Entre modèle et outil pour l'acquisition de la langue de spécialité: Le prototype R.A.F.A.L.E.S (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité)

ZAMPA Virginie, RABY Françoise
Laboratoire des Sciences de l'éducation,
Université Pierre-Mendès-France Grenoble 2

Résumé :
Le travail que nous présentons est le produit d'une recherche interdisciplinaire impliquant l'Intelligence Artificielle, la psychologie cognitive et la didactique des langues. LSA (Analyse Sémantique Latente) est un outil informatique qui a été appliqué à l'acquisition du langage (Landauers et Dumais, 1997). Nous avons utilisé LSA pour élaborer un prototype appelé RAFALES (Recueil Automatisé Favorisant l'Acquisition d'une Langue de Spécialité). RAFALES est un outil informatique d'aide à l'acquisition d'une langue étrangère inspiré à la fois des théories de Vygotsky sur la zone proximale de développement (1968) et de la théorie de l'input développée par Krashen (1985).
Après avoir exposé les bases théoriques de LSA et de RAFALES, nous présentons la procédure que nous avons adoptée pour valider la Proximité Optimale d'Acquisition des sujets apprenant l'anglais juridique comme langue de spécialité.

Abstract :
This work is the product of an interdisciplinary research involving Artificial Intelligence, cognitive psychology and applied linguistics. LSA (Latent Semantic Analysis) is a computer tool which later became applied to language acquisition (Landauers et Dumais, 1997). We have been using LSA to elaborate a prototype : RAFALES (an automatized text collection favouring LSP acquisition). RAFALES is a computer tool designed to help foreign langage acquisition that borrows from Vygotsky's theory of Proximal Zone of Development (1968) and Krashen's Input Theory (1985). After exposing LSA and RAFALES theoretical basis, we are detailing the procedure through which we are going to validate subjects' Optimum Acquisition Proximity (POA) of Law English.

Mots clés :
Language acquisition, comprehension, cognitive psychology, computerized text analysis, English for Specialists.

Introduction

Dans les années 80-90, les travaux menés dans le cadre de l'analyse formelle automatisée de la langue intéressent de plus en plus les chercheurs en ALE (Acquisition des Langues Etrangères), en particulier ceux qui développent des programmes d'enseignement des

langues assisté par ordinateur. En effet, que ce soit en éducation, en linguistique ou en psychologie de nombreuses applications informatiques ont vu le jour créant, en même temps, de nouveaux champs de recherche liés à l'outil informatique : les technologies éducatives, le travail collaboratif assisté par ordinateur, la linguistique computationnelle et la linguistique de corpus.

Sur le plan théorique, LSA (Latent Semantic Analysis), un modèle associationniste, que nous présenterons plus en détail dans la deuxième partie de l'article, ramène à l'ancien débat acquisition/acquisition, comme l'indique explicitement l'article fondateur de Dumais et Landauer (1997). En effet, leur référence à Platon reprend le paradoxe maintes fois évoqué par Chomsky : comment se fait-il que nous connaissions et exprimions tant de choses quand les stimuli fournis par notre environnement ne suffisent pas à expliquer cette richesse ? La réponse de Chomsky (2000) fut computationnelle et innéiste, celle des connexionnistes ou de LSA sera computationnelle et non-innéiste.

Latent Semantic Analysis : une approche sémantique de l'acquisition.

LSA a été créé par les laboratoires Bellcors en 1989. Au départ, LSA était un outil de recherche documentaire (Deerwester et al., 1990) mais, en fonction de ses performances, son utilisation a été étendue au filtrage d'informations (Foltz et Dumais, 1992), à l'évaluation automatique de copies (Foltz, 1996 ; Lemaire et Dessus, 2001) et à la modélisation de l'acquisition (Landauer et Dumais, 1997).
Nous allons, dans un premier temps, expliquer comment LSA fonctionne, puis nous détaillerons certaines de ses applications.

La méthode

LSA analyse un large corpus de textes par le biais d'une analyse statistique et représente le sens de chaque mot, paragraphe et texte, par un vecteur, dans un espace de grandes dimensions.

LSA est un outil automatique ne nécessitant que très peu l'intervention d'un humain. En effet, la seule intervention de l'humain réside dans la fabrication du corpus qu'il va lui donner, c'est-à-dire dans la sélection des textes qui vont permettre à LSA de créer ses "connaissances du domaine". Par exemple, si nous voulons travailler en anglais des sciences sociales, nous aurons recours à un expert de ce domaine pour fabriquer ce corpus ; si, par contre, nous travaillons sur une langue "dans sa globalité", nous utiliserons un corpus contenant des textes divers et variés (livres, encyclopédie, article, etc.).

L'expert du domaine va donc fournir à LSA ce corpus de textes. LSA l'analyse par le biais d'une analyse statistique, et crée ainsi un espace sémantique de grande dimension (environ 300) qui contient tous les mots, paragraphes et textes. Cet espace sémantique est construit en prenant en compte le nombre de fois où chaque mot apparaît dans chaque partie de texte (les paragraphes) ; les mots outils tels que les articles et les pronoms ne sont pas pris en compte. Par exemple, si le corpus est formé de 300 paragraphes contenant au total 2000 mots différents, nous obtiendrons une matrice 300 x 2000. Chaque mot est alors représenté par un vecteur à 300 dimensions et chaque

paragraphe par un vecteur à 2000 dimensions. Ce sont ces vecteurs qui représentent le sens des mots.
Le sens d'un mot est donc donné par tous les mots qui sont proches de lui dans les différents paragraphes où il apparaît.
Puis cette matrice est réduite. C'est dans cette réduction que réside la puissance de LSA. En effet, c'est ce processus qui induit les similarités sémantiques entre mots. Tous les vecteurs sont réduits par une méthode proche d'une décomposition aux valeurs singulières. Cela permet de ne garder que les dimensions ayant les valeurs les plus élevées. Le nombre de dimensions est très important ; il doit se situer entre 100 et 300 afin d'obtenir les meilleurs résultats dans le domaines des langues (Landauer et Dumais, 1997).

Cette réduction est au cœur de la méthode car elle extrait les relations sémantiques : si un mot, par exemple *mouse*, co-occurre avec des mots, par exemple *cat*, *cheese*, qui co-occurrent avec un second mot, par exemple *mice*, et que le premier mot ne co-occurre pas avec des mots tels que : *rice*, *television*, qui ne co-occurrent pas non plus avec le second mot, alors les deux mots sont considérés comme proches.

Les similarités entre mots ou paragraphes sont calculées à partir des cosinus entre les vecteurs les représentant. Une mesure de similarité sémantique a une valeur comprise entre –1 et 1.
Cette méthode est très puissante : un mot peut être considéré comme proche sémantiquement d'un autre mot sans jamais apparaître dans le même texte. De la même façon, deux documents peuvent être proches sans avoir aucun mot en commun. Une intéressante fonctionnalité de cette méthode est que l'information sémantique ne provient que du niveau lexical. Il n'est pas nécessaire de représenter la théorie du domaine par un réseau sémantique ou une formule logique.

Les différentes validations

Nous allons détailler quelques recherches dans trois domaines : la recherche documentaire, car elle a été le point de départ de la création de LSA, puis l'évaluation de copie et, pour finir, l'acquisition de connaissances.

Au départ, LSA a été créé comme outil de recherche documentaire. Un problème dans le champ de la recherche documentaire est de retrouver un texte à partir d'une liste de mots-clés. En effet, avec les problèmes de polysémie, de synonymie, d'inflexion, retrouver le texte qui contient uniquement un ou plusieurs des mots-clés n'est pas facile. Par exemple, le livre de Steinbeck "*Of mice and men*" ne peut pas être retrouvé en donnant les mots-clés *mouse* et *man* avec un outil classique mais le sera avec LSA alors qu'aucun de ces mots n'apparaît, sous cette forme, dans le titre. LSA permet un gain en efficacité d'environ 36 %.

Un autre domaine d'application est l'évaluation de copies. Dans plusieurs de ces recherches (Foltz, 1996 ; Kintsch, 2001; Lemaire et Dessus, 2001; Wiemer-Hastings, 1999, Wolfe, 1998) il est demandé aux sujets d'écrire une synthèse à partir de textes d'un domaine donné. Les copies sont classées par des juges humains. Il leur est demandé de juger l'adéquation entre la copie et les textes. En parallèle, LSA s'entraîne avec les

textes et classe les copies en fonction de la proximité entre elles et chacun des textes. Les résultats de LSA sont comparables à ceux des humains. La corrélation entre les juges humains et LSA est proche de 0.6 pour la totalité des études, ce qui est similaire à la corrélation entre les corrections de deux humains pour l'évaluation d'un même texte.

Une expérimentation (Landauer et Dumais, opus cité) consistait à construire un espace sémantique général à partir d'un large corpus de textes en anglais, puis à le tester sur la partie synonymes du TOEFL (Test of English as a Foreign Language), qui est composé de 80 questions. A partir d'un mot donné, il faut identifier parmi 4 mots, celui qui est le plus proche sémantiquement. LSA a passé le test en choisissant parmi les 4 mots celui pour lequel il y a la plus grande similarité entre son vecteur et celui du mot donné. LSA a obtenu un score de 51.5 alors que la moyenne des étudiants étrangers admis dans les universités américaines est de 51.6. A notre connaissance, il s'agit du premier système capable d'effectuer un exercice standard sans avoir recours à des connaissances sémantiques supplémentaires.

LSA : un modèle d'acquisition de connaissances

A partir de ces résultats dans les domaines de la recherche documentaire et du classement de copies, LSA est donc considéré comme un modèle d'acquisition de la langue en psychologie cognitive.

Landauer et Dumais (opus cité), ont comparé l'acquisition du vocabulaire par LSA et par un enfant américain allant à l'école et sachant lire. Ces auteurs estiment qu'un tel sujet, entre 2 et 20 ans, lit environ 3500 mots par jour, et apprend entre 7 et 15 mots par jour au cours de cette période. Ils ont donc simulé ce modèle, en fournissant à LSA un nombre similaire de textes. LSA apprend ainsi 10 mots par jour pour obtenir une performance similaire à celle d'un humain à l'âge de 20 ans (définie par la performance du TOEFL, exercice décrit avant).

Nous allons utiliser LSA au sein de notre prototype pour modéliser les connaissances du domaine et les connaissances de l'élève mais aussi pour sélectionner les textes les plus appropriés pour optimiser l'acquisition du vocabulaire de spécialité.
Dans le prototype que nous avons développé, nous sélectionnons les textes qui nous semble les plus appropriés afin d'optimiser l'acquisition du vocabulaire de spécialité. Nous avons choisi de travailler en acquisition d'une langue naturelle, mais, signalons qu'il est aussi possible de travailler sur des langues non naturelles telles que les jeux par exemple. En effet, des recherches sur l'acquisition de jeux tels que khala ou tic-tac-toe, ont montré que l'acquisition est meilleure quand elle est fondée sur des informations sémantiques et non uniquement sur des informations syntaxiques (Lemaire 1998, Zampa et Lemaire à paraître).

Présentation du prototype RAFALES

Notre prototype RAFALES est un tuteur intelligent et, comme tous les tuteurs intelligents, il comporte trois modules (Wenger, 87) : la base de connaissances du domaine, la base de connaissances de l'élève et le module pédagogique. Chacun de ces modules utilise LSA ce qui permet de n'avoir qu'un seul formalisme pour modéliser toutes les connaissances.

Nous utilisons donc LSA pour modéliser les connaissances du domaine et de l'élève et cela pour différentes raisons. En premier lieu, des expérimentations ont déjà testé LSA en tant que modèle de représentation des connaissances (Landauer et Dumais, 1997). En second lieu, LSA permet une construction automatique de la base de connaissances du domaine, sans avoir recours à des humains ; cette dernière se construit tout simplement, en fournissant des textes à LSA.

Pour construire le module des connaissances du domaine de RAFALES nous fournissons à LSA deux types de textes : des textes de la langue générale ainsi que des textes appartenant au discours spécialisé. Dans le cadre de notre expérimentation, nous travaillons sur l'acquisition du droit constitutionnel américain. Nous avons choisi ce domaine pour plusieurs raisons. La première est qu'il correspond à notre domaine de compétence de professeur d'anglais de spécialité puisque nous disposons d'une double compétence en anglais et sciences politiques. La deuxième est que nous estimons judicieux de choisir un domaine universitaire de spécialité dans lequel le lexique de spécialité est le moins possible transparent entre l'anglais et le français ; autrement dit, en dépit d'une forte transparence graphique due à l'origine latine de nombreux mots, il existe, souvent, une nette distinction conceptuelle, qui émane de l'histoire. Souvent, la graphie peut être presque identique alors que le concept change radicalement. Le mot *radical* est d'ailleurs un bon exemple de ces fausses proximités !

Notre base de connaissances du domaine contient 1 013 174 mots, pour la partie base de connaissances en anglais général, répartis dans huit œuvres complètes et 1 123 362 mots répartis dans six cent soixante dix sept textes, pour la partie base de connaissances de la langue de spécialité. La plupart des textes proviennent de la Toile, et le reste nous a été fourni par divers collègues enseignant l'anglais juridique auxquels nous avons demandé de l'aide, *via* la SAES , pour constituer notre corpus.

Avec RAFALES nous nous situons dans le cadre de l'acquisition d'un lexique par l'exposition des sujets à des textes. L'hypothèse selon laquelle l'acquisition d'une seconde langue exige avant toute chose une quantité minimum d'exposition à la langue, est maintenant admise par les chercheurs en acquisition des langues. Nous testons donc cette hypothèse dans le domaine de la compréhension et de l'acquisition du vocabulaire en leur faisant lire des textes. Au cours de l'expérimentation, la lecture est le seul travail qui leur est demandé ; toute activité de type méta-linguistique portant sur le domaine de l'anglais juridique ou de la langue de spécialité est bannie.
La tâche de nos sujets consiste donc à lire des textes sélectionnés par RAFALES dans la base de connaissances du domaine en tenant compte du modèle de l'élève. Ce dernier est initialisé avec des textes de la langue générale (nous estimons qu'un élève de premier cycle a déjà été exposé à environ 1 000 000 de mots dans une langue étrangère au cours de sa scolarité), puis au fur et à mesure de l'utilisation du prototype, nous mettons à jour le modèle de l'élève en ajoutant les textes qu'il a lus.

Comme nous l'avons mentionné, le sujet apprend en lisant, mais nous pensons que le processus d'acquisition peut être accéléré en sélectionnant les bons textes, c'est-à-dire les plus appropriés au sujet en tenant compte de ses connaissances. Le problème est

donc de savoir quel texte a la plus grande chance d'élargir l'espace sémantique de l'élève. Il est évident que si l'on donne des textes trop proches ou trop éloignés de ce que l'élève a déjà acquis, il n'acquerra que peu ou pas de connaissances supplémentaires. Il faut donc fournir à l'élève des connaissances qui ne sont ni trop éloignées ni trop proches de ce qu'il connaît déjà. En référence aux théories de Vygotsky (1968) avec la notion de *zone proximale de développement* et celle de Krashen (1985) *the Input Hypothesis,* nous sommes en mesure de définir une proximité optimale d'acquisition (POA), grâce à LSA. Afin de valider (ou d'invalider) notre valeur de POA, nous avons fait une expérimentation.

Expérimentation du prototype

L'expérience s'est déroulée sur 5 séances, chacune durant à peu prés une demi-heure avec 10 sujets répartis dans 2 groupes. Le premier groupe lit les textes de la POA, le second les textes les plus éloignés de ce qu'il connaît déjà. Nous voulons ainsi savoir si le groupe auquel nous fournissons les textes les mieux situés progresse mieux que l'autre. Au départ, nous aurions souhaité pouvoir constituer quatre groupes incluant également un groupe aléatoire et un groupe auquel nous aurions fourni des textes situés juste dans la zone de connaissances du modèle des étudiants. Malheureusement, nous n'avons pas suffisamment de volontaires et nous avons dû nous limiter à deux groupes. De ce fait, notre travail n'a pas encore de statut expérimental, à proprement parlé, mais la pré-expérimentation à laquelle nous procédons nous permettra de savoir si les outils expérimentaux que nous avons élaborés sont pertinents ou non pour tester la validité de la POA.

Les hypothèses

A travers notre expérimentation nous essayons de valider différentes hypothèses :

- L'acquisition est optimale quand la sélection des textes se fait avec la POA.
- LSA permet de fabriquer un modèle de l'apprenant qui peut être testé.

Nous avons une hypothèse supplémentaire qui sera validée (ou invalidée) par des sujets différents : les experts:

- les réponses faites aux différents exercices par LSA et par des experts du domaine sont similaires.

Pour valider nos hypothèses notre plan d'expérience comporte trois types de variables.
La variable dépendante est la note aux exercices. Elle peut être divisée en plusieurs variables qui correspondent aux différents domaines de compréhension traités : compréhension générale, vocabulaire général, vocabulaire spécialisé.

La variable indépendante correspond à la manière dont les textes sont sélectionnés. Elle comporte deux modalités : textes les plus éloignés, textes ayant une POA.

Des variables contrôlées qui sont : l'homogénéité des groupes (pour cela nous avons utilisé un test de placement) et la quantités de mots lus par séance (2000 mots environ à chaque séance).

L'exercice de placement ou positionnement

Une séance préliminaire a consisté à faire faire aux étudiants un exercice de compréhension sur le modèle des exercices de positionnement afin de les répartir dans deux groupes équivalents.
Le texte source est un extrait d'un livre d'histoire et raconte l'invention de l'imprimerie (280 mots). Les questions sont des QCM à quatre propositions ; l'étudiant doit justifier sa réponse par une citation du texte. Les questions portent sur la nature du document puis sur la compréhension générale. Ensuite viennent des questions portant sur les idées explicites et implicites contenues dans le texte. Les inférences produites doivent conduire à une ré-élaboration du texte. L'ensemble du test exige de passer du stade de la construction à celui de l'intégration (Kintsch, 1988 ; Coirier & al, 1996 ; Fayol, 1997). Les résultats obtenus dans ce genre de tests permettent une ventilation assez large des étudiants dans les groupes. Les étudiants ont tout le temps nécessaire à la complétion de l'exercice.

Lors de la première séance, les sujets faisaient un exercice de vocabulaire, ensuite ils lisaient un texte et répondaient à des questions de compréhension sur ce dernier, puis ils lisaient des textes[1] fournis par RAFALES en fonction de leur groupe expérimental et finissaient en refaisant le même exercice de vocabulaire qu'en début de séance.
Les séances deux, trois et quatre, se déroulaient de la manière suivante : les sujets répondaient à l'exercice de vocabulaire, puis ils lisaient les textes fournis par RAFALES et terminaient en refaisant l'exercice de vocabulaire.
Lors de la dernière séance, les sujets répondaient à l'exercice de vocabulaire, lisaient les textes fournis par RAFALES, lisait le texte sur lequel portaient les questions de compréhension et y répondaient puis ils terminaient avec l'exercice de vocabulaire.

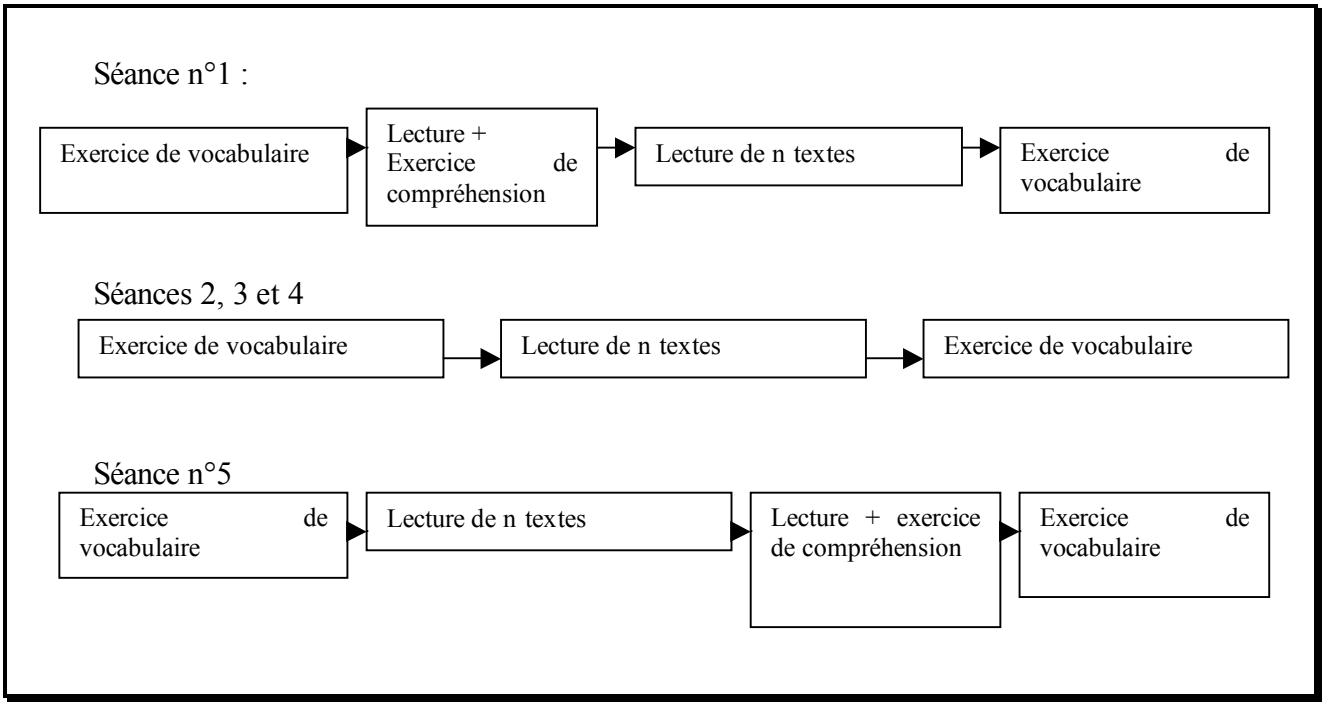


Figure 1 : schéma de déroulement de l'expérimentation

[1] Les textes ont été donnés sur papier. En effet, dans un premier temps, nous cherchons à valider un prototype et le fait d'effectuer l'expérimentation sur machine n'apporte rien.

Les exercices de compréhension

Nous avons deux exercices de compréhension, un en début de première séance et un en dernière séance. Ils portent sur les textes les plus centraux de la base de connaissances dans la langue de spécialité. Nous avons choisi de prendre les deux textes les plus centraux car ils contiennent les mots les plus fréquents de la langue de spécialité. Ils ont la même longueur. Ils sont construits sur le même modèle construction-intégration que le test de positionnement mais, cette fois-ci, ils s'inscrivent dans le domaine de l'anglais du droit constitutionnel. Il s'agit de deux commentaires d'arrêts de la Cour Suprême. L'un concerne la question de savoir si les employés d'un état sont protégés par une loi clef : la loi sur la discrimination à l'encontre des américains handicapés. L'autre concerne la possibilité pour des citoyens non affiliés à un parti de participer à la désignation d'un candidat dans le cadre des primaires. Le travail de compréhension porte sur des anaphores et des inférences allant du niveau lexical et syntaxique au niveau sémantique. Les commentaires d'arrêts se prêtent particulièrement bien au travail d'intégration modélisé par Kintsch. En effet, à mesure que l'étudiant avance dans le commentaire il est conduit à inhiber certaines informations qui jouent le rôle de 'distracteurs'. Par exemple, tel arrêt, telle décision, tel texte de loi, avancés au début du commentaire par telle ou telle partie estant en justice, va céder le pas à un autre arrêt plus récent ou à une loi plus ancienne dont le poids juridique sera jugé plus fort. Il sera donc inhibé sur le modèle de l'information contradictoire ou inattendue (Tapiero & Otero, 1999). Pour bien comprendre la décision finale et le commentaire, l'étudiant doit garder en mémoire de travail chaque proposition pertinente. Il doit se souvenir de son contenu sémantique, et de l'explication de son poids ou de sa faiblesse juridique. Le processus fait appel au travail de synthèse sémantique chez Walker et Meyer (1980) au travers duquel le texte lu fait l'objet d'une constante ré-évaluation à un niveau supérieur d'emboîtement et dépend fortement des connaissances du sujet dans le domaine. Ici, les deux textes sélectionnés par LSA comme les plus centraux posent un problème intéressant. Les exercices proposés ont été conçus de manière à rendre les tâches de compréhension aussi équivalentes que possible : même format, mêmes types d'inférence, même type de structure et de nature du discours. En revanche, pour un étudiant français la connaissance du domaine présente un grand écart : la question de l'intégration par le travail des handicapés et des minorités est une question qui se pose en France, tant sur le plan politique que social ou juridique, tandis que les processus conduisant à la désignation d'un candidat aux postes de gouverneurs, sénateurs ou attorney lui sont, *à priori*, étrangers. Aussi, si nous comparons les deux phrases ci-dessous, nous pouvons faire l'hypothèse qu'elles présentent des difficultés identiques sur le plan lexical, mais que la deuxième sera plus difficile à traiter à cause de la connaissance du domaine.

Texte1

Earlier this week, the court ordered lower courts to restudy rulings that said states and their agencies must abide by a 1963 federal law that requires employers to give men and women equal pay for equal work.

Texte2

The justices will hear arguments in the case in April. A decision is expected by July. Before 1996, California allowed only voters who were members of a political party to vote in

that party's primary to nominate candidates for the general election.

Les exercices de vocabulaire

Les 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 selon les modes de la synonymie, l'antinomie, la carte conceptuelle, l'absence de relation, ou le mot inconnu (voir infra, p.22).

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. Notons que les mots de spécialité sont presque tous également des mots appartenant à la langue générale, donc souvent polysémiques comme *bound*, *trial*, *suit*.

Nous avons également sélectionné au hasard, toutes les cinq lettres, des mots de la langue générale, dans le dictionnaire général du CNRS[2]. Nous ne nous préoccupons pas des effets de polysémie puisque LSA postule que, grâce aux effets de contexte et de proximité, la polysémie ne gêne pas l'acquisition du langage. Cependant, lorsque nous analyserons les résultats, nous distinguerons les trois catégories : mots appartenant uniquement au vocabulaire général, mots appartenant uniquement au vocabulaire de spécialité et mots appartenant aux deux vocabulaires, afin de savoir si les mots appartenant uniquement à la langue de spécialité s'acquièrent mieux que ceux appartenant au double registre. En effet, Tapiero et Otero (1999) ont montré que l'information la mieux retenue était l'information la plus surprenante comme des mots totalement étrangers ou des idées contraires aux hypothèses contenues dans le reste du texte (*inconsistent information*). En ce cas, les mots appartenant strictement au lexique de spécialité devraient être mieux rappelés. Mais seront-ils mieux acquis ? C'est ce que devrait vérifier les exercices de vocabulaire. Par ailleurs, LSA stipule que l'acquisition se fait par l'établissement de liens de proximité à travers des proximités de contextes. Un mot appartenant aux deux lexiques (général et de spécialité) devrait alors être plus facilement acquis du fait de la variété des contextes dans lequel il apparaît.

Pour l'exercice de début de séance, la consigne était donné en français

> Nous allons vous fournir une série de mots cibles. Pour chacun d'eux nous vous donnerons une liste de 5 mots et vous devrez indiquer le type de relation qui les unit au mot cible. Il y a quatre types de relation : même sens, sens contraire, mot d'un même domaine, et pas de relation.
>
> Vous pouvez aussi signaler que vous ne connaissez pas le mot en cochant la case "mot inconnu".

[2] Dictionnaire anglais en ligne du CNRS : http://dico.isc.cnrs.fr/dicocgi/anglais/areq

Pour chacune des relations de même sens, même domaine et sens contraire, que vous aurez pu établir, vous voudrez bien juger aussi de sa force (+ pour une relation forte ou – pour une relation faible).
Attention : Pour chacun des mots de la liste vous ne devez cocher qu'une seule case.

De même l'exemple était aussi donné en français pour les mêmes raisons

Réussir	Même sens		Même domaine		Sens contraire		Pas de relation	Mot inconnu
	+	-	+	-	+	-		
Ascaridiose								**X**
Erreur						**X**		
Echouer					**X**			
Examen			**X**					
Eau							**X**	

Livre	Même sens		Même domaine		Sens contraire		Pas de relation	Mot inconnu
	+	-	+	-	+	-		
Bande dessinée		**X**						
Ecrivain			**X**					
Bouquin	**X**							
Illustrateur				**X**				
Disserter				**X**				

Pour chaque mot nous vous fournirons un tableau tel que ceux ci-dessus que vous devrez remplir (faites une croix pour donner votre réponse).
Remarque : Les relations sont indépendantes de la nature grammaticale (verbe, nom, adjectif, etc.) des mots.

Pour l'exercice de fin de séance la consigne était la suivante :

Vous allez maintenant refaire le même exercice de vocabulaire qu'en début de séance.
En effet, nous voulons savoir si vos lectures ont modifié ou non la nature et/ou la force de certaines relations.
Le but de votre travail n'est pas de vous rappeler les réponses que vous avez déjà données, mais plutôt de réfléchir aux relations entre les mots.

Les textes sont sélectionnés grâce aux deux méthodes suivantes :

- Sélection des textes les plus éloignés :

LSA calcule la proximité entre chaque texte de l'espace du domaine et tous les textes de l'espace de l'élève. Pour chacun des textes de l'espace du domaine nous faisons la moyenne des proximités avec les textes de la base de l'élève. Puis nous sélectionnons les n textes ayant les moyennes de proximités les plus basses et n'appartenant pas déjà au modèle de l'élève.

Exemple : soient e1 à e4 les textes contenus dans le modèle de l'élève, et soient d1 à d10 les textes contenus dans l'espace des connaissances du domaine.
Nous calculons un tableau des proximités et ajoutons la ligne M qui correspond à la moyenne des proximités :

	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*	*d8*	*d9*	*d10*
e1	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1*
e2	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1*
e3	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1*
e4	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1*
M	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.7*	*.8*	*.9*	*1*

Dans le cas où nous sélectionnons deux textes (n=2), nous choisissons ici les textes d1 et d2 car les moyennes de leurs proximités avec les textes du modèle de l'élève sont les plus faibles.

Remarque : la proximité est une valeur donnée sur une échelle allant de -1 à 1. Une valeur de 1 correspond à un texte très proche (proximité sémantique d'un texte avec lui-même) et une valeur proche de 0 correspond à des textes n'ayant aucun rapport.

Sélection des textes les plus proches, correspondant à la POA

La POA correspond aux textes les plus proches après exclusion des textes appartenant à un périmètre que nous jugeons "trop proche" des connaissances de l'élève. Nous sélectionnons ainsi les textes dont les moyennes des proximités sont à un écart-type de la proximité maximale***.***

Les différentes analyses à faire

Nous n'avons pas encore fini d'analyser les résultats, mais nous allons expliquer les différentes analyses que nous réalisons et dans quel but.

Validation de la première hypothèse : la POA permet une optimisation de l'acquisition.

Tout d'abord, nous allons vérifier les effets de notre prototype sur l'acquisition au travers de l'évolution des "notes" aux exercices de vocabulaire (cette notion de note est explicité dans la section « validation de la seconde hypothèse »). Cette vérification se fera à deux niveaux: au niveau de chaque séance et au niveau de l'ensemble des séances. Nous comparerons, tout d'abord, les notes obtenues à l'exercice de vocabulaire en début de séance à celles de fin de séance. Ensuite, nous observerons l'évolution de ces "notes" entre le premier test de la première séance et le dernier test de la dernière séance. Enfin nous analyserons les résultats obtenus aux deux tests de compréhension situés en première et en dernière séance.

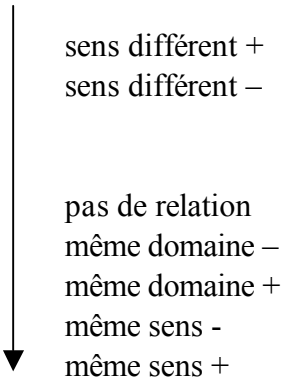
Nous allons comparer les résultats des deux groupes aux différents exercices ainsi que leurs évolutions. Nous allons aussi étudier comment évoluent les trois grandes catégories

de mots : mots appartenant à l'anglais juridique et à la langue générale ; mots appartenant seulement à l'anglais juridique, mots n'appartenant pas à l'anglais juridique.

Validation de la seconde hypothèse : LSA peut créer un modèle de l'apprenant

Pour vérifier notre seconde hypothèse, LSA va être mis dans les même conditions que le sujet, c'est-à-dire que LSA va simuler le sujet ; il va donc effectuer les exercices en ayant les mêmes connaissances que celles données par le modèle de l'apprenant. Par exemple, lorsque nous allons valider le modèle de l'apprenant du sujet 1, LSA effectue le premier exercice en n'ayant dans sa base de connaissances que le 1er texte lu par l'apprenant ainsi que les textes additionnels correspondant aux connaissances en anglais général et spécialisé qu'un élève moyen de DEUG est supposé avoir acquis. Nous avons grossièrement estimé cela à 1 000000 mots pour l'anglais général et aux 25 textes les plus centraux de la base de connaissances de la langue de spécialité. En prenant ce nombre de 25 nous obtenons le même nombre de fois la réponse "mot inconnu" par les sujets et par LSA.
Puis quand LSA effectuera le 2e exercice il aura en plus dans ses connaissances les n textes lu par le sujet entre les deux exercices. Et ainsi de suite.
La base de connaissances de LSA sera initialisée avant la simulation de chaque sujet.
Pour chaque sujet nous comparerons l'évolution des notes de LSA à celle des sujets. Si cette évolution est comparable et si les notes sont proches nous pourrons valider notre hypothèse.

Nos exercices de vocabulaire ne permettent pas de classer les réponses des sujets en terme de vrai/faux. Nous allons donc comparer les réponses des sujets aux réponses des experts. De plus nous cherchons à évaluer la progression entre le début et la fin de la séance. Afin de pouvoir comparer ces résultats, il faut pouvoir attribuer un score pour chacune des réponses des sujets. Ce score est calculé en faisant l'écart entre sa réponse et la norme. La norme correspond à la moyenne des réponses des experts.
Les réponses données par les sujets sont des relations de proximité entre deux mots. Nous plaçons toutes les réponses sur un continuum allant d'une distance de –4 à une distance de 4.



Donner le type de relation et la force entre deux mots correspond à donner une valeur de proximité sémantique. Nous fabriquons donc une échelle permettant de placer toutes nos réponses possibles. Nous centrons cette échelle sur 0 qui correspond au fait que les deux mots n'ont pas de relation. Puis nous mettons à 4 les mots de même sens avec une relation forte (ce qui correspond en quelque sorte à une relation de synonymie). Une

relation “sens différent ” avec une relation forte (qui correspond à une relation d’antonymie), a une valeur de –4 sur cette échelle. Nous avons donc les relations de synonymie et d’antonymie qui se trouvent à une égale distance de “pas de relation”. Puis nous plaçons sur ce continuum, à 1, les mots appartenant à un même domaine mais ayant une relation faible, et à 2 les mots appartenant au même domaine et ayant une relation forte.

Pour chaque question le score de l'étudiant est égal à la valeur absolue de la différence entre la valeur de sa réponse et la norme.

Exemple de calcul de la norme et de score du sujet :
L’expert 1 dit que la relation est de type même domaine avec une force faible, ce qui correspond donc à un score de 1.
L’expert 2 évalue cette même relation comme une relation de même sens faible, ce qui correspond à un score de 3.
L’expert 3 dit qu’il s’agit d’une relation de même domaine forte, ce qui correspond à un score de 4.
La norme ici est égale à : $(1 + 3 + 4)/3 = 2,5$

Si l’étudiant indique qu’il s’agit d’une relation forte de sens différent, ce qui correspond à un score de –4, son score sera égal à : $|2,5 – (-4)| = 6,5$

Cas particuliers :
- Quand un expert ne connaît pas un mot ou qu’il n’a pas répondu nous ne tenons pas compte de sa réponse (ou non réponse) pour le calcul de la norme.
- Quand l’étudiant ne connaît pas un mot, nous considérons que la distance entre sa réponse et la norme est égale à la valeur maximale, soit 8.

Une fois que nous avons calculé le score de l’étudiant au pré et post-exercice, nous évaluons son acquisition au cours de la séance. Cette acquisition correspond à l’évolution de la valeur absolue de l’écart entre le pré-exercice et la norme et le post-exercice et la norme. Elle est calculée de la manière suivante : score post-exercice – score pré-exercice.
Avec cette méthode de calcul, une acquisition maximale a une valeur de –8 (le sujet ne connaît pas le mot lors du pré-exercice (score de 8) et donne la même valeur que la norme lors du post-exercice (score de 0) ; 0-8 = -8) et une acquisition minimale a une valeur de 8 (le sujet donne la même valeur que la norme lors du pré-exercice et coche la réponse “mot inconnu” lors du post-exercice).

Vérification de la troisième hypothèse : Les réponses faites par LSA et les experts sont similaires.

Nous avons soumis ces exercices de vocabulaire à douze enseignants d’anglais, experts du domaine, et, afin de valider notre troisième hypothèse, nous avons comparé leurs réponses à celles données par LSA.

Conclusion

Nos travaux en sont encore au stade d'une pré-expérimentation et les résultats sont en cours de traitement. Nous essayons de mener un travail de type recherche-développement, c'est-à-dire d'élaborer un prototype, et de le tester afin de revenir sur nos hypothèses théoriques et didactiques. Ce type de recherche est peu fréquent en didactique des langues et en sciences de l'éducation, et, dans ce domaine, nous ne pouvons nous appuyer sur des travaux antérieurs ; c'est pourquoi nous avons besoin d'une phase importante de pré-expérimentation[3].

S'il est vrai que le cadre théorique relève d'une recherche fondamentale sur les processus d'acquisition du langage, notre souci est bien de déboucher sur des applications didactiques concernant l'acquisition d'une langue étrangère de spécialité. L'idée qu'il serait possible de constituer un logiciel de type RAFALES dans un domaine de spécialité, d'introduire ce programme dans des centres de langues ou des médiathèques pour que les étudiants puissent venir tester leurs connaissances et lire les documents fournis par la base de données comme convenant le mieux à leurs progrès, a semblé à nos collègues une idée prometteuse. Les travaux de Lemaire et Dessus (2001) sur la correction automatique des copies vont dans le même sens : alléger le travail de l'enseignant, faciliter le travail personnel de l'étudiant.

Françoise Raby est professeur agrégé d'anglais, diplômée de sciences politiques, maître de conférences en études anglaises à l'IUFM de Grenoble et chercheur au Laboratoire des Sciences de l'Education de Grenoble. Université Pierre Mendes France. Depuis 1992 ses travaux portent sur les TICE et visent à élaborer une ergonomie de la formation appliquée à l'apprentissage des langues étrangères.
http ://www.upmf-grenoble.fr/sciedu/fraby
Francoise.Raby@upmf-grenoble.fr

Virginie Zampa est doctorante et A.T.E.R en Sciences de l'Education au Laboratoire de Grenoble, après avoir obtenu un DEA en informatique "interface homme machine et ingénierie éducative" au Mans. Depuis la maîtrise ses travaux portent sur les EIAO d'acquisition des langues étrangères.
Virginie.Zampa@upmf-grenoble.fr

[3] C'est pour cette raison que nous avons présenté ce programme de recherche, avant même d'en connaître les résultats, au sous-groupe « droit » du GERAS et lors de l'atelier sur les TICE de la SAES. Nous remercions vivement les membres du sous-groupe pour l'accueil qu'ils nous ont réservé. Les résultats seront communiqués lors du prochain GERAS.

Références

Chomsky, N. *New Horizons in the Study of Language*. Cambridge: CUP, 2000.

Coirier, P., **Gaonac'h**, D**. Passerault,** J.-M**. (1996).** *sycholinguistique textuelle* **.** Paris : Armand Colin.

Deerwester, S.T.; **Dumais**, G.W.; **Launder**, T.K.; **Harshmann**, R. «Indexing by Latent Semantic Analysis. » *Journal of the American Society for Information Science* 41, 391-407 1990.

Fayol, M. (1997) Des idées aux textes. Psychologie cognitive de la production orale, verbale et écrite. Paris, PUF.

Foltz, P.W. «Latent Semantic Analysis for Text-Based Research. » Behavior Research *Methods, Instruments & Computers* 28.2, 197-202, 1996.

Foltz, P.W.; **Dumais**, S.T. «Personalized Information Delivery : An Analysis of Information Filtering Methods. » *Communications of the ACM* 35.12 (1992): 51-60.

Kintsch, W. *Comprehension : A Paradigm for Cognition*. New York: CUP, 1988.

---. «Metaphor Comprehension : A Computationnal Theory.»" *Psychonomic Bulletin and Review* 7.2,257-66, 2000.
---. «Predication.»" *Cognitive Sciences* 25.2 , 2001.

Krashen, S *The Input Hyptothesis : Issues and Implications*. London: Longman, 1985.

Landauer, T.K., **Dumais**, S.T. «A Solution to Plato's Problem : The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge.» *Psychological Review* 1.04 , 211-40, 1997

Landauer, T.K., **Laham**, D. , **Rehedr**, B., **Schreiner**, M.E «How Well Can Passage Meaning Be Derived without Using Word Order ? A Comparison of Latent Semantic Analysis and Humans.» *19th annual meeting of the Cognitive Science Society*. Ed. P. Langley M.G. Shafto. Mawhwah, N.J.: Erlbaum, 412-17, 1997.

Lemaire, B. «Models of High-Dimensional Semantic Spaces.» *International Workshop on Multistrategy Learning*. Desenzano, Italie, 1998.

Lemaire, B. «Tutoring System Based on Latent Semantic Analysis. » *AIED'99*. Ed. M. Vivet S.P. Lajoie. Le Mans: IOS Press, 527-34, 1999.

Lemaire, B., **Dessus**, P. «A System to Assess the Semantic Contend of Student Essays. » *Journal of Educational Computing Research* 24/3, 305-320, 2001

Tapiero, I., **Otero**, J. «Distinguishing between Text-Based and Situation Models in the Processing of Inconsistent Information : Elaboration Versus Tagging.» *The Construction of Mental Representation During Reading* in S.R. Goldman H. van Oostendorp (eds), Hillsdale, N.J.: Lawrence Erbaulm,341-65, 1999.

Vygotsky, L.S., Thought and language, 1934, A. Kozulin, Trans. Cambridge, MA: The MIT Press, 1968.

Walker, C.H., **Meyer**, B.J. « Integrating Different Types of Information in Text». *Journal of Verbal Learning and Verbal Behavior* 19, 263-75, 1980.

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

Wiemer-Hastings, P. **Wiemer-Hastings**, K et **Graesser**, A.C.. Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis, In S.P. Lajoie, M. Vivet (Eds*) Artificial Intelligence in Education (proceedings of the AIED'99 Conference)*, pages 535-542, IOS Press, 1999

Wolfe, M.B.W, **Schreiner**, M.E, Rehder, B., **Laham**, D., **Foltz**, P.W., **Kintsch**, W., **Landauer**, T.K. « Learning from Text : Matching Readers and Texts by Latent Semantic Analysis. » Discourse Processes 25, 309-36, 1998.

Zampa, V. **Lemaire**, B. (à paraître) Latent Semantic Analysis for Student Modeling, Journal of intelligent Information Systems, special issue on Education applications.