Conception et expérimentation d'un manuel électronique de langue étrangère utilisant LSA

Virginie Zampa

LSE
Université Pierre-Mendès-France, BP 47
38040 Grenoble Cedex
Virginie.Zampa@upmf-grenoble.fr
www.upmf-grenoble.fr/sciedu/vzampa/

RÉSUMÉ. *Cet article présente la conception et l'expérimentation d'un prototype nommé RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) utilisant l'analyse de la sémantique latente (LSA). LSA peut être considéré, d'une part comme un outil de représentation du sens des mots à partir de l'analyse automatique de grands corpus et d'autre part, comme un modèle cognitif de l'acquisition de connaissances à partir de textes. Dans RAFALES nous utilisons LSA pour modéliser les connaissances du domaine, pour modéliser le profil de l'apprenant, pour sélectionner et ordonner les textes à fournir à l'apprenant et ainsi créer un manuel personnalisé et évolutif.*

ABSTRACT. *This paper presents design and experimentation of a prototype named RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) which is based on LSA (Latent Semantic Analysis). LSA can be seen as a tool for representing the meaning of words from automatic analysis of larges corpus but also as a cognitive model of learning. RAFALE use LSA for modeling the domain knowledge, for modelling the learner, and for selecting and ordering the texts that are presented to the learner ; thus creating a personnalised and evolutive book.*

MOTS-CLÉS : *RAFALES, LSA, POA, acquisition en langue.*

KEY WORDS : *RAFALES, LSA, POA, language learning .*

1. Introduction

Pour acquérir des connaissances, essentiellement au niveau du vocabulaire, dans une langue étrangère de spécialité, l'apprenant a le plus souvent recours à un enseignant ou à des manuels. L'apprentissage avec un enseignant entraîne des contraintes d'horaire, de lieu, etc. Quant au manuel, un inconvénient majeur réside dans le fait que l'enseignement n'y est pas personnalisé (il ne varie pas en fonction du profil de l'apprenant et ne répond pas forcément à ses attentes), qu'il n'est pas évolutif (le choix des textes est fait une seule fois, avant l'impression) et que la sélection de ses textes n'est pas automatique mais dépend d'experts humains.

Dans cet article, nous présentons un prototype nommé RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) que nous avons conçu et développé ainsi que son expérimentation. Ce prototype tente d'apporter des pistes de recherche pour l'individualisation, l'évolutivité et l'automatisation des manuels.

RAFALES est un prototype d'acquisition en langue par expositions à des textes. Cette exposition se fait par le biais de la lecture des textes qu'il sélectionne. L'unique tâche de l'apprenant est la lecture. Nous reprenons l'approche constructiviste du langage. Des travaux en psychologie cognitive montrent que la majorité des mots sont acquis par la lecture [LAN 97]. De plus, un courant de recherche important en didactique des langues, privilégie l'exposition à la langue dans l'apprentissage d'une langue seconde, l'apprentissage des règles étant secondaire [KRA 81]. L'apprenant, exposé à des textes, va petit à petit affiner le sens des mots grâce aux occurrences conjointes de ces mots avec d'autres. Par exemple, sans le lui définir explicitement, l'apprenant va acquérir le sens du mot « ordinateur » parce que ce mot apparaît avec d'autres comme «souris», « programmer », « logiciel », dans les textes qu'il lit. Dans cette représentation, nous voyons que le sens d'un mot est défini par l'ensemble des mots qui lui sont proches, tout comme l'a défini Saussure en linguistique [SAU 93]. Cependant, depuis Platon et le fameux paradoxe de l'induction, il est connu que ces simples cooccurrences répétées ne suffisent pas à expliquer la connaissance que nous avons des mots. Les modèles psychologiques parviennent difficilement à expliquer ce phénomène autrement que par des hypothèses innéistes. Une explication possible est que ce n'est pas simplement la cooccurrence répétée d'un mot avec d'autres qui permet l'acquisition du sens du mot, mais plutôt l'ensemble des cooccurrences de tous les mots au fil des textes. Landauer et Dumais (97) montrent, par une simulation de l'apprentissage entre 2 et 20 ans que 75 % de la connaissance sémantique d'un mot provient de la lecture de textes ne le contenant pas.

Dans RAFALES le choix des textes n'est pas neutre, leur sélection est le point central. Cette sélection, étroitement liée au profil de l'élève et aux connaissances du domaine, se fait par le biais d'une analyse sémantique. Pour ceci nous avons besoin d'un analyseur sémantique capable de fournir des proximités entre des textes. C'est pourquoi nous avons utilisé *latent semantic analysis* (LSA).

2. RAFALES

RAFALES utilise LSA pour modéliser les connaissances du domaine et de l'apprenant ainsi que pour sélectionner les stimuli à lui fournir. LSA est donc central dans notre prototype. Nous allons présenter le modèle cognitif sous-jacent à LSA, son fonctionnement ainsi que quelques expérimentations et validations intéressantes pour notre recherche. Puis nous expliquerons le fonctionnement de notre prototype.

2.1. Latent Semantic Analysis

2.1.1. La méthode

LSA s'appuie sur une représentation multidimensionnelle des mots de 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 correspond à la proximité de sens de ces mots. Le modèle d'apprentissage prend donc un ensemble de textes en entrée et prédit les proximités sémantiques qui vont résulter de la lecture de ces textes.

LSA analyse l'ensemble des textes sources 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 avec lui. Ainsi, les mots vélo et bicyclette sont considérés sémantiquement proches car ils apparaissent tous deux avec des mots comme guidon, pédaler, etc., et ils n'apparaissent qu'occasionnellement avec des mots comme bouillir, ordinateur, 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 d'un 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 la matrice d'occurrences. Il s'agit d'une matrice dont les lignes sont des unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes, des mots. 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 dimensions. Ce nombre est important car une réduction à un espace trop grand ne ferait pas suffisamment émerger les liaisons sémantiques entre mots, et un espace trop petit conduirait à une trop grande perte d'informations. Ce nombre de dimension est issu de tests empiriques, [DEE 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 dimensions *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 la façon 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.

2.1.2. Quelques applications et validations

Au départ, LSA a été développé comme outils de recherche d'information [DUM 94], [DUM 97]. Avec les problèmes de choix de mots-clés liés à la synonymie, à la polysémie et à l'inflexion, il est aisé de postuler que la recherche devrait se faire sur le sens des mots clés et non uniquement sur leur « forme ».

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 à 3 500 mots par jour et un apprentissage de 7 à 15 mots nouveaux par jour. Avec une même exposition, qui correspond au corpus de 4,6 millions de mots tirés d'une encyclopédie, LSA apprend 10 mots par jour. LSA passe les QCM de synonymie du TOEFL, en choisissant, parmi les quatre mots, 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 postulant à l'entrée 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 [RED 98] ou un domaine particulier comme celui traité dans un cours [DES 00a]. Elles peuvent être dans un langage et non une langue naturelle. C'est le cas par exemple avec l'apprentissage des jeux tels que le tic-tac-toe [LEM 98] ou kalah [LEM 99].

LSA a aussi été utilisé de diverses manières dans des EIAH. Des travaux ont ainsi porté sur l'évaluation de copies. Apex, par exemple [DES 99], [DES 00b], est un système d'aide à la préparation des examens, dans lequel l'étudiant rédige une copie qu'il peut soumettre à l'évaluation puis modifier, etc. D'autres travaux ont porté sur la notation de copies, certains utilisant des copies de référence. Il s'agit pour l'apprenant de rédiger une « synthèse ». La corrélation entre les juges humains et LSA est, le plus souvent, proche de la corrélation entre les juges. D'autres travaux ont porté sur la modélisation de l'apprenant [ZAM 01a], [ZAM 01b] et la détection des erreurs dans une copie en langue étrangère [ZAM 01a].

2.2. RAFALES : fonctionnement et initialisation pour l'expérimentation

Le prototype RAFALES a pour but d'optimiser, c'est-à-dire d'accélérer et de cibler, l'acquisition dans une langue étrangère de spécialité. Comme nous l'avons déjà signalé, RAFALES utilise LSA pour modéliser les connaissances du domaine, pour modéliser le profil de l'apprenant et pour sélectionner les stimuli. Le fait d'utiliser LSA dans les trois modules permet de n'avoir qu'un seul formalisme pour modéliser et comparer les connaissances. De plus LSA permet de construire ces modélisations automatiquement, sans avoir recours à des humains. Dans notre expérimentation, nous travaillons sur l'apprentissage de l'anglais juridique.

2.2.1. La base de connaissances du domaine

La base de connaissances correspond à une grande base de données formée uniquement de textes sur lesquels aucun traitement préalable (lemmatisation, etc.) n'a été réalisé. Ces textes, et les connaissances qui en résultent sont représentés dans un espace sémantique multidimensionnelle. À la différence du modèle de l'apprenant, cette base est conçue avant la première utilisation et elle n'évolue pas. De plus la base de connaissances du domaine se divise en deux parties : la base de connaissances de la langue générale (BCLG) et la base de connaissances de la langue de spécialité (BCLS). Les textes de la BCLG permettent à LSA d'acquérir des connaissances sur cette langue et en particulier sur un grand nombre d'expressions et de mots usuels. La BCLS donne à LSA les connaissances sur le sens précis des mots de ce vocabulaire spécifique. Les textes donnés à l'apprenant sont uniquement des textes de la base de connaissances de la langue de spécialité.

2.2.2. Le profil de l'apprenant

Le profil de l'apprenant doit, selon notre hypothèse personnaliser l'apprentissage et, de ce fait, le rendre plus efficace. Comme nous l'avons déjà présenté, LSA a été utilisé pour modéliser l'apprentissage humain. Nous l'avons donc utilisé pour modéliser le profil de l'apprenant. Cette modélisation est initialisée en fonction du public, puis elle évolue au fur à mesure des séances. À la fin de chaque séance, les textes lus sont ajoutés. Le profil de l'apprenant, tout comme la base de connaissances du domaine, ne contient que des textes.

Nous faisons l'hypothèse que les liens sémantiques établis par LSA à partir des textes qui lui sont fournis sont proches de ceux que l'apprenant va créer. De ce fait, nous supposons que les connaissances issues de ses relations seront sensiblement équivalentes pour LSA et pour l'apprenant. En fournissant à LSA la même base de textes qu'à l'apprenant, nous souhaitons « estimer » les connaissances, c'est-à-dire les liens sémantiques entre les mots, qu'il a acquises et ainsi remplir la fonction prédictive de Self (1987).

2.2.3. Le module pédagogique

Dans le module pédagogique nous utilisons les comparaisons de proximités sémantiques pour sélectionner les textes à fournir à l'apprenant. LSA calcule les proximités sémantiques entre le profil de l'apprenant et chaque texte de la BCLS. Nous considérons que la distance entre un texte de la BCLS et le profil de l'élève est égale à la moyenne des distances entre ce texte et tous les textes du profil.

Nous pensons que le processus d'apprentissage peut être accéléré en sélectionnant la bonne séquence de textes, c'est-à-dire la plus appropriée à l'apprenant en tenant compte de ses connaissances supposées modélisées par son profil. Le problème est donc de savoir quels textes sont les plus susceptibles d'agrandir l'espace sémantique de l'élève. Notre module pédagogique sélectionne, grâce à LSA, les textes de la base de connaissances de la langue de spécialité, qui sont ni trop proches ni trop éloignés de ce que l'apprenant connaît déjà à priori, mais qui se situent à la proximité optimale d'acquisition (POA).

2.3. Son fonctionnement

Dans RAFALES, le module pédagogique sélectionne dans le module de connaissances du domaine les textes les mieux adaptés à l'apprenant afin d'optimiser l'acquisition en langue étrangère de spécialité, en tenant compte du profil de l'élève.

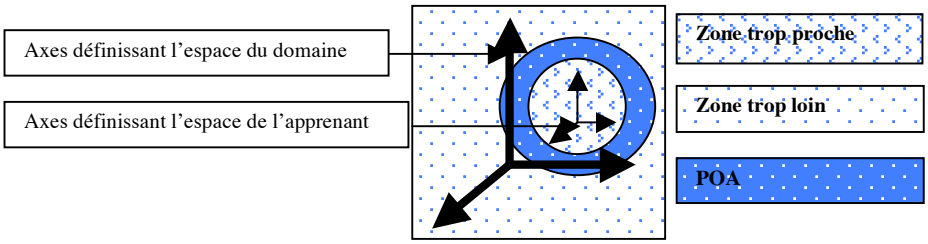

Figure 1. Visualisation de la POA

Le profil de l'apprenant est un sous-espace de l'espace des connaissances du domaine. Le module pédagogique sélectionne dans l'espace du domaine les textes qui ne sont ni trop proches, ni trop éloignés de l'espace des connaissances de l'apprenant. Cette zone dans laquelle le module pédagogique sélectionne les textes est la POA.

RAFALES fournit ces textes à l'apprenant qui les lit. Puis RAFALES met à jour le modèle de l'apprenant en lui ajoutant les textes qui viennent d'être lus, ainsi l'espace de connaissances de l'apprenant s'agrandit. Le module pédagogique de RAFALES sélectionne d'autres textes en fonction de ce nouveau profil de l'apprenant, et les donne à l'élève, etc. La boucle s'interrompt quand l'espace des connaissances estimées de l'apprenant est identique à celui du domaine.

3. Expérimentation et résultats

Un de nos buts avec cette expérimentation est de vérifier l'existence de la POA. Nous avons ainsi quatre groupes. Le groupe « loin » lit les textes les plus éloignés du modèle de l'apprenant, le groupe « proche » les plus proches, le groupe « aléatoire » des textes sélectionnés aléatoirement, enfin le groupe « POA ». Nous avons fixé cette proximité à un écart type des textes les plus proches. Nous avons choisi de travailler avec cette notion d'écart type qui n'est pas une valeur fixe. A chaque utilisation de RAFALES les proximités sémantiques entre le profil de l'apprenant et BCLS deviennent plus faible puisque le profil tend à devenir identique à la BCLS. Une distance fixe n'a donc pas la même signification en début et en fin d'utilisation du prototype.

3.1. Initialisation des modélisations

3.1.1. Le profil de l'apprenant

Dans notre expérimentation, nous avons un profil par groupe expérimental, mais l'initialisation est identique pour tous. Elle se fait avec des textes en anglais général et quelques textes de la langue de spécialité. Les textes de la langue générale comportent environ 1 000 000 mots (estimation très approximative du nombre de mots, en langue anglaise, auxquels nos sujets ont été exposés au cours de leur scolarité). Les textes de la langue de spécialité sont les 25 textes les plus centraux de la BCLS. Nous avons choisi de prendre des textes centraux car ils contiennent les mots les plus fréquents de la langue de spécialité et ceci est une estimation des connaissances de nos sujets. Nous avons décidé d'en prendre 25, car avec ce nombre, LSA et les sujets répondent « mot inconnu » le même nombre de fois à nos tests. Il s'agit donc d'une initialisation des profils très « grossière ».

3.1.2. La base de connaissances du domaine

Dans notre expérimentation, la BCLG contient environ 1 000 000 mots appartenant à huit œuvres complètes. Ces oeuvres font partie du domaine public, mais sont relativement récentes. La BCLS contient un peu plus de 1 100 000 mots. Il s'agit de textes de lois, de comptes-rendus de procès, etc.

3.2. Le plan expérimental

3.2.1. Les sujets

Quarante-deux sujets ont passé l'expérience dans les temps impartis : 19 étudiants de licences et maîtrise de langue étrangère et 23 stagiaires d'IUFM. Pour

homogénéiser leur répartition nous avons utilisé leurs notes (étudiants) ou leur classement (stagiaires). Nous obtenons ainsi trois groupes de 11 et un de 9 sujets.

Nous avons aussi fait passer les tests qui permettent d'évaluer l'acquisition du vocabulaire, à 25 experts du domaine. Leurs réponses nous permettent d'étalonner les résultats, et ainsi d'avoir une norme, égale à la moyenne de leurs réponses.

3.2.2. Les conditions de passation et les variables

Il y a deux méthodes de passation : version papier et version électronique. Quelle que soit la condition, les étudiants ont deux semaines pour faire les cinq séances. Pour chaque séance, les sujets disposent des consignes de début et de fin de séance, du test de vocabulaire de début et de fin et des textes à lire, fournis dans l'ordre.

Notre plan d'expérience comporte une variable dépendante qui correspond à l'évolution entre le pré-test et le post-test par rapport à la norme fournie par les experts et une variable indépendante à quatre modalités qui correspond à la distance entre les textes fournis à l'apprenant et son profil.

3.2.3. Les tests utilisés

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

Les sujets passent les tests de vocabulaire en début et fin de chaque séance. Au sein d'une même séance, les deux tests de vocabulaire sont identiques mais ils varient d'une séance à l'autre. Chaque test comporte 30 tableaux où 1 signifie « même sens », 2 « même domaine », 3 « sens différent », 4 « pas de relation » et ? « mot inconnu ». Le sujet indique s'il s'agit d'une relation forte (+) ou faible (-). Dans l'exemple ci-contre, la croix indique que le sujet juge qu'il y a une relation forte de même sens, ce qui correspond à une relation de synonymie, entre les mots *clip* et *blow*.

3.3. Les Résultats des sujets

Les valeurs des réponses codées vont de 0 à 4 où 0 correspond à aucune relation entre les mots et 4 à une relation forte (synonymie ou antinomie). De plus nous n'analysons que 20 tableaux sur 30 (ceci afin de limiter les effets de primauté et de récence). Nous avons ainsi 100 couples de mots par tests.

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 d'évolution ne tenant pas compte du sens de la modification, c'est-à-dire sans regarder si la réponse du sujet se rapproche ou s'éloigne de la norme. Cette évolution est indépendante (coefficient de corrélation à – 0,14) de la présence des mots testés dans les textes lus, ce qui confirme les résultats présentés précédemment [LAN 97].

Une seconde analyse porte sur les évolutions par rapport à la norme. L'évolution est calculée de la manière suivante : E = |(pré-test – norme)| – |(post-test – norme)|

Ainsi, l'évolution est supérieure à zéro quand, entre le pré-test et le post-test, la réponse donnée par l'étudiant se rapproche de la norme. Avec cette analyse, trois groupes obtiennent des évolutions négatives. En effet, les moyennes des évolutions pour chacun des quatre groupes, sur les 500 couples testés sont les suivantes :

aléatoire	Proche	Loin	POA
-0,0099	-0,0098	-0,0067	0,0216

Tableau 1. *Moyennes des évolutions des sujets des quatre groupes expérimentaux*

Une analyse avec un t de Student, calculé de manière unilatérale indique qu'il y a une différence significative entre le groupe POA et les trois autres groupes. Ces tests sont calculés à partir des moyennes par groupes des 500 couples de mots. Les résultats sont les suivants :

POA et …	Aléatoire	Proche	Loin
Probabilité	0,023	0,019	0,038
t	2,26	2,35	2,08

Tableau 2. *t de Student des différences entre la POA et les autres conditions*

4. Conclusion

Malgré le faible nombre de sujets, la durée limitée de l'expérimentation, nous constatons une différence significative entre les évolutions des sujets en fonction des groupes expérimentaux. En effet, les sujets qui ont lu les textes se trouvant à la POA sont les seuls à avoir des résultats qui se rapprochent de ceux des experts. Ces résultats se limitent à l'acquisition du vocabulaire, il n'est en aucun cas possible d'inférer quant à l'acquisition de connaissances, même s'il existe, comme nous l'avons déjà signalé, des éléments de réponse théorique.

Avec notre expérimentation, nous avons vérifié qu'une distance située à «un écart type» de la plus petite distance entre le profil de l'élève et les textes du domaine, permet une meilleure acquisition de vocabulaire, qu'une distance trop grande ou trop petite ou sélectionnée aléatoirement. Cette distance correspond à notre proximité optimale d'acquisition dans l'expérimentation. Nous signalons toutefois que cette distance à été choisie de façon empirique. Malgré tout, il serait important de vérifier si ces résultats sont reproductibles, puis il serait intéressant d'affiner cette proximité et de contrôler si elle est identique quelle que soit la langue étrangère de spécialité et quel que soit le profil de l'apprenant.

Le fait d'utiliser LSA dans les trois modules permet de n'avoir qu'un seul formalisme et ainsi d'obtenir les distances entre profil de l'apprenant et domaine des connaissances facilement. De plus notre prototype est entièrement réutilisable, avec très peu d'intervention humaine. En effet pour changer le domaine d'application il suffit de changer la partie de la base de connaissance qui porte sur le domaine de la langue de spécialité.

Enfin, comme nous l'avons constaté, la lecture permet d'acquérir des mots, ceci est vérifié avec les variations des relations estimées par le sujet entre des couples de mots. De plus ces modifications ne sont pas forcément liées au fait de rencontrer ces mots dans les textes. Ceci rejoint des résultats déjà présentés.

Remerciements

Nous remercions Philippe Dessus, Benoît Lemaire et Erica Devries pour les commentaires d'une version précédente de cet article.

5. Bibliographie

[DEE 90] DERWESTER S., DUMAIS S.T., FURNAS G.W., LANDAUER T.K., and HARSHMANN R. "Indexing by Latent Semantic Analysis". *Journal of the American Society for Information Science*, vol 41, p.391-407.

[DES 99] DESSUS P., LEMAIRE B. "Apex, un système d'aide à la préparation d'examens". *Sciences et Technologies Educatives.* Vol 6, n°2, p.409-415.

[DES 00a] DESSUS P. "Construction de connaissances par exposition à un cours avec LSA." *In Cognito*, vol 18, p.27-34.

[DES 00b] DESSUS P., LEMAIRE B., VERNIER A. "Free-text assessment in a virtual campus" in K. Zreik (ed), *proc . third internationnal conference on human system learning* (CAPS'3). Paris, Europia, p.61-76.

[DUM 94] DUMAIS S.T. "Latent Semantic Indexing (LSI) and TREC-2". In D. Harman (Ed.), *The Second Text RE-trieval Conference* (TREC2), National Institute of Standards and Technology Special Publication vol 500, n°215, p.105-116.

[DUM 97] DUMAIS S. T. "Using Latent Semantic Indexing for information retrieval, information filtering and other things", *Cognitive Technology Conference*.

[KRA 81] KRASHEN S.D., *Second language acquisition and second language learning*. Oxford Pergamon press.

[LAN 97] LANDAUER T.K, DUMAIS S.T., "A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge", *Psychological Revue*, 104, p.211-240.

[LEM 98] LEMAIRE B., "Models of high-dimensional semantic spaces" *Proc. 4th int. Workshop on multistrategy learning* (MSL 98). Desenzano, Italie.

[LEM 99] LEMAIRE B., "Tutoring systems based on Latent Semantic Analysis". In S. Lajoie, M. Vivet (eds) *Artificial Intelligence in Education*. Amsterdam, IOS press. p. 527-534.

[RED 98] REDINGTON M., CHATER, N. "Connectionist and statistical approaches to language acquisition : A distributional perspective". *Language and Cognitive Processes*, 13-2/3.

[SAU 93] DE SAUSSURE F. *Saussure's third course of lecture in general linguistics*. Oxford, Pergamon press.

[SEL 87] SELF J., "Student Models : what use are they ?" in P Ercoli and R, Lewis eds. *Artificial Intelligence tools in Education*. Amsterdam, North-Holland.

[ZAM 01a] ZAMPA V. et LEMAIRE B., "Latent Semantic Analysis for user modelling", *Journal of intelligent information systems.* Vol 18 n°1. p.15-30.

[ZAM 01b] ZAMPA V. et RABY F., "Entre modèle d'acquisition et outil pour l'apprentissage de la langue de spécialité : Le prototype R.A.F.A.L.E.S", *Asp* n°31-33, p.163-179.