans lesquelles l'étudiants rédigeait une synthèse. La corrélation entre LSA et les juges humains est équivalente à celle entre juges humains. D'autres travaux ont porté sur la modélisation de l'apprenant (Zampa et Lemaire, 01 ; Zampa et Raby, 01) et la détection des erreurs dans une copie en langue étrangère (Zampa et Lemaire, 01).

3. Le prototype RAFALES

RAFALES est un Recueil de lecture qui se crée et évolue en fonction de l'utilisateur c'est-à-dire en prenant en compte son niveau et ses lectures. Dans RAFALES, l'unique tâche de l'apprenant est la lecture. Ce choix n'est pas neutre. En effet, des travaux indiquent que les enfants entre 2 et 20 ans apprennent en moyenne 10 mots nouveaux par jour, soit 3650 par an. Or seulement 100 mots par an sont acquis par instruction directe. De plus, la plupart des mots n'apparaissent pas à l'oral (l'oral ne correspond qu'à moins d'un quart de l'écrit). Ceci indique donc que la plupart des mots sont acquis par la lecture.

Nous pensons que cette acquisition peut être « optimisée » en fournissant à l'apprenant des textes qui sont à la proximité optimale d'acquisition (POA) de son profil. Cette POA correspond à une tentative de modélisation de la zone proximale de développement (Vygotsky, 1985).

RAFALES possède l'architecture d'un tuteur intelligent (Wenger, 87). Il comporte trois modules : le profil de l'apprenant, la base de connaissances du domaine étudié et le module pédagogique. Dans les trois modules, nous utilisons LSA. De ce fait, la base de connaissances du domaine étudié ainsi que le profil de l'apprenant sont fabriqués uniquement à partir de textes et sont des espaces sémantiques à 300 dimensions.

3.1. Fonctionnement

Le fonctionnement est simple. Le module pédagogique sélectionne en fonction des connaissances de l'apprenant les textes de la base de connaissances à lui fournir. Le profil de l'apprenant est un sous-espace de l'espace des connaissances du domaine étudié. Le module pédagogique sélectionne les textes qui sont ni trop proches ni trop éloignés. En d'autres termes, il sélectionne les textes de la base de connaissances qui se trouve à la POA du profil de l'apprenant.

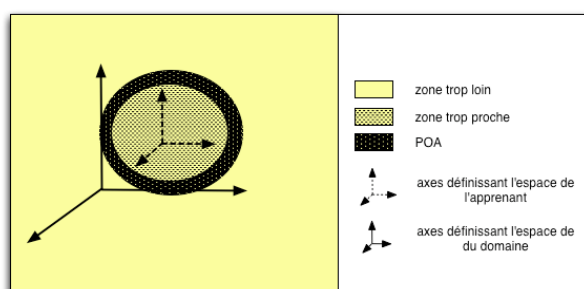


Figure 1 : les différents espaces et la POA

RAFALES fournit ces textes à l'apprenant qui les lit. Quand les textes sont lus, RAFALES met à jour le profil de l'apprenant en lui ajoutant ces textes et en recompilant l'espace. Le module pédagogique de RAFALES choisit les textes de la session suivante en fonction de ce nouveau profil de l'apprenant, et les donne à l'élève, etc. La boucle s'interrompt quand l'espace de l'apprenant est identique à celui du domaine.

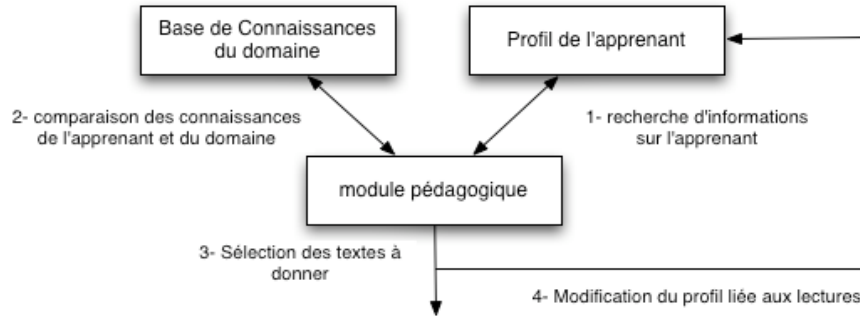


Figure 2 : architecture et fonctionnement de RAFALES

3.2. Initialisations pour l'expérimentation

Dans notre expérimentation principale du prototype, le domaine étudié est l'anglais juridique et plus particulièrement le droit constitutionnel américain.

La base de connaissances du domaine étudié comporte deux parties : une pour la langue étrangère « générale » et une pour la langue de spécialité. La base de connaissance de la langue étrangère générale contient environ 1 000 000 mots issus de huit oeuvres complètes. Ces oeuvres font partie du domaine public mais sont relativement récentes. Quant à la base de connaissance de la langue de spécialité, elle contient un peu plus de 1 000 000 de mots issus de textes de loi, de comptes-rendus de procès, etc.

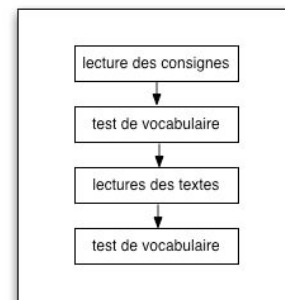
Le profil de l'apprenant est initialisé en fonction de nos sujets. Quarante-deux sujets ont passé notre expérimentation dans les temps impartis. Il s'agit de 19 étudiants de licence et maîtrise de langue étrangère et 23 stagiaires d'IUFM. Ils ont été répartis dans quatre groupes expérimentaux. Pour homogénéiser leur répartition, nous avons utilisé leurs notes (étudiants) ou leur classement au CAPES (stagiaires). Le profil de l'apprenant de départ est identique pour les quatre groupes. Il contient 1 000 000 de mots pour la langue générale et les 25 textes les plus centraux de la langue étrangère de spécialité. En effet, nous considérons que le million de mots correspond à ce à quoi ils ont été exposés au cours de leurs études, mais ils n'ont que très peu de connaissances dans la langue de spécialité (d'où les 25 textes).

3.3. Expérimentation et résultats

3.3.1. Conditions de passation et tests

Les sujets avaient deux semaines pour faire les cinq séances. Pour chaque séance, les sujets disposaient des consignes de début et de fin de séance, du test de vocabulaire de début et de fin et des textes à lire.

Les quarante-deux sujets étaient répartis dans quatre groupes expérimentaux : un groupe lisait les textes les plus éloignés de son profil, un les textes les plus proches, un des textes sélectionnés aléatoirement et enfin le dernier groupe lisait les textes à la POA.



3.4. Les tests utilisés

En début et fin de chaque séance, les sujets passaient un test de vocabulaire. Au sein d'une même séance ce test est identique, mais il varie d'une séance à l'autre.

Dans ces tableaux « 1 » signifie « même sens », 2 « même domaine », 3 « sens différent », 4 « pas de relation » et ? « mot inconnu ». De plus, le sujet indique s'il s'agit d'une relation forte (+) ou faible (-). Dans l'exemple du tableau ci-contre, la croix indique que le sujet juge que les mot clip et blow entretienne une relation forte et de même sens. Ce qui correspond à une relation de synonymie.

clip	1		2		3		4	?
	+	-	+	-	+	-		
blow	X							
cut								
magazine								
stroke								
film								

3.5. Analyse des résultats

Les réponses sont codées sur une échelle allant de 0 à 4 où 0 correspond à aucune relation entre les deux mots et 4 à une relation forte (synonymie ou antinomie). De plus nous n'analysons que vingt tableaux sur trente. Nous ne prenons pas en compte les 5 premiers et les cinq derniers afin de limiter les effets de primauté et de récence. Nous avons ainsi 100 couples de mots par test, soit 200 par séance, soit 1000 par sujet.

Nos tests ne permettent pas une correction en termes de vrai / faux. De ce fait, nous avons établi une norme ou réponse normale à partir des réponses de vingt-cinq experts du domaine. Pour chaque couple de mots, la norme correspond à la moyenne des vingt-cinq réponses des experts.

3.5.1. Analyse quantitative

Une première analyse indique qu'une partie des réponses données par les sujets diffère entre le pré-test et le post-test. Ces moyennes vont de 24% (groupe aléatoire) à 33,75% (groupe loin). Mais il s'agit de modification ne regardant pas si la réponse du sujet se rapproche ou s'éloigne de la réponse normale. Toutefois cette évolution est intéressante car elle est indépendante (coefficient de corrélation à $-0,14$) de la présence dans les textes lus des mots testés. Ceci signifie que la lecture a des effets même sur les mots qui ne sont pas lus, ce qui confirme les résultats de Landauer et Dumais (1997).

3.5.2. Analyse qualitative

Une seconde analyse porte sur les évolutions par rapport à la norme. Cette évolution est calculée de la manière suivante : $E = |\text{pré-test} - \text{norme}| - |\text{post-test} - \text{norme}|$. Il s'agit donc de prendre la valeur absolue de l'écart au pré-test et de lui soustraire la valeur absolue de l'écart au post-test. Nous avons choisi de travailler avec les valeurs absolues des écarts car nous pensons que c'est la longueur de l'écart qui compte et non son sens. Ainsi, nous considérons par exemple, quand la norme est à 2 (c'est-à-dire même domaine), qu'une réponse 0 (c'est-à-dire pas de relation entre les deux mots) est équivalente à une réponse 4 (c'est-à-dire synonymie ou antinomie) car, pour les deux l'écart à la norme vaut 2. Avec cette méthode de calcul, l'évolution est supérieure à zéro quand, entre le pré-test et le post-test, la réponse donnée par le sujet se rapproche de la norme. Ainsi si un sujet répond 1 au pré-test et 3 au post-test alors que la norme est à 4 son évolution sera égale à $|1-4| - |3-4| = 3 - 1 = 2$.

Les résultats de cette analyse indiquent que trois groupes ont des évolutions négatives, c'est-à-dire que leurs lectures font « régresser » leurs connaissances. Les moyennes des évolutions pour chacun des quatre groupes, sur les 500 couples de mots testés sont les suivantes :

POA	Proche	loin	aléatoire
0.0215	- 0.0098	- 0.0067	- 0.0099

Tableau 1 : Moyennes des évolutions sur l'ensemble des cinq séances

Une analyse avec un test de Kolmogorov et Smirnov sur 500 évolutions par groupe (moyenne des évolutions des sujets pour chacun des 100 couples de mots des cinq séances), sur les groupes pris deux à deux indiquent que les sujets du groupe POA obtiennent des évolutions significativement différentes de celles des autres groupes, comme le montre le tableau de résultats ci-dessous.

	Aléatoire	loin	Proche
POA	D = 0.102 p = 0.0110	D = 0.126 p = 0.0007	D = 0.1 p = 0.0135
Aléatoire		D = 0.074 p = 0.1294	D = 0.044 p = 0.7184
loin			D = 0.09 p = 0.0348

Tableau 2 : tests de Kolmogorov et Smirnov sur les évolutions des groupes pris 2 à 2

Il semble donc y avoir un effet de la distance sur l'évolution des réponses des sujets. Le groupe POA est celui qui favorise le plus les évolutions, il est le seul à avoir une moyenne des évolutions positives et cette différence est significative.

Une troisième analyse porte sur l'évolution entre le premier pré-test (celui de la première séance) et le dernier post-test (celui de la cinquième séance). En effet, puisque nous considérons que nos cinq tests sont équivalents (Zampa, 03), l'évolution sur l'ensemble de l'expérimentation correspond à l'évolution entre le premier pré-test et le dernier post-test. Les moyennes des écarts entre les réponses des sujets et celles des experts sont données dans le tableau suivant.

	POA	aléatoire	loin	proche
Pré-test 1	0.954	0.966	0.859	0.926
Post-test 5	0.846	0.870	0.888	0.819
Évolution	0.108	0.96	- 0.29	0.107

Tableau 3 : moyennes des écarts aux experts au pré-test1 et post-test5 et évolutions

Les différences entre les quatre groupes lors du premier pré-test ne sont pas significatives. Et, bien que trois groupes sur quatre voient les écarts aux experts diminuer (évolution > 0) cette différence n'est significative que pour le groupe POA comme le montre le tableau des tests de Kolmogorov et Smirnov suivant.

POA	aléatoire	loin	Proche
D = 0.20 p = 0.0366	D = 0.16 p = 0.1545	D = 0.16 p = 0.1545	D = 0.18 p = 0.0783

Tableau 4 : ks-test entre les écarts au pré-test1 et au post-test5

Il semblerait ainsi que le seul groupe dont l'évolution est positive et significative sur l'ensemble de l'expérimentation soit le groupe POA. Pour le groupe proche, nous pouvons parler d'une tendance puisque la valeur du p est à 0.0783.

4. Justification de l'utilisation de LSA dans RAFALES

Notre choix d'utiliser LSA dans RAFALES reste à justifier.

4.1. Des choix méthodologiques

Pour commencer, utiliser LSA dans les trois modules de RAFALES nous permet de n'avoir qu'un seul formalisme.

De plus, en utilisant LSA, RAFALES peut être utilisé dans différents domaines uniquement en changeant les textes qui composent les bases de connaissances. En effet, pour passer de l'acquisition de l'anglais juridique à l'anglais médical il suffit de supprimer les textes de la base de connaissances de spécialité (1 million de mots dans notre expérience) et à remplacer par des textes du nouveau domaine. Il n'y a aucune connaissance à coder manuellement, le processus de sélection des textes reste identique.

De plus, LSA est automatique, c'est-à-dire qu'il est possible de sélectionner les textes sans avoir recours à des experts. Les experts du domaine d'apprentissage ne doivent être présents que pour valider la construction initiale de la base de connaissances.

4.2. Des choix didactiques

Puisqu'il a déjà été validé comme un modèle de l'acquisition d'une langue, nous pensons qu'il peut permettre de simuler les apprentissages et ainsi modéliser l'apprenant au cours de son apprentissage. Toutefois, nous reconnaissons que cette modélisation de l'apprenant, telle qu'elle est définie pour l'instant et telle que nous l'avons présentée, est loin d'être parfaite. En effet, dès qu'un texte est fourni à l'apprenant, il est considéré comme totalement lu et compris. Le modèle actuel ne tient pas compte du facteur humain.

LSA peut aussi permettre de tenter une modélisation de la zone proximale de développement avec la proximité optimale d'acquisition. Mais, contrairement à la définition de Vygotski, ceci se réalise sans médiation humaine dans notre prototype. En effet, il ne s'agit pas de résoudre des problèmes mais d'acquérir une langue étrangère de spécialité, c'est-à-dire essentiellement acquérir du vocabulaire et des concepts. De ce fait, nous avons transposé et interprété la ZDP dans ce cadre précis. Nous avons ainsi interprété le fait que :

- la ZDP définit la distance entre ce que l'apprenant est capable de faire seul et ce qu'il est capable de faire avec une aide externe
- au-delà de la ZDP, l'apprenant ne peut réussir même avec l'aide d'autrui,

par :

- l'apprenant à des acquis (il maîtrise un vocabulaire et des notions de base dans la langue de spécialité), il existe une zone favorisant l'acquisition
- au-delà de cette zone, les connaissances sont trop éloignées de ce qu'il connaît et il ne peut rien apprendre.

Nous avons ainsi défini une POA, c'est-à-dire une distance ni trop grande ni trop petite pour sélectionner les textes à fournir à l'apprenant. En effet, si les textes sont trop proches de ce qu'il connaît, il n'apprendra que peu ou pas et si les textes sont trop éloignés, il ne les comprendra pas et n'acquerra pas non plus de connaissances. Grâce à LSA, nous obtenons des distances sémantiques entre textes, nous pouvons ainsi déterminer à quelle distance du profil de l'apprenant se situe chacun des textes de la base de connaissances de la langue de spécialité étudiée. Nous avons fixé empiriquement cette POA à un écart-type du texte le plus proche du profil (Zampa, 03).

Dans notre expérimentation nous obtenons des résultats qui indiquent que la POA que nous avons fixée favorise l'acquisition par rapport aux autres modalités. Mais cette POA fixée à un écart-type du texte le plus proche n'est sans doute pas la meilleure. Il est probable que cette distance diffère en fonction du niveau initial des apprenants dans la langue étrangère et dans le domaine de spécialité. Il est aussi possible que cette distance varie selon le domaine étudié. Il reste donc à expérimenter RAFALES dans d'autres domaines et avec des apprenants de niveaux différents, afin de pouvoir affiner cette notion de POA et regarder s'il existe une distance indépendante du domaine étudié et du niveau des apprenants.

Bibliographie

Derwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshmann, R., 1990. "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science*, 41 : 391-407.

De Saussure, F. 1993. *Saussure's third course of lecture in general linguistics*. Oxford : pergamon press.

Dessus, P., 1990. Construction de connaissances par exposition à un cours avec LSA. *In Cognito* 18 : 27-34.

Dessus, P., Lemaire, B. 1999. Apex, un système d'aide à la préparations des examens. *Sciences et Technologies Educatives*. 6(2) : 409-415.

Dessus, P., Lemaire, B., Vernier, A. 2000. "Free-text assessment in virtual campus". in K. Zreik (ed), *proc third international conference on human system learning (CAPS'3)*. Paris, Europa, .61-76.

Dumais, S.T. 1994. Latent Semantic Indexing (LSI). In D. Harman (Ed), *The second text REtrieval conference (TREC2)*, National Institute of Standards and Technology Special Publication vol 500, n°215 : 105-116.

Dumais, S.T. 1997. Using Latent Semantic Indexing for information retrieval, information filtering and other things. *Cognitive Technology Conference*.

Foltz, P.W., Laham, D., Landauer, T.K., 1999. "The Intelligent Essay Assessor : applications to educational technology". *Intercative Multimedia Electronic Journal of computer Enhanced Learning*, 1(2).

Krashen, S.D., 1981. *Second language acquisition and second language learning*. Oxford : Pergamon press.

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.

Lemaire, B. 1998. Models of high-dimensional semantic spaces. proc . *4th int. Workshop on multistrategy learning (MSL 98)*. Desenzano, Italie.

Lemaire, B. 1999. Tutoring systems based on Latent semantic Analysis. In S. Lajoie, M. Vivet (eds) *Artificial Intelligence in Education*. Amsterdam, IOS press. 527-534.

Redington, M. Chater, N. 1998. Connectionist and statistical approaches to language acquisition : a distributional perspective *Language and Cognitive Processes*, 13 : 129-191

Rehder, B., Schreiner, M.E., Wolfe, M.B., Laham, D., Landauer, T.K., Kintsch, W. 1998. "Using Latent Semantic Analysis to assess knowledge : some technical considerations". *Discourse Processes*, 25 : 337-354.

Vygotsky, L.S., 1985. *Pensée et Langage*. Paris : éditions sociales. (édition originale 1962. Cambridge : MIT press.)

Wenger, E. 1987. *Artificial Intelligence and Tutoring Systems*, Morgan Kaufman

Zampa, V. Lemaire, B. 2001. "Latent Semantic Analysis for user modelling". *Journal of intelligent information systems*. 18(1) : p.15-30.

Zampa, V. Raby, F. 2001. "Entre modèle d'acquisition et outil pour l'apprentissage de la langue de spécialité : le prototype R.A.F.A.L.E.S." *Asp (Anglais de spécialité)* 31-33 : 163-179.

Zampa, V. 2003. *Les outils dans l'enseignement : conception et expérimentation d'un prototype pour l'acquisition par exposition à des textes*. thèse de doctorat, Université Pierre-Mendès-France, Grenoble.