Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité

Virginie Zampa

Apprentissage des Langues et Systèmes d'Information et de Communication

alsic.org ou alsic.u-strasbg.fr
Volume 8, 2005
article mis en ligne en septembre 2005
pp. 135-146
Recherche

Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité

Virginie ZAMPA
CNRS, laboratoire LaCo, Poitiers, France

Résumé : *Cet article présente l'utilisation de l'analyse sémantique latente (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 Étrangère 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 l'analyse sémantique latente, 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 poussée à utiliser cette analyse dans Rafales.*

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 connaissance 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 Étrangère de Spécialité) qui est fondé sur l'utilisation de l'analyse sémantique latente.

Nous allons dans un premier temps exposer les principes de cette analyse 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. L'analyse sémantique latente (LSA)

Cette analyse est un modèle statistique développé par les laboratoires *Bellcore* en 1989, comme outil de recherche documentaire [Derwester90]. 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

L'analyse de la sémantique latente (*Latent Semantic Analysis,* LSA) s'appuie sur une représentation multidimensionnelle de la signification des mots dans la langue. Pour cette analyse, un mot correspond à un graphème. Il n'y a pas de lemmatisation préalable. Il existe une liste des "mots outils" (liste facilement modifiable) qui ne sont pas pris en compte. Grâce à une analyse statistique, le sens de chaque graphème 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 graphèmes 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.

L'analyse de la sémantique latente s'appuie sur l'ensemble des textes source pour en représenter les graphèmes dans un espace sémantique multidimensionnel. Cette analyse statistique (présentée plus loin) permet de faire ressortir les relations sémantiques entre graphèmes ou entre textes. Deux graphèmes peuvent être considérés sémantiquement proches s'ils sont utilisés dans des contextes similaires. Le contexte d'un graphème est ici défini comme l'ensemble des graphèmes qui apparaissent conjointement à lui dans un paragraphe. Ainsi, les mots vélo et bicyclette sont considérés comme sémantiquement proches puisqu'ils apparaissent tous 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. L'analyse de la sémantique latente se fait en deux étapes. Dans un premier temps, la matrice d'occurrences est construite. Il s'agit d'une matrice dont les lignes représentent les unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes les graphèmes. L'élément (i,j) de la matrice correspond ainsi au nombre d'occurrences du graphème 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 [Derwester90]. 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 graphème 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 déterminer 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 graphèmes, entre deux paragraphes ou entre un graphème 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, l'analyse de la sémantique latente a été développée comme outil de recherche d'information. 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". Des expérimentations [Dumais97] ont mis en évidence un gain sur la pertinence des textes sélectionnés allant de 16 à 30 % entre la recherche traditionnelle par mots-clés dans une large base de données et les textes sélectionnés par similarité sémantique avec ces mêmes mots-clés (les textes sélectionnés sont ceux ayant une similarité maximale avec la requête).

Un second domaine d'application est l'apprentissage. Ce modèle a été testé par Landauer et Dumais [Landauer97]. Ils ont simulé l'acquisition du vocabulaire entre 2 et 20 ans. Pour cela, ils ont "entraîné" leur modèle informatique de LSA avec une encyclopédie électronique de plus de 30 000 articles. Ils ont ensuite testé ses capacités de reconnaissance sémantique en lui faisant passer les 80 items de synonymie du TOEFL (*Test Of English as a Foreign Language*), où il s'agit d'apparier un mot cible avec le mot sémantiquement le plus proche, choisi parmi quatre. Le modèle LSA obtient un résultat (64,4 % de bonnes réponses) comparable à la moyenne des sujets non anglophones admis dans les universités américaines (64,5 %). LSA peut ainsi être considéré comme un modèle plausible de l'acquisition de connaissances à partir de données textuelles.

Un troisième domaine d'application concerne l'acquisition de connaissances. Ces acquisitions peuvent concerner les langues [Redington98] ou un domaine particulier comme celui traité dans un cours [Dessus00]. Elles peuvent concerner un langage et non une langue naturelle. C'est la cas par exemple avec l'apprentissage de jeux tels que le *tic-tac-toe* ou le *kalah* [Lemaire99].

L'analyse sémantique latente est aussi utilisée 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) de [Dessus99] et [Dessus00]. D'autres travaux ont porté sur la notation de copies [Foltz99] dans lesquelles l'étudiant 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 [ZampaRaby01] et la détection des erreurs dans une copie en langue étrangère [ZampaLemaire01].

3. Le prototype Rafales

Rafales est un recueil de lectures 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 [Landauer97] indiquent que les enfants entre 2 et 20 ans apprennent en moyenne 10 mots nouveaux par jour. Or seulement quelques centaines de mots par an sont acquis par instruction directe. De plus, une majorité des mots n'apparaissent qu'à 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" [Vygotski34].

Rafales possède l'architecture d'un tuteur intelligent [Wenger87]. 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 l'analyse sémantique latente. 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 (cf. figure 2). 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 ne sont ni trop proches ni trop éloignés. En d'autres termes, il sélectionne les textes de la base de connaissances qui se trouvent à la POA du profil de l'apprenant (cf. figure 1).

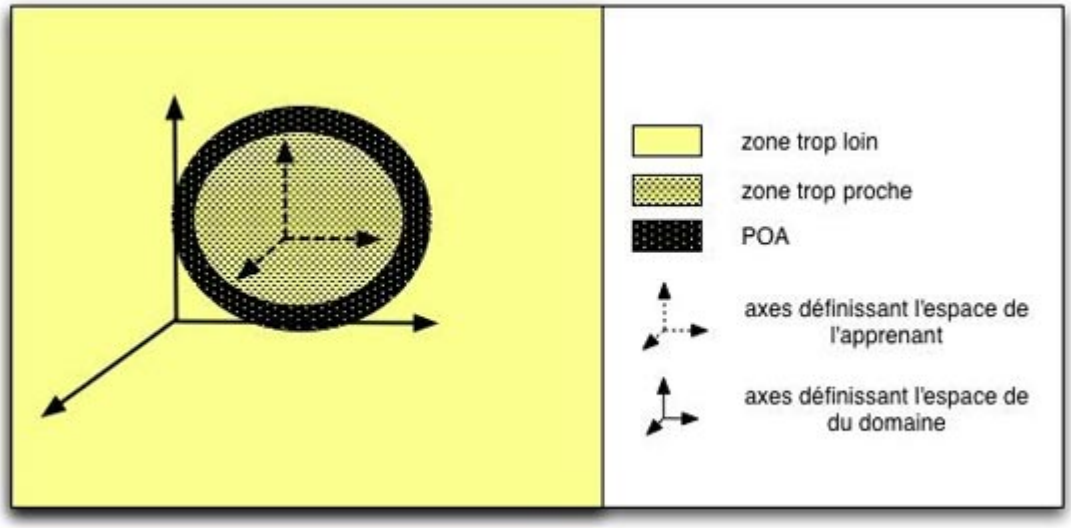

Figure 1 - Les différents espaces et la POA.

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.

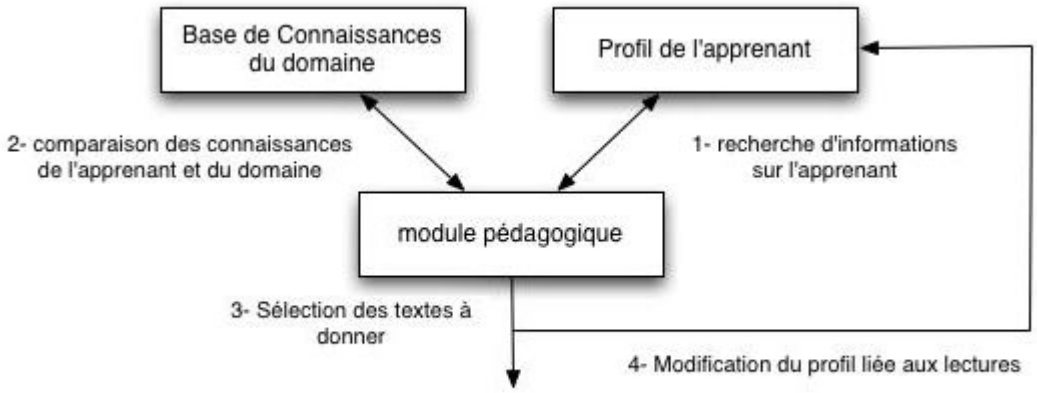

Figure 2 - Architecture et fonctionnement de *Rafales*.

3.2. Conditions 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 million de mots issus de huit œuvres complètes. Ces œuvres font partie du domaine public mais sont relativement récentes. Quant à la base de connaissances de la langue de spécialité, elle contient un peu plus de 1 million de mots issus de textes de loi, de comptes-rendus de procès, etc.

Le profil de l'apprenant est initialisé en fonction des sujets ; 42 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 (*Institut Universitaire de Formation des Maîtres*). Ils ont été répartis dans quatre groupes expérimentaux. Pour homogénéiser leur répartition, nous avons utilisé leurs notes (pour les étudiants) ou leur classement au Capes (*Certificat d'Aptitude au Professorat de l'Enseignement Secondaire*) (pour les stagiaires). Le profil de l'apprenant de départ est identique pour les quatre groupes. Il contient 1 million 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.

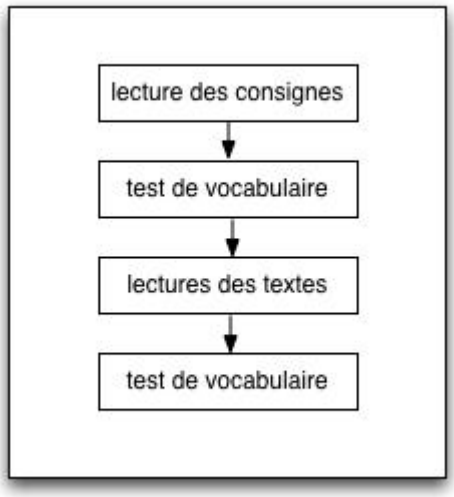

Figure 3 - Déroulement d'une séance.

Les 42 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.

3.3.2. 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. Chaque test comporte 30 tableaux, pour un exemple voir le tableau 1.

Ces tests de vocabulaire ont été conçus d'une manière à la fois empirique et systématique. Nous choisissons chaque fois un mot qui va faire l'objet du travail de proximité. Puis nous sélectionnons, toujours au hasard, des mots de la langue de spécialité ou de la langue générale qui vont lui être confrontés. Le vocabulaire de spécialité a été sélectionné au hasard d'un manuel d'anglais juridique qui comporte une liste de vocabulaire de spécialité à la fin de chaque chapitre. Nous avons suivi l'ordre des chapitres et pris les mots de cinq en cinq. De manière empirique, nous avons exclu les mots morphologiquement transparents entre le français et l'anglais comme par exemple constitution qui est un terme transparent du point de vue de la théorie du droit, même si les réponses apportées par chaque constitution ou loi constitutionnelle aboutissent à des concepts divergents du point de vue de l'extension des prédicats. Nous avons également sélectionné au hasard, toutes les cinq lettres, des mots de la langue générale, dans le dictionnaire général en ligne du CNRS.

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

Tableau 1 - Exemple de tableau fourni dans les tests. 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 du tableau 1, la croix indique que le sujet juge que les mots clip et blow entretiennent une relation forte et de même sens, ce qui correspond à une relation de synonymie.

3.3.3. 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 20 tableaux sur 30. Nous ne prenons pas en compte les cinq premiers et les cinq derniers afin de limiter les effets de primauté et de récence. Nous avons ainsi 100 couples de mots par test (20 tableaux * 5 couples de mots), soit 200 par séance (100 en pré-test et 100 en post-test), 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 25 experts du domaine. Pour chaque couple de mots, la norme correspond à la moyenne des 25 réponses des experts.

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 modifications ne prenant pas en compte le sens de la variation (rapprochement ou éloignement 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 [Landauer97].

Analyse qualitative

Une seconde analyse porte sur les évolutions par rapport à la norme. Cette évolution est calculée de la manière suivante : E = |pré-test - norme| - |post-test - 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 données dans le tableau 2.

POA	Proche	Loin	Aléatoire
0.0215	- 0.0098	- 0.0067	- 0.0099

Tableau 2 - Moyennes des évolutions sur l'ensemble des cinq séances.

Une analyse avec un test de Kolmogorov et Smirnov (test non paramétrique utilisé sur des variables ordinales) 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 3. 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.

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

Tableau 3 - Tests de Kolmogorov et Smirnov sur les évolutions des groupes pris 2 à 2. D correspond à la valeur du test et p à la significativité.

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 [Zampa03], 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 4. 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 5 des tests de Kolmogorov et Smirnov.

	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 4 - Moyennes des écarts aux experts au pré-test 1 et post-test 5 et évolutions.

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 5 - Significativité des tests de Kolmogorov et Smirnov entre les écarts au pré-test 1 et au post-test 5.

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 l'analyse sémantique latente dans *Rafales* reste à justifier.

4.1. Des choix méthodologiques

Pour commencer, utiliser l'analyse sémantique latente dans les trois modules de *Rafales* nous permet de n'avoir qu'un seul formalisme.

De plus, grâce à l'analyse sémantique latente, *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 de les 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 d'acquisition par exposition à des textes, 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 est donc identique pour les membres d'un même groupe, il ne tient donc pas compte des différences interindividuelles.

LSA peut aussi tenter une modélisation, ce que nous avons appelé "proximité optimale d'acquisition", qui se fonde sur les travaux de Vygotski concernant la "zone proximale de développement" (ZDP). Mais, contrairement à la définition de Vygotski, ceci se réalise sans médiation humaine dans notre prototype (notons toutefois que dans le cas de l'utilisation du prototype par l'enseignant pour sélectionner les textes à fournir à ses étudiants, il y a bien médiation humaine, mais ce n'est pas le cas dans l'expérimentation que nous présentons). En effet, il ne s'agit pas de résoudre des problèmes mais d'acquérir une langue étrangère de spécialité, c'est-à-dire essentiellement d'acquérir du vocabulaire et des concepts.

De ce fait, nous avons transposé et interprété la ZDP dans ce cadre précis. Dans l'approche de Vygotski, la ZDP est définie comme 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. Dans notre approche, nous reformulons ainsi les choses : i) l'apprenant a des acquis (il maîtrise un vocabulaire et des notions de base dans la langue de spécialité) ; ii) il existe une zone favorisant l'acquisition ; iii) 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 l'analyse sémantique latente, nous obtenons des distances 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 [Zampa03].

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.

Références

Les liens externes étaient valides à la date de publication.

[Derwester90]
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 Informartion Science*. pp. 391-407.

[Dessus99]
Dessus, P. & Lemaire, B. (1999). "Apex, un système d'aide à la préparations des examens". *Sciences et Technologies Educatives (STE)*, vol. 6, 2. pp. 409-415.

[Dessus00]
Dessus, P., Lemaire, B. & Vernier, A. (2000). "Free-text assessment in virtual campus". In Zreik, K. (dir.). *Proceedings- — third internationnal conference on human system learning* (CAPS'3). Paris. pp. 61-76.

[Dumais97]
Dumais, S.T. (1997). "Using Latent Semantic Indexing for information retrieval, information filtering and other things". *Cognitive Technology Conference*.

[Foltz99]
Foltz, P.W., Laham, D. & Laundauer, T.K., (1999). "The Intelligent Essay Assessor: applications to educational technology". *Intercative Multimedia Electronic Journal of Computer Enhanced Learning*, vol. 1, http://imej.wfu.edu/articles/1999/2/04/printver.asp

[Landauer97]
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 Revue,* n° 104. pp. 211-240.

[Lemaire99]
Lemaire, B. (1999). "Tutoring systems based on Latent semantic Analysis". In Lajoie, S.& Vivet, M. (dir.). *Artificial Intelligence in Education*. Amsterdam: IOS press. pp. 527-534.

[Redington98]
Redington, M. & Chater, N. (1998). "Connectionist and statistical approaches to language acquisition: a distributional perspective". *Language and Cognitive Processes*, vol. 13. pp. 129-191.

[Vygotski34]
Vygotski, L.S. (1934). *Pensée et Langage*. (Trad. par Sève, F.). Paris : La Dispute (1997).

[Wenger87]
Wenger, E. (1987). *Artificial Intelligence and Tutoring Systems*. Morgan Kaufman.

[ZampaLemaire01]
Zampa, V. & Lemaire, B. (2001). "Latent Semantic Analysis for user modelling". *Journal of intelligent information systems, Special Issue on Intelligent Information Systems and Education Applications*, vol. 18, 1. pp. 15-30.

[ZampaRaby01]
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)*, vol. 31-33. pp. 163-179.

[Zampa03]
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.

À propos de l'auteure

Virginie ZAMPA est actuellement chercheur post-doctorat au CNRS, dans le laboratoire LaCo (Langage et Cognition) à Poitiers. Elle a soutenu une thèse en sciences de l'éducation après avoir obtenu un DEA d'informatique. Sa recherche porte sur les environnements informatiques d'apprentissage humain et sur l'utilisation de LSA (*latent semantic analysis*) dans ces derniers.
Courriel : virginie.zampa@ext.univ-poitiers.fr
Toile : http://www.upmf-grenoble.fr/sciedu/vzampa
Adresse : CNRS, Laboratoire LaCo, Poitiers, France.

Ce texte fait partie des textes de la journée Atala 2005 qui font l'objet d'un numéro spécial d'Alsic.

Référence de l'article :
Zampa, V. (2005). "Utilisation de l'analyse sémantique latente pour tenter d'optimiser l'acquisition par exposition à une langue étrangère de spécialité". *Apprentissage des langues et systèmes d'information et de communication (ALSIC)*, vol. 8, n° 1. pp. 135-146. http://alsic.u-strasbg.fr/v08/zampa/alsic_v08_09-rec5.htm, mis en ligne le 15/09/2005.

ALSIC| Sommaire | Consignes aux auteurs | Comité de rédaction | Inscription