Les modèles vectoriels de la mémoire sémantique : description, validation et perspectives

Cédrick Bellissens, Pierre Thérouanne, Guy Denhière

Deux modèles vectoriels de la mémoire sémantique : Description, théorie et perspectives.

Par

Cédrick Bellissens*, Pierre Thérouanne et Guy Denhière*****

*Laboratoire Cognition et Activités Finalisées, Université de Paris VIII, 2, rue de la Liberté, 93526 St Denis, Cedex 02. cedrick.bellissens@univ-paris8.fr.

**Laboratoire de Psychologie Expérimentale et Quantitative, Université de Nice Sophia-Antipolis, 24, Avenue des Diables Bleus, 06357 Nice. Pierre.Therouanne@unice.fr.

***Laboratoire de Psychologie Cognitive, UMR 6146, CNRS et Université d'Aix-Marseille I, 3, place Victor Hugo, 13331 Marseille cedex 3. denhiere@up.univ-mrs.fr.

A paraître : Le Langage et l'Homme, Décembre 2004

Titre courant : Deux modèles vectoriels de la mémoire sémantique.

Résumé :

Les modèles "Latent Semantic Analysis" (LSA ; Landauer et Dumais, 1997) et "Hyperspace Analog to Language" (HAL ; Lund et Burgess, 1996) peuvent être qualifiés d'abstractifs (Tiberghien, 1997) car ils modélisent le résultat d'une généralisation sémantique sur un ensemble d'épisodes d'apprentissage. Ces modèles analysent statistiquement la distribution de mots dans de larges corpus textuels pour élaborer un espace sémantique dans lequel chaque mot est représenté par un vecteur. L'objectif de cet article est de décrire ces deux modèles en les comparant. Nous montrons que ces modèles sont capables de prédire des résultats d'amorçage sémantique et qu'ils peuvent être couplés à des modèles de compréhension tels que Construction-Intégration (Kintsch, 1988) pour simuler les activités psychologiques d''accès à la signification, de prédication et de construction d'une représentation mentale cohérente d'un texte.

Mots-clés : Compréhension, Généralisation, Mémoire sémantique, Modèles abstractifs, Vecteurs.

Title: Two abstractive models of semantic memory: Description, validation and perspective.

Summary:

"Latent Semantic Analysis" (LSA ; Landauer et Dumais, 1997) and "Hyperspace Analog to Language " (HAL ; Lund et Burgess, 1996) models can be called abstractive models (Tiberghien, 1997) because they model a semantic generalization over a number of learning episodes. LSA and HAL statistically analyze distribution of terms belonging to a large textual corpus to elaborate a semantic space in which each term is represented by a vector. The main goal of this note is to describe these two models by comparing them each other. We show that LSA and HAL are able to predict priming experiment data and to be combined with a comprehension model like Construction-Integration (Kintsch, 1988) to simulate in a realistic manner, signification access, predication process and construction of a consistent mental representation of a text.

Keywords: Abstraction, Comprehension, Generalization, Semantic memory, Vectors.

La recherche en psychologie cognitive de la mémoire et du langage a pour objectif de déterminer quelles sont les structures mnésiques intervenant dans le traitement et la compréhension du langage, considérant que les processus impliqués dans le langage ne sont pas spécifiques mais généraux et contraints par la mémoire (Denhière & Baudet, 1992 ; Kintsch, 1998 ; Myers & O'Brien, 1998 ; Noordman & Vonk, 1998 ; Albrecht & Myers, 1998 ; Sanford & Garrod, 1998). Parmi les structures mnésiques impliquées, les représentations des connaissances stockées en mémoire à long terme sont considérées comme influençant notablement la construction d'une représentation mentale cohérente d'un texte lu (Caillies, Denhière & Jhean-Larose, 1999 ; Kintsch, 1988).

Depuis plus d'une dizaine d'années, la recherche en psychologie cognitive de la mémoire et du langage a généré de nouveaux modèles de la représentation des connaissances. La conception générale de ces nouveaux modèles est de représenter les connaissances sous la forme de vecteurs dans un espace de grande dimension, ce qui sous-entend une théorie de l'apprentissage qui caractérise ce que Tiberghien (1997, 2002) qualifie de modèles abstractifs. Les modèles abstractifs se différencient radicalement des modèles en réseaux sémantiques et des modèles associatifs d'un point de vue théorique, car ils formalisent l'hypothèse selon laquelle la mémoire sémantique émerge de l'encodage successif des résultats d'apprentissages épisodiques ; et d'un point de vue méthodologique, car ils ne nécessitent pas la constitution d'ontologies, ni la construction de normes d'association, bien que cette option ne soit pas rejetée par tous les auteurs (Steyvers, Shiffrin, & Nelson, in press). La mémoire est ici considérée non plus comme un réseau constitué de nœuds concepts et de liens étiquetés ou pondérés mais comme un espace à plusieurs dimensions dans lesquelles sont représentés des vecteurs concepts.

L'objectif de cet article est de décrire ces modèles et d'en évaluer l'utilité pour la recherche sur la compréhension et la mémoire. Dans la famille des modèles abstractifs, les

modèles que nous avons choisi de décrire sont Latent Semantic Analysis (LSA ; Landauer & Dumais, 1997) et Hyperspace Analog to Language (HAL ; Lund & Burgess, 1996) car ils nous paraissent les plus typiques. S'ils sont les plus typiques, ces deux modèles ne sont cependant pas les « tout premiers » modèles vectoriels et abstractifs, cette originalité pouvant être attribuée aux modèles fonctionnels à appariement global tels que SAM (Raijmaker & Shiffrin, 1980), CHARM (Eich-Metcalfe, 1985) MINERVA (Hintzman, 1986) et TODAM (Murdock, 1993). Ces modèles représentent les informations en mémoire sous la forme de vecteurs et font émerger la mémoire sémantique de la généralisation de traces épisodiques. Ils peuvent rendre compte d'un grand nombre de phénomènes, comme le rappel, la reconnaissance, la catégorisation ou l'abstraction de prototypes (voir Tiberghien, 1997). Ils fournissent également des règles d'encodage et de récupération, mais ils ne précisent ni la nature des représentations, ni d'où proviennent les coordonnées des vecteurs qui codent l'information.

HAL et LSA représentent également les concepts par des vecteurs. Cependant, les entrées sont constituées par un large corpus de textes auquel est appliqué un traitement statistique qui permet d'obtenir un espace sémantique. La construction d'un espace sémantique nécessite traditionnellement la définition d'un ensemble d'axes, de dimensions, et l'évaluation de la position de chaque concept sur chaque axe à partir de réponses fournies par des individus (Osgood, Suci et Tannenbaum, 1957 ; McRae, de Sa et Seidenberg, 1997). À la différence de cette tradition, les modèles HAL et LSA ne déterminent pas a priori les dimensions de l'espace sémantique. La sémantique de HAL et LSA résulte d'un calcul qui prend en entrée des productions langagières, sous leur forme la plus naturelle, c'est-à-dire le discours. La construction d'un tel espace ne prend donc pas comme principe de placer les concepts dans un cadre sémantique, défini a priori, dont procèderait le discours mais de déterminer le cadre sémantique a posteriori, comme étant le résultat d'une analyse de l'usage

des mots dans le discours. Il s'agit de considérer la mémoire sémantique comme une généralisation des événements spécifiques stockés en mémoire épisodique.

Nous décrirons, dans un premier temps les caractéristiques de ces modèles et, en particulier, la manière dont ils transforment un large corpus de textes en un espace sémantique. Dans un deuxième temps, nous tenterons de montrer que ces modèles sont effectivement des modèles abstractifs qui permettent de faire émerger des relations entre concepts en utilisant un mécanisme de généralisation de l'usage des mots dans le discours. Enfin, dans un troisième temps nous évaluerons théoriquement l'opportunité qu'offrent ces modèles à la recherche sur la mémoire et la compréhension du langage.

1. CARACTERISTIQUES PRINCIPALES DES MODELES LSA ET HAL

La méthodologie générale employée pour développer les représentations des concepts à partir d'un environnement langagier est semblable dans les deux modèles. Premièrement, il s'agit de rassembler un corpus de textes. Le corpus doit satisfaire plusieurs exigences si l'on veut qu'il soit le plus représentatif de l'environnement langagier auquel sont confrontées les personnes. Dans le but de construire un espace saisissant un vocabulaire exhaustif, il est nécessaire de constituer un corpus de taille conséquente. De plus, les thèmes abordés dans les textes doivent être assez nombreux pour que chaque mot puisse se rencontrer dans la majorité de ses contextes d'usage. Dans un deuxième temps, les mots du corpus sont représentés dans un tableau qui inscrit leur distribution dans le corpus étudié. Enfin, à partir de ce tableau sont constitués les vecteurs qui vont représenter les mots dans un espace de grande dimension.

1.1. L'ANALYSE SEMANTIQUE LATENTE (LSA)

L'Analyse Sémantique Latente est la procédure automatique proposée par Landauer et Dumais (1997) pour construire un espace vectoriel. Cette procédure s'applique à un vaste corpus de textes et comporte trois étapes au cours desquelles le corpus de textes est progressivement transformé en un espace vectoriel de plusieurs centaines de dimensions. Le corpus de textes comporte deux types de séparateurs, le saut de paragraphes et l'espace entre deux mots. Le paragraphe est considéré comme la suite de caractères comprise entre deux sauts de paragraphe et le mot est la suite de caractères comprise entre deux espaces.

La première étape de la procédure consiste à représenter le corpus sous la forme d'une matrice de cooccurrences. La deuxième, consiste à appliquer à cette matrice une analyse factorielle appelée décomposition en valeurs singulières pour obtenir un espace. La dernière étape consiste à éliminer, parmi les dimensions de l'espace résultant de la décomposition en valeurs singulières, un certain nombre de dimensions, considérées comme non pertinentes.

Première étape : Représentation du corpus textuel sous la forme d'un tableau

Les corpora utilisés comme entrées du modèle sont généralement constitués de textes littéraires ou encyclopédiques (voir annexe). Leur taille doit être suffisamment grande pour que chaque mot du vocabulaire puisse être exploité dans la plupart de ses contextes d'usage. Après constitution du corpus, le nombre de fois que chaque mot apparaît dans chaque paragraphe est comptabilisé. Les fréquences de cooccurrence entre mots et paragraphes sont calculées. Ces fréquences sont inscrites dans un tableau. En colonne, se trouve chaque paragraphe et en ligne, chaque mot. À l'intersection d'une colonne et d'une ligne, chaque cellule contient la fréquence de cooccurrence d'un mot et d'un paragraphe.

Deuxième étape : La décomposition en valeurs singulières

La décomposition en valeurs singulières est une méthode générale de décomposition linéaire d'une matrice en composantes principales indépendantes. Comme une analyse en composantes principales, cette méthode permet de dégager d'un ensemble de données - ici des fréquences de cooccurrence - un nombre de facteurs sans corrélation entre eux et rendant chacun compte de la variance de l'ensemble des données. Si n facteurs rendent compte de la totalité de la variance des fréquences de cooccurrence, alors les données peuvent être représentées dans un espace à n dimensions, chaque dimension correspondant à un facteur.

Le tableau comportant les mots en lignes et les contextes en colonnes forme une matrice rectangulaire, X_{m*c} dans laquelle m est le nombre de lignes et c le nombre de colonnes. Cette matrice rectangulaire X_{m*c} est décomposée en trois matrices dont elle est le produit, U_{m*n}, Σ_{n*n} et V_{c*n} :

$$X = U \Sigma V^T$$

La matrice Σ_{n*n} est une matrice diagonale avec n colonnes et n lignes, dont les cellules de la diagonale contiennent "les valeurs singulières". La matrice mot, U, est une juxtaposition de m lignes comportant n valeurs. Les n valeurs de chaque ligne sont les coordonnées d'un vecteur représenté dans un espace à n dimensions associé à un mot du corpus. Chaque mot est donc représenté dans un espace à n dimensions. Les coordonnées d'un vecteur « mot » ne doivent pas être considérées comme la représentation numérique du type d'environnement verbal dans lequel le mot tend à apparaître. Les vecteurs sont représentés dans n dimensions qui demeurent parfaitement abstraites et indépendantes des contextes. En effet, la décomposition en valeurs singulières met à jour des facteurs issus de la variabilité des environnements verbaux des mots du corpus et ces facteurs sont donc indépendants des environnements verbaux. Le processus de construction de l'espace sémantique correspond en

fait à la formalisation de l'hypothèse selon laquelle la mémoire sémantique émerge d'un ensemble d'épisodes d'apprentissage.

À l'issue de cette étape, la similarité sémantique entre deux mots peut alors être calculée. Cette similarité sémantique est estimée par le calcul du cosinus de l'angle que forment les vecteurs représentant ces mots dans l'espace à n dimensions. Soient le vecteur u et le vecteur v, le cosinus de l'angle θ par u et v est

$$\cos\theta = \frac{u \cdot v}{\|u\| \, \|v\|}$$

Deux vecteurs identiques forment un angle nul dont le cosinus est égal à 1. Deux vecteurs perpendiculaires forment un angle droit dont le cosinus est égal à 0 et deux vecteurs opposés forment un angle plat dont le cosinus est égal à –1. La similarité sémantique varie donc de –1 à 1.

Troisième étape : réduction du nombre de dimensions

Toutes les dimensions dégagées de la décomposition en valeurs singulières ne sont pas pertinentes. Les dimensions associées aux valeurs singulières les plus faibles n'expliquent qu'une très faible part de la variance des données d'origine. Si ces dimensions n'étaient pas éliminées, le modèle ferait des erreurs d'estimation de la similarité sémantique. Comme les dimensions sont abstraites, il n'existe pas de critères d'élimination des dimensions non pertinentes. En conséquence, le nombre de dimensions éliminées doit être déterminé de manière empirique.

La méthode employée par Landauer et Dumais (1997) consiste à manipuler le nombre de dimensions de l'espace jusqu'à ce que les résultats du modèle soient le plus fortement corrélés aux résultats attendus ou à des résultats humains dans une tâche appropriée donnée. Lorsque deux vecteurs sont proches dans LSA, on remarque qu'intuitivement tout individu est

susceptible de considérer que les deux mots donnés sont sémantiquement proches. Cette propriété a été évaluée en comparant les données de similarité sémantique du modèle à des résultats de jugements humains. Un test standard de vocabulaire issu du TOEFL (Test of English as a Foreign Language) a été utilisé. Le TOEFL est un test standard utilisé pour évaluer la compétence linguistique des étudiants désirant étudier aux Etats-Unis. Pour comparer les résultats humains à ceux de LSA, les auteurs ont utilisé un corpus de 4,5 millions de mots, constitué de 30473 contextes comportant chacun environ 500 caractères. Le test de vocabulaire du TOEFL comporte 80 items dont la partie question est constituée d'un seul mot. Quatre réponses sont proposées au candidat. Parmi ces réponses, le candidat doit choisir le meilleur synonyme. Pour simuler la performance humaine dans LSA, la similarité sémantique entre le mot inducteur et chaque alternative est calculée. Le meilleur synonyme correspond à la réponse associée au vecteur le plus corrélé au vecteur associé au mot inducteur. Alors qu'ils augmentent progressivement le nombre de dimensions de l'espace sémantique, les auteurs calculent, pour chaque ordre de l'espace, la corrélation entre les résultats du modèle et les résultats obtenus par des étudiants passant le test de vocabulaire. La meilleure corrélation est obtenue pour 300 dimensions. Dans la plupart des applications, le nombre de facteurs pertinents peut passer de 30000 à 300. Ceci signifie que les vecteurs passent de 30000 coordonnées dans un espace à 30000 dimensions à 300 coordonnées dans un espace à 300 dimensions.

1.2. HYPERSPACE ANALOG TO LANGUAGE (H.A.L.)

Les sources du corpus auquel est confronté le modèle HAL sont différentes de celles constituant le corpus du modèle LSA. Plutôt que d'exploiter des textes littéraires ou encyclopédiques, Burgess et ses collègues ont recueilli un ensemble de textes en langue anglaise

sur les forums informatiques. Ce corpus respecte les exigences de taille[1] et de richesse thématique. De plus, les textes composant le corpus sont par nature "conversationnels" et bruités par la présence de mots familiers, de fautes d'orthographes ou grammaticales, de structures syntaxiques incorrectes, de parties de discours relativement peu structurées ou incohérentes, et donc plus proches du langage parlé quotidien.

À partir du corpus constitué, est construite une matrice de valeurs de cooccurrence entre les mots du corpus. Opérationnellement, deux mots sont considérés comme cooccurrents lorsqu'ils apparaissent dans une même fenêtre flottante. La taille de cette fenêtre est de quelques occurrences à gauche et à droite du mot considéré. Une telle taille a été choisie afin de préserver la localité de la référence tout en minimisant les effets des différentes constructions syntaxiques (Burgess, Livesay et Lund, 1998). Les valeurs de corrélation les plus élevées entre les distances sémantiques calculées au moyen de HAL et des temps de décision lexicale dans une expérience d'amorçage sémantique sont obtenues avec une fenêtre de 8 mots, puis avec des fenêtres de 4 et 10 mots (Lund et Burgess, 1996). Lors de chaque déplacement de cette fenêtre d'une occurrence dans le corpus, les valeurs de cooccurrence sont enregistrées de la façon suivante : la valeur de cooccurrence entre deux mots est maximale lorsque les 2 mots sont immédiatement adjacents pour décroître de façon linéaire avec la distance entre ceux-ci à l'intérieur de la fenêtre.

Ensuite, les valeurs ainsi recueillies sont utilisées pour élaborer une matrice de cooccurrence qui possède en tant qu'axes la totalité du vocabulaire pris en considération. Ce vocabulaire est constitué des 70 000 mots les plus fréquents dans le corpus. La valeur d'une cellule de la matrice de HAL est égale à la somme des cooccurrences pour une paire orientée de mots : l'information contenue dans une ligne représente la somme des cooccurrences pour les

[1] Le corpus est composé initialement de 160 millions d'occurrences (Lund et Burgess, 1996), ce nombre étant porté à 300 millions d'occurrences par la suite (Burgess et Lund, 1997 ; Burgess, Livesay, Lund,

mots apparaissant avant le mot considéré, alors que celle contenue dans une colonne représente la somme des cooccurrences pour les mots apparaissant après le mot considéré.

Considérons par exemple la phrase suivante analysée à l'aide d'une fenêtre flottante dont la taille est égale à 5 occurrences de part et d'autre du mot considéré : "Le chat dort sur le canapé."

Prenons le mot "canapé" ; dans la matrice, la ligne correspondant à "canapé" contient l'information qui cooccurre et précède "canapé". L'occurrence du mot "le" juste avant le mot "canapé" prend une valeur de cooccurrence maximale de 5 dans la mesure où aucun mot ne les sépare. De plus, la première occurrence de "le" dans la phrase prend une valeur de co-occurrence de 1 parce que 4 mots les séparent. La sommation de 5 et 1 donne la valeur 6 qui est enregistrée dans la cellule.

Insérer TABLEAU I

Pour chaque mot, la ligne et la colonne de la matrice qui lui sont associés sont assemblées et résultent dans un vecteur de longueur égale à 2n dans un espace à 2n dimensions, n étant le nombre de mots différents contenus dans la matrice. Ainsi, dans le modèle HAL, chaque élément d'un vecteur correspond à la valeur de cooccurrence du mot qu'il représente avec un autre mot. Dans l'exemple simplifié présenté ci-dessus, le vecteur de cooccurrences représentant le mot "canapé" serait (0, 2, 3, 6, 4, 0, 0, 0, 0, 0). Dans certaines simulations (par exemple, Lund et Burgess, 1996), seuls les 200 éléments les plus informatifs du vecteur sont retenus dans la mesure où ils sont responsables de la quasi-totalité de la variance du vecteur.

1998)

La similarité sémantique entre deux mots est évaluée en calculant la distance métrique entre les deux vecteurs correspondants de la façon suivante[2]:

$$\text{distance} = \sqrt[r]{\sum \left(|x_i - y_i|\right)^r}$$

Le vecteur correspondant à un mot contient donc l'ensemble de l'expérience de ce dernier avec tous les autres mots. C'est pourquoi ce sont les contextes dans lesquels les mots apparaissent qui déterminent leur signification. Initialement qualifiée de similarité sémantique (Lund et Burgess, 1996), cette distance sera ensuite appelée distance contextuelle entre deux vecteurs contextuels dans un espace contextuel à de nombreuses dimensions. Ce choix est motivé par la nature contextuelle de la signification des mots (Burgess et Lund, 1997).

2. DIFFERENCES ET POINTS COMMUNS ENTRE LES DEUX MODLES

Les modèles LSA et HAL diffèrent sur plusieurs points. On peut légitimement se poser la question de la portée pratique de ces différences, car à première vue, elles semblent définir des choix théoriques tranchés.

2.1. LES DIFFERENCES

Premièrement, dans l'un et l'autre modèle, la définition opérationnelle de la cooccurrence diffère : elle est cooccurrence entre mots qui appartiennent à une même fenêtre dans le modèle HAL, elle est cooccurrence entre un mot et un paragraphe dans le modèle LSA. Deuxièmement, la matrice d'origine utilisée dans le modèle LSA a pour axes (a) l'ensemble du vocabulaire et (b) l'ensemble des paragraphes, alors que l'ensemble du vocabulaire constitue les deux axes de la matrice du modèle HAL. Troisièmement, dans LSA, une décomposition en

[2] La valeur du coefficient de corrélation entre les distances métriques calculées au moyen de HAL et des temps de décision lexicale dans une expérience d'amorçage sémantique est optimum avec une valeur de *r* égale à 1 (Lund et Burgess, 1996).

valeurs singulières est réalisée sur la matrice d'origine afin d'extraire des facteurs indépendants rendant compte de la variance de la distribution des mots dans le corpus textuel. Ces facteurs indépendants sont considérés comme des dimensions sémantiques latentes. Ainsi, le modèle LSA peut être assimilé à une approche subsymbolique, dans la mesure où les caractéristiques d'un vecteur renvoient à des dimensions abstraites qui ne sont pas en soi interprétables. En revanche, le modèle HAL ne relève pas à proprement parler d'une approche subsymbolique, dans la mesure où chaque caractéristique d'un vecteur renvoie à un mot, et donc à un symbole, dont l'interprétation est relativement transparente.

Ainsi la différence majeure entre les deux modèles réside dans le codage de l'information. Dans HAL, un vecteur est constitué du codage de l'ensemble des contextes lexicaux dans lequel le mot peut être rencontré. Dans LSA, un vecteur est constitué du codage de la généralisation de l'ensemble des contextes textuels dans lesquels le mot peut être trouvé. Cependant, la généralisation n'est pas absente du calcul opéré par le modèle HAL. En effet, si HAL n'intègre pas dans le codage des vecteurs une quelconque généralisation, mais reste au contraire très proche de la distribution du vocabulaire dans une fenêtre flottante, le calcul de la similarité entre vecteurs implique nécessairement cette généralisation.

2.2. LES DEUX MODELES FORMULENT-ILS, MALGRE TOUT, LES MEMES PREDICTIONS ?

Dans cette section, nous nous intéresserons à une tentative de reproduction de résultats d'amorçage sémantique obtenus par Moss, Ostrin, Tyler et Marslen-Wilson (1995) que nous avons réalisée (Bellissens et Denhière, 2002) après McDonald et Lowe (1998). Cette tentative nous paraît importante car elle pourrait démontrer que LSA et HAL, sont bien tous

les deux des modèles qui généralisent l'information spécifique des contextes d'usage des mots du corpus pris en entrée.

Moss et al. (1995) ont construit des paires de mots amorce-cible en manipulant la Relation Sémantique entre l'amorce et la cible (présente ou absente), le Type de Relation Sémantique (catégorielle vs fonctionnelle) et le degré d'Association établi grâce à des jugements d'association[3]. Nous renvoyons le lecteur à la lecture de cet article pour obtenir des détails sur la construction du matériel. La tâche consistait en une décision lexicale portant sur le second mot de chaque paire, lequel était présenté une seconde après la présentation auditive d'une amorce. Les facteurs Relation Sémantique et Force Associative exerçaient un effet significatif sur le temps de réaction. L'effet d'amorçage était significativement plus grand pour les items associés que pour les items moins associés. Cependant, la variation du Type de Relation Sémantique n'exerçait pas d'effet significatif sur le temps de réaction.

À l'aide d'un modèle semblable à HAL, McDonald et Lowe (1998) ont tenté de reproduire les résultats de l'expérience de Moss et al. (1995). La simulation consistait à calculer le cosinus entre les vecteurs associés à chaque membre d'une paire d'items, la valeur du cosinus étant prise comme estimation de la similarité entre les deux items.

Dans une première partie, les résultats de la simulation reproduisaient effectivement les résultats obtenus par Moss et al. (1995). En effet, en condition Sémantiquement Relié la similarité était significativement plus élevée qu'en condition Sémantiquement non Relié, 0,28 et 0,12 respectivement, ($F(1,92) = 73,6$, $p < .01$). Cet effet d'amorçage sémantique était significativement plus important en condition Fortement Associé qu'en condition Faiblement Associé, 0,19 et 0,11 respectivement, ($F(1,92) = 4,6$, $p < .05$). Cependant, la seconde partie des résultats de la simulation prédisait des différences non obtenues par Moss et al. (1995).

En effet, le cosinus entre coordonnés catégoriels était significativement plus grand qu'entre l'amorce et la cible des paires fonctionnelles, 0,24 et 0,15 respectivement, ($F(1,92) = 23,2$, $p < .01$). La différence de similarité entre les conditions Sémantiquement Relié et Non Relié était significativement plus grande en condition Coordonnés Catégoriels qu'en condition Paires Fonctionnelles, 0,23 et 0,08 respectivement, ($F(1,92) = 17,3$, $p < .01$). Ainsi, le modèle de MacDonald et Lowe prédisait un effet du facteur Type de Relation Sémantique, qui implique un temps de décision lexicale significativement plus long en condition Coordonnés Catégoriels qu'en condition Paire Fonctionnelle et un plus grand effet d'amorçage sémantique en condition Coordonnés Catégoriels qu'en condition Paire Fonctionnelle, prédictions qui ne sont pas vérifiées par les résultats expérimentaux.

McDonald et Lowe (1998) ont expliqué ces derniers résultats par le fait que le modèle HAL codait l'information de manière spécifique et qu'à ce titre, l'ordre usuel des mots dans le corpus utilisé par le modèle influençait directement le calcul de la similarité. Selon eux, deux mots entretenant une relation fonctionnelle ont tendance à se suivre dans un corpus. De ce fait, le cosinus de l'angle que forme leur vecteur respectif tendrait vers zéro.

Nous avons tenté de reproduire ces résultats (Bellissens et Denhière, 2002) en utilisant LSA. Au contraire de McDonald et Lowe (1998), nous considérions que l'ordre des mots ne constituait pas une raison suffisante pour expliquer les résultats obtenus. Nous formulions l'hypothèse selon laquelle la relation fonctionnelle et la relation catégorielle étaient formées par la généralisation d'épisodes spécifiques définis par le partage de l'extension entre l'un et l'autre mot de la paire, indépendamment de leur succession. En effet, par définition, la relation catégorielle repose nécessairement sur un large partage de l'extension des propriétés de l'un et l'autre mot de la paire, contrairement aux paires de mots reliés fonctionnellement. Ainsi,

[3] Par exemple, dans la condition "relation fonctionnelle", beach et sand (plage sable) sont jugés comme étant plus associés que broom et floor (balai sol). Dans la condition "relation catégorielle",

comme la similarité dans LSA ne résulte pas de l'ordre des mots dans le corpus mais d'un plus ou moins large partage de l'extension des mots, les résultats obtenus à l'aide de LSA seront analogues à ceux obtenus avec HAL qu'à la condition qu'ils ne soient pas déterminés par l'ordre des mots.

Les résultats obtenus étaient les suivants. Tout d'abord, LSA permettait également de prédire les effets d'amorçage obtenus dans l'expérience de Moss et al. (1995) : le cosinus était significativement plus élevé en condition Sémantiquement Relié qu'en condition Non Sémantiquement Relié, 0,39 et 0,10 respectivement, ($F(1,27) = 168,4$, $p < .01$). De plus, le cosinus était significativement plus élevé en condition Associé qu'en condition Faiblement Associé, 0,28 et 0,22 respectivement, ($F(1,27) = 9,3$, $p < .01$). L'effet d'amorçage était significativement plus important en condition Associé qu'en condition Faiblement Associé, 0,36 et 0,23 respectivement, ($F(1,27) = 11,2$, $p < .01$).

De plus, comme dans la simulation de McDonald et Lowe (1998), les coordonnés catégoriels étaient jugés plus similaires que les paires fonctionnelles, 0,27 et 0,22 respectivement, ($F(1,27) = 5,5$, $p < .05$). L'effet d'amorçage prédit était significativement plus grand pour les coordonnés catégoriels que pour les paires fonctionnelles, 0,34 et 0,26 respectivement, ($F(1,27) = 5,0$, $p < .05$).

2.3. DE LA GENERALISATION COMME EXPLICATION

Les patrons de résultats obtenus grâce aux deux modèles étaient donc semblables : ils reproduisaient les différences significatives obtenues par Moss et al. (1995) et, en plus, ils prédisaient un effet du facteur Type de Relation Sémantique, résultat qui n'était pas obtenu par Moss et al. Dans le cadre spécifique de leur modèle de type HAL, McDonald et Lowe (1998) suggéraient la possibilité que la prise en compte de l'ordre des mots dans le corpus

---

brother et sister (frère sœur) sont plus associés que prince et duke (prince duc ; d'après Moss et al.

d'étude menait à ce que l'angle de vecteurs associés à des mots habituellement adjacents tende à être un angle droit, le cosinus d'un angle droit étant égal à zéro. Cette explication n'est pas suffisante : LSA, qui ne prend pas en compte l'ordre des mots dans la construction de l'espace vectoriel, conduit à l'obtention des mêmes résultats. L'explication que nous proposons est la suivante. La relation catégorielle découle d'un large partage de l'extension de l'un et l'autre mot impliqués dans cette relation. En fait, l'extension comme le partage des extensions sont déterminés, au moins dans LSA, par la généralisation des contextes d'usage des mots considérés. Il en résulte que le modèle HAL généralise de la même manière que LSA, alors que le codage de l'information diffère dans l'un et l'autre modèle.

Ainsi, les deux modèles prédisent les mêmes résultats, résultats ne sont pas obtenus dans l'expérience de Moss et al. (1995). Il est possible que l'intervalle inter-stimulus (ISI) utilisé par Moss et al. (1995) soit trop long (1000 ms) pour obtenir les résultats prédits par les modèles. Il est également possible que la généralisation de l'extension de chacun des mots d'une paire ne permette pas de prédire un effet d'amorçage. De futures recherches devront répondre à cette question importante.

Il reste que la généralisation semble être le mécanisme de base de ces modèles de la mémoire sémantique. Ce mécanisme permettant de faire émerger des relations catégorielles est évoqué pour expliquer la construction d'une représentation mentale cohérente d'un texte en cours de lecture (Kintsch & van Dijk, 1978 ; van Dijk & Kintsch, 1983), mais également pour expliquer le processus de prédication qui consiste à représenter en mémoire la signification d'une proposition (Bellissens, Thiesbonenkamp & Denhière, 2002 ; Kintsch, 2000, 2001). Cette communauté de processus permet de comprendre la raison pour laquelle des modèles abstractifs peuvent être couplés à des modèles du traitement du discours, notamment au modèle de Construction-Intégration proposé par Kintsch (1988, 1998).

3. GENERALISATION ET COMPREHENSION : LE COUPLAGE CI-LSA

Kintsch, Patel, et Ericsson (1999) et Kintsch (1998) posent les bases d'un modèle de la compréhension qui combine LSA et Construction-Intégration (Kintsch, 1988). Construction-Intégration est un modèle de la compréhension de textes dans lequel un texte est traité de manière cyclique. Un cycle de traitement comporte deux phases. La première, la phase de construction, a pour finalité la construction d'une représentation d'un segment de texte, résultant de l'activation associative des plus proches voisins, en mémoire à long terme, des propositions et des concepts présents dans le segment. La seconde, la phase d'intégration, simule l'intervention du contexte discursif dans le traitement du segment de texte et permet de supprimer les éléments redondants ou incohérents activés au cours de la première phase. Dans le cadre CI-LSA, LSA est utilisé comme modèle de mémoire associative. Les éléments activés au cours de la phase de construction appartiennent donc à une partie active de la mémoire à long terme. Cette partie active est en relation constante avec une partie accessible de la mémoire à long terme qui n'apparaît pas en mémoire de travail à court terme. Les auteurs considèrent la mémoire à long terme comme un vaste réseau associatif sur lequel opèrent des fonctions relatives à la mémoire de travail à court terme et la mémoire de travail à long terme (Ericsson et Kintsch, 1995). La mémoire de travail à court terme intervient dans l'activation temporaire d'un sous-réseau et permet de maintenir actifs des nœuds directement associés à d'autres nœuds en mémoire de travail à long terme.

Pour Kintsch (1998), les connaissances peuvent être représentées sous la forme d'un réseau de propositions interconnectées. La signification d'un nœud dans le réseau est donnée par sa localisation dans le réseau, c'est-à-dire par la force d'association qui le lie aux autres nœuds du réseau. Les concepts ne possèdent pas une signification fixée et permanente. Au contraire, chaque fois qu'un concept est utilisé, sa signification est construite en mémoire de travail par l'activation d'un sous-réseau particulier de propositions dans le voisinage du nœud

concept. Grâce à LSA, la construction d'un tel sous-réseau est possible en demandant au modèle de donner les plus proches voisins d'un mot particulier.

3.1. L'ALGORITHME DE PREDICATION

L'utilisation de LSA comme réseau de connaissances peut poser un problème pour la construction des propositions sémantiques dans le réseau. En effet, LSA donne en sortie un espace sémantique de grande dimension dans lequel chaque mot du corpus d'étude est représenté sous la forme d'un vecteur. La question est de savoir comment construire une proposition sémantique en utilisant un tel modèle. On peut, à partir de LSA, calculer un vecteur proposition, en calculant la somme des vecteurs associés au prédicat et aux arguments de la proposition, c'est-à-dire en additionnant deux à deux les coordonnées correspondantes aux vecteurs considérés. Cependant Kintsch (2000, 2001) remarque que le calcul de la somme ne tient pas compte des fonctions différentes jouées par le prédicat et l'argument d'une prédication minimale. En effet, selon la définition de Le Ny (1979, 2002 ; voir aussi François et Denhière, 1997), la prédication consiste essentiellement à dire quelque chose de quelque chose et le prédicat est caractérisé par le fait que dans sa signification même, il comporte des places vides qui attendent d'être remplies par des arguments appropriés (Le Ny & Franquart-Declercq, 2002). Alors que prédicats et arguments ne sont pas interchangeables (propriété de non-réversibilité), le calcul de la somme ne différencie pas les fonctions de prédicat et d'argument. L'interdépendance du prédicat et de ses arguments, c'est-à-dire les propriétés de non-réversibilité et de contrainte différente ont été étudiées par Glucksberg, McGlone et Manfredi (1997). Les auteurs ont expérimentalement montré, que non seulement le prédicat et l'argument d'une proposition métaphorique, telle que "mon avocat est un requin", n'étaient pas réversibles mais également que les contraintes exercées par le prédicat et l'argument pouvaient être différenciées. Cette démonstration s'appuie sur une théorie formulée par les auteurs selon

laquelle le processus de compréhension d'une proposition consiste à attribuer des propriétés du prédicat à l'argument. La prédication est un processus de catégorisation ad hoc (Barsalou, 1983) durant lequel l'argument est pris comme exemplaire de la catégorie définie par le prédicat et résulte de la généralisation de l'extension du prédicat et de l'argument.

C'est dans ce cadre que l'on peut envisager l'utilisation de LSA pour simuler le processus de prédication. Kintsch (2000, 2001) propose un algorithme de prédication consistant à récupérer dans l'espace LSA les plus proches voisins du prédicat qui sont également les plus proches voisins de l'argument de la proposition. Par exemple, considérons les propositions suivantes : VOLER(avion) ; VOLER(oiseau) ; VOLER (bandit). Dans un espace construit à partir d'un corpus appelé "Textenfants.2" et essentiellement constitué de textes pour enfants, de contes et de productions enfantines, on applique l'algorithme de prédication à ces trois exemples (voir Denhière & Lemaire, 2004). Le principe de cet algorithme implique de sélectionner, dans LSA, les plus proches voisins d'un prédicat, ici VOLER, et de conserver les voisins qui sont également les plus proches de l'argument de la proposition. Pour construire la proposition VOLER(avion), on calcule le cosinus entre "avion" et les plus proches voisins de "VOLER". Par exemple, les plus proches voisins de "VOLER" qui sont également les plus proches de "avion" sont "Vol, Ailes, Envole, Plané, et Pilote" (voir Tableau 2).

Insérer TABLEAU II

Les plus proches voisins de "VOLER" étant le plus proches de "oiseau" sont "Vol, Ailes, Envole, Envergure, Plumage" et les plus proches de "bandit" sont " Prend, Popularité, Saisit,

Arrache". La méthode décrite par Kintsch (2001), consiste à construire un réseau auto-inhibiteur de $\underline{n}$ nœuds qui représentent les $\underline{n}$ vecteurs $\underline{v_i}$ simulant le prédicat, l'argument et les voisins. La pondération des liens entre prédicat et voisins, entre argument et voisins et entre prédicat et argument est égale aux cosinus des angles formés par les vecteurs $\underline{v_i}$. La pondération des liens entre voisins est négative et sa valeur absolue est égale à la moyenne des pondérations positives. La valeur initiale d'activation de tous les nœuds est égale à 1. Après stabilisation, on obtient une valeur finale d'activation a_i pour chaque nœud. Le vecteur prédication est alors égal à la combinaison linéaire des vecteurs prédicat, argument et des voisins. On a :

avec P le vecteur prédication, $\underline{v_i}$ les vecteurs prédicat, argument et voisins et $\underline{a_i}$ la valeur finale d'activation du nœud représentant $\underline{v_i}$ dans le réseau.

En utilisant cet algorithme, Kintsch (2000) reproduit la non-réversibilité de la prédication. Par ailleurs, en utilisant le même algorithme de prédication, nous avons réussi à reproduire les résultats obtenus par Glucksberg et al. (1997) permettant de montrer les contraintes différentes exercées par le prédicat et l'argument (Bellissens, Thiesbonenkamp, et Denhière, 2002).

L'algorithme de prédication permet de reproduire les résultats obtenus par Glucksberg et al. (1997) car LSA permet de catégoriser les mots d'un corpus en généralisant les différents contexte d'usage sur la base d'un chevauchement entre leurs extensions et que l'algorithme de prédication donne un rôle prépondérant au prédicat dans la construction de la signification de la prédication. Il est ensuite possible de calculer les distances sémantiques entretenues par un vecteur prédication et les autres éléments de l'espace sémantique. Dans le cadre CI-LSA, ces distances fournissent la pondération du réseau de propositions interconnectées construit par le modèle.

3.2. LA RECUPERATION AUTOMATIQUE D'INFERENCES ET L'EMERGENCE DE MACROPROPOSITIONS

Le couplage de LSA et CI permet de simuler un processus de prédication et d'activation associative des plus proches associés d'une proposition ou d'un concept. Le cadre CI-LSA permet également une avancée considérable dans l'explication de la récupération automatique d'informations non explicitement mentionnées dans un texte qui sont appelées classiquement "inférences" (Kintsch, 1993). La récupération de ce type d'information a pour but l'établissement de la cohérence locale et globale d'un texte lorsque celui-ci ne contient pas explicitement cette information. Sans la base de connaissances fournie par LSA, la cohérence locale entre deux segments de texte est établie lors de la construction de la base de texte à l'aide de règles de liaison par chevauchement d'arguments (Kintsch et van Dijk, 1978). Si aucune règle de cette sorte ne peut s'appliquer, l'établissement de la cohérence locale entre segments de texte s'effectue en contraignant de manière ad hoc le réseau, c'est-à-dire en ajoutant arbitrairement des nœuds permettant de faire le lien entre deux segments. Avec la base de connaissances dans CI-LSA, la liaison entre deux segments est implicite et se produit par chevauchement sémantique entre deux segments : soit ils partagent des éléments activés et directement associés, soit ils comportent des éléments activés qui sont indirectement associés. De la sorte, des informations peuvent émerger dans le réseau du seul fait de leur forte connexité alors même qu'elles ne sont pas explicitement mentionnées dans le texte.

Kintsch (1993) a proposé une classification des inférences applicable à l'étude du rôle des connaissances dans la compréhension de texte. L'auteur part du principe que la notion classique d'inférence dans la compréhension du discours est très difficile à cerner (McKoon & Ratcliff, 1992 ; Caillies, Denhière & Kintsch, 2002). Il propose alors de définir les inférences non pas en fonction de la nature des informations à expliciter mais en fonction des processus

qui génèrent ces explicitations (Guthke, 1991). L'auteur propose de considérer deux facteurs indépendants expliquant les processus inférentiels, un facteur de transformation de l'information : augmentation vs. réduction, et un facteur de contrôle de la transformation : automatique vs. contrôlé.

Les transformations contrôlées de l'information diffèrent des transformations automatiques par le nombre d'indices de récupération que contient une information en cours de traitement. Plus un segment de texte en cours de traitement contient d'indices, plus un traitement automatique est possible. Le processus automatique d'augmentation d'information peut être compris comme le résultat de la récupération automatique d'information en mémoire à long terme (Guthke, 1991). La réduction automatique d'information peut être décrite comme la disparition progressive de certains éléments moins importants pour la compréhension d'un texte ou comme une généralisation sémantique d'information qui se produit dès que nécessaire à la lecture d'un texte (Mross, 1989). Dans le cadre CI-LSA, ces transformations automatiques de l'information s'envisagent aisément.

La généralisation d'information peut être considérée comme une réduction du nombre d'informations apportées par le texte dans la mesure où, nécessitant un partage de propriétés entre deux informations (O'Reilly et Rudy, 2000), elle catégorisera ces deux informations. En fait, la généralisation des informations apportées par le texte est une abstraction, d'où la nécessité de se doter d'un modèle abstractif. Au cours de la phase de construction, CI-LSA généralise la signification du prédicat et des arguments d'une proposition. Au cours de la phase d'intégration, deux propositions sémantiquement proches, c'est-à-dire partageant un même voisinage associatif, renforceront leurs liens. Au-delà du traitement d'un segment de texte particulier, plusieurs phrases peuvent partager une proposition ou un ensemble de propositions proches dans la base de connaissances. Soit ces segments sont adjacents et les propositions partagées sont transportées d'un cycle de traitement à l'autre en mémoire de

travail à court terme. Soit les segments en question ne sont pas adjacents et dans ce cas, malheureusement, le modèle de prévoit rien. Cependant, on peut envisager dans ce dernier cas, un processus de généralisation de l'ensemble des phrases partageant une ou plusieurs propositions, conduisant à l'émergence d'une structure de macropropositions appartenant à un même modèle de situation. L'intérêt d'un tel processus est qu'à la surface du réseau associatif, le modèle de situation émergeant pourrait jouer le rôle de structure stable d'indices de récupération, activables chaque fois qu'un segment de texte comporterait une information associée, c'est-à-dire entrant en cohérence avec ce que le traitement du texte a permis de comprendre. Une telle structure de récupération formerait la base d'une mémoire de travail à long terme et serait dépendante à la fois de l'organisation des connaissances dans LSA et des cycles d'activation du réseau (voir Bellissens, 2002).

4. CONCLUSION

L'approche vectorielle des représentations considère la mémoire sémantique comme un espace à plusieurs dimensions dans lequel un mot est représenté par un vecteur. La représentation de la signification est donc distribuée sur les éléments du vecteur et ce sont les coordonnées du vecteur qui déterminent la signification d'un mot. Ces modèles construisent des espaces dont l'organisation permet de prédire des effets d'amorçage. Au-delà de ce type de prédictions, ces modèles peuvent être utilisés dans des simulations de plus grande envergure en les associant à des modèles de traitement. Envisagée ainsi, la nature des représentations est analogue à celle qui est supposée dans les modèles connexionnistes distribués pour lesquels la signification d'un mot correspond à un patron d'activation d'unités élémentaires qui sont des traits sémantiques (voir Masson, 1995 ; McClelland et Kawamoto, 1986 ; McRae et al., 1997). En effet, un ensemble de coordonnées dans un espace à n

dimensions est l'équivalent d'un patron d'activation sur un ensemble de n unités dans un réseau connexionniste. Autrement dit, les éléments du vecteur correspondent aux traits sémantiques utilisés dans la plupart des autres modèles distribués de la représentation de la signification. Cependant, contrairement à la plupart de ces modèles, les dimensions des espaces sémantiques de LSA et de HAL ne sont pas définies par le chercheur ou à l'aide de réponses empiriques et, de manière plus générale, elles ne sont pas déterminées par les contraintes exercées par le chercheur sur le modèle (voir par exemple McRae et al., 1997). Ce point de vue appelle deux réserves.

Premièrement, pour réellement considérer que la signification soit représentée de manière distribuée, il faudrait que la signification associée à un mot émerge dans un réseau d'unités distribuées. Or ce n'est pas le cas. Que ce soit dans HAL ou dans LSA, un vecteur représente la signification potentiellement associée à un mot et dans les applications de ces modèles, une représentation sémantique émerge dans un contexte constitué d'unités qui représentent chacune un vecteur particulier (voir par exemple, Kintsch, 1998). Cet état de fait semble être une nécessité pour le fonctionnement de ces modèles. Très clairement, les espaces sémantiques résultent de l'analyse statistique de la distribution d'unités localisées dans un environnement textuel donné. On considèrera donc que les modèles HAL et LSA sont des modèles hybrides de la représentation sémantique.

La seconde réserve concerne le fait que les modèles connexionnistes distribués, et particulièrement les modèles basés sur une architecture de Hopfield, ne peuvent être entraînés à partir d'un corpus composé d'un vocabulaire trop étendu. En effet, si un modèle connexionniste est confronté à de nombreux mots différents lors de la phase d'apprentissage, de nombreux bassins d'attraction ne correspondant à aucun mot se forment et le réseau ne parvient pas à identifier correctement un mot (Masson, 1995). À l'inverse, il n'existe aucune limite théorique au nombre d'occurrences et de mots différents auxquels peuvent être

confrontés les modèles HAL et LSA. Ces modèles statistiques peuvent construire des représentations à partir d'un environnement langagier quantitativement et qualitativement proche de celui auquel ont été confrontés les individus d'un âge donné. Cette caractéristique leur confère une supériorité indéniable par rapport aux autres modèles de la mémoire sémantique.

Les modèles LSA et HAL développent leurs connaissances à partir de la distribution des mots dans un corpus textuel. Certains auteurs (Glenberg et Robertson, 2000) n'acceptent pas la validité psychologique de cette approche et considèrent que ces modèles posent le problème de l'enracinement du symbole (Harnard, 1990, voir également Barsalou, 1999 ; Glenberg et Robertson, 2000). Selon ces auteurs, un enfant n'ayant jamais été baigné dans la langue ne pourra jamais comprendre les mots d'un dictionnaire, même s'il le parcourt longuement. Il pourra par exemple savoir qu'il existe une relation entre les symboles "banane" et "fruit" parce que le mot "fruit" se trouve dans la définition du mot "banane". Cependant, il sera incapable de donner une signification à ces symboles, du fait que les mots sont caractérisés par la relation arbitraire entre le signifiant et le signifié. En l'absence de quelques associations formées par exemple entre des connaissances perceptives et les symboles, ces derniers ne pourront jamais acquérir une signification pour l'enfant. De telles associations étant nécessaires pour constituer un point de départ dans l'acquisition du langage, une représentation exhaustive et/ou générale des connaissances ne peut reposer exclusivement sur un modèle de type LSA ou HAL dans lequel les connaissances sont acquises à partir de relations entretenues par des symboles arbitraires et des environnements constitués de symboles arbitraires.

Cette limite est explicitement acceptée par certains tenants des modèles vectoriels (voir Burgess, 2000 ; Kintsch, 2002). Cependant, HAL et plus particulièrement LSA ont le mérite de mettre en avant un mécanisme assez général de l'acquisition de la signification des mots qui

est la généralisation ou la catégorisation. Il serait donc intéressant dans une recherche future de simuler directement le développement des connaissances en appliquant une stratification cumulative à un corpus constitué de textes à destination des enfants de niveaux scolaires différents et d'étudier l'effet de cette manipulation sur l'organisation de l'espace sémantique (Denhière, Lemaire, Bellissens, Jhean, à paraître).

BIBLIOGRAPHIE

ALBRECHT, J.E. & MYERS, J.L. - (1998) Accessing distant text information during reading: Effects of Contextual cues. Discourse Processes, 26, 87-107.

BARSALOU L.W. - (1999) Perceptual symbol systems, Behavioral and Brain Sciences, 22, 577-60.

BARSALOU L.W. - (1983) Ad hoc categories. Memory & Cognition, 11, 211-227.

BELLISSENS C. - (2002) Mémoire de travail à long terme et compréhension de texte : Expérimentations et simulations, Thèse de doctorat, Octobre 2002, Aix-en-Provence.

BELLISSENS C., THIESBONENKAMP J., DENHIÈRE G. - (2002) Property attribution in metaphor comprehension: simulations of topic and vehicle contribution within the LSA-CI-framework, Poster presented at the 12th annual meeting of the Society for the Text and Discourse, Chicago.

BELLISSENS, C., & DENHIÈRE, G. - (2002) Word order or environment sharing: A comparison of two semantic memory models. Current Psychology Letter, 9, 43-60.

BURGESS C. - (2000) Theory and operational definitions in computational memory models: A response to Glenberg and Robertson, Journal of Memory and Language, 43, 402-408.

BURGESS C., LIVESAY K., LUND K. - (1998) Explorations in context space: Words, sentences, discourse, Discourse Processes, 25, 211-257.

BURGESS C., LUND K. - (1997) Modeling parsing constraints with high-dimensional context space, Language and Cognitive Processes, 12, 177-210.

CAILLIES S., & DENHIÈRE G. - (2001) The interaction between textual structures and prior knowledge: Hypotheses, data and simulations, European Journal of Psychology of Education, 16, 1, 17-31.

CAILLIES S., DENHIÈRE G., & JHEAN-LAROSE, S. - (1999) The intermediate effect: Interaction between prior knowledge and text structure. In H. van Oostendorp & S. R. Goldman

(Eds), The construction of mental representation during reading (pp. 151-168). Mahwah, NJ: Lawrence Erlbaum Associates.

CAILLIES S., DENHIÈRE G., KINTSCH W. - (2002) The effect of prior knowledge on understanding from text: Evidence from primed recognition, European Journal of Cognitive Psychology, 14, 267-286.

DENHIERE, G. & BAUDET, S. - (1992). Lecture, compréhension de texte et science cognitive. Paris : Presses Universitaires de France.

DENHIERE, G., LEMAIRE, B. - (2004) A computational model of a child semantic memory. 26th Annual Meeting of the Cognitive Science Society, Chicago, August 5-8.

DENHIERE, G., LEMAIRE, B., BELLISSENS, C., & JHEAN, S. - (à paraître) A semantic space for modeling a child semantic memory. In W. Kintsch, & T., Landauer, (Eds.), LSA: A road to meaning. Mahwah, NJ: Lawrence Erlbaum Associates.

EICH-METCALFE J. - (1985) Levels of processing, encoding specificity, elaboration, and CHARM, Psychological Review, 91, 1-38.

ERICSSON K.A., KINTSCH W. - (1995) Long-term working memory, Psychological Review, 102, 211-245.

FRANÇOIS J., DENHIERE, G. - (1997) Sémantique Linguistique et Psychologie Cognitive, Aspects Théoriques et Expérimentaux, Grenoble : Presses Universitaires de Grenoble.

GLENBERG A. M., ROBERTSON D.A. - (2000) Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning, Journal of Memory and Language, 43, 379-401.

GLUCKSBERG S., MCGLONE M.S., MANFREDI D.A. - (1997) Property attribution in metaphor comprehension, Journal of Memory and Language, 36, 50-67.

GUTHKE T. - (1991) Psychologische Untersuchungen zu inferenzen beim satz- and textverstehn, Dissertation doctorale non publiée, Humboldt Universität, Berlin.

HARNARD S. - (1990) The symbol grounding problem, Physica D, 42, 335-46.

HINTZMAN D.L. - (1986) "Schema abstraction" in a multiple-trace memory model, Psychological Review, 93, 411-428.

KINTSCH W. - (1988) The use of knowledge in discourse processing: A construction-integration model, Psychological Review, 95, 163-182.

KINTSCH W. - (1993) Information accretion and reduction in text processing: Inferences, Discourses Processes, 16, 193-202.

KINTSCH W. - (1998) Comprehension: A paradigm for cognition, Cambridge, University Press.

KINTSCH W. - (2000) Metaphor comprehension: A computational theory, Psychonomic Bulletin and Review, 7, 257-266

KINTSCH W. - (2001) Predication, Cognitive Science, 25, 173-202

KINTSCH W. - (2002) On the notions of theme and topic in psychological process models of text comprehension, in M. LOUWERSE et W. VAN PEER (Edit.), Thematics: Interdisciplinary Studies, Amsterdam, Benjamins, 157-170

KINTSCH W., PATEL V.L., ERICSSON K.A. - (1999) The role of long-term working memory in text comprehension, Psychologia, 42, 186-198.

KINTSCH W., VAN DIJK T.A. - (1978) Toward a model of text comprehension and production, Psychological Review, 85, 363-394.

LANDAUER T.K., DUMAIS S.T. - (1997) A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, Psychological Review, 104, 211-240.

LE NY J.F. - (2002) Prédicat. In G. Tiberghien, H. Abdi, J-P. Desclès, N. Georgieff, M. Jeannerod, J-F. Le Ny, P. Livet, J. Pynte, & G. Sabah (Eds.) Dictionnaire des sciences cognitives, Paris, Armand Colin.

LE NY J.F. - (1979) La Sémantique Psychologique, Paris, Presses Universitaires de France.

Le Ny J.F. & Franquart-Declercq, C.- (2002) Signification des verbes, relations verbe/patient et congruence sémantique. Le Langage et l'Homme, vol. XXXVII, N° 2, 9-26.

LUND K., BURGESS C. - (1996) Producing high-dimensional semantic spaces from lexical co-occurrence, Behavior Research Methods, Instrumentation, and Computers, 28, 203-208.

MASSON M.E.J. - (1995) A distributed memory model of semantic priming, Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 3-23.

MCCLELLAND J.L., KAWAMOTO A.H. - (1986) Mechanisms of sentence processing: Assigning roles to constituents, in D.E. RUMELHART, J.L. MCCLELLAND, and the PDP RESEARCH GROUP (Edit.), Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 2, Cambridge, MIT Press, 272-325.

MCDONALD S., LOWE W. - (1998) Modelling Functional Priming and the Associative Boost, Proceedings of the 20th Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, 675-680.

MCKOON G., RATCLIFF R. - (1992) Inference during reading, Psychological Review, 99, 440-466.

MCRAE K., DE SA V.R., SEIDENBERG M.S. - (1997) On the nature and scope of featural representations of word meaning, Journal of Experimental Psychology: General, 126, 99-130.

MOSS H.E., OSTRIN R.K., TYLER L.K., MARSLEN-WILSON, W.D. - (1995) Accessing different types of lexical semantic information: Evidence for priming, Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 863-883.

MROSS E.F. - (1989) Macroprocessing in expository text comprehension, Dissertation doctorale non publiée, University of Colorado, Boulder.

MURDOCK B.B. - (1993) TODAM2: A model for the storage and retrieval of item, associative, and serial-order information, Psychological Review, 100, 183-203.

MYERS J.L., O'BRIEN E.J. - (1998) Accessing the discourse representation during reading, Discourse Processes, 26, 131-157.

NOORDMAN, L.G.M., & VONK, W. - (1998) Memory-based processing in understanding causal information. Discourse Processes, 26, 191-212

O'REILLY R.C., RUDY J.W. - (2000) Computational principles of learning in the néocortex and hippocampus, Hippocampus, 10, 389-397.

OSGOOD C.E., SUCI G.J., TANNENBAUM P.H. - (1957) The measurement of meaning, Urbana, University of Illinois Press.

RAAIJMAKERS, J. G. W., & SHIFFRIN, R. M. - (1980) SAM: A theory of probabilistic search of associative memory. In Bower, G. H. (Ed.), The Psychology of Learning and Motivation, Vol. 14, 207-262. New York: Academic Press.

SANFORD, A.J. & GARROD, S.C. - (1998) The role of scenario mapping in text comprehension, Discourse Processes, 26, 159-190.

STEYVERS, M., SHIFFRIN, R.M., & NELSON, D.L. - (in press) Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. Healy (Ed.), Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer. Washington, DC: American Psychological Association.

TIBERGHIEN G. - (1997) La mémoire oubliée, Liège : Mardaga.

TIBERGHIEN, G. - (2002) Abstractif (modèles) In G. Tiberghien, H. Abdi, J-P. Desclès, N. Georgieff, M. Jeannerod, J-F. Le Ny, P. Livet, J. Pynte, & G. Sabah (Eds.), Dictionnaire des Sciences Cognitives. Paris : Armand Colin.

VAN DIJK, T.A., & KINTSCH, W. - (1983) Strategies of discourse comprehension. New York: Academic Press.

TABLEAU I. - Matrice obtenue en appliquant une fenêtre de co-occurrence de 5 termes à la phrase "Le chat dort sur le canapé".

Matrix for the sentence "Le chat dort sur le canapé" applying a 5-words co-occurrence window.

	canapé	chat	dort	le	sur
canapé	0	2	3	6	4
chat	0	0	0	5	0
dort	0	5	0	4	0
le	0	3	4	2	5
sur	0	4	5	3	2

TABLEAU II.- Exemple d'application de l'algorithme de prédication à trois propositions comportant le même prédicat. (1: cosinus entre prédicat et voisins ; 2: cosinus entre argument et voisins)

Application example of the predication algorithm to three propositions with the same predicate. (1: cosine between predicate et neighbours ; 2: cosine between argument et neighbours)

Proposition	VOLER(avion)	1	2	VOLER(oiseau)	1	2	VOLER(bandit)	1	2
Voisin 1	Vol	.67	.56	Vol	.67	.20	Prend	.18	.30
Voisin 2	Ailes	.60	.21	Ailes	.60	.22	Attaque	.16	.20
Voisin 3	Envol	.52	.30	Envol	.52	.28	Popularité	.15	.21
Voisin 4	Plané	.51	.30	Envergure	.51	.26	Saisit	.14	.23
Voisin 5	Pilote	.51	.88	Plumage	.50	.39	Arrache	.14	.32

ANNEXE –

Les bases de données Françaises sur le site http://lsa.colorado.edu

Dans l'article de Bellissens et Denhière (2002) relaté dans cet article, le matériel de Moss et al. (1995) a été traduit afin de comparer les résultats obtenus par LSA en Anglais aux résultats de LSA en Français. Nous avons constaté que les résultats avec le matériel traduit sont identiques aux résultats obtenus avec le matériel Anglais. Ceci est conforme à l'hypothèse selon laquelle les vecteurs de LSA résultent d'un apprentissage par généralisation de l'usage des mots dans un large corpus. On suppose donc que, d'une langue donnée à une autre, à la condition que ces deux langues soient ancrées dans des cultures proches, les représentations les plus abstraites sont implicitement communes à différents corpus si ceux-ci sont assez larges. De la même manière, les structures mnésiques sous-jacentes à la généralisation des informations encodées sont sans doute analogues d'un individu à l'autre que celui-ci parle une langue ou une autre. Pour appréhender ce type d'hypothèse, produire des LSA en langue Française ou en d'autres langues devient nécessaire. Dans cette section sont donc présentées les caractéristiques statistiques de l'adaptation Française de LSA. Ces corpus ont fait l'objet d'une analyse LSA dont le résultat est consultable sur le site http://lsa.colorado.edu. Les bases de données présentes sur le site sont les suivantes :

Français-Livres (300 facteurs) :

Cette base de données est constituée de textes provenant de différentes sources : les bases électroniques GALLICA et ABU qui regroupent des textes littéraires et scientifiques du XVIIIe et XIXe siècles et des textes littéraires contemporains.

Français-Monde (300 facteurs) :

Cette base de données comporte l'intégralité des articles de presses parus dans le quotidien Le Monde au cours de l'année 1993 ainsi que l'ensemble des articles contenu dans l'encyclopédie historique de la guerre 39-45 éditée par Le Monde.

Français-Total (300 facteurs) :

Cette base regroupe les deux précédentes. Les caractéristiques statistiques des bases de données sont présentées dans le tableau suivant :

	Pages	Occurrences	Caractères	Paragraphes
Monde 1993	12 380	8 078 441	42 998 560	20 114
Monde 39-95	7 594	4 880 669	25 993 542	12 228
Gallica	1 859	1 411 778	6 200 704	1 307
ABU	5 085	3 920 255	21 341 878	11 025
Divers	602	416 548	2 094 213	2 748
Total	28 558	19 480 405	102 117 693	49 893

Les titres des articles et des rubriques ainsi que les sous-titres ont été supprimés des bases.

TABLEAU II.- Exemple d'application de l'algorithme de prédication à trois propositions comportant le même prédicat. (1: cosinus entre prédicat et voisins ; 2: cosinus entre argument et voisins).

Application example of the predication algorithm to three propositions with the same predicate. (1: cosine between predicate et neighbours ; 2: cosine between argument et neighbours)

Proposition	VOLER(avion)	1	2	VOLER(oiseau)	1	2	VOLER(bandit)	1	2
Voisin 1	Vol	.67	.56	Vol	.67	.20	Prend	.18	.30
Voisin 2	Ailes	.60	.21	Ailes	.60	.22	Attaque	.16	.20
Voisin 3	Envol	.52	.30	Envol	.52	.28	Popularité	.15	.21
Voisin 4	Plané	.51	.30	Envergure	.51	.26	Saisit	.14	.23
Voisin 5	Pilote	.51	.88	Plumage	.50	.39	Arrache	.14	.32