Acquisition de connaissances à partir de sélections automatiques de textes : tentative d'opérationnalisation de la Zone Proximale de Développement

Virginie Zampa

Acquisition de connaissances à partir de sélections automatiques de textes : tentative d'opérationnalisation de la Zone Proximale de Développement

Virginie ZAMPA

LSE - Université Pierre-Mendès-France ; BP47, F-38040 Grenoble Cedex 9,FRANCE
Tel : 04 76 82 77 64 ; Virginie.Zampa@upmf-grenoble.fr ; http://www.upmf-grenoble.fr/sciedu/vzampa

Résumé

Cet article présente un manuel scolaire numérique et son expérimentation auprès d'étudiants de second cycle universitaire et de stagiaires d'IUFM. RAFALES (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité) est capable de sélectionner automatiquement une séquence de textes visant à optimiser l'apprentissage. Le module pédagogique sélectionne les textes à fournir à l'apprenant en fonction des distances sémantiques fournies par LSA (Latent Semantic Analysis) entre son profil et la base de connaissances de la langue de spécialité étudiée. Nous pensons qu'il existe une distance sémantique optimale pour favoriser l'apprentissage, nous l'appelons POA (Proximité Optimale d'Acquisition). Cette POA est une tentative d'opérationnalisation de la zone proximale de développement.

Mots-clés : manuel numérique, proximité optimale d'acquisition, zone proximale de développement, LSA.

Abstract

This paper presents a numerical schoolbook and our experiments with it involving undergraduate students and IUFM (lecturer training school) students. RAFALES (french acronym for an automatic book favoring the acquisition of a foreign speciality language) is able to automatically select a text sequence that aims at optimizing the learning. The pedagogic module selects the texts that are given to the learner according to the semantic distances provided by LSA (Latent Semantic Analysis). This is the distance between the learner model and the knowledge database of the speciality language of concern. We claim that it exists an optimal distance that favors the learning; we call it OAP (Optimal Acquisition Proximity). This OAP represent an attempt to put in practice the theory of the zone of proximal development.

Keywords: electronic textbook, POA, LSA, zone of proximal development.

Introduction

Pour acquérir des connaissances en langue étrangère de spécialité les apprenants ont le plus souvent recours soit à un enseignement classique (avec enseignant) soit à des manuels scolaires. L'apprentissage avec un enseignant entraîne des contraintes d'horaires, de lieux, etc. Quant aux manuels scolaires, un inconvénient majeur réside dans le fait qu'ils ne sont pas personnalisés (ils ne varient pas en fonction de l'apprenant et ne correspondent pas toujours à ses attentes), qu'ils ne sont pas évolutifs (le choix des textes est fait une seule fois, avant l'impression du manuel) et que la sélection n'est pas automatique et 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, développé et expérimenté. Ce prototype tente une opérationnalisation de la zone proximale de développement [1]. De plus il tente d'apporter des pistes de recherche concernant l'individualisation, l'évolutivité et l'automatisation des manuels numériques.

Mais son utilisation ne se limite pas aux apprenants. Comme tous les manuels scolaires, il peut être utilisé par l'enseignant qui prépare son cours. En effet, RAFALES contient une grande base de données textuelles qui peut être exploitée par l'enseignant. De plus, si l'enseignant fournit un profil d'apprenant, RAFALES peut sélectionner pour lui les textes les plus adaptés à ses étudiants.

RAFALES est un manuel numérique d'acquisition en langue par exposition à des textes. Cette exposition se réalise au travers de la lecture des textes qu'il sélectionne pour l'apprenant. Ainsi l'unique tâche de l'apprenant est la lecture. Nous reprenons l'approche constructiviste du langage. Des travaux en psychologie cognitive montrent que la plupart des mots sont acquis par la lecture [2]. 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 [3]. L'apprenant, exposé à des textes va affiner petit à petit le sens de chaque mot grâce à ses différents contextes d'apparition. Ainsi, par exemple, sans lui définir explicitement, l'apprenant va acquérir le sens du mot « piscine » parce que ce mot apparaît avec d'autres comme « nager », « eau », « maillot de bain », dans les textes qu'il lit. Dans cette représentation nous voyons que le sens d'un mot est donné par ses occurrences conjointes, c'est à dire par les mots qui lui sont proches, tout comme l'a défini Saussure [4] en linguistique. Cependant depuis Platon et le fameux paradoxe de l'induction, il est indéniable

que ces simples cooccurrences répétées d'un mot avec d'autres ne suffisent pas à expliquer la connaissance que nous avons de ces 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 [2] montrent que 75% de la connaissance sémantique d'un mot proviendrait de la lecture de textes ne contenant pas ce mot.

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

Nous allons dans un premier temps présenter LSA, puis nous présenterons l'architecture et le fonctionnement de RAFALES, nous détaillerons ensuite le principe de sélection de texte (tentative d'opérationnalisation de la ZPD). Enfin nous analyserons nos retours d'expérimentation.

Latent Semantic Analysis

LSA a été développé par les laboratoires Bellcore en 1989, comme outil de recherche documentaire [5]. Mais très vite, grâce à ses performances, son utilisation s'est étendue à d'autres domaines comme nous allons le voir.

La méthode

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

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

Le principe est le suivant. LSA construit dans un premier temps la matrice d'occurrences. Il s'agit d'une matrice dont les lignes représentent les unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes les mots (plus précisément les graphies). L'élément *(i,j)* de la matrice correspond ainsi au nombre d'occurrences du mot *j* dans le paragraphe *i*. L'étape suivante consiste à réduire ces dimensions à environ 200. Ce nombre est important car une réduction à un espace trop grand ne fait pas suffisamment émerger les liaisons sémantiques entre les mots, et un espace trop petit conduit à une trop grande perte d'informations. Ce nombre de dimensions est issu de tests empiriques [5]. Cette réduction est réalisée par le biais d'une décomposition aux valeurs singulières. La réduction à *n* dimensions va consister à ne conserver que les *n* premières de ces valeurs pour reconstituer une matrice approchée, de dimension *n*. Chaque mot et chaque paragraphe, traité de la même façon dans cette procédure, est ainsi représenté par un vecteur à *n* dimensions.

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

Quelques applications et validations

Au départ, LSA a été développé comme outil de recherche d'information [6 ; 7]. Avec les problèmes de choix de mots-clés liés à la polysémie, aux inflexions et à la synonymie, il est aisé de postuler que la recherche devrait se faire sur le sens des mots et non sur leur « forme ». D'ailleurs des expérimentations avec LSA ont mis en évidence un gain allant de 16 à 30%.

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

admis dans les universités américaines (64,5%). Le comportement de LSA est donc un modèle intéressant de l'apprentissage du vocabulaire chez les humains.

Un troisième domaine d'application concerne l'acquisition de connaissances. Ces acquisitions peuvent concerner les langues [8] ou un domaine particulier comme celui traité dans un cours [9]. Elles peuvent concerner un langage et non une langue naturelle. C'est le cas par exemple avec l'apprentissage des jeux tels que le tic-tac-toe [10] ou kalah [11].

LSA est aussi utilisé de diverses manières dans des EIAH (environnement informatique d'apprentissage humain). Des travaux ont ainsi porté sur l'évaluation de copies, c'est le cas du système APEX (Aide à la Préparation aux EXamens) [12 ; 13]. D'autres travaux ont porté sur la notation de copies [14] 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 [15 ; 16] et la détection des erreurs dans une copie en langue étrangère [15].

RAFALES

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

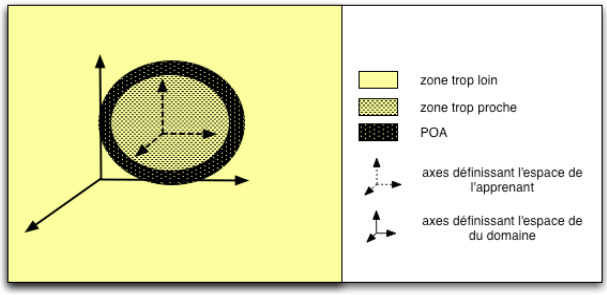
Nous pensons que cette acquisition peut être « optimisée » en fournissant à l'apprenant des textes qui sont à la proximité optimale d'acquisition (POA) de son profil. Cette sélection des textes est effectuée par le module pédagogique.

RAFALES possède l'architecture d'un tuteur intelligent [17]. Il comporte trois modules : le profil de l'apprenant, la base de connaissances du domaine étudié et le module pédagogique. Dans les trois modules, nous utilisons LSA afin de n'avoir qu'un seul formalisme. 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.

Fonctionnement

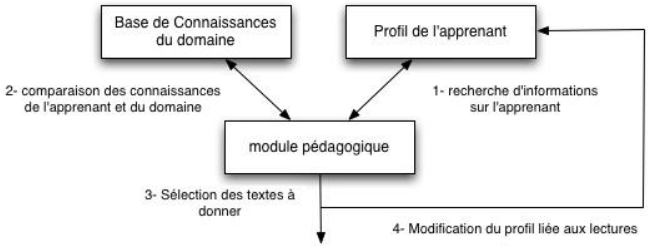
Le fonctionnement est simple. Le module pédagogique sélectionne en fonction des connaissances de l'apprenant les textes de la base de connaissances à lui fournir. Le profil de l'apprenant est un sous-espace de l'espace des connaissances du domaine étudié. Le module pédagogique sélectionne les textes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

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.

Initialisations pour l'expérimentation



Dans notre expérimentation principale du prototype, le domaine étudié est l'anglais juridique et plus particulièrement le droit constitutionnel américain.
La base de connaissances du domaine étudié comporte deux parties : une pour la langue étrangère « générale » et une pour la langue de spécialité. La base de connaissance de la langue étrangère générale contient environ 1 000 000 mots issus de huit oeuvres complètes. Ces oeuvres font partie du domaine public mais sont relativement récentes. Quant à la base de connaissance de la langue de spécialité, elle contient un peu plus de 1 000 000 de mots issus de textes de loi, de comptes-rendus de procès, etc.

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

peu de connaissances dans la langue de spécialité (d'où les 25 textes).

POA et ZPD

La définition donnée par Vygotsky de la Zone Proximale de Développement est la suivante : « c'est la distance entre le niveau de développement actuel tel qu'on peut le déterminer à travers la façon dont l'enfant résout des problèmes seul et le niveau potentiel tel qu'on peut le déterminer à travers la façon dont l'enfant résout des problèmes lorsqu'il est assisté par l'adulte ou collabore avec d'autres enfants plus avancés. ». En d'autres termes, la ZPD correspond à ce qui peut constituer la prochaine étape du développement. Mais pour Vygotsky la ZDP nécessite une interaction entre individus grâce à une activité conjointe avec le médiateur.

Dans notre module pédagogique, nous avons tenté une opérationnalisation de la ZPD [1] au travers de notre Proximité Optimale d'Acquisition, mais ceci se réalise sans médiation humaine. Dans notre prototype il ne s'agit pas de résoudre des problèmes mais d'acquérir une langue étrangère de spécialité, c'est à dire essentiellement acquérir du vocabulaire et des concepts. De ce fait, nous avons transposé et interprété la ZDP dans ce cadre précis. Nous avons ainsi interprété le fait que :

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

par :

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

Nous avons ainsi défini une POA, c'est-à-dire une distance ni trop grande ni trop petite pour sélectionner les textes à fournir à l'apprenant. En effet, si les textes sont trop proches de ce qu'il connaît, il n'apprendra que peu ou pas et si les textes sont trop éloignés, il ne les comprendra pas et n'acquerra pas non plus de connaissances.

Grâce à LSA, nous avons des distances sémantiques entre textes, nous pouvons ainsi déterminer à quelle distance du profil de l'apprenant se situe chacun des textes de la base de connaissances de la langue de spécialité étudiée. Nous avons fixé empiriquement cette POA à un écart-type du texte le plus proche du profil [18].

Expérimentation et Résultats

Les conditions de passation de l'expérimentation

Il y avait deux méthodes de passation pour l'expérimentation : version papier et version électronique. Quelle que soit la méthode, 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, le tout fourni dans l'ordre.

Les sujets qui passaient la version papier avaient tous les documents dès le début. Ils avaient donc en main les cinq séances complètes, avec chaque séance dans une pochette différente et dans chaque pochette les tests et textes adéquats. Les sujets qui passaient la version électronique recevaient les séances une par une, ils recevaient la séance suivante dès qu'ils avaient envoyé leurs réponses aux tests de la séance précédente.

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

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.

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

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

Les résultats

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

récence. Nous avons ainsi 100 couples de mots par test, soit 200 par séance, soit 1000 par sujet.

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

Une première analyse indique qu'une partie des réponses données par les sujets diffère entre le pré-test et le post-test. Ces moyennes vont de 24% (groupe aléatoire) à 33,75% (groupe loin). Mais il s'agit de modification ne tenant pas compte du sens de modification, c'est-à-dire ne regardant pas si la réponse du sujet se rapproche ou s'éloigne de la réponse normale. Toutefois cette évolution est intéressante car elle est indépendante (coefficient de corrélation à – 0,14) de la présence dans les textes lus des mots testés. Ceci signifie que la lecture a des effets même sur les mots qui ne sont pas lus, ce qui confirme les résultats de Landauer et Dumais [2].

Une seconde analyse porte sur les évolutions par rapport à la norme. C'est elle qui permet de valider (vs invalider) notre tentative d'opérationnalisation de la ZPD. 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. En effet, les moyennes des évolutions pour chacun des quatre groupes, sur les 500 couples de mots testés sont les suivantes :

POA	Proche	loin	aléatoire
0.0215	- 0.0098	- 0.0067	- 0.0099

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

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

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

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

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

Une troisième analyse porte sur l'évolution entre le premier pré-test (celui de la première séance) et le dernier post-test (celui de la cinquième séance). En effet, puisque nous considérons que nos cinq tests sont équivalents [18], l'évolution sur l'ensemble de l'expérimentation correspond à l'évolution entre le premier pré-test et le dernier post-test. Les moyennes des écarts entre les réponses des sujets et celles des experts sont données dans le tableau suivant.

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

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

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

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

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

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

Conclusion

Notre Recueil Automatique a pour but de Favoriser l'Acquisition d'une Langue Etrangère de Spécialité. Lors de son expérimentation, la langue de spécialité était le droit constitutionnel américain et les sujets étaient des étudiants de second cycle universitaire de

langue (Langue Etrangère Appliquée et Langue et Civilisation Etrangères).

Dans RAFALES nous avons choisi de n'utiliser qu'un seul formalisme : LSA, pour représenter les connaissances et tenter d'opérationnaliser la ZPD au sein du module pédagogique.

L'expérimentation a permis aux sujets du groupe POA de se rapprocher des réponses des experts de façon significative. Ceci a été évalué en comparant les écarts aux réponses des experts entre le début et la fin de l'expérimentation. Cette différence n'est significative qu'avec ce groupe et il est possible de parler de tendance avec le groupe proche. Ces résultats sont en accords avec d'autres expériences [19]. En effet, il est préférable de fournir des connaissances trop proches, même si elles n'apportent que peu de connaissances nouvelles, que de présenter des connaissances trop éloignées, qui de ce fait ne seront pas (ou mal) comprises par l'apprenant.

Notons toutefois que, quel que soit le groupe, l'écart aux experts diminue entre le premier pré-test et le dernier post-test. Suite aux lectures, quelles qu'elles soient, les apprenants se rapprochent tous des réponses des experts. Ceci n'est pas significatif avec les groupes loin et aléatoire car les écarts-type sont grands. De plus l'expérimentation n'a lieu que sur cinq séances, soit une lecture totale de 10 000 mots (tout type de mot compté), ce qui est peu pour mesurer une acquisition de vocabulaire.

Le groupe POA est le seul qui obtienne une évolution positive sur les 500 couples de mots. C'est-à-dire qu'il est le seul à se rapprocher des réponses des experts. De plus, les différences entre ce groupe et les trois autres sont significatives ($p < 0.05$). Il semble donc qu'il existe un effet de la distance sémantique entre le profil et les textes lus, sur l'évolution des connaissances sémantiques. La POA que nous avons fixée à un écart-type des textes les plus proches, n'est sûrement pas la meilleure distance, mais elle semble tout de même mieux adaptée que les trois autres conditions que nous avons ici, c'est à dire « loin », « proche » et « aléatoire ». Elle correspond à une première tentative d'opérationnalisation de la ZPD en utilisant LSA. Il serait intéressant de voir si cette distance de un écart-type reste plus efficace avec une autre langue de spécialité et/ou avec des sujets ayant un niveau de départ différent.

Références

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

[2] 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* 104 : 211-240.

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

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

[5] 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*, 41 : 391-407.

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

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

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

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

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

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

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

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

[14] 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*, 1(2).

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

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

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

[18] 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.

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