

D'une hiérarchie figée de concepts vers une hiérarchie évolutive de notions de base

Alain Joubert, Mathieu Lafourcade

¹LIRMM – UM2, 161, rue Ada – 34392 Montpellier Cédex 5 – France

{joubert, lafourcade}@lirmm.fr

Abstract

In the construction of a thematic representation system of knowledge in Natural Language Processing, it became current to use a thesaurus. Essentially, a thesaurus consists of a set of concepts, often organized in an arborescent structure : it is a fundamental, but relatively fixed tool. But concepts (and language!) are evolving, and this, more and more quickly, particularly in technical fields. We propose a system which makes possible for the notion of concept to evolve by the introduction of the "Basic Notions". Those, necessarily definite on the vector space of the concepts of the thesaurus, constitute another generating system of the space of thematic representation of knowledge. Contrary to the concepts of the thesaurus, the basic notions evolve progressively with the analysis of new texts. We discuss the optimal value of the dimension of the space of representation generated by the basic notions, then of the determination of the acceptations allowing to express them. Lastly, we consider the differentiation between basic notions of general space and those of a specialized field.

Key words : Natural Language Processing, conceptual vectors, basic notions, thesaurus, thematic distance

Résumé

Dans la construction d'un système de représentation thématique des connaissances en Traitement Automatique du Langage Naturel, il est devenu courant d'utiliser un thésaurus. Par essence, un thésaurus est constitué d'un ensemble de concepts, souvent organisé en une structuration arborescente : c'est un instrument fondamental, mais relativement figé. Or les notions (et la langue !) évoluent, et ce, de plus en plus rapidement, en particulier dans les domaines techniques. Nous proposons un système qui permet de faire évoluer la notion de concept par l'introduction des « notions de base ». Celles-ci, définies nécessairement sur l'espace vectoriel des concepts du thésaurus, constituent un autre système générateur de l'espace de représentation thématique des connaissances. Contrairement aux concepts du thésaurus, les notions de base évoluent au fur et à mesure de l'analyse de nouveaux textes. Nous discutons de la valeur optimale de la dimension de l'espace de représentation généré par les notions de base, puis de la détermination des acceptions permettant de les exprimer. Enfin, nous envisageons la différenciation entre notions de base de l'espace généraliste et celles d'un domaine spécialisé.

Mots clés : Traitement Automatique du Langage Naturel, vecteurs conceptuels, notions de base, thésaurus, distance thématique

1. Introduction de la problématique

Notre système de représentation thématique des connaissances générales (Lafourcade et Sandford, 1999) est fondé sur une représentation vectorielle qui repose initialement sur le thésaurus Larousse (Larousse, 1999). La modélisation par un espace vectoriel est utilisée depuis longtemps en Recherche d'Informations, par exemple (Salton et MacGill, 1983). Le thésaurus utilisé ici qui possède une structure arborescente, hiérarchisée en 4 niveaux plus la racine, comporte 873 concepts feuilles. Ces concepts forment un système générateur de

l'espace vectoriel noté C873, censé modéliser notre représentation thématique¹ des connaissances générales (par opposition aux connaissances spécifiques restreintes à un domaine particulier). Cette approche vectorielle, s'appuyant sur un ensemble de concepts prédéterminé, est celle préconisée par Chauché (1990).

Les 873 concepts du thésaurus utilisé sont donnés a priori ; constituent-ils le « meilleur » système générateur ? Il est d'autant plus légitime de se poser cette question que d'autres thésaurus ont été publiés. Le plus ancien est probablement le Roget's Thesaurus (Kipfer, 2001) dont la première édition date du milieu du XIX^e siècle, organisé en une structure arborescente qui compte dans sa version la plus récente 1075 concepts feuilles regroupés en 15 classes. Ce thésaurus a servi de base à différents travaux, par exemple Yarowsky (1992).

Il est à remarquer qu'il est possible de trouver des relations entre certains des vecteurs de base de l'espace C873 : par exemple, entre les concepts 1_EXISTENCE et 2_INEXISTENCE existe une relation d'antonymie. Toutefois, sur C873, les vecteurs représentant ces deux concepts sont orthogonaux (ce sont des vecteurs de base !) et la décomposition d'une acception quelconque reste unique dans le thésaurus considéré. Cette interdépendance entre concepts a parfois été exploitée, comme par exemple dans le modèle LSA (Deerwester et al. 1990).

De manière à la fois intuitive et apprise, chaque individu humain possède son propre système générateur qui lui permet de se faire sa propre représentation des connaissances ; de plus, ce système individuel évolue dans le temps. On peut considérer que le système générateur du thésaurus Larousse correspond en fait à celui d'un individu type. Le système générateur individuel ne correspond vraisemblablement jamais à celui du thésaurus utilisé (n'est-ce pas aussi ce qui forme notre liberté de pensée ?).

En considérant uniquement les connaissances générales, c'est-à-dire en faisant abstraction de toute spécialisation (la question sera abordée au chapitre V), peut-on trouver un « meilleur » système générateur de l'espace vectoriel modélisant la représentation thématique des connaissances que celui défini par les concepts du thésaurus utilisé ?

2. Principe de la méthodologie : définition des notions de base

De façon incrémentale, au fur et à mesure de l'analyse de textes, on rencontre des mots. Soit k le nombre de termes différents rencontrés (la plupart des dictionnaires généralistes actuels possèdent environ 80.000 entrées). La grande majorité des termes étant polysémiques, ces k termes différents possèdent k' acceptions et génèrent donc k' vecteurs sur l'espace vectoriel C_{873} (dans l'état actuel de nos expérimentations, k' dépasse 400.000) (Schwab et al. 2004). De plus, il est nécessaire de pondérer ces k' vecteurs en fonction de leur fréquence d'apparition dans les textes étudiés. En effet, certaines acceptions de termes se rencontrent beaucoup plus fréquemment que d'autres et elles jouent donc un rôle plus important dans notre connaissance². Il semble naturel que cette fonction de pondération, nécessairement croissante, soit une fonction logarithmique de la fréquence d'apparition ; effectivement, il est nécessaire

¹ Il s'agit d'une représentation *thématique* des connaissances car notre système de représentation vectorielle des concepts ne fera pas d'association, par exemple, entre un personnage possesseur et un objet possédé s'ils n'appartiennent pas au même domaine : le lien entre Tintin et Milou ne pourra se faire qu'au travers de la notion de bande dessinée et des concepts qui lui sont associés. Pour créer un lien direct entre Tintin et Milou, il faudrait tenir compte des co-occurrences de ces deux termes, éventuellement pondérées par les distances les séparant dans les arbres d'analyse morpho-syntaxique des phrases concernées.

² En fonction du corpus de textes étudié, en particulier s'il s'agit de dictionnaires, il faudra toutefois se méfier de termes généraux très fréquemment utilisés dans les libellés des définitions, tels que « action », « partie de » ...

de réaliser un amortissement de l'impact de cette fréquence. Il paraît également raisonnable de pondérer chaque occurrence de ces k' vecteurs en fonction de la profondeur de l'acception correspondante dans l'arbre d'analyse syntaxique du texte étudié ; en effet, il est logique de penser que plus un mot est « perdu » dans les profondeurs d'une phrase, moins il est important pour le sens global de cette phrase. Il semble naturel que cette fonction de pondération, nécessairement décroissante, puisse être une exponentielle négative de la profondeur.

De manière plus formelle, si nous appelons p la profondeur d'un terme t dans l'arbre d'analyse syntaxique du texte étudié, la norme du vecteur v représentant l'acception correspondante (après désambiguïsation entre les éventuelles différentes acceptions de t) sera :

$$\|v\| = e^{-\alpha p} \quad \text{où } \alpha \text{ est un coefficient de pondération.}$$

Sandford (1998) propose que $\|v\| = (1/2)^p$, ce qui est équivalent à l'expression précédente avec $\alpha = \text{Log}(2)$.

La $i^{\text{ème}}$ occurrence de cette acception a du terme t sera représentée sur C_{873} par un vecteur $v_{a,i}$ dont la norme sera :

$$\|v_{a,i}\| = e^{-\alpha p_{a,i}} \quad \text{où } p_{a,i} \text{ désigne la profondeur dans l'arbre d'analyse de la } i^{\text{ème}} \text{ occurrence de l'acception } a.$$

Afin de tenir compte de la fréquence d'apparition des acceptions, il paraît raisonnable d'envisager une sommation (logarithmique) des différents vecteurs représentant chaque occurrence d'une même acception. Ainsi, le vecteur v_a représentant l'acception a sur l'ensemble des textes traités aura pour norme :

$$\|v_a\| = \text{Log} (\sum_i f(\|v_{a,i}\|)),$$

car, comme cela est expliqué plus haut, nous souhaitons que $\|v_a\|$ soit fonction du logarithme de la fréquence d'apparition de l'acception a .

$\|v_a\|$ étant une norme, elle doit en vérifier les propriétés :

$$1^\circ / \|v_a\| \geq 0$$

$$2^\circ / \|v_a\| = 0 \Leftrightarrow v_a \text{ est le vecteur nul}$$

$$3^\circ / \|v_{a,i} + v_{a,j}\| \leq \|v_{a,i}\| + \|v_{a,j}\|$$

$$\text{qui se traduit ici par : } \|v_a\| \leq \sum_i \|v_{a,i}\|.$$

$$\text{S'il n'y a qu'une seule occurrence de l'acception } a, \text{ alors } \|v_a\| = \|v_{a,1}\|.$$

$$\text{S'il y a plusieurs occurrences de l'acception } a, \text{ alors } \|v_a\| < \sum_i \|v_{a,i}\|.$$

Ces différentes conditions conduisent à envisager :

$$\|v_a\| = \text{Log} (\sum_i e^{\|v_{a,i}\|}).$$

Dans un but de simplification, en tenant compte de la proximité thématique, les k' vecteurs v_a obtenus peuvent se regrouper en n nuages, avec $n \ll k'$. Les n vecteurs barycentres de ces n nuages forment un système générateur d'un espace vectoriel B_n . Nous regroupons ainsi les termes thématiquement proches d'un même concept. Bien que la méthode soit différente, notre objectif est à rapprocher de celui développé par Landauer et Dumais (1997) pour la méthode Latent Semantic Analysis (LSA). En considérant un nombre n relativement grand, probablement de plusieurs centaines à quelques milliers, cet espace B_n peut être quasiment confondu avec C_{873} , à condition que les textes étudiés ne se restreignent pas à un domaine spécifique et balayent l'ensemble des connaissances générales.

À quoi correspondent ces n vecteurs ? Ce sont les « notions de base » déduites de l'analyse des textes. La figure 1 explicite le principe de notre méthodologie.

Se pose alors la question de la discrétisation de ces n nuages à partir des k' vecteurs correspondant aux acceptions.

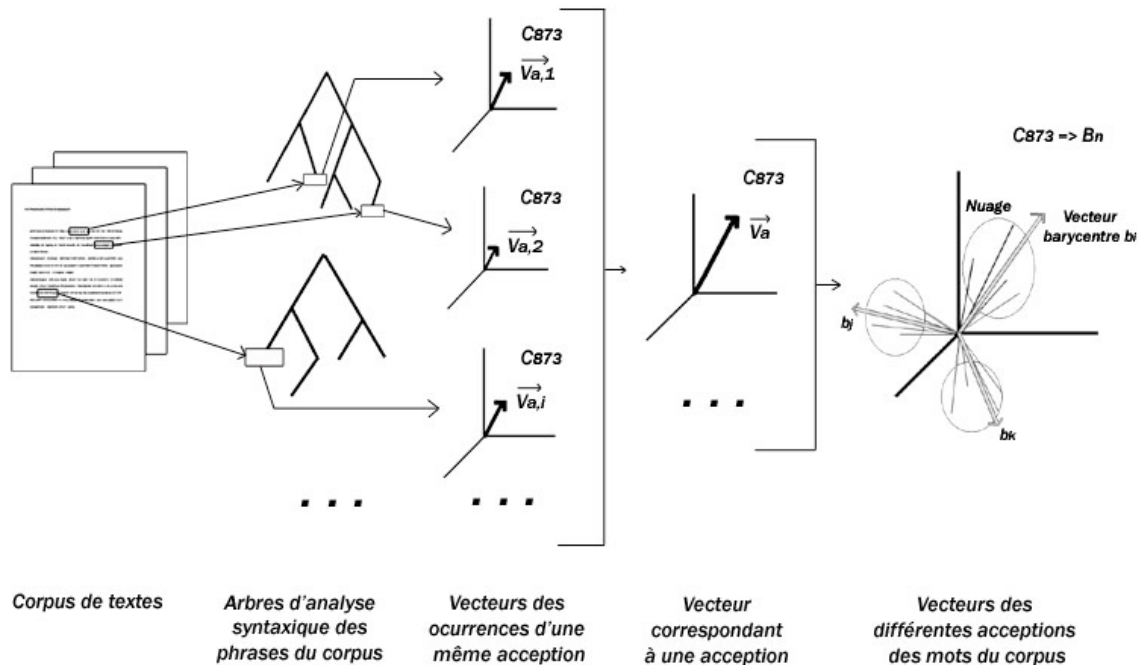


Fig.1 : Ce schéma illustre la méthode utilisée pour définir les notions de base. L'étude des textes permet d'obtenir les arbres d'analyse morpho-syntaxique des phrases les constituant. Cette analyse est réalisée grâce à l'outil SYGFRAN (pour le Français) développé avec SYGMART, présenté initialement par Chauché (1984). Chaque occurrence d'une acception se traduit par un vecteur sur C_{873} dont la norme est fonction de sa profondeur dans l'arbre d'analyse. En sommant (logarithmiquement) les vecteurs des différentes occurrences d'une même acception, on obtient le vecteur correspondant à cette acception sur C_{873} . Les vecteurs des différentes acceptions peuvent se regrouper en nuages dont les barycentres définissent les notions de base.

3. Existe-t-il une valeur optimale pour le nombre de notions de base ?

D'un côté extrême, il est possible de poser $n = k'$ en considérant chaque acception comme un concept ; il serait alors envisageable de tout représenter sans ambiguïté. Mais, d'une part, un espace vectoriel de dimension 400.000, bien que réalisable dans l'état actuel de la technologie des ordinateurs au niveau volume de stockage, pourrait conduire à des temps de traitement difficilement acceptables. D'autre part, est-ce vraiment « sans ambiguïté » ? Le fait que toute acception ne soit « décomposable » que sur un seul concept ne permet pas, en particulier, de mettre en œuvre la notion de distance angulaire entre vecteurs. Entre deux vecteurs concepts-acceptations quelconques, la distance angulaire serait invariablement égale à $\pi/2$. Il est manifeste que cette solution revenant à considérer que les distances entre acceptions sont toutes identiques ne correspond en rien à la réalité : intuitivement, s'il est possible d'établir une relation thématique entre deux acceptions, leur distance semble inférieure à celle qui sépare deux acceptions qui ne possèdent aucun point commun ; par exemple, s'il existe une relation de co-hyponymie (exemple : entre *CHAT* et *CHIEN*) ou une relation

d'hyponymie / hyponymie (exemple : entre *CHAT* et *MAMMIFÈRE*), la distance entre les acceptions est manifestement inférieure à celle qui sépare deux acceptions pour lesquelles aucune relation n'apparaît clairement (exemple : entre *CHAT* et *INEXISTENCE*, dont les champs sémantiques n'ont apparemment que peu de points communs).

Il serait toutefois envisageable de définir une distance ultramétrique entre concepts, comme initialement présentée par Schwab (2001), qui est calculée en fonction du chemin minimal entre les deux concepts considérés dans la structure arborescente du thésaurus. Cette distance pourrait même être améliorée par une pondération tenant compte de l'ordre des concepts dans leurs fratries, dans les cas où un tel ordre représente une information pertinente. Mais il faudrait pour cela disposer d'un thésaurus. Or, en construisant B_n , nous voulons nous affranchir de C_{873} et de l'aspect contraint et figé d'un thésaurus fixe défini *a priori*. La solution pourrait consister à considérer que l'espace B_n n'est pas homogène et donc à découper l'espace B_n en régions. Chacun des n vecteurs appartenant à une région, la distance entre deux vecteurs dépendrait de leur appartenance ou non à la même région. Il pourrait même être envisagé de regrouper certaines régions en super-régions³ ; seule l'expérience pourrait montrer si une telle structuration en plusieurs niveaux se manifeste clairement.

À l'opposé, il est manifeste qu'une valeur trop faible du nombre des concepts limite de façon draconienne la discrétisation. Par exemple, en remontant simplement d'un niveau dans la structure hiérarchique du thésaurus Larousse, et en ne considérant donc que les 95 concepts de niveau 3 (les 873 concepts générateurs de C_{873} sont de niveau 4), tous les animaux sont regroupés sous le concept *C3:ANIMAUX*, et il serait alors difficile de les discriminer, même en considérant leurs caractéristiques ou leurs comportements. Une trop grande simplification des concepts, et donc une réduction trop importante de leur nombre, conduit à un maillage beaucoup trop lâche de l'espace, généraliste ou spécialisé, que l'on souhaite modéliser⁴.

Cette solution qui conduit à une représentation thématique trop imprécise dans le cas d'une analyse fine, pourrait toutefois être utilisée en classification de textes. En effet, pour une phrase ou un paragraphe donné, elle permet de manière rapide de déterminer le domaine thématique concerné.

Seule l'expérience peut donner un ordre de grandeur du nombre optimal de notions de base. Dans le cadre des connaissances générales, il semble logique, au vu des thésaurus existants, que cet ordre de grandeur soit d'environ un millier. Les travaux menés par (Lafourcade et al. 2002) semblent montrer qu'un nombre de quelques milliers conduirait, sans trop alourdir le système, à de meilleurs résultats.

4. Comment exprimer les notions de base ?

Au fur et à mesure de l'analyse de nouveaux textes, les nuages de vecteurs évoluent : c'est « l'évolution des notions », avec éventuellement des phénomènes de différenciation ou de regroupement. Ceci est à opposer à la définition des 873 concepts de base qui sont immuables. Il est tout de même à remarquer que ces notions de base évolutives sont exprimées en fonction des 873 concepts fixes (n'est-il pas rassurant d'avoir des repères fixes, tout en ayant la possibilité d'évoluer ?).

³ À titre d'illustration, ceci pourrait être comparé à la structure spatiale de l'Univers. Les galaxies se regroupent en amas de galaxies, eux-mêmes se regroupant en super-amas : notre Galaxie appartient à l'Amas Local, lui-même compris dans le super-amas de la Vierge.

⁴ En simplifiant à l'extrême, ne rencontrerait-on pas un monde manichéen dans lequel ne subsisteraient probablement que deux notions : le bien et le mal ? ... ou le spirituel et le matériel ? ...

Cet espace B_n qui dépend fortement des textes rencontrés correspond en fait à l'espace vectoriel de représentation thématique des connaissances pour un individu. Chacun d'entre nous possède ses propres références et ses propres définitions : les espaces B_n , ainsi que leurs systèmes générateurs, des différents individus se ressemblent, mais ne sont pas nécessairement absolument identiques.

Pour chacune de ces n notions de base, notée b_i , il est possible de trouver le terme le plus proche, en utilisant la notion de distance définie sur C_{873} . Sous la condition qu'il existe un terme « suffisamment » proche de b_i , celui-ci exprimera cette notion de base. En fait, il s'agit non pas de trouver le terme le plus proche, mais l'acceptation de terme la plus proche de b_i . Il peut se poser alors la question de son expression afin d'éviter tout risque d'ambiguïté. De plus, il est indispensable de prendre en compte la fréquence de l'acceptation candidate pour exprimer la notion de base : une acceptation trop peu fréquente peut-elle être un « bon » concept ? (par exemple, n'est-elle pas trop désuète ? ou trop spécialisée ?) L'acceptation la plus fréquente, ou plus exactement celle dont le vecteur a la norme la plus grande dans le nuage (voir la fig.1), ne constitue pas nécessairement le « meilleur » concept, si elle est relativement éloignée du barycentre du nuage qu'elle est censée représenter.

Mais alors, comment avoir une idée de la taille du domaine concerné par b_i ? Cette différenciation b_i - b_j dépend majoritairement de la valeur de n et donc de la définition des nuages, c'est-à-dire des paramètres de la discrétisation des nuages. Comme nous l'avons vu plus haut, il est bien évident que plus on voudra de précision, plus les nuages auront une étendue réduite, et plus ils seront nombreux : plus les notions de bases seront fines et précises, plus elles seront nombreuses (c'est le classique compromis précision-simplicité). Si les nuages sont nombreux, chacun d'eux aura un poids relativement faible ; ils seront donc susceptibles d'évoluer plus rapidement lors de l'analyse de nouveaux textes (c'est le compromis précision-stabilité). En conséquence, plus les notions de base seront nombreuses, plus il risque d'être difficile de trouver des termes non ambigus « suffisamment » proches pour les exprimer.

5. Notions de base généralistes vs notions de base spécialisées

Il paraît manifeste que, même en considérant un système générateur de B_n comportant plusieurs milliers de notions de base, il serait délicat de discriminer des termes qui d'un point de vue généraliste sont très voisins, même si dans un domaine spécialisé ils possèdent des significations différentes. Il est indispensable de disposer d'un thésaurus spécialisé sur le domaine concerné. Ceci a été montré par Lafourcade et al. (2002).

Il semble illusoire d'envisager que l'on puisse disposer d'un thésaurus qui serait spécialisé sur tous les domaines⁵. Il faudra donc considérer soit que l'on reste sur une représentation thématique des connaissances générales, et alors la discrétisation des nuages sera "homogène", soit que l'on se spécialise sur un domaine particulier. Dans ce dernier cas, Lafourcade et al. (2002) ont montré que l'on ne pouvait pas faire l'économie d'un thésaurus général : il faudra donc conserver les n notions de base généralistes auxquelles viendront s'ajouter les notions de base spécialisées. Cela signifie qu'il faudra cerner, dans l'espace général, le domaine spécialisé. Ce discernement peut se faire en fonction de la fréquence

⁵ Pour s'en convaincre, il suffit de considérer la classification décimale de Dewey, utilisée en recherche thématique documentaire, dont la structure arborescente possède 5 à 6 niveaux (en plus de la racine), ce qui représente donc approximativement 10^5 à 10^6 notions différentes. Ce nombre est à comparer aux 400.000 acceptations rencontrées dans les dictionnaires.

observée des acceptions : plus une acception est fréquente, plus grande est sa probabilité d'appartenir au domaine spécialisé considéré. La norme de chacun des vecteurs dépendant de la fréquence d'apparition, celle des vecteurs du domaine spécialisé sera nécessairement plus grande que celle des vecteurs du domaine généraliste. De plus, en raison de leur proximité thématique (ne serait-ce que par le domaine concerné !), les acceptions du domaine spécialisé conduiront à des vecteurs qui seront plus proches les uns des autres que ceux du domaine généraliste.

La discrétisation des nuages devra donc se faire en fonction de leur taille, de leur densité spatiale, mais également en fonction de la fréquence dans le corpus étudié des vecteurs les constituant. Dans le domaine spécialisé, les vecteurs correspondant aux notions de base seront plus proches les uns des autres, au sens de la définition de la distance angulaire sur C_{873} , que les vecteurs correspondant aux notions de base généralistes : le maillage sur le domaine spécialisé est beaucoup plus fin que sur le domaine généraliste.

La figure 2 illustre cette différenciation entre le domaine spécialisé et le domaine généraliste.

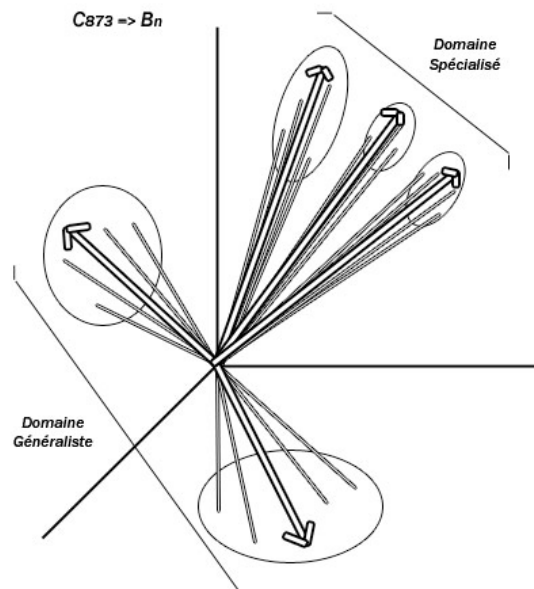


Fig.2 : Illustration de la différence entre domaine spécialisé et domaine généraliste

Dans la région de l'espace correspondant au domaine spécialisé, les nuages de vecteurs ont une étendue plus réduite, ils sont plus denses et la norme des vecteurs les constituant est plus importante que celle des vecteurs situés dans le domaine généraliste.

6. Conclusion

Quel que soit le thésaurus utilisé pour représenter thématiquement des connaissances, il est défini *a priori* : par essence, il possède une structure fixe et l'évolution des concepts qui le constituent ne peut se faire que par un processus modificatif relativement « lourd ». Un système de représentation thématique reposant uniquement sur un thésaurus est par conséquent figé. Or l'évolution du langage et l'apparition de nouvelles notions sont des phénomènes de plus en plus fluctuants. Il est donc indispensable de disposer d'un système qui puisse évoluer facilement : c'est le cas avec les notions de base dont la construction s'appuie toutefois sur l'ossature que constituent les concepts. En effet, au fur et à mesure de l'analyse

de nouveaux textes, il y a évolution des vecteurs correspondant aux mots rencontrés ; les nuages de vecteurs sont alors modifiés dans leurs formes et éventuellement dans leur nombre, ce qui conduit en conséquence à l'évolution des vecteurs représentant les notions de base.

Références

- Chauché J. (1984). « Un outil multidimensionnel de l'analyse du discours ». *Proceedings of the 22nd conference on Association for Computational Linguistics*, Stanford California : 11-15.
- Chauché J. (1990). « Détermination sémantique en analyse structurale : une expérience basée sur une définition de distance ». *TA Information*, vol. 31, (1) : 17-24.
- Deerwester S., Dumais S., Landauer T., Fumas G., Harshman R. (1990). « Indexing by latent semantic analysis ». *Journal of the American Society of Information Science*, 416 (6) : 391-407.
- Kipfer B.A. (2001, [1852]). *Roget's International Thesaurus*. sixth edition, Harper Resource.
- Lafourcade M., Sandford E. (1999). « Analyse et désambiguïsation lexicale par les vecteurs sémantiques », TALN'1999, Cargèse, France : 351-356.
- Lafourcade M., Prince V., Schwab D. (2002). « Vecteurs Conceptuels et Structuration émergente des Terminologies ». *Traitement Algorithmique des Langues*, vol. 43, (1) : 43-72.
- Landauer T., Dumais S. (1997). « A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction and representation of knowledge ». *Psychological Review*, 104(2) : 211-240.
- Larousse (1999). *Thésaurus Larousse – des idées aux mots, des mots aux idées*. Larousse.
- Salton G., MacGill M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
- Sandford E. (1998). « Augmentation lexicale sémantique et désambiguïsation lexicale sémantique : application à la traduction automatique du français vers le tahitien ». Thèse de doctorat, Université de Montpellier II.
- Schwab D. (2001). « Vecteurs conceptuels et fonctions lexicales : application à l'antonymie ». Mémoire de DEA, Université de Montpellier II.
- Schwab D., Lafourcade M., Prince V. (2004). « Hypothèses pour la construction et l'exploitation conjointe d'une base lexicale sémantique basée sur les vecteurs conceptuels ». *JADT 2004*, Louvins-la-Neuve, Belgique.
- Yarowsky D. (1992). « Word-sense disambiguation using statistical models of Roget's categories trained on large corpora ». *COLING'92*, Nantes : 454-460.