

Évaluations comparées de deux méthodes d'acquisitions lexicale et ontologique : Jeux De Mots vs Latent Semantic Analysis

Virginie ZAMPA - LIDILEM Université Grenoble3

Mathieu LAOFURCADE – LIRMM Université Montpellier2

Mot-clés : LSA, JeuxDeMots, corpus, ontologie, acquisition de relations entre termes

I. Introduction

Pour travailler en traitement automatique du langage naturel (TALN) qu'il s'agisse de faire de la correction automatique, de l'indexation de document, etc. il est indispensable d'avoir des informations lexicales. Il s'agit ainsi de construire une ontologie, de notre point de vue, ensemble de connaissances a-domaine, représentant des connaissances à la fois sur le monde et sur la langue. Par exemple, le mot « neige » pourra être associé à « froid » et « blanc » comme caractéristiques relevant du monde, mais également à « lourde » ou « poudreuse » comme phénomène linguistique pour nommé des états possibles. En TALN ces deux aspects sont primordiaux pour la compréhension.

Les informations constituant les ontologies peuvent être collectées et traitées en utilisant différentes méthodes. Nous présentons ici deux approches : l'analyse de la sémantique latente (LSA) et le projet JeuxDeMots (JDM).

Avec LSA, il suffit de fournir des textes, les relations entre mots ou textes sont issues de traitements réalisés automatiquement. Le choix du corpus : ce qu'il contient, sa longueur, le type de langage utilisé, etc. est donc primordial. Il s'agit ainsi d'une analyse automatique permettant d'obtenir des proximités sémantiques entre des mots, entre deux textes, entre un mot et un texte. Ces proximités ont une valeur comprise entre -1 et 1 (valeur expliquée plus en avant), mais en aucun cas les relations ne sont typées.

Avec le projet JDM, il s'agit de faire participer un grand nombre de personnes en leur proposant une application ludique accessible sur le web. À partir d'une base de termes préexistante, ce sont ainsi les joueurs qui vont construire le réseau lexical, en fournissant des associations qui ne sont validées que si elles sont proposées par au moins une paire d'utilisateurs. De plus, ces relations typées sont pondérées en fonction du nombre de paires d'utilisateurs qui les ont proposées.

Ces deux méthodes diffèrent donc par de multiples aspects dont notamment :

- la constitution du corpus : dans LSA les corpus sont interchangeable (il suffit de modifier la base de textes) alors que dans JDM il est lié aux joueurs ;
- la fabrication des associations entre termes : dans LSA les associations sont issues des contextes dans lesquels les termes apparaissent, alors que dans JDM, il s'agit d'associations libres et spontanées (pouvant potentiellement être fausses) ;
- le niveau de langue utilisés : dans LSA il dépend du corpus retenu mais reste essentiellement soutenu contrairement à JDM dans lequel le niveau de langue varie de soutenu à populaire.

Nous allons donc, dans un premier temps présenter ces deux méthodes, leurs utilisations et validations ainsi que leurs limites. Dans un second temps, nous les comparerons et enfin nous regarderons ce que leur combinaison pourrait nous apporter.

II. jeux de mots

JeuxDeMots est un jeu en ligne créé en 2007 (<http://www.jeuxdemots.org>) qui a pour but la construction d'un réseau lexical. Ce dernier est composé de termes et de relations étiquetées et pondérées. Les relations peuvent être ontologiques (hyper/hyponymie, partie/tout, caractéristiques typiques, etc), lexicales (synonymes, antonymes, locutions, etc.) ou libres. Par exemple, pour le terme neige, les relations les plus actives dans le réseau lexical de JDM sont les suivantes :

relation	poids	mot
association libre	350	froid
	320	ski
	290	hiver
	260	blanc
locution	220	bonhomme de neige
association libre	190	montagne
locution	170	boule de neige
association libre	140	flocon
locution	130	blanc comme neige
magn		glace
idée associée	120	
locution		flocon de neige
magn		glacier
hyponyme	110	poudreuse
locution	100	neige poudreuse
magn		névé
association libre	90	boule
locution		neige artificielle
magn		neige carbonique
		neiges éternelles
association libre	80	luge
hyponyme		neige carbonique
partie de		flocon
magn		tempête de neige
association libre	70	eau
hyponyme		carbonique
partie de		eau
locution		glace
		chute de neige
association libre	60	bonhomme
locution		flocons
caractéristique typique		noël
		oeufs en neige
		tempête de neige
		blanc

Ainsi, la structure du réseau lexical que nous cherchons à obtenir se fonde sur les notions de nœuds et de relations entre nœuds, comme présentées ci-dessus [Polguère 06]. D'une façon générale, chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions lexicales, telles que présentées par [Mel'čuk et al. 95].

Ce type de réseau lexical constitue une ontologie. En effet, il comporte des objets faisant référence à des objets du monde ou à des objets linguistiques. Les relations entre ces objets sont explicitées comme nous venons de le voir ci-dessus. De plus, un point important de notre approche, est l'ajout d'une pondération sur chaque relation traduisant implicitement son intensité.

1. Présentation de la méthode

En pratique, les validations des propositions sont faites par concordance entre paires de joueurs. Ce processus de validation rappelle celui utilisé par [von Ahn & Dabbish 04] pour l'indexation d'images ou plus récemment par [Lieberman et al. 07] pour la collecte de "connaissances de bon sens". À notre connaissance, il n'a jamais été mis en œuvre dans le domaine des réseaux lexicaux.

Une partie se déroule entre deux joueurs, en asynchrone, et est fondée sur la concordance de leurs propositions. Lorsqu'un premier joueur que nous appellerons (A) débute une partie, une consigne concernant un type de compétence est affichée (synonymes, contraires, domaines, etc.), ainsi qu'un terme T tiré aléatoirement dans une base de mots. Par exemple il est demandé au joueur de donner des « idées associées » au terme « télésiège » comme le montre la copie d'écran ci-dessous. Ce joueur (A) a alors un temps limité (60 secondes) pour répondre en donnant des propositions correspondant, selon lui, à la consigne appliquée au terme T.



Figure 1 : partie en cours

Remarque : les mots à droite (« neige », « ski », « piste », etc.) correspondent aux mots que l'utilisateur vient de rentrer.

Ce même mot, avec cette même consigne, est proposé par la suite à un autre joueur que nous appellerons (B) ; le processus est identique. Afin d'accroître l'aspect ludique, pour toute réponse commune dans les propositions de (A) et (B), ces deux joueurs gagnent un certain nombre de points (cf figure 2).

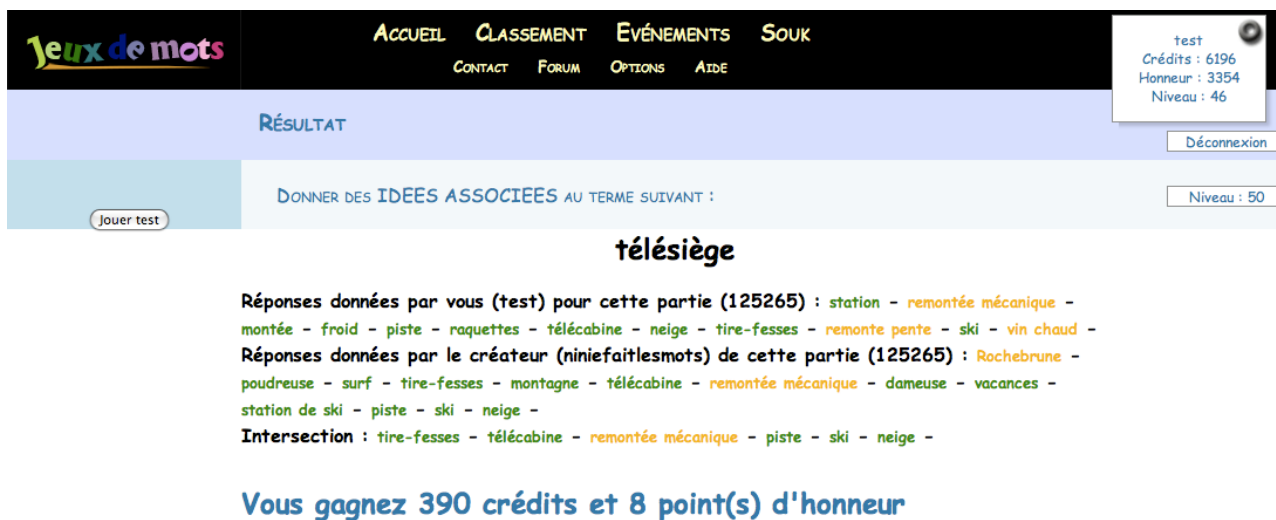


Figure 2 : résultat de la partie

Pour le terme cible T, nous mémorisons les réponses communes aux joueurs (A) et (B). Nous ne mémorisons pas les réponses proposées uniquement par l'un des deux joueurs. Cela permet la construction d'un réseau lexical reliant les termes par des relations typées et pondérées, validées par paires de joueurs. Ces relations sont typées par la consigne imposée aux joueurs ; elles sont pondérées en fonction du nombre de paires de joueurs qui les ont proposées. Initialement, les nœuds sont constitués des termes de notre base de départ, mais celle-ci peut s'accroître ; effectivement, si les deux joueurs (A) et (B) d'une même partie proposent un terme initialement inconnu, alors ce terme est ajouté à notre base.

2. Limites

Malgré tout, cette approche comporte certaines limites :

Premièrement, il s'agit d'un vocabulaire actif, forcé par la consigne. Le vocabulaire est dit actif au sens vocabulaire faisant partie du discours de tous les jours du joueur. Une grande partie du vocabulaire passif (connu mais non régulièrement usité) reste donc non proposée. De plus, il est forcé par la consigne, car le joueur donne ses réponses en fonction de ce qui lui est demandé et non ce qui lui vient en premier à l'esprit à part dans le cas de la consigne « association libre ».

Deuxièmement, certains joueurs ont recours à des ressources externes telles que wikipédia ou des dictionnaires de synonymes. Cela introduit un biais car il ne s'agit pas du vocabulaire actif, mais cela permet aussi d'introduire dans le réseau du vocabulaire plus soutenu. Ceci compense partiellement la limite précédente.

Troisièmement, les associations sont fortement liées à l'actualité. Le réseau évolue ainsi avec le temps et est représentatif des relations sémantiques à un instant t. Par exemple, actuellement, le terme « Amérique » sera très fortement en association avec « Obama » ce qui n'était pas forcément le cas il y a quelques temps.

Enfin, les connaissances sont issues d'un échantillon non représentatif de la population. En effet, toutes les tranches d'âge ne sont pas représentées, ni toutes les classes sociales.

III.LSA

LSA a été brevetée en 1988 et publiée en 1990.

1. Présentation de la méthode

Le but de LSA est de représenter dans un espace multidimensionnel (300 dimensions) les mots de la langue. Grâce à une analyse statistique, le sens de chaque mot est caractérisé par un vecteur. La proximité de sens entre deux mots correspond à la proximité entre les vecteurs.

Pour construire cet espace LSA prend un ensemble de textes en entrée et construit une matrice d'occurrences qui est réduite par le biais d'une analyse statistique. Ceci permet ainsi de faire ressortir les relations sémantiques entre mots ou entre textes.

Grâce à cette méthode, deux mots peuvent être considérés sémantiquement proches même s'ils n'apparaissent jamais conjointement dans un texte. Il suffit qu'ils soient utilisés dans des contextes similaires. Le contexte d'un mot est ici défini comme l'ensemble des mots qui apparaissent conjointement avec lui. Ainsi, les mots vélo et bicyclette sont considérés sémantiquement proches car ils apparaissent tous deux avec des mots comme randonnée, guidon, pédaler, etc., et ils n'apparaissent qu'occasionnellement avec des mots comme bouillir, imprimante, vase, etc.

Cette notion de cooccurrence est statistique : la méthode fonctionne si un nombre suffisant de textes est utilisé. Il ne s'agit pas simplement d'un comptage d'occurrences, il faut aussi disposer

d'une procédure pour établir les liaisons entre mots. Cette procédure est la réduction de la matrice d'occurrences. Le principe est donc le suivant. Dans un premier temps LSA construit la matrice d'occurrences. Il s'agit d'une matrice dont les lignes sont des unités textuelles (l'unité généralement utilisée est le paragraphe) et les colonnes, des mots. L'élément (i,j) de la matrice correspond ainsi au nombre d'occurrences du mot j dans le paragraphe i . L'étape suivante va consister à réduire ces dimensions à environ 200 dimensions. Ce nombre est important car une réduction à un espace de trop grande dimension ne ferait pas suffisamment émerger les liaisons sémantiques entre mots que nous recherchons, et un trop petit nombre de dimensions conduirait à une trop grande perte d'informations. Ce nombre adéquat de dimension est issu de tests empiriques, dans le cas de l'anglais il se situe entre 100 et 300 [Deerwester et al. 90]. Cette réduction des dimensions est réalisée par le biais d'une décomposition aux valeurs singulières. La réduction à n dimensions va consister à ne conserver que les n premières de ces valeurs pour reconstituer une matrice approchée, de dimensions n^2 . 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 étant construit, il faut choisir la façon de mesurer la proximité entre deux éléments. Les tests empiriques réalisés 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. La valeur 1 indique ainsi une très forte proximité sémantique entre les termes.

2. Utilisations, validations, limites

LSA a déjà été utilisé et validé dans différents domaines tels que :

- la recherche d'information. LSA permet de limiter les problèmes de choix de mots-clés liés à la synonymie, à la polysémie et à l'inflexion. La recherche se fait sur le sens des mots clés et non uniquement sur leur «forme» [Dumais 94], [Dumais 97].
- l'acquisition de connaissances. Ces acquisitions peuvent concerner les langues [Landauer & Dumais 97], [Redington & Chater 98], les jeux tels que le tic-tac-toe [Lemaire 98] ou kalah [Lemaire 99].
- les EIAH (environnement informatique pour l'apprentissage humain). Dans ces environnements LSA a été utilisé pour modéliser les connaissances, pour évaluer des copies ou résumés, etc. [Dessus et al. 00], [Zampa & Lemaire 02], [Zampa & Raby 01], [van Bruggen et al. 04], [Dikli 06].
- les modélisations cognitives lors de la compréhension des métaphores [Lemaire & Bianco 03] ou lors de la rédaction de résumés [Lemaire et al. 05].
- la mesure de la cohérence textuelle [Miller 03].
- la compréhension de texte [Kintsch 00], [Kintsch 01] [Kintsch 02]
- la modélisation de la mémoire [Kintsch et al. 99], [Denhière & Lemaire 04], [Howard & Kahanna 02], [Howard & Kahanna 07]

Mais, même si LSA a été utilisé et validé dans de nombreux domaines et provoque actuellement un certain engouement, il montre certaines limites.

Tout d'abord les connaissances sur les mots dépendent fortement du corpus utilisé pour construire l'espace de connaissance. Par exemple, si le corpus ne contient que des textes issus d'un journal tel que « le monde diplomatique », le vocabulaire sera relativement soutenu, ne contiendra pas un certain nombre de mots couramment utilisés.

De plus, la nature de la relation entre les mots n'est pas spécifiée, ce qui peut poser problème selon l'utilisation voulue. Il peut s'agir d'un synonyme, d'un mot du même domaine, d'une partie, d'une spécification, etc. Par exemple, dans les mots les plus proches de « salade » se trouvent des mots tels que « laitue » (qui est un hyponyme), « vinaigrette » et « persil » (qui sont thématiquement liés), « soupe » (thématiquement lié ou co-hyponyme de « plat »).

Enfin la catégorie morpho-syntaxique n'est pas fournie ni traitée ainsi porte (substantif) et porte (verbe conjugué) sont considérés comme le même mot : seule la graphie compte. De ce fait, il n'y a qu'un seul vecteur pour représenter ces deux termes dans l'espace sémantique. Les proximités sémantiques de chaque terme pouvant être du bruit pour l'autre.

IV. comparaison de ces deux approches

Dans cette section, nous comparons pour un mot cible donné trois types de résultats :

- les associations produites par LSA sur un « gros » corpus (ce corpus contient 101123 graphies différentes et 189726 paragraphes).
- les associations produites par LSA sur une sous-partie du réseau de JDM (10 000 mots-cibles les plus fréquents et leurs relations)
- les associations issues du réseau JDM.

Nous allons ici nous appuyer sur des exemples précis : tout d'abord le mot « examen ».

		LSA			JDM	
	gros corpus		Sous-partie JDM			
rang	mot	proximité	mot	proximité	mot	activation
1	examen	1.000	examen	1.000	concours	150
2	recel	0.928	bepc	0.988	baccalauréat	140 + 90
3	abus	0.893	examen final	0.984	test	120
4	écroué	0.886	examen blanc	0.984	passer un examen	110
5	escroquerie	0.856	réussir un examen	0.984	note	100
6	factures	0.841	examen médical	0.984	examiner	90
7	blaes	0.841	passer un examen	0.984	épreuve	90
8	courroye	0.829	évaluation	0.876	contrôle	80
9	méry	0.819	études supérieures	0.868	diplôme	80
10	complicité	0.813	fin d'étude	0.868	médical	80
11	supplétif	0.809	études de droit	0.868	médecine	80
12	filippini	0.805	diplôme de fin	0.868	université	80
13	instruction	0.801	longues études	0.868	examen médical	80
14	varces	0.800	études longues	0.868	bac	70
15	incarcéré	0.787	faire des études	0.866	école	70
16	fossorier-longuet	0.784	études	0.865	études	70
17	judge	0.783	brevet	0.865	évaluation	70
18	mis	0.783	diplôme	0.827	santé	70
19	névache	0.774	certificat	0.826	éducation	70
20	escroqueries	0.770	master	0.808	brevet	70

Tout d'abord, nous pouvons constater l'influence du corpus. En effet, le « gros corpus » contient entre autres, une année du « monde », de ce fait le terme examen renvoie au mot composé « mise en examen » et non à examen au sens concours ou évaluation. Les mots les plus proches renvoient à plusieurs noms propres (Blaes, Courroye, Méri, Filippini, Varces, Fossorier-Longuet, Névache) qui ont marqué l'actualité cette année-là. Dans le cadre de la constitution d'une ontologie, ces termes sont moyennement pertinents. Il s'agit ainsi d'un exemple typique des biais liés à un corpus d'actualité. Avec un corpus différent, contenant par exemple, les JO et BO relatifs au ministère de l'enseignement les mots les plus proches sémantiquement de « examen » auraient sans doute été tout autre. Définir un corpus idéal (représentatif des connaissances d'un humain) reste utopique et même un corpus équilibré (textes de niveaux de vocabulaire différents, traitant de sujets différents, etc.) reste particulièrement délicat. Il faudrait être capable d'évaluer ce à quoi les gens sont exposés (discours et textes) depuis leur naissance et d'établir un profil d'un individu moyen (sachant que ce dernier n'existe sans doute pas).

Dans le cas de JDM, le corpus est d'une certaine manière constitué par l'ensemble des joueurs. De ce fait, il représente aussi un biais car les joueurs ne sont certainement pas représentatifs de la population. Tout d'abord puisqu'il s'agit d'un jeu, le public est majoritairement constitué de jeunes adultes. De plus, il s'agit d'un projet de recherche présenté lors de conférences et touchant un public essentiellement universitaire. Ce qui explique les résultats présentés avec le mot « examen ».

Prenons maintenant le terme « sida ».

LSA				
rang	Gros corpus		Sous-partie JDM	
	mot	proximité	mot	proximité
1	sida	1.000	sida	1.000
2	virus	0.948	mst	0.952
3	infection	0.900	vih	0.925
4	contamination	0.888	hiv	0.915
5	immunodéficience	0.880	maladie	0.861
6	sexuellement	0.877	contagion	0.854
7	transmissibles	0.872	maladie infantile	0.850
8	dépistage	0.856	maladie grave	0.850
9	infections	0.855	infantile	0.850
10	vih	0.849	varicelle	0.850
11	infectées	0.844	maladie mortelle	0.849
12	infectieuses	0.841	maladie bénigne	0.849
13	vaccin	0.837	hépatite	0.849
14	anti-vih	0.836	choléra	0.848
15	contaminée	0.831	quarantaine	0.847
16	immuno	0.828	infection	0.844
17	contaminé	0.825	rougeole	0.844
18	opportunistes	0.822	incurable	0.842
19	séropositif	0.813	peste	0.842
20	contaminés	0.803	bénigne	0.841

D'une façon générale, le vocabulaire obtenu par LSA semble plus riche que celui acquis via JeuxDeMots. Par exemple, un terme comme « immunodéficience » apparaît en 5^e position et n'existe pas dans le réseau lexical de JDM.

En effet, via le jeu, les connaissances acquises sont des associations plus immédiates, ceci est lié au fait que les joueurs ne sont pas experts du domaine et que le temps est limité. Certaines associations, évidentes, apparaissent très rapidement dans le réseau lexical du jeu. C'est le cas par exemple de l'association sida-maladie qui apparaît en 5^e position dans le sous-corpus. Mais cette association est beaucoup plus faible dans le gros corpus (proximité de 0,711) car elle est trop évidente pour être donnée explicitement dans les textes. Ce phénomène semble d'autant plus marqué que le terme cible est spécifique.

Toutefois, dans LSA la segmentation des textes est faite à partir des caractères de séparation (blanc, virgules, etc) ce qui interdit l'apparition de termes composés. Ce n'est évidemment pas le cas dans JeuxDeMots où les joueurs ont toute liberté dans le choix des termes qu'ils suggèrent. L'ajout d'un prétraitement dans LSA qui consisterait à repérer les termes composés pose plusieurs difficultés, la principale est qu'il faut alors disposer d'une telle liste de termes, or c'est justement ce que nous cherchons à faire ici. Les repérer automatiquement, par des moyens statistiques, avec suffisamment de précision reste difficile.

Enfin, un dernier point concernant les biais du réseau issu de JDM, réside dans le fait que la connaissance acquise par le système est celle des joueurs. Il est donc possible d'obtenir des connaissances erronées, par exemple l'association de synonymie courtisane-paysanne

V. Conclusion et perspectives : PtiClic ... la combinaison de ces deux approches

Nous avons montré dans cet article que LSA tout comme JDM présentent des biais et ne proposent qu'une couverture partielle du vocabulaire. De plus, ces méthodes ont chacune des parts distinctes de bruits (association qui devrait être plus faible) et de silence (association qui devrait être plus forte).

Une idée intéressante serait d'arriver à combiner les deux méthodes afin que chacune compense les défauts de l'autre. LSA fournirait des suggestions qui seraient validées par des utilisateurs. Ces derniers n'étant pas des experts, cette validation peut se faire via le principe d'accord par paires d'utilisateurs comme dans JeuxDeMots.

Ceci pourrait se faire sous la forme d'un site contributif où des utilisateurs valideraient/invalideraient les propositions d'associations fournies par LSA de façon bénévole. Mais ce type de méthode n'a pas un succès durable ce qui pourrait nuire à notre réseau. De façon naturelle, nous nous dirigeons donc vers un jeu dérivé de JeuxDeMots permettant de valider indirectement des termes proposés par LSA.

Nous esquissons ici, un tel jeu, en cours de développement : PtiClic. Nous présentons d'abord comment se déroule une partie, et motivons les choix de conception adoptés dans un second temps.













Un premier joueur se voit proposer un terme cible et un nuage de mots produit par LSA à partir de ce terme cible. La tâche du joueur est ici de sélectionner un sous-ensemble de ces termes (de 1 à 10) en cliquant dessus, mais également à l'aide d'un jeu de menus de définir la consigne qui sera proposée au second joueur. La consigne est simplement la sélection du type de relation en jeu ainsi que le nombre maximum de termes devant être choisis. Il est par exemple possible de définir les consignes suivantes :

- sélectionner jusqu'à 5 synonymes de <cible> ;
- sélectionner jusqu'à 4 contraires (anonymes) de <cible> ;
- sélectionner jusqu'à 6 termes qui ne sont pas des parties de <cible>
- sélectionner jusqu'à 4 termes qui sont des types de <cible> ;

ondée	vent	impluvium	orage
eau de pluie	Brest	gris	pébroque
précipitations	tempéré	précipiter	cumulo-stratus
éclaircie	prévisions météorologiques	champignon	saucé
nuages	gouttes	giboulées	saison

Figure 3 : mots proposés par LSA pour le terme « pluie » avec les sélections du joueur 1

Lors de la partie, le second joueur doit suivre la consigne en cliquant sur les termes du nuage de mots qui lui semblent pertinents. Le résultat est ensuite affiché et les points gagnés sont calculés par comparaison entre les ensembles de termes fournis par les deux joueurs. Chaque terme trouvé par le second joueur donne 2 points et chaque terme inapproprié enlève 1 point. Les points sont gagnés de façon symétrique par les deux joueurs. Si les réponses de deux joueurs sont identiques, un bonus de 3 points est attribué.

délit d'initié	voler	annulé	plume  
criminel 	larcin	cambrioleur 	escroc 
mouche 	Edvard Munch	aile	délit
oiseau 	frégate (oiseau) 	crapule 	fric-frac
butin	mésange 	papillon 	tire-laine 




 les réponses du joueur 2
 choix du joueur 2 mais pas du joueur 1
 choix du joueur 1 mais pas du joueur 2

Figure 4 : résultat suite à la partie du joueur 2 pour le terme « vole ».

L'intérêt du premier joueur évidemment est de trouver la combinaison de consigne et d'ensemble de termes ayant la plus forte probabilité de maximiser ses gains. Par contre, il n'a pas de contrôle sur le contenu du nuage de mots et sur le mot cible, et doit parfois faire au mieux.

Pour le second joueur, les termes du nuage de mots sont présentés dans un ordre aléatoire. L'affichage du nuage de mots est donc différent de celui que s'était vu proposer le premier joueur. N'importe quel ordre pouvant présenter un biais, une présentation aléatoire a tendance à statistiquement réduire l'impact de la présentation sur la sélection.

PtiClic est composante de JDM agissant sur le même réseau lexical. Contrairement à ce dernier, c'est un jeu en monde clos pour les utilisateurs (le joueur sélectionne parmi des propositions mais ne peut en faire). Ce choix de conception permet d'obtenir des associations sur des termes relevant du vocabulaire passif sélectionnés par LSA, termes qui n'auraient pas spontanément été proposés par les joueurs. L'ajout de PtiClic dans JeuxDeMots permet de réduire le bruit (des termes mal orthographiés ou des confusions de sens) ainsi que le silence (de nouveaux termes sont introduits grâce à LSA). PtiClic permet ainsi de consolider les relations de la base et de densifier le réseau lexical.

VI. Références :

[von Ahn & Dabbish 04] von Ahn L., Dabbish L. (2004) *Labelling Images with a Computer Game*. In ACM Conference on Human Factors in Computing Systems (CHI), pp. 319-326.

[van Bruggen et al., 04]. van Bruggen, J., Sloepp, P., van Rosmalen, P., Brouns, F., Vogten, H., Koper, R., Tattersall, C. (2004) *Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning*, British Journal of Educational Technology 35(6), 729-738.

[Denhière et al., 04] Denhière G., Lemaire, B. (2004) *A Computational Model of a Child Semantic Memory*, Proc. 26th Annual Meeting of the Cognitive Science Society (CogSci'2004), 297-302.

[Dessus et al. 00] Dessus, P., Lemaire, B., Vernier, A. (2000). *Free-text assessment in a Virtual Campus* . In K. Zreik (Ed.), Proc. Third International Conference on Human System Learning (CAPS'3). Paris : Eurovia, 61-76.

[Diki 03]. Dikli, S. (2006) *An Overview of Automated Scoring of Essays*. The Journal of Technology, Learning and Assessment 5(1).

[Dumais 94] Dumais S.T. (1994) *Latent Semantic Indexing (LSI) and TREC-2*. In D. Harman (Ed.), The Second Text RE-trieval Conference (TREC2), National Institute of Standards and Technology Special Publication vol 500, n°215, p.105-116.

[Dumais 97] Dumais S. T. (1997). *Using Latent Semantic Indexing for information retrieval, information filtering and other things*, Cognitive Technology Conference.

[Howard & Kahanna, 02] Howard, M.W. & Kahanna, M.J. (2002) *When does Semantic Similarity Help Episodic Retrieval?*, Journal of Memory and Language 46, 85-98.

[Howard & Kahanna, 07] Howard, M.W. & Kahanna, M.J. (2007) *Semantic Structure and Episodic Memory*, In T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (Eds). Handbook of Latent Semantic Analysis, Mahwah: Lawrence Erlbaum Associates.

[Kintsch 00]. Kintsch, W. (2000). *Metaphor comprehension: A computational Theory* . Psychonomic Bulletin & Review, 7-2, 257-266.

[Kintsch 01]. Kintsch, W. (2001). *Predication* . Cognitive Science, 25-2, 173-202.

[Kintsch 02]. Kintsch, W. (2002). *On the notions of theme and topic in psychological process models of text comprehension*, In M. Louwerse & W. van Peer (Eds) *Thematics: Interdisciplinary Studies*, Amsterdam: Benjamins, 151-170.

[Kintsch et al. 99]. Kintsch, W., Patel, V.L., & Ericsson, K. A. (1999) *The role of long-term working memory in text comprehension*. Psychologia 42, 186-198.

[Kipfer 01]. Kipfer B.A. (2001). *Roget's International Thesaurus*, sixth edition, Harper Resource (First Edition : 1852)

[Landauer & Dumais 97] Landauer, T. K., Dumais, S. T. (1997). *A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge* . Psychological Review, 104, 211-240.

[lapata 05] Lapata M., Keller F. (2005) *Web-based Models for Natural Language Processing*. In ACM Transactions on Speech and Language Processing, vol.2, n°1, pp. 1-30.

[Lemaire 98] Lemaire, B. (1998). *Models of High-dimensional Semantic Spaces*. Proc. 4th Int. Workshop on Multistrategy Learning (MSL 98). Desenzano, Italie.

[Lemaire 99] 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.

[Lemaire & Bianco 03] Lemaire, B. & Bianco, M. (2003). *Contextual Effects on Metaphor Comprehension: Experiment and Simulation*. Proc. of the 5th International Conference on Cognitive Modeling (ICCM'2003), Bamberg, Germany.

[Lemaire et al. 05] Lemaire, B., Mandin, S., Dessus, P. & Denhière, G. (2005). *Computational cognitive models of summarization assessment skills*, in Proceedings of the 27th Annual Meeting of the Cognitive Science Society (CogSci' 2005). Stresa, Italy, July 21-23, 1266-1271.

[Lieberman et al. 07]. Lieberman H., Smith D.A. and Teeters A. (2007) *Common Consensus: a web-based game for collecting commonsense goals*, International Conference on Intelligent User Interfaces (IUI'07), Hawaii, USA.

[Mel'čuk et al. 95]. Mel'čuk I.A., Clas A., Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire*, Editions Duculot AUPELF-UREF.

[Miller et al. 90]. Miller G.A., Beckwith R., Fellbaum C., Gross D. and Miller K.J. (1990) *Introduction to WordNet: an on-line lexical database*. In: International Journal of Lexicography 3 (4), pp. 235-244.

[Miller 03] Miller T. (2003). *Latent semantic analysis and the construction of coherent extracts*. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, and Nikolai Nikolov, editors, Proc. of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing), pages 270–277, September 2003.

[Polguère 06]. Polguère A. (2006) *Structural properties of Lexical Systems: Monolingual and Multilingual Perspectives*. Proceedings of the Workshop on Multilingual Language Resources and Interoperability (COLING/ACL 2006), Sydney, pp. 50-59.

[Redington & Charter 98] Redington M., Chater, N. (1998). *Connectionist and statistical approaches to language acquisition: A distributional perspective*. Language and Cognitive Processes, 13-2/3, 29-91.

[Robertson & Spark 76]. Robertson S. et Spark Jones K. (1976) *Relevance weighting of search terms*, Journal of the American Society for Information Science, n° 27, pp. 129-146.

[Salton 68] Salton G. (1968) *Automatic Information Organization and Retrieval*, Mac Graw Hill, NY.

[Véronis 01]. Véronis J. (2001) *Sense tagging: does it make sense?* Corpus linguistics' 2001 Conference, Lancaster, U.K.

[Zampa & Lemaire 02] Zampa V. Lemaire B., (2002). *Latent Semantic Analysis for user modelling*, Journal of intelligent information systems. Vol 18 n°1. p.15-30.

[Zampa & Raby 01] Zampa V., Raby F., *Entre modèle d'acquisition et outil pour l'apprentissage de la langue de spécialité: Le prototype R.A.F.A.L.E.S (Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère de Spécialité)*, Asp (Anglais de Spécialité) n°31-33, p 163-179.