Modèles cognitifs issus de l'analyse de la sémantique latente

Benoît Lemaire, Philippe Dessus

MODELES COGNITIFS ISSUS DE L'ANALYSE DE LA SEMANTIQUE LATENTE

Benoît LEMAIRE, Philippe DESSUS

Laboratoire des sciences de l'éducation, Bât SHM,
Université Pierre-Mendès-France, BP 47, 38040 Grenoble Cedex 9
Mél : {Prenom.Nom@upmf-grenoble.fr}
Toile : http:www.upmf-grenoble.fr/sciedu/

Résumé

L'objet de cet article est de présenter l'analyse de la sémantique latente (Latent Semantic Analysis*), un modèle cognitif de la représentation et de l'acquisition du sens des mots, ainsi que de la compréhension de textes. Nous présentons quatre types de modélisations cognitives fondées sur LSA, ainsi que les expérimentations qui les valident : — représentation de connaissances, — acquisition du vocabulaire, — compréhension de textes, — évaluation automatique de textes. Nous terminons en suggérant une extension de ce modèle à des connaissances non issues de textes.*

***Mots clés** : Modélisation cognitive, simulation informatique, compréhension de textes, acquisition du vocabulaire, évaluation automatique de textes.*

Abstract

This paper aims at presenting Latent Semantic Analysis (LSA), a cognitive theory of representation and acquisition of word meaning and text comprehension. First we describe the model, then we present several experiments for validating the model along four categories: knowledge representation, vocabulary acquisition, text comprehension, text assessment. At the end, we suggest an extension of that model to deal with non-textual knowledge.

***Keywords**: Cognitive modelling, computer simulation, text comprehension, vocabulary acquisition, automated text assessment.*

1. Introduction

LSA (pour *Latent Semantic Analysis* ou analyse de la sémantique latente) est un programme informatique qui simule l'acquisition de connaissances à partir de l'analyse entièrement automatique de grands corpus de textes. Ces connaissances sont représentées sous la forme de vecteurs dans un espace de très grande dimension. Les connaissances, ainsi que diverses performances simulées à partir de ce mécanisme de représentation, sont comparables à celles de sujets humains lors de tests standardisés. Ainsi, LSA est à la fois vu comme un modèle d'acquisition et de représentation des connaissances (Landauer & Dumais, 1997) et comme une machine qui apprend, ce qui a l'avantage de rendre simulables et généraux les modèles cognitifs qui en sont issus. Trop souvent, en effet, les validations de modèles cognitifs s'appuient sur des représentations sémantiques *ad hoc*, couvrant des domaines très limités.

Après une présentation générale de LSA, nous passerons en revue les différentes modélisations cognitives qu'il sous-tend : — la représentation de connaissances ; — les mises à jour de la mémoire selon l'exposition à des textes (acquisition de connaissances) ; — les mécanismes qui utilisent cette représentation de connaissances (compréhension et évaluation de textes) ; — l'extension à des connaissances non issues de textes.

2. Présentation générale de LSA

2.1. Modèle mathématique

Le fonctionnement de LSA est issu d'un principe de linguistique distributionnelle (Harris, 1975) : le sens d'un mot peut être défini statistiquement, à partir de l'ensemble des contextes (*i.e.*, paragraphes, phrases, textes) dans lesquels ce mot apparaît. Par exemple, le mot *avion* va apparaître souvent conjointement à des mots comme *décoller*, *aile*, *aéroport*, et rarement conjointement à des mots comme *sous-bois* ou *cerises*. Cependant, cette information statistique sur le contexte d'un mot *M* n'est pas suffisante pour en définir le sens, puisqu'elle ne dit rien quant aux liens sémantiques avec tous les autres mots n'apparaissant jamais conjointement à *M*. Par exemple, le contexte statistique de *lampadaire*

(*éclairé*, *allumer*, *lumière*, etc.) fournit une information insuffisante sur le sens de ce mot. En effet, si *lampadaire* n'apparaît jamais conjointement à *abat-jour*, nous n'aurons aucune information sur le lien sémantique entre ces mots (ce que Grefenstette, 1994, nomme les affinités de second ordre). Il faut pour cela un mécanisme permettant de croiser les informations de co-occurrence propres à chaque mot. Or, *abat-jour* doit être considéré proche de *lampadaire* parce qu'il est co-occurrent de mots comme *éclairé*, *allumer*, *lumière*, etc. qui eux-mêmes sont co-occurrents de *lampadaire*. Ce sont ces enchaînements de liens de co-occurrence, à plusieurs niveaux, qui permettent une représentation correcte du sens des mots. En d'autres termes, LSA repose sur la définition suivante : deux mots sont similaires s'ils apparaissent dans des contextes similaires. Deux contextes sont similaires s'ils comportent des mots similaires. Cette récursivité croisée exige un mécanisme particulier, bien plus complexe qu'un simple comptage d'occurrences, que nous allons décrire.

Il est nécessaire de réduire la masse énorme d'informations constituée de tous les contextes rencontrés. En effet, l'apprentissage n'est pas qu'un processus d'accumulation, il doit nécessairement inclure une phase de généralisation afin de modéliser d'une manière économique un grand nombre de stimuli. LSA construit donc une matrice de co-occurrences, constituée du nombre d'apparitions de chaque mot dans chaque contexte, sans tenir compte de leur ordre. L'unité de contexte utilisée est le paragraphe, qui possède l'avantage de n'être ni trop restreint comme la phrase, ni trop important comme la page, tout en étant facilement détectable par un programme informatique. Cette matrice volumineuse est ensuite réduite, de façon à faire apparaître les liens exprimés par la définition récursive précédente. La procédure mathématique utilisée est la décomposition aux valeurs singulières (Deerwester *et al*., 1990), une généralisation de l'analyse factorielle, qui permet de représenter chaque mot et chaque paragraphe par un vecteur de très grande dimension, de l'ordre de plusieurs centaines. Le nombre de dimensions optimal (autour de 300) a été estimé empiriquement pour la langue anglaise (Landauer & Dumais, 1997) sans que l'on puisse le justifier théoriquement. Des valeurs plus grandes conduisent à prendre en compte trop de bruit et des valeurs plus petites aboutissent à une trop grande perte d'informations. Contrairement aux analyses factorielles classiques, les axes ne sont pas étiquetés. Ce ne sont donc pas les vecteurs eux-mêmes qui font l'objet d'analyses, mais les relations qu'ils entretiennent entre eux.

LSA utilise deux types de mesures. La première permet d'estimer la similarité entre deux mots ou deux groupes de mots, à partir du cosinus entre les angles des vecteurs correspondants. C'est donc une mesure entre –1 (similarité minimale) et 1 (similarité maximale). La seconde mesure caractérise la connaissance que LSA a sur un mot ou sur un groupe de mots, à partir de la longueur du vecteur associé. Cette mesure, beaucoup moins utilisée dans la littérature, dépend de la fréquence des mots et de la diversité des contextes dans lesquels ils apparaissent.

2.2. Modèle cognitif

On peut maintenant chercher à passer du modèle mathématique à un modèle cognitif. Pour cela, nous nous appuierons sur une autre manière de concevoir LSA, due à Landauer (2002). Celle-ci part du postulat que la signification d'un groupe de mots (un paragraphe par exemple) est fonction de la signification des mots qui le composent ainsi que du contexte dans lequel il se trouve.

$$S_{\text{paragraphe}} = f(S_{\text{mot}_1}, S_{\text{mot}_2}, \ldots, S_{\text{mot}_n}, \text{Contexte})$$

Ce postulat est simplifié dans LSA de la façon suivante : le contexte est omis, et la fonction est additive. On aboutit donc à la formule suivante :

$$S_{\text{paragraphe}} = S_{\text{mot}_1} + S_{\text{mot}_2} + \ldots + S_{\text{mot}_n}$$

Notre apprentissage de la signification des mots résulterait de l'exposition à des millions de telles équations, notamment depuis que nous savons lire. Ce processus cognitif d'acquisition peut donc être modélisé par la résolution d'un système d'équations. La décomposition aux valeurs singulières est une méthode de résolution d'un tel système, que les mathématiciens nomment « mal conditionné ». Ainsi, LSA peut donc être vu comme un modèle cognitif. Plusieurs types de modélisations sont envisageables (*voir figure 1*), qui correspondent aux cinq questions suivantes :

1. LSA est-il un bon modèle de *la représentation* du sens des mots ? En d'autres termes, peut-on représenter de manière acceptable les significations lexicales uniquement à partir de l'analyse d'un grand nombre de textes, et notamment sans s'appuyer sur la syntaxe ?
2. LSA est-il un bon modèle de *l'acquisition* humaine du sens des mots ? Les taux d'acquisition du vocabulaire par LSA en fonction du nombre de mots du corpus correspondent-ils à ceux des sujets humains qui apprennent en lisant ?
3. LSA est-il un bon modèle de *la compréhension de textes* et, en particulier, du traitement des métaphores ou des inférences ?

4. LSA est-il un bon modèle de *l'évaluation des connaissances* ? Peut-on mesurer les connaissances de sujets à partir des textes qu'ils produisent ?
5. LSA est-il un bon modèle *de l'acquisition de connaissances à partir de textes* ? Le modèle est-il généralisable à d'autres connaissances que celles du vocabulaire ?

Nous allons passer en revue ces différentes modélisations, et nous les résumons dans la figure 1 et le tableau 1 ci-dessous. La figure 1 représente graphiquement les différents processus cognitifs modélisés par LSA (en italiques) listés ci-dessus et les différents types d'inputs nécessaires (en encadré). Le tableau 1 explicite ces informations. L'item 1 (représentation) n'a pas été repris dans le tableau, car il correspond à la description du fonctionnement général de LSA.

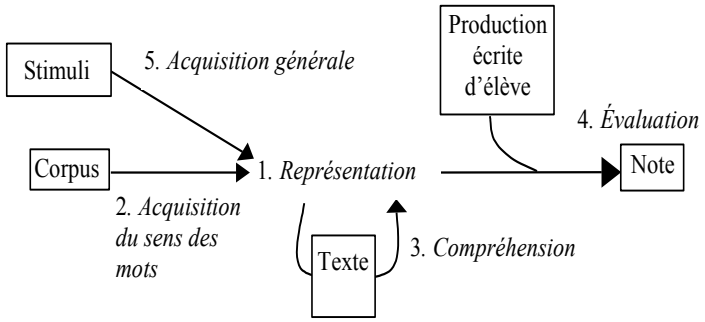

Figure 1 *— Représentation graphique des différentes modélisations utilisées avec LSA*

Tableau 1 *— Aperçu des différentes modélisations utilisant LSA et caractéristiques des inputs et outputs.*

Objet	Input	Output
2. Acquisition du sens des mots	Un corpus de textes, représentant les inputs auxquels un enfant est soumis	Vecteurs associés à chaque mot dans un espace sémantique
3. Compréhension de textes	Espace sémantique et un nouveau texte (phrase, productions d'élèves, métaphore)	Le nouveau texte, représenté par un vecteur dans l'espace sémantique
4. Évaluation des connaissances	Espace sémantique et productions écrites d'élèves	Note exprimant le degré de proximité entre la production d'élève et les textes didactiques
5. Acquisition de connaissances non issues de textes	Stimuli représentés par des séquences d'unités lexicales	Vecteurs associés à chaque unité lexicale dans un espace sémantique

3. Modèle de représentation du sens des mots

Un champ important de la recherche sur le traitement automatique de la langue naturelle utilise des méthodes statistiques. Pour la plupart, ces méthodes sont issues de la recherche documentaire (Salton & McGill, 1983) et utilisent un modèle de codage vectoriel : des objets comme des mots ou des documents sont représentés dans un espace multidimensionnel et certaines opérations (addition vectorielle, multiplication par un scalaire) permettent de rendre compte de la proximité entre objets (Memmi, 2000). LSA traite, en input, un grand corpus textuel, et permet d'établir des proximités sémantiques intermots en output.

L'expérience la plus souvent citée dans ce domaine est due à Landauer et Dumais (1997). Elle a consisté à vérifier si les proximités sémantiques entre mots obtenues par LSA à partir de l'analyse d'un volumineux corpus de textes étaient comparables à celles estimées par les humains. Le corpus utilisé provenait de la version numérique de l'encyclopédie *Grolier's Academic American Encyclopedia*. Il totalisait 4,6 millions de mots : 30 473 paragraphes et 60 768 mots différents. Chacun de ces mots a été représenté par un vecteur à 300 dimensions. Le test utilisé est celui du TOEFL (*Test of English as a Foreign Language*) qui possède, entre autres, 80 questions de synonymie. Chaque question est composée d'un mot et de quatre mots proches. La tâche du sujet est de trouver, parmi les quatre, le véritable synonyme. Landauer et Dumais ont simulé cette tâche en choisissant le mot dont le vecteur était le plus proche du mot initial. Le

modèle a obtenu le score de 51,5, ce qui est très voisin du score moyen (51,6) obtenu par les étudiants étrangers non-anglophones qui souhaitent s'inscrire dans les universités américaines. Ce score n'a pas été comparé avec celui de natifs anglais, mais il demeure néanmoins un résultat intéressant. En effet, pour la première fois à notre connaissance, un modèle est capable de réussir un test standardisé général à partir de l'analyse automatique d'un grand nombre de textes (depuis, Turney, 2001, a obtenu des résultats meilleurs avec une méthode différente).

Comme nous l'avons signalé précédemment, LSA peut également associer des vecteurs à des groupes de mots, en additionnant les vecteurs des mots qui les constituent. On passe alors d'une représentation sémantique des mots à une représentation sémantique des textes. La question se pose donc de savoir si la représentation qui était valide pour les mots l'est encore pour les textes et notamment si cette simple sommation de vecteurs est pertinente. Rehder *et al*. (1998) ainsi que Zampa (1999) ont montré que la représentation sémantique produite par LSA permettait de discriminer des textes selon leur difficulté avec les mêmes résultats que les humains. Cependant, nous verrons dans la partie 5 que ce passage des représentations des mots à celles des textes par simple sommation ne permet pas de modéliser des processus cognitifs plus complexes.

Notons tout de suite que la syntaxe est absente du traitement des textes par LSA. Chaque paragraphe est traité comme un ensemble de mots et non comme une séquence ordonnée. Ni l'ordre des mots ni, *a fortiori*, d'autres éléments d'ordre syntaxique ne sont donc pris en compte. Nous discuterons de ce point plus en détail à la fin de cet article. Un modèle concurrent de LSA, HAL (*Hyperspace Analogue to Language*), tient compte de l'ordre des mots en codant dans la matrice la distance entre les mots dans le paragraphe (Burgess, Livesay, & Lund, 1998 ; Burgess & Lund, 1997 ; Lund & Burgess, 1996). Cependant, HAL ne procède pas à une réduction de dimensions. Une autre tentative pour modéliser des éléments de la syntaxe est dû à Kintsch (2001), qui tient compte de la nature des mots (prédicat ou argument). Cette modélisation est décrite dans la partie 5 de cet article.

Plus encore, LSA n'effectue aucun traitement morphologique : chaque séquence de lettres constitue un mot différent. Chaque forme fléchie correspond donc à un mot : *cheval* et *chevaux* sont ainsi deux mots distincts. Cependant, les formes différentes d'un même mot vont être associées à des vecteurs proches dès lors que ces formes apparaissent dans des contextes similaires. Ainsi, le cosinus entre *cheval* et *chevaux* va être relativement élevé. Ces deux termes sont ainsi associés, non pas en raison de leur lemme commun, mais en raison de leur apparition dans des contextes proches. La sémantique supplée ici l'absence de la morphologie.

Cette partie concernait une utilisation statique des représentations sémantiques de mots. La suivante s'intéresse toujours aux mêmes représentations, mais comme outputs du processus d'acquisition du vocabulaire, dans une perspective développementale.

4. Modèle de l'acquisition du vocabulaire

Nous nous intéressons ici à la manière dont LSA peut rendre compte de l'acquisition du vocabulaire. Les modèles computationnels de l'acquisition du vocabulaire partent des postulats suivants (Rapaport & Ehrlich, 2001, p. 5) : « le sens d'un mot *peut* être déterminé quel que soit le contexte, peut être *révisé* et raffiné lors de nouvelles occurrences, et *converge* vers une définition semblable à celle du dictionnaire, si le contexte et les occurrences sont suffisants » (ce sont les auteurs qui soulignent). Selon les mêmes auteurs, il existe quatre techniques permettant, dans un système automatique, d'apprendre le sens de mots nouveaux : 1°) le chercher dans un dictionnaire, 2°) le demander à un autre système, 3°) utiliser une taxonomie ou des schémas pour trouver un synonyme de ce mot, 4°) utiliser le contexte dans lequel ce mot est utilisé, sans aucune aide extérieure. C'est cette dernière alternative qu'utilise LSA.

Nous nous intéressons dans cette partie à la construction même de cette représentation et sa capacité à décrire les processus d'acquisition des significations lexicales à partir de la lecture. Selon Landauer et Dumais (1997), la majeure partie du lexique est acquise par la lecture pour deux raisons : le lexique oral n'est qu'une faible partie du lexique total et peu de mots sont acquis par instruction directe (de 6 % à 10 % pour Nagy *et al*., 1987). La plupart des mots que nous connaissons ont donc été acquis par la lecture, or LSA apprend également à partir de textes. On peut donc se poser la question de l'adéquation entre le modèle et les humains.

Pour illustrer le mécanisme d'apprentissage tel que le modélise LSA, supposons qu'un sujet ne connaisse pas le sens du mot *M* et qu'il ne l'ait jamais rencontré. A la lecture d'un paragraphe contenant ce mot, il va probablement en déduire un sens grossier du mot, c'est-à-dire qu'il va le situer (positionner le vecteur dans le modèle LSA) par rapport aux autres mots qu'il connaît. Au fur et à mesure de ses lectures (contenant ou ne contenant d'ailleurs pas ce mot, comme on le verra par la suite), le sujet va progressivement affiner le sens du mot *M*, et donc le situer par rapport aux autres mots de plus en plus précisément, en convergeant vers le sens que le sujet attribue au mot — idéalement celui du dictionnaire.

Plusieurs expérimentations ont montré que le taux moyen d'acquisition chez les enfants est d'environ 10 mots nouveaux (formes fléchies) par jour. On peut s'en convaincre rapidement en divisant grossièrement le nombre de mots connus au

sortir de l'adolescence (de 40 000 à 100 000, selon Kail & Fayol, 2000) par une durée de 20 ans (environ 7 000 jours). C'est une valeur moyenne qui cache bien évidemment des disparités en fonction de l'âge. Pour comparer ce taux avec celui de LSA, Landauer et Dumais (1997) ont estimé que les enfants sont exposés en moyenne à 3 500 mots écrits par jour. LSA atteint rapidement le niveau obtenu à l'issue du test du TOEFL lorsqu'on le soumet à une telle fréquence d'exposition à partir d'une encyclopédie. Notons cependant que ce test final concernait des sujets non anglophones et qu'il est difficile d'extrapoler cette mesure à l'apprentissage de la langue maternelle. Soulignons également qu'il faut moins de quatre années simulées pour que LSA atteigne des performances atteintes par les sujets en une vingtaine d'années, avec, de plus, d'autres inputs. Les ordres de grandeur permettent cependant de considérer avec beaucoup d'intérêt ce modèle.

Le modèle apporte également une réponse au fameux paradoxe de la « pauvreté du stimulus » (*poverty of stimulus*). Platon déjà s'était interrogé sur le fait que notre observation d'un nombre limité d'événements nous conduit cependant à des comportements valides dans une infinité de situations. Dans le cas de l'acquisition du langage, Chomsky (1985) remarque que le rapide apprentissage de la grammaire par un enfant ne peut être expliqué uniquement par son exposition au langage des adultes. Pour Chomsky, la solution à ce paradoxe nécessite de considérer le caractère inné du langage. Sans rentrer dans le débat (voir Pullum & Scholz, 2002), voyons maintenant comment les données de l'expérience précédente répondent à ce paradoxe. Le taux de 10 mots nouveaux appris par jour suite à l'exposition à 3 500 mots est en effet fort élevé. Landauer et Dumais (1997) estiment qu'un enfant rencontre un mot non connu par paragraphe de 70 mots. C'est-à-dire que, sur une journée, un enfant aura rencontré 50 mots nouveaux et qu'il en aura appris 10. Cette estimation est cohérente avec le taux (mots appris)/(mots rencontrés) de 15 % donné par Swanborn et de Glopper (1999). Des expériences de laboratoire ont tenté de simuler ce mécanisme en insérant judicieusement des mots inconnus ou des non-mots dans des paragraphes afin de les faire lire à des enfants (Elley 1989 ; Jenkins *et al*., 1984 ; Nagy *et al*., 1985 ; tous cités par Landauer & Dumais, 1997). Après contrôle de l'apprentissage, on n'aboutit qu'à environ un quart du taux réel : seuls 2,5 mots/jour sont acquis. Cela signifie donc que des mots sont appris en dehors de l'exposition à des paragraphes contenant ces mots. En d'autres termes, les enfants apprennent des mots en lisant des paragraphes qui ne contiennent pas ces mots.

Pour estimer ce phénomène, Landauer et Dumais ont recherché dans l'apprentissage d'un mot par LSA la part provenant des paragraphes contenant ce mot (effet direct) et la part de ceux ne contenant pas ce mot (effet indirect). En reprenant le test du TOEFL décrit précédemment, ils ont cherché un modèle mathématique exprimant z, la valeur normalisée de réussite au score, en fonction de T, nombre total de paragraphes analysés, et de S, nombre de paragraphes analysés contenant le mot donné, pour chaque question. Ce modèle est :

$$z = 0{,}128(\log 0{,}076 \times T)(\log 31{,}910 \times S)$$

Les prédictions de ce modèle corrèlent très fortement avec le z observé (r = 0,98). En supposant qu'un sujet a lu 25 000 paragraphes (c'est-à-dire environ 3,8 millions de mots), Landauer et Dumais ont cherché quel était le gain apporté par la lecture d'un 25 001^{e} paragraphe pour les mots s'y trouvant (effet direct), ainsi que pour les mots ne s'y trouvant pas (effet indirect). Il résulte de cette simulation que le gain moyen par effet direct est de 0,0007 mots acquis par mot rencontré, c'est-à-dire 0,0007 x 3 500 mots lus/jour = 2,5 mots appris/jour, ce qui est cohérent avec les résultats des expériences de laboratoire décrites précédemment. Le gain moyen par effet indirect est de 0,15 mots par paragraphe de 70 mots, c'est-à-dire de 0,15 x 50 = 7,5 mots appris/jour. On retrouve bien le total de 10 mots appris/jour du modèle LSA et des données de la littérature. Un mot serait donc appris pour 25 % en lisant des textes le contenant et pour 75 % en lisant des textes ne contenant pas ce mot.

Il convient cependant d'être prudent avec cette expérience qui a largement recours à l'approximation pour déterminer les valeurs initiales et qui, du fait de la méthode, ne tient pas compte de l'acquisition par analyse morphologique. Il n'en reste pas moins que LSA apporte ainsi une solution au paradoxe de la pauvreté du stimulus sans recourir à une explication innéiste, puisque toutes les significations sont construites uniquement à partir de l'exposition à la langue, sans aucune connaissance préalable sur la langue.

La partie suivante traite de la compréhension de textes (phrases, textes didactiques, métaphores). À cette fin, ils sont comparés sémantiquement à un corpus de référence.

5. Modèle de la compréhension de textes

Comprendre un texte, c'est en construire un modèle mental, c'est-à-dire faire des connexions entre les idées exposées dans le texte et les connaissances pertinentes préalables (Kintsch, 1998). LSA peut justement permettre de représenter ces deux types de connaissances. Nous allons décrire ici trois études qui ont simulé la compréhension de textes. Tout d'abord, celle de Foltz *et al*. (1998), qui ont utilisé LSA pour établir une mesure de la cohérence interphrases de textes, mesure qu'ils ont reliée au niveau de compréhension de ces textes lus par des sujets. Ensuite, nous décrirons le travail

de Rehder et ses collègues (Rehder *et al*., 1998 ; Wolfe *et al*., 1998), qui ont proposé une méthode prédisant le niveau de compréhension d'étudiants lisant un texte. Enfin, nous exposerons les travaux de Kintsch (2001), qui propose une implémentation avec LSA de son modèle Construction-Intégration. Cette implémentation a notamment été testée dans des tâches de compréhension de métaphores.

Le niveau de compréhension d'un texte lu est lié à sa cohérence. Foltz, Kintsch et Landauer (1998) ont utilisé les capacités de LSA à évaluer le degré de relation sémantique entre phrases adjacentes pour mesurer la cohérence textuelle. Ils sont partis de deux études précédentes (Britton & Gülgöz, 1991 ; McNamara *et al.*, 1996) évaluant les performances de compréhension de sujets selon le niveau de cohérence, manipulé, des textes qu'ils lisaient. LSA compare les phrases contiguës de chaque texte deux à deux, la moyenne de ces comparaisons donnant un score global de cohérence. Les résultats montrent que ces scores de cohérence sont similaires à ceux obtenus par d'autres méthodes manuelles. Ces scores moyens de cohérence corrèlent aussi très fortement (r = 0,99) avec la compréhension des lecteurs. Bien évidemment, il existe d'autres moyens de mesurer la cohérence de textes (par exemple le comptage des connecteurs) mais, par la seule prise en compte du sens des mots, LSA a des performances voisines de méthodes plus sophistiquées.

On a aussi utilisé les possibilités de LSA de mesurer la connaissance pour proposer une méthode pouvant prédire l'acquisition de connaissances à partir de la lecture de textes (Rehder *et al*., 1998 ; Wolfe *et al*., 1998). Il s'agit de simuler l'apprentissage d'étudiants en mettant en adéquation deux mesures, obtenues par LSA : celle de la difficulté des textes et celle des connaissances préalables des étudiants. Cette étude se passe en trois temps : une série de cinq textes sur le fonctionnement du cœur, de difficulté croissante, a été traitée par LSA, qui peut les ranger par difficulté croissante. Ensuite, les étudiants ont répondu à un questionnaire et rédigé un texte libre, texte qui a été ajouté à l'espace vectoriel des cinq textes. Cette étape permet d'évaluer les connaissances préalables des sujets. Une corrélation importante a été mesurée entre la note attribuée par LSA pour ces textes libres et celles données par des juges humains ou celles obtenues via des questionnaires de connaissances (r = 0,63 à r = 0,74). Enfin, les sujets ont lu un des cinq textes et ont validé leurs connaissances acquises à la lecture de ce texte par la rédaction d'un nouveau texte libre et la réponse à un nouveau questionnaire. Les auteurs montrent que le gain prétest-posttest, selon le niveau des sujets, suit une courbe de Gauss : les sujets qui profitent le plus de la lecture des textes sont ceux ayant une connaissance suffisante du domaine, tout en n'étant pas trop élevée.

Kintsch (2000) a initié une perspective de travail intéressante, qui permet de prendre en compte des informations syntaxiques pour simuler la compréhension de métaphores. La compréhension d'une métaphore nécessite d'établir un lien entre la topique (l'objet décrit par la métaphore) et le véhicule (les termes qui le décrivent). La compréhension de telles métaphores met donc en jeu, pour LSA, trois catégories de mots : le sens de la topique, du véhicule et de mots de référence, qui indiquent de quelle manière la métaphore a été comprise. Par exemple, la métaphore *mon avocat est un requin* va être comparée avec les vecteurs de mots de référence comme *justice*, *crime*, *requin*, *poisson* et *avidité*. De la même manière, la topique (ici *avocat*) va être aussi être comparée à ces mots de référence. La métaphore sera jugée comprise s'il existe une différence suffisante entre ces comparaisons. Par exemple, le vecteur de la métaphore doit être plus proche du mot de référence *avidité* que la topique. Kintsch a cependant remarqué que le vecteur associé à la métaphore ne peut simplement résulter de l'addition des vecteurs de la topique et du véhicule. Une des raisons est liée à l'absence de traitement syntaxique. En effet, pour LSA, les deux métaphores suivantes sont équivalentes : *mon chirurgien est un boucher* et *mon boucher est un chirurgien*. Or il existe, dans ces expressions métaphoriques, un déséquilibre entre le prédicat et l'argument qu'il faut prendre en compte. Kintsch a recours, pour cette opération de prédication, à son modèle de Construction-Intégration (Kintsch, 1988). La mise en œuvre de ce modèle est complexe et nécessite un logiciel spécifique (Mross & Roberts, 1992). Nous n'exposons ici qu'une version simplifiée, jugée suffisante par son auteur.

Il s'agit de créer un vecteur définissant la métaphore qui rende compte du fait que le véhicule est le prédicat et que la topique est l'argument. Pour cela, on recherche, parmi tous les voisins sémantiques possibles du prédicat, ceux qui sont également voisins de l'argument. On va donc conserver, parmi les n mots les plus proches du prédicat, les cinq qui sont également proches de l'argument. La valeur de n va dépendre des exemples utilisés. Si la valeur de n est trop grande, on va récupérer des mots trop peu voisins, si elle est trop petite, on n'accédera pas à un des sens du mot initial qui aurait pu être celui recherché pour la métaphore. Des valeurs plus grandes (de 500 à 1 500) sont nécessaires puisque, dans une métaphore, le prédicat et l'argument sont parfois sémantiquement très éloignés. Cet ensemble de mots (le prédicat, l'argument et les cinq additionnels) constitue, par simple sommation des vecteurs les définissant, le vecteur définissant la métaphore. Une extension de cette méthode a montré des corrélations intéressantes avec les performances de sujets humains (Lemaire *et al*., 2001).

La partie suivante, en partant du traitement de mêmes inputs, s'intéresse à la modélisation de l'évaluation de connaissances.

6. Modèle de l'évaluation des connaissances

LSA a également été utilisé pour modéliser le jugement d'enseignants évaluant des copies d'étudiants. On part du principe que la note donnée dans un tel cadre est fonction de l'adéquation entre des connaissances traitées dans la copie et des connaissances de référence (par exemple un cours, une encyclopédie). Ce principe est compatible avec les données de la psychologie de l'évaluation scolaire (Noizet & Caverni, 1978), qui décrivent l'évaluation comme une activité de comparaison entre une production scolaire et un modèle de référence. Précisons encore que la tâche demandée aux étudiants est en général plus proche d'un résumé ou d'une note de synthèse que d'une dissertation. LSA peut ainsi contribuer à plusieurs champs de recherche que sont la conception de tuteurs intelligents et celui de l'évaluation automatique informatisée (*Computer-Assisted Assessment*). On peut distinguer deux manières d'évaluer automatiquement les connaissances acquises pendant un cours (Dessus, Lemaire, & Vernier, 2000) :

- *les évaluations informatisées basées sur des traits de surface*, nous faisons référence ici aux travaux initiés par Page (1966, cité par Chung & O'Neil, 1997) et complétés par Burstein *et al*. (1998). La mesure de paramètres quantitatifs de copies (comme la longueur moyenne des mots d'une copie, la longueur totale de l'essai, etc.) s'est avérée liée à des évaluations qualitatives humaines (corrélations autour de 0,8). Le principal inconvénient est que, comme elles sont uniquement fondées sur des traits de surface de la copie, elles ne prennent pas du tout en compte les connaissances incluses ;
- *les évaluations informatisées se basant sur le contenu du cours*, la plupart des évaluations de ce type utilisent un codage multidimensionnel de l'information présente dans le texte produit, codage ensuite comparé à un corpus représentant des connaissances de base (cours, encyclopédie, etc.). Certains de ces logiciels utilisent LSA comme moteur, bien que d'autres méthodes aient été développées (Larkey, 1998 ; McKnight & Walberg, 1998).

Les implémentations de LSA dans le domaine de l'évaluation automatique sont toutes basées sur le même principe général, à partir duquel on peut observer certaines variantes. Ce principe consiste à utiliser la capacité de LSA à représenter le contenu d'un texte. Ainsi, sont successivement représentés dans un espace multidimensionnel un corpus de référence (en général un cours, une encyclopédie) et un essai (copie d'élève). Une comparaison entre l'essai et le corpus de référence est alors réalisée, et une note donnée, note qui est liée à la proximité entre l'essai et le corpus de référence. Parmi les tuteurs basés sur LSA, deux d'entre eux présentent les variantes suivantes.

IEA (*Intelligent Essay Assessor*), conçu par Foltz *et al*. (1999) a la particularité de délivrer deux types de scores à la copie : 1°) *le score « holistique »,* qui compare successivement le texte à noter à une série de copies notées au préalable par un jury. La note de la copie sera celle de la série de copies avec laquelle elle entretient la plus grande proximité. Une évaluation de ce calcul de score a été faite à partir de 190 copies de biologie. Elle montre une corrélation de 0,80 entre les scores des évaluateurs humains et ceux calculés par IEA. 2°) *le score « étalon-or »* (*gold standard*), qui compare le texte à noter avec une copie-modèle idéale, réalisée par exemple par l'enseignant. La comparaison peut être globale ou bien faite paragraphe par paragraphe, de manière à vérifier si l'élève traite correctement chaque notion.

Apex (Dessus & Lemaire, 1999 ; Dessus & Lemaire, 2002 ; Dessus, Lemaire & Vernier, 2000 ; Lemaire & Dessus, 2001) procède différemment, en ne s'appuyant pas sur des copies-types, mais sur le cours de l'enseignant, préalablement découpé en deux niveaux de hiérarchie : les *thèmes principaux* et les *notions* qui se rattachent à chaque thème. Un corpus de textes en français est traité conjointement à cette copie, de manière à ajouter des connaissances de la langue, qui augmente ainsi la précision des comparaisons sémantiques. Trois types d'évaluation sont possibles. *Au niveau du contenu*, en appariant la copie avec chacune des notions du thème choisi, ce qui permet d'évaluer la manière dont le contenu a été couvert. *Au niveau du plan*, en appariant chaque paragraphe de la copie avec chaque notion du thème choisi, ce qui permet à l'étudiant d'appréhender le plan général de sa copie. *Au niveau de la cohérence textuelle*, en mesurant successivement la proximité sémantique de deux phrases contiguës de la copie, ce qui permet de détecter une microcohérence insuffisante, due à des ruptures brutales de cohérence interphrases (*voir partie précédente pour le lien avec la compréhension*). Un test a permis de montrer une corrélation moyenne de 0,59 ($p < 0,01$) entre les notes partielles d'un enseignant et celles données par *Apex*. Plus récemment Gounon et Lemaire (2002) ont amélioré cette valeur (0,62) en partitionnant au préalable le texte de la copie en unités sémantiquement cohérentes.

Jusqu'à présent, les connaissances traitées par LSA étaient issues de textes. Nous allons montrer que ce dernier peut également rendre compte de l'acquisition de connaissances non fondées sur le langage, mais sur d'autres domaines.

7. Modèle de l'acquisition de connaissances non textuelles

7.1. Liens avec les théories associationnistes

LSA rend opératoires les théories associationnistes qui postulent depuis Aristote un fonctionnement cognitif fondé sur l'association d'éléments irréductibles. Aristote a proposé quatre formes d'association : la similarité, la différence, la contiguïté temporelle et la contiguïté spatiale. Par la suite, d'autres auteurs, comme Mill, ont étendu cette liste avec notamment des associations résultant de la répétition. LSA fonctionne à partir d'associations de contiguïté entre unités lexicales. Ces contiguïtés peuvent être d'ordre temporel, comme dans l'apprentissage à partir de la lecture, ou spatial, comme dans l'apprentissage à partir d'éléments visuels. LSA calcule alors de nouvelles associations de similarité. Par la suite, ce sont ces deux types d'association qui vont être en jeu : des associations de *similarité* pour modéliser ce que le sujet connaît et des associations de *contiguïté* qui décrivent ce à quoi le sujet est nouvellement exposé. Ces dernières vont alors affecter les premières, et déterminer ainsi de nouvelles valeurs de similarité. En d'autres termes, la mémoire serait formée d'associations de similarité qui sont constamment mises à jour à partir de stimuli formés d'associations de contiguïté.

D'une manière générale, on peut dire que LSA apprend à partir de stimuli constitués de séquences d'unités lexicales. LSA acquiert des similarités sémantiques entre unités lexicales ou entre stimuli, dans un espace sémantique multidimensionnel. En d'autres termes, LSA apprend le positionnement relatif de chaque unité lexicale par rapport aux autres unités lexicales. On peut alors s'interroger sur la généralisation de ce modèle : peut-on décrire tout processus d'apprentissage par des inputs sous forme de séquences d'unités lexicales et par des outputs sous forme de proximités entre mots dans un espace sémantique ? Probablement non, mais il convient alors de circonscrire le champ d'application du modèle.

7.2. Deux exemples : diagnostic médical et jeu d'échecs

Prenons quelques exemples pour illustrer ce propos. Un médecin est confronté par sa pratique à des cas de malades. Chaque cas peut être décrit par une séquence d'unités lexicales de la forme (`fièvre forte`, `vomissements`, `maux de tête`, etc.). Notons que, bien que nous ayons recours à la langue pour dénommer ces unités lexicales, le lexique du domaine est différent du lexique de la langue puisque certains termes y sont absents et que d'autres sont des composés de termes existants. Les fréquences d'apparition sont également très différentes. Il faut aussi ajouter les unités lexicales qui caractérisent la reconnaissance, immédiate ou différée, de la pathologie (`méningite` par exemple). Cette séquence d'unités lexicales constitue l'input du processus d'apprentissage, de manière analogue aux textes auxquels les sujet qui apprennent une langue sont exposés. Ce que le médecin va apprendre au fur et à mesure de son expérience, c'est à associer de plus en plus finement les symptômes et les pathologies, ce qui revient dans le modèle LSA, à positionner le vecteur de chaque pathologie comme étant le plus proche du vecteur moyen des symptômes. Le diagnostic médical consiste alors à retrouver la pathologie la plus proche des symptômes observés.

De la même manière, l'exposition à un grand nombre de parties d'échecs est probablement un facteur important de l'apprentissage de ce jeu. Chaque partie d'échecs peut être considérée comme une séquence d'états successifs de l'échiquier, décrit à partir d'unités lexicales du type `Fa1`, `Rb8`, etc. L'apprentissage consiste alors à positionner de plus en plus finement ces états de l'échiquier de manière à retrouver une partie antérieure similaire à la partie courante. Les travaux de de Groot (1965) ont montré que c'est surtout cette capacité de mémorisation qui distingue les grands maîtres des experts, plus que leur aptitude au calcul. De plus, cette similarité est évidemment sémantique puisque ce ne sont pas les positions des pièces qui sont mémorisées par les grands maîtres, mais des indicateurs de plus haut niveau, comme par exemple, « faiblesse de la dernière rangée », « colonne ouverte » ou « roi protégé ».

7.3. Généralisation du modèle

Dans le cadre de ce modèle, apprendre un concept *C* représenté par une unité lexicale bien précise, c'est donc situer *C* par rapport aux autres entités qui le définissent. Ces entités peuvent être elles-mêmes des concepts ou des variables de contexte. On dira que le concept est appris s'il est entouré des mêmes voisins que dans un espace sémantique idéal, représentant les significations communément admises dans la langue (issues d'un dictionnaire ou d'une encyclopédie). Par exemple, le concept « maire » est entouré, dans un espace idéal construit à partir de l'analyse d'un grand nombre de textes de la langue, de voisins comme « municipal », « adjoint », etc. Si dans un espace sémantique particulier, ce même concept est entouré de voisins comme « ordinateur », « mémoire », etc., le concept ne sera pas considéré comme appris. Zampa et Lemaire (2002) ont proposé un algorithme général d'identification de significations erronées à partir de l'analyse des productions d'un sujet relativement à un espace idéal.

Les inputs de ce processus général d'acquisition de connaissances doivent être constitués de séquences d'unités lexicales. Ces unités lexicales constituent le lexique du domaine et sont donc en nombre fini. Cette restriction pose problème pour l'extension à d'autres domaines. En effet, lorsque les inputs sont de type continu, par exemple des

valeurs numériques ou des courbes, le modèle ne peut s'appliquer directement et il convient de lui adjoindre une phase de discrétisation. Les valeurs continues sont alors transformées en classes qui constituent chacune une unité lexicale pour LSA. Une deuxième étape de discrétisation opère ensuite pour séparer les contextes entre eux. LSA structure en effet les inputs à deux niveaux de granularité : les unités lexicales, elles-mêmes regroupées en contextes.

Les inputs peuvent également représenter des images. La discrétisation consiste alors à extraire de l'image des indicateurs pertinents. Cette phase est classique dans le domaine du traitement d'informations spatiales. Des travaux récents en vision montrent l'intérêt d'une représentation fondée sur la similarité, avec des problématiques similaires de réduction de dimensions (Edelman, 1998). Une fois le concept situé plus ou moins correctement dans l'espace, le sujet peut retrouver *C* à partir d'un contexte donné, en recherchant le vecteur le plus proche de celui caractérisant le contexte. Landauer (2002) a récemment tenté d'appliquer LSA au domaine de la reconnaissance visuelle. Le « vocabulaire » de la vision serait constitué des différentes cellules de la rétine, les différentes saccades oculaires formant les séquences lexicales. Ce travail reste pour l'instant théorique.

Prenons un exemple dans le domaine des jeux de stratégie : l'apprentissage du jeu du morpion (dont l'objectif est d'aligner 3 symboles sur une grille 3x3) relève en partie de l'exposition à des configurations de jeu. Chaque configuration de jeu est une unité de contexte, comme le sont les paragraphes pour les textes. Les différents lexèmes constitutifs des unités de contexte (autrement dit les mots) sont les suivants : `X1`, `X2`, ..., `X9` pour dénoter chacune des positions du joueur sur la grille, `O1`, `O2`, ..., `O9` pour les positions de l'adversaire, ainsi que des lexèmes caractérisant l'état, gagnant ou perdant, de la grille. Par exemple la grille suivante peut être représentée par : `X1 X2 O4 O5 O6 X7 X9 perdant`.

X	X	
O	O	O
X		X

Figure 2 *— Un exemple de configuration de jeu (morpion)*

Les configurations partielles sont également mémorisées, le lexème perdant ou gagnant n'étant associé qu'en fin de partie. L'espace sémantique est alors composé des configurations de parties rencontrées par le sujet. La stratégie de jeu consiste à rechercher, dans cette mémoire, des configurations sémantiquement similaires à la grille courante et à jouer en conséquence. Lemaire (1998) a simulé cette stratégie sur plusieurs centaines de parties et a montré que le modèle a de meilleurs résultats avec une telle représentation qu'avec une représentation de type purement lexical.

8. Discussion

Les différents travaux recensés plus haut montrent l'intérêt de LSA pour rendre compte de la représentation, de la compréhension et de l'acquisition de connaissances humaines à partir de textes. Cependant, certains problèmes subsistent. Ils sont de trois ordres : — le rôle du type de corpus utilisé en input ; — l'absence de prise en compte des informations syntaxiques, lors du traitement ; — le caractère relatif des relations sémantiques produites par LSA.

La taille et la nature du corpus textuel traité par LSA en input jouent bien évidemment un rôle primordial dans la qualité de l'analyse sémantique. À notre connaissance, son influence n'a pas été encore rigoureusement testée, et le choix de sa taille a plutôt dépendu des capacités de traitement de l'ordinateur utilisé. On ne peut se contenter de préconiser l'utilisation, comme le fait Perfetti (1998), d'un corpus le plus grand possible, car il faut également considérer sa validité. Par exemple, dans le domaine de l'éducation, les corpus à traiter — comme les cours (Dessus, 2000) ou les interactions élève-machine (Wiemer-Hastings, Graesser, & Harter, 1999) — ont une taille relativement faible. De plus, il est également utile de considérer la nature du corpus, et l'on se heurte au dilemme suivant : si l'on ajoute de nombreux textes hors du domaine, on risque de diluer les connaissances du cours proprement dit ; si l'on ajoute des textes du domaine, on risque de modifier plus ou moins profondément les connaissances de la base. Dans une perspective développementale, la solution consiste probablement à ce que le corpus corresponde le plus fidèlement possible aux textes auxquels ont été exposés les sujets. Ce problème de constitution de corpus fiables, pour chaque âge, est crucial pour la poursuite de recherches avec LSA.

Une caractéristique importante de LSA est l'absence de traitement syntaxique : les mots sont traités de manière ensembliste au sein des paragraphes. Les différents résultats que nous avons décrits dans cet article tendraient donc à montrer que la sémantique est surtout portée par les mots, *indépendamment même de leur ordre*. On passe directement du niveau lexical au niveau sémantique, contrairement au schéma classique selon lequel la syntaxe est le soubassement nécessaire de la sémantique. C'est en fait l'effet statistique qui dispense LSA d'un traitement syntaxique : lorsque les énoncés sont courts, LSA ne peut plus dériver d'information sémantique suffisante. Par exemple, la phrase *Les Alliés bombardèrent les Allemands en 1945* ne peut être distinguée de la phrase *Les Allemands bombardèrent les*

Alliés en 1945. En revanche, lorsque ces énoncés apparaissent dans des textes plus longs, LSA peut alors discriminer ce type de phrases. C'est cette limite qui a conduit Kintsch (2001) à réintroduire des éléments syntaxiques (prédicat et argument) lorsqu'il a appliqué le modèle au traitement de métaphores isolées de type « *A* est un *B* » (*voir plus haut*).

Cette absence de traitement syntaxique peut être justifiée par certaines théories récentes. Bever et Townsend (2002) supposent que la compréhension humaine de phrases se réalise en deux phases : dans un premier temps, une analyse grossière d'un sens « vraisemblable » de la phrase, en utilisant les formes syntaxiques canoniques et, dans un deuxième temps, une synthèse, une reconstruction de la structure syntaxique de la phrase, cohérente avec la forme et le sens. Cette théorie, bien que non encore testée empiriquement, est compatible avec de nombreuses données de la psycholinguistique montrant que l'analyse syntaxique, cognitivement coûteuse, donc lente paraît souvent non nécessaire pour comprendre le sens global d'une phrase simple, la syntaxe fonctionnant, comme l'indique Landauer (2002), comme un code de réduction d'erreurs. Ainsi, LSA simulerait cette première analyse indépendante de la syntaxe, suffisante pour avoir une idée du sens général de la phrase.

Le troisième problème est lié à la manière dont LSA représente les mots et les contextes dans l'espace multi-dimensionnel. LSA ne peut rendre compte que de la quantité *relative* de connaissance contenue dans un texte, et non de sa quantité *absolue*. Cela est notamment dû à l'impossibilité d'étiqueter les différents axes. Rehder *et al*. (1998) ont mis au jour ce problème. Ils ont montré qu'il n'était pas toujours possible de différencier, avec LSA, des sujets novices de ceux ayant une connaissance importante d'un domaine, puisque la distance de leurs productions respectives au texte de référence pouvait être identique : le cosinus de l'angle formé par leur vecteur et le vecteur du texte de référence peut avoir la même valeur. Toutefois, cette limitation peut être contournée, comme le montrent Rehder *et al*., notamment en réduisant à une seule dimension l'information contenue dans les textes.

Si l'on a depuis longtemps prouvé que les modèles multidimensionnels pouvaient être des modèles adéquats de la mémoire sémantique (Redington & Chater, 1998), il reste que le nombre de dimensions auxquelles un espace doit être représenté ne fait pas l'objet de consensus. Pour LSA, ce nombre a été fixé empiriquement entre 100 et 300 par Dumais (1991), lors de tests de recherche d'informations, et répliqué par La Cascia, Sethi et Sclarof (1998). En ce qui concerne les tests de synonymie du TOEFL, Landauer et Dumais (1997) montrent que le nombre de dimensions optimal est de l'ordre de trois cents. Teplovs et van Aalst (1998) ont utilisé, pour des corpus réduits, une vingtaine de dimensions. Il n'en reste pas moins que l'impact du nombre de dimensions auxquelles le corpus initial est réduit n'a pas fait l'objet, à notre connaissance, de tests systématiques.

Nous avons montré, dans cet article, que LSA pouvait être utilisé pour modéliser quelques processus cognitifs comme représenter le sens de mots, acquérir du vocabulaire, évaluer des connaissances à partir de textes, comprendre un texte ou des métaphores. Qu'il soit clair que l'équivalence des performances entre humains et modèle ne présage en rien une identité des mécanismes sous-jacents : il serait ainsi surprenant que notre cerveau réalise une réduction de matrice. Par ailleurs, des pans entiers de la linguistique sont négligés par LSA, comme la syntaxe ou la pragmatique. Or, c'est aussi l'intérêt de ce modèle que d'être épuré, puisqu'il permet de cerner précisément les limites de cette sémantique conceptuellement pauvre mais cognitivement plausible.

Remerciements

Nous remercions vivement Jacques Baillé, Maryse Bianco et Catherine Pellenq pour leurs commentaires d'une version précédente de cet article.

Références bibliographiques

[Bever et Townsend, 2002] Bever, T. G., & Townsend, D. J. (2002). Quelques phrases sur notre conscience perceptive des phrases. In E. Dupoux (Ed.), *Les langages du cerveau*. Paris : Jacob, pp. 147-159.

[Britton et Gülgöz, 1991] Britton, B. K., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329-345.

[Burgess et al., 1998] Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes, 25*, 211-257.

[Burgess et Lund, 1997] Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*. *12*(2/3), 177-210.

[Burstein et al., 1998] Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. *Paper presented at the NCME Symposium on Automated Scoring*.

[Carver, 1990] Carver, R. P. (1990). *Reading rate : A review of research and theory*. San Diego: Academic Press.

[Chomsky, 1985] Chomsky, N. (1985). *Règles et représentations*. Paris : Flammarion.

[Chung et O'Neil, 1997] Chung, G., & O'Neil, G. (1997). *Methodological approaches to online scoring of essays*. Los Angeles: Center for the Study of Evaluation, CRESST, Rapport technique n ° 461.

[De Groot, 1965] De Groot, A. D. (1965). *Thought and choice in chess*. La Hague: Mouton.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

[Dessus, 2000] Dessus, P. (2000). Construction de connaissances par exposition à un cours avec LSA. *In Cognito*, *18*, 27-34.

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

[Dessus et Lemaire, 1999] Dessus, P., & Lemaire, B. (1999). Apex, un système d'aide à la préparation des examens. *Sciences et Techniques Educatives, 6*(2), 409-417.

[Dessus et Lemaire, 2002] Dessus, P., & Lemaire, B. (2002). Using production to assess learning: an ILE that fosters Self-Regulated Learning. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems (ITS 2002)*. Berlin : Springer, pp. 772-781.

[Edelman, 1998] Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences, 21*(4), 449-467.

[Elley, 1989] Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly, 24*(174-187).

[Foltz et al., 1998] Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 285-307.

[Foltz et al., 1999] Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. *Proc. ED-MEDIA '99*. Seattle.

[Gounon et Lemaire, 2002] Gounon P., & Lemaire B. (2002). Semantic comparison of texts for learning environments. *Proc. 8e IberoAmerican Conference on Artificial Intelligence (IBERAMIA'2002).* Seville.

[Grefenstette, 1994] Grefenstette, G. (1994). Corpus-derived first, second and third-order word affinities. *Proc. 6th EURALEX*. Amsterdam.

[Harris, 1975] Harris, Z. (1975). Structure distributionnelle. In M. Arrivé & J.-C. Chevalier (Eds.), *La grammaire, lectures*. Paris : Klincksieck, pp. 249-257.

[Jenkins et al., 1984] Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal, 21*(4), 767-787.

[Kail et Fayol, 2000] Kail, M., & Fayol, M. (2000). *L'acquisition du langage* (Tome 1, le langage en émergence). Paris : P.U.F.

[Kintsch, 1988] Kintsch, W. (1988). The role of knowledge in discourse comprehension: A Construction-Integration model. *Psychological Review, 95*(2), 163-182.

[Kintsch, 1998] Kintsch, W. (1998). *Comprehension, a paradigm for cognition*. Cambridge: Cambridge University Press.

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

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

[La Cascia et al., 1998] La Cascia, M., Sethi, S., & Sclaroff, S. (1998). Combining textual and visual cues for content-based image retrieval on the World Wide Web. *IEEE Workshop on Content-Based Access of Image and Video Libraries*.

[Landauer, 2002] Landauer, T. K. (2002). On the computational basis of learning and cognition : Arguments from LSA. *The Psychology of Learning and Motivation*, *41*, 43-84.

[Landauer et Dumais, 1997] 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.

[Larkey, 1998] Larkey, L. S. (1998). Automatic essay grading using text categorization techniques. *Proc. SIGIR'98*. Melbourne.

[Lemaire, 1998] Lemaire, B. (1998). Models of high-dimensional semantic spaces. *Proc. 4th Int. Workshop on Multistrategy Learning (MSL '98)*. Desenzano.

[Lemaire et al., 2001] Lemaire, B., Bianco, M., Sylvestre, E., Noveck, I. (2001). Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente. In H. Paugam-Moisy, V. Nyckees & J. Caron-Pargue (Eds.), *La cognition entre individu et société (ARCo'2001)*. Paris : Hermès, pp. 309-320.

[Lemaire et Dessus, 2001] Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research, 24*(3), 305-320.

[Lowe, 2000] Lowe, W. (2000). What is the dimensionality of human semantic space ? *Proc. 6th Neural Computation and Psychology Workshop*.

[Lund et Burgess, 1996] Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research, methods, instruments and computers, 28*(2), 203-208.

[McKnight et Walberg, 1998] McKnight, K. S., & Walberg, H. J. (1998). Neural network analysis of student essays. *Journal of Research and Development in Education, 32*(1), 26-31.

[McNamara et al., 1996] McNamara, D., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better ? Interactions of text coherence, background knowlegde, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43.

[Memmi, 2000] Memmi, D. (2000). *Le modèle vectoriel pour le traitement de documents*. Grenoble : Université Joseph-Fourier, Cahiers du laboratoire Leibniz n° 14.

[Mross et Roberts, 1992] Mross, E. F., & Roberts, J. O. (1992). *The Construction-Integration model: A program and manual* (n° 92-14). Boulder: Université du Colorado.

[Nagy et al., 1985] Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, 20, 223-253.

[Nagy et al., 1987] Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meaning from context during normal reading. *American Educational Research Journal, 24*(2), 237-270.

[Noizet et Caverni, 1978] Noizet, G., & Caverni, J.-P. (1978). *Psychologie de l'évaluation scolaire*. Paris : P.U.F.

[Page, 1966] Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238-243.

[Perfetti, 1998] Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes, 25*(2/3), 363-377.

[Pullum et Scholtz, 2002] Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19*(1/2), 9-50.

[Rapaport et Ehrlich, 2001] Rapaport, W. J., & Ehrlich, K. (2001). A computational theory of vocabulary acquisition. In L. Iwanska & S. C. Shapiro (Eds.), *Natural Language processing and Knowledge Representation*. Menlo Park: AAAI Press/MIT Press, pp. 347-378.

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

[Rehder et al., 1998] Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*(2/3), 337-354.

[Salton et McGill, 1983] Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

[Stockburger, 1999] Stockburger, D. (1999). Automated grading of homework assignments and tests in introductory and intermediate statistics courses using active server pages. *Behavior Research Methods, Instruments, and Computers, 31*(2), 252-262.

[Swanborn et de Glopper, 1999] Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research, 69*(3), 261-285.

[Teplovs et van Aalst, 1998] Teplovs, C., & van Aalst, J. (1998). Latent Semantic Analysis of CSILE/KF databases. *Poster presented at NCE'98*. Vancouver.

[Turney, 2001] Turney, P. (2001) Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In L. de Raedt & P. Flach (Eds.), *Proc. 12th European Conference on Machine Learning (ECML-2001)*. Fribourg, pp. 491-502.

[Wiemer-Hastings et al., 1999] Wiemer-Hastings, P., Graesser, A. C., & Harter, D. (1999). The foundations and architecture of Autotutor. In H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Intelligent Tutoring Systems (ITS '98)*. Berlin : Springer, pp. 334-343.

[Wolfe et al., 1998] Wolfe, M. B. W., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 309-336.

[Zampa, 1999] Zampa, V. (1999). Automatic text selection by LSA. *Young Researchers Track of the Artificial Intelligence in Education International Conference (AI-ED'99)*, Le Mans.

[Zampa et Lemaire, 2002] Zampa, V., & Lemaire, B. (2002). Latent Semantic Analysis for Student Modeling. *Journal of Intelligent Information Systems*, *18*(1), 15-30.