# Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux

Didier Schwab, Lim Lian Tze, Mathieu Lafourcade

▶ **To cite this version:**

Didier Schwab, Lim Lian Tze, Mathieu Lafourcade. Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux. TALN: Traitement Automatique des Langues Naturelles, Jun 2007, Toulouse, France. pp.293-302. lirmm-00200889

**HAL Id: lirmm-00200889**

**https://hal-lirmm.ccsd.cnrs.fr/lirmm-00200889v1**

Submitted on 21 Dec 2007

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Les vecteurs conceptuels, un outil complémentaire aux réseaux lexicaux.

Didier Schwab[1] Lim Lian Tze[1] Mathieu Lafourcade[2]
[1] Computer-Aided Translation Unit (UTMK)
School of Computer Sciences, Universiti Sains Malaysia
Penang, Malaysia
[2] TAL-LIRMM, Université Montpellier II-CNRS
161 rue ada, 34392 Montpellier Cedex 5 - France
didier@cs.usm.my, liantze@cs.usm.my, lafourcade@lirmm.fr

**Résumé.** Fréquemment utilisés dans le Traitement Automatique des Langues Naturelles, les réseaux lexicaux font aujourd'hui l'objet de nombreuses recherches. La plupart d'entre eux, et en particulier le plus célèbre *WordNet*, souffrent du manque d'informations syntagmatiques mais aussi d'informations thématiques (« *problème du tennis* »). Cet article présente les vecteurs conceptuels qui permettent de représenter les idées contenues dans un segment textuel quelconque et permettent d'obtenir une vision continue des thématiques utilisées grâce aux distances calculables entre eux. Nous montrons leurs caractéristiques et en quoi ils sont complémentaires des réseaux lexico-sémantiques. Nous illustrons ce propos par l'enrichissement des données de *WordNet* par des vecteurs conceptuels construits par émergence.

**Abstract.** There is currently much research in natural language processing focusing on lexical networks. Most of them, in particular the most famous, *WordNet*, lack syntagmatic information and but also thematic information (« *Tennis Problem* »). This article describes conceptual vectors that allows the representation of ideas in any textual segment and offers a continuous vision of related thematics, based on the distances between these thematics. We show the characteristics of conceptual vectors and explain how they complement lexico-semantic networks. We illustrate this purpose by adding conceptual vectors to *WordNet* by emergence.

**Mots-clés :** *WordNet*, Vecteurs Conceptuels, informations lexicales, informations thématiques.

**Keywords:** *WordNet*, conceptual vectors, lexical information, thematic information.

# 1 Introduction

Originellement issus des travaux de Ross Quillian sur la psycholinguistique à la fin des années 60 (Quillian, 1968), les réseaux lexicaux sont toujours aujourd'hui au centre des recherches en Traitement Automatique des Langues Naturelles. Ils sont utilisés dans de nombreuses tâches (désambiguisation lexicale (Mihalcea *et al.*, 2004)) ou applications du domaine (traduction automatique avec les réseaux multilingues comme Papillon (Mangeot-Lerebours *et al.*, 2003) ou (Knight & Luk, 1994), recherche d'informations ou classification de textes (Harabagiu & Chai, 1998)). La plupart de ces réseaux, et spécifiquement le plus célèbre d'entre eux *WordNet* (Fellbaum, 1988), souffrent du manque d'informations syntagmatiques mais aussi d'informations concernant le domaine d'usage des termes ou du moins les termes thématiquement associés. Il n'y a ainsi aucune relation directe entre des termes comme ‹*teacher*›-‹*student*› (‹*enseignant*›-‹*étudiant*›) et ‹*boat*›-‹*port*› (‹*bateau*›-‹*port*›). Ce phénomène a été nommé « *Problème du tennis* » [(Fellbaum, 1988), p. 10] lorsqu'il a été remarqué qu'il fallait chercher les équivalents de ‹*balle*›, ‹*raquette*› et ‹*court*› à différents endroits de la hiérarchie.

Depuis quelques années, l'équipe de traitement automatique des langues (TAL) du LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier) travaille sur une formalisation de la projection de la notion linguistique de champ sémantique dans un espace vectoriel, les vecteurs conceptuels. Ils permettent de représenter les idées contenues dans un segment textuel quelconque et permettent d'obtenir une vision continue des thématiques utilisées grâce aux distances calculables entre eux.

Dans cet article, nous présentons les vecteurs conceptuels et en particulier leur version émergente. Nous montrons leurs caractéristiques et en quoi ils sont complémentaires des réseaux lexico-sémantiques. Nous illustrons ce propos par une expérience menée à Penang en Malaisie qui a consisté à enrichir les données de *WordNet* de vecteurs conceptuels par émergence.

# 2 Réseaux lexico-sémantique : l'exemple de *WordNet*

**Principe et lacunes** *WordNet* est une base de données lexicale pour l'anglais développée sous la direction de George Armitage Miller par le *Cognitive Science Laboratory* de l'université de Princeton (États-Unis d'Amérique). Il se veut représentatif du fonctionnement de l'accès au lexique mental humain.

*WordNet* est organisé en ensembles de synonymes appelés synsets. À chaque synset correspond un concept. Le sens des termes est décrit dans *WordNet* par trois moyens : (1) leur *définition* ; (2) le *synset* auquel ce sens est rattaché ; (3) les *relations lexicales* qui unissent entre eux les synsets. On trouve parmi ces relations, l'hyperonymie, la méronymie et l'antonymie.

La version 2 de *WordNet* compte 152059 termes ce qui constitue une couverture relativement large de la langue anglaise. Dans les premières versions de *WordNet*, les relations lexicales ne connectent que les termes de même morphologie. Il y a donc une hiérarchie pour les noms, une pour les adjectifs, une pour les verbes et enfin une dernière pour les adverbes.

Dans (Harabagiu *et al.*, 1999), les auteurs de *WordNet* (alors à sa version 1.6) relèvent six faiblesses dans la construction de leur réseau : (1) le manque de liens entre les hiérarchies ; (2) le nombre limité de relations entre termes traitant du même sujet ; (3) le manque de relations morphologiques ; (4) l'absence de relations thématiques ; (5) l'absence de certains sens de mots ;

(6) le manque d'uniformisation et de cohérence dans les définitions. Si les points 3, 5 et 6 ne nous intéressent pas dans cet article, nous allons montrer l'apport des vecteurs conceptuels pour la résolution des autres, tous trois formant le problème du tennis.

**Expériences cherchant à résoudre le problème du tennis** Dans cet article, nous nous intéresserons uniquement à la version 2.1 de *WordNet* qui était la dernière disponible au moment où nous avons réalisé nos expériences. Une nouvelle version (3.0) est sortie en Décembre 2006 mais elle ne semble pas comporter de réelles améliorations par rapport à la version précédente pour ce qui nous intéresse ici.

Depuis la version 2, des relations comme *derivationally related form* (formes dérivationnelles) permettent de lier des adjectifs à des verbes ou des adjectifs à des noms. De même, les synsets peuvent se voir attribuer un domaine d'usage. Toutefois, ces données semblent encore en nombre trop restreint pour être suffisamment pertinentes. Des relations typiques comme ‹*teacher*›-‹*student*› (‹*enseignant*›-‹*étudiant*›) ‹*boat*›-‹*port*› (‹*bateau*›-‹*port*›) ou ‹*doctor*›-‹*hospital*› (‹*docteur*›-‹*hôpital*›), pourtant souvent indispensables à une tâche de désambiguïsation lexicale, ne s'y trouvent toujours pas et le nombre restreint d'indications thématiques comme l'est le domaine ne permet pas de compenser ce défaut. Plusieurs solutions ont été proposées pour résoudre tout ou partie de ce problème.

Avec *eXtended WordNet*, (Harabagiu *et al.*, 1999) propose de désambiguïser l'ensemble des définitions de *WordNet* de façon semi-automatique. L'idée est, pour chaque définition, de dire quel est le sens utilisé pour chacun des termes. On peut ensuite comparer deux synsets et évaluer leur similarité. Nous verrons que nous utilisons ces informations pour fabriquer les vecteurs conceptuels de cette expérience. D'autres eux aussi rajoutent des informations aux synsets. Ainsi, (Agirre *et al.*, 2001) ajoutent des signatures lexicales issues de corpus taggés ou du Web. En revanche, d'autres cherchent plutôt à augmenter le nombre d'arcs existants. (Stevenson, 2002), par exemple, combine différentes métriques pour créer des arcs entre synsets à partir de leur définition et d'un thésaurus. (Ferret & Zock, 2006) utilisent eux un réseau de coocurrences pour extraire des relations typiques comme celles présentées dans un paragraphe précédent.

On le voit, toutes ses propositions ont en commun d'appartenir en particulier au domaine du discret. La nôtre est d'introduire une représentation continue des idées contenues dans le réseau, les vecteurs conceptuels.

# 3 Les vecteurs Conceptuels

Nous présentons ici les points fondamentaux à comprendre sur les vecteurs conceptuels. Nous revenons sur le mode de construction classique des vecteurs conceptuels, c'est-à-dire tels qu'ils ont été étudiés au LIRMM depuis 1997[1], à partir d'un ensemble de concepts choisis *a priori*. Nous expliquons dans cette partie certaines notions de base qui nous seront utiles pour présenter ensuite la construction par émergence, c'est à dire sans concepts prédéfinis.

**Principe Généraux** Nous représentons les aspects thématiques des segments textuels (documents, paragraphes, syntagmes, etc.) par des vecteurs conceptuels, une formalisation de la

[1] Voir les articles de l'équipe dans les précédentes éditions de cette conférence ou (Schwab, 2005).

projection de la notion linguistique de champ sémantique dans un espace vectoriel. À partir d'un ensemble de notions élémentaires dont nous faisons l'hypothèse, les concepts[2], il est possible de construire des vecteurs dont chaque composante correspond à un concept et est positive. Par exemple, le vecteur de l'item lexical ‹*vie*›, qui fusionne tous les sens de ‹*vie*›, peut être projeté sur les concepts suivants (les *CONCEPT*⌈*intensité*⌋ sont ordonnés par valeurs décroissantes de l'intensité) : $V$‹*vie*› = (*VIE*⌈*0.7*⌋, *NAISSANCE*⌈*0.48*⌋, *ENFANCE*⌈ *0.46*⌋, *MORT*⌈*0.43*⌋, *VIEILLESSE*⌈*0.41*⌋, ...).

La construction des vecteurs conceptuels se fait à partir de définitions extraites de diverses sources (dictionnaires, listes de synonymes, indexations manuelles, ...). Cette méthode d'analyse construit, à partir de vecteurs conceptuels déjà existants et de nouvelles définitions, de nouveaux vecteurs.

**Distance angulaire** La comparaison entre deux vecteurs se fait grâce à la distance angulaire $D_A$. Pour deux vecteurs conceptuels $A$ et $B$, $D_A(A,B) = arccos(Sim(A,B))$ où $Sim$ est $Sim(X,Y) = \cos(\widehat{X,Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|}$. Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et en pratique la mesure de l'angle entre les deux vecteurs. Empiriquement, nous estimons que pour une distance $D_A(X,Y) \leq \frac{\pi}{4}$ (45°), X et Y sont thématiquement proches et partagent plusieurs concepts. Pour $D_A(X,Y) \geq \frac{\pi}{4}$, la proximité thématique est considérée comme faible et aux alentours de $\frac{\pi}{2}$ (90°), X et Y n'ont aucune relation. Nous obtenons, par exemple, les angles suivants :

$D_A$(V(‹*fourmilier*›), V(‹*fourmilier*›))=0 (0°)
$D_A$(V(‹*fourmilier*›), V(‹*mammifère*›))=0.36 (21°)
$D_A$(V(‹*fourmilier*›), V(‹*animal*›))=0.45 (26°)
$D_A$(V(‹*fourmilier*›), V(‹*quadrupède*›))=0,42 (24°)
$D_A$(V(‹*fourmilier*›), V(‹*train*›))=1.18 (68°)
$D_A$(V(‹*fourmilier*›), V(‹*fourmi*›))=0,26 (15°)

Le premier résultat a une interprétation directe, ‹*fourmilier*› ne peut être plus proche d'autre chose que de lui même. Le fait qu'un ‹*fourmilier*› soit un ‹*mammifère*› explique le deuxième résultat. Un ‹*fourmilier*› n'a que peu de rapport avec un ‹*train*› ce qui explique l'angle plus important. Dans le dernier exemple, l'angle peu important entre ‹*fourmilier*› et ‹*fourmi*› se comprend si on se rappelle que $D_A$ est une distance thématique et non une distance ontologique. L'examen de la définition de fourmilier, « *mammifère qui se nourrit de fourmis* », explique le résultat.

**Le voisinage thématique, une vision continue de la thématique** La fonction de voisinage thématique permet de connaître les items lexicaux voisins d'un item lexical donné. On définit $\mathcal{V}$ la fonction de voisinage qui renvoie les $k$ items les plus proches en termes de distance angulaire $D_A$ d'un texte $Z$ dans une base vectorielle. Soit
$|\mathcal{V}(D_A, Z, k)| = k \qquad \forall X \in \mathcal{V}(D_A, Z, k), \quad \forall Y \notin \mathcal{V}(D_A, Z, k), \quad D_A(\mathrm{X}, \mathrm{Z}) \leq D_A(\mathrm{Y}, \mathrm{Z})$
Par exemple, les 7 termes proches et ordonnés par distance thématique croissante du nom ‹*mort*› peuvent être :

$\mathcal{V}(D_A$, ‹*mort*›, 7) = (‹*mort*› 0) (‹*meurtre*› 0.367) (‹*tueur*› 0.377) (‹*âge de la vie*› 0.481) (‹*tyrannicide*› 0.516) (‹*tuer*› 0.579) (‹*mort :adj*› 0.582)

La méthode de voisinage peut être utilisée lors de l'apprentissage des vecteurs conceptuels pour vérifier la cohérence globale de la base ou en phase d'exploitation pour trouver le meilleur mot à utiliser dans un énoncé. Ainsi, elle constitue un nouvel outil pour accéder aux mots et à leur sens, complémentaire à ceux décrits dans (Zock, 2002) comme la forme, la morphologie ou la navigation dans un grand réseau lexical. La fonction de voisinage permet ainsi une navigation

[2]Dans notre expérimentation sur le français nous utilisons (Larousse, 1992) qui défini 873 concepts.

dans le domaine du continu contrairement aux réseaux sémantiques qui ne permettent qu'une navigation discrète.

**Somme vectorielle** Soient $X$ et $Y$ deux vecteurs, leur *somme vectorielle normée* V est définie par : $\vartheta^2 \rightarrow \vartheta \ : V = X \oplus Y \quad | \quad V_i = \frac{X_i+Y_i}{\|X+Y\|}$ où $\vartheta$ est l'ensemble des vecteurs conceptuels, $V_i$ (resp $X_i$, $Y_i$) représente la i-ème composante du vecteur V (resp. X, Y).

La somme vectorielle normée de deux vecteurs donne un vecteur équidistant en termes d'angle des deux premiers vecteurs. Il s'agit en fait d'une moyenne des vecteurs sommés. En tant qu'opération sur les vecteurs conceptuels, on peut donc voir la somme vectorielle normée comme l'union des idées contenues dans les termes.

Soient $X$ et $Y$ deux vecteurs, leur *produit terme à terme normalisé* $V$ est défini par : $\vartheta^2 \rightarrow \vartheta \ : V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i}$ L'opérateur $\otimes$ peut être interprété comme un opérateur d'intersection entre vecteurs. Si l'intersection entre deux vecteurs est le vecteur nul, alors ils n'ont rien en commun. Du point de vue des vecteurs conceptuels, cette opération permet donc de sélectionner les idées communes à un ensemble de termes.

**Construction des vecteurs par émergence** L'approche par émergence s'affranchit de tout thésaurus et vecteurs de concept comme base de départ. Seule $d$ la taille du vecteur est fixée *a priori*. Le mode de construction des vecteurs est identique au modèle classique à la différence que si un des vecteurs entrant dans la somme est inexistant, car non encore calculé, alors ce vecteur est tiré au hasard. Le processus de calcul est itéré jusqu'à convergence de chaque vecteur.

Comme nous le montrons de façon plus détaillée dans (Lafourcade, 2006), il y a un certain nombre d'avantages à utiliser ce modèle. Le premier d'entre eux est de pouvoir choisir librement la quantité de ressources que l'on souhaite utiliser en choisissant la taille des vecteurs de façon appropriée. Pour donner une idée de l'importance de ce choix, une base de 500000 vecteurs de dimension 1000 fait environ 2Go, de taille 2000, 4Go, ... Comme il ne serait pas alors ni raisonnable ni facile de définir une jeu de concept de la taille choisie, autant chercher une approche nous permettant de nous en passer. De plus, ce qui peut sembler un pis-aller ou au mieux un compromis, s'avère un avantage car la densité lexicale dans l'espace des mots calculés par émergence est bien plus constante que dans un espace où les concepts sont précalculés. En effet, les ressources (les dimensions de l'espace) ont tendance à être harmonieusement distribuées en fonction de la richesse lexicale.

# 4 Modélisation hybride du sens : vecteurs conceptuels et réseaux lexicaux

## 4.1 Apport des réseaux lexicaux aux vecteurs conceptuels

Les distances utilisées sur les vecteurs, comme le montre (Besançon, 2001), mettent en exergue les composantes communes et/ou les composantes distinctes. Si nous utilisons en particulier la distance angulaire, c'est que ses caractéristiques mathématiques, sa simplicité à comprendre et à interpréter linguistiquement ainsi que son efficacité en termes de temps de calcul en font

un bon outil. Quelle que soit la distance choisie, utilisée sur ce type de vecteur (représentant des idées, des concepts plutôt que des termes cooccurrents), elle est d'autant plus faible que les vecteurs des objets lexicaux qui en sont les arguments sont dans un champ sémantique proche (en isotopie selon la terminologie de Rastier (Rastier, 1985)).

Dans le cadre d'une analyse sémantique comme celle qui nous intéresse ici, nous l'utilisons pour tirer profit des informations mutuelles contenues dans les vecteurs conceptuels pour faire de la désambiguïsation lexicale sur des mots qui ont des sens situés dans un champ sémantique proche. Ainsi, « *Zidane a marqué un but* » peut être désambiguïsée grâce aux idées communes concernant le sport tandis que « *L'avocat a plaidé à la cour* » peut l'être grâce à celles concernant la justice. De même, en ce qui concerne les rattachements prépositionnels, les vecteurs peuvent permettre dans « *Il voit la fille avec un télescope.* » de rattacher « *avec un télescope* » au verbe ‹*voir*› grâce aux idées communes sur la vision.

En revanche, les vecteurs conceptuels ne peuvent pas aider à résoudre des cas où les termes mis en jeu sont dans des champs sémantiques différents. On remarquera même qu'une analyse ne reposant que sur eux peut conduire à de gros contre-sens. Par exemple, dans la phrase « *L'avocat a mangé un fruit* », ‹*avocat*› ne peut être interprété que comme le fruit et non comme l'auxiliaire de justice. Ces limites des vecteurs conceptuels ont été expérimentalement montrées pour l'analyse sémantique sur des algorithmes à fourmis dans (Lafourcade & Guinand, 2006).

Il aurait fallu que des connaissances comme « *un avocat est un être humain* » et « *un être humain mange* » puissent être identifiées, ce qui n'est donc pas possible avec des vecteurs conceptuels seuls. Les vecteurs conceptuels seuls ne sont ainsi pas suffisants pour exploiter certaines instances de fonctions lexicales dans les textes et un réseau lexical peut donc aider à pallier ces manques. Des publications antérieures ont montré la nécessité de cette approche hybride : (Schwab *et al.*, 2002) pour les antonymies, (Lafourcade & Prince, 2003) pour les génériques et les hyperonymes. (Schwab, 2005) étend cette constatation à toute relation susceptible d'aider à la résolution d'une analyse sémantique.

## 4.2 Apport des vecteurs conceptuels aux réseaux lexicaux

S'ils bénéficient d'une précision certaine, le rappel des réseaux est bien moins fort. Il est, en effet, difficile de penser que l'on pourrait représenter toutes les relations entre les termes. En effet, comment considérer deux termes qui sont dans le même champ sémantique ? Ils peuvent très bien ne pas se trouver dans le réseau car ils ne seraient pas forcément reliés par un des arcs "classiques". Envisager l'introduction d'arcs de type *champ sémantique*, poserait à nos yeux deux problèmes dus au caractère flou et flexible de cette relation :

– le premier est lié à l'idée de la relation que se fait le concepteur de la base, à quel moment considère-t'il que deux synsets sont dans le même champ sémantique ? Dans un cas défavorable, on aurait très peu de ces arcs tandis que dans un cas opposé, on pourrait se trouver avec une explosion combinatoire du nombre d'arc ;
– le second problème, plus fondamental, est lié à la représentation elle même. Comment envisager de représenter par un élément discret une relation floue donc du domaine du continu ?

Ainsi, le domaine du continu offert par les vecteurs conceptuels offre des flexibilités que le domaine du discret offert par les réseaux ne peut donner. Il permet de pouvoir rapprocher des mots sur des idées peu importantes mais pourtant communes à deux objets.

Avec cette approche hybride - vecteurs conceptuels, réseau lexical - nous proposons de combiner des informations de nature complémentaire. Les vecteurs conceptuels et l'opération de distance thématique par leur nature peuvent pallier le faible rappel intrinsèque aux réseaux lexicaux tandis que ces derniers peuvent permettre de désambiguïser les cas qui sont dans un champs sémantique différent contrairement aux vecteurs conceptuels. Les défauts des uns sont ainsi compensés par les qualités des autres ce qui fait des vecteurs conceptuels et des réseaux lexicaux des outils complémentaires.

# 5 Expérience sur *WordNet* : utilisation des données

## 5.1 Exploitation des définitions

Le projet *eXtended WordNet* (Mihalcea & Moldovan, 2001) est mené à la *Southern Methodist University* de Dallas au Texas et vise deux objectifs : (1) désambiguïser l'ensemble des termes utilisés dans les définitions des synsets, c'est-à-dire indiquer quels sont les synsets employés dans la définition ; (2) Transformer ces définitions en forme logique pour permettre plus facilement les calculs.

Ces données ont été réalisées de façon semi-automatique en utilisant les informations du réseau[3], des distances entre définitions ou bien les informations sur le domaine. Ces données sont en partie contrôlées à la main et le taux de précision de plus de 90%.

Pour les vecteurs conceptuels, nous avons utilisé ces données sous forme logique car elles permettent de repérer les éléments les plus importants de la définition, en particulier le genre. Le calcul se fait ainsi sur un arbre en dépendances fabriqué à partir de cette définition prétraitée pour enlever le métalangage difficilement exploitable pour une analyse thématique. Dans nos explications, nous allons prendre pour exemple la forme logique de la définition de fourmi.

*ant :NN(x1) -> social :JJ(x1) insect :NN(x1) live :VB(e1, x1, x3) in :IN(e1, x2) organized :JJ(x2) colony :NN(x2)*

Elle est organisée en 3 ensembles : x1 = {*social*, *insect*}, x2 = {*organised*, *colony*} et e1 = {*live*}. Ce dernier ainsi que *in* permettent de hiérarchiser les ensembles. Le vecteur de chacun des ensembles est calculé en faisant la somme vectorielle de l'élément le plus porteur de sens de cet ensemble (verbes, VB ; noms, NN ) et de la moitié des adjoints (adverbes, RB ; adjectifs, JJ). Le calcul du vecteur global se fait ensuite par somme vectorielle pondérée des différents ensembles dans l'arbre en commençant par la partie la plus basse. Ce mode de calcul permet de considérer de façon prépondérante le genre sur les autres termes de la définition et de façon plus générale les têtes sur leurs dépendants syntaxiques. La figure 1 synthétise ce calcul. Aucun prédicat n'étant dans l'ensemble x3, il n'apparaît pas sur le schéma.

## 5.2 Exploitation des relations

L'exploitation des relations se fait à deux niveaux : (1) pour la construction des vecteurs, elles permettent de fabriquer de manière complémentaire aux définitions le vecteur d'un synset ; (2) pour éviter les phénomènes de regroupement d'ensembles distincts.

[3]Par exemple, pour une définition aristotélicienne (en genre et différences), si le genre a un sens qui est aussi un hyperonyme du synset défini, on considère que ce sens est celui utilisé dans la définition.

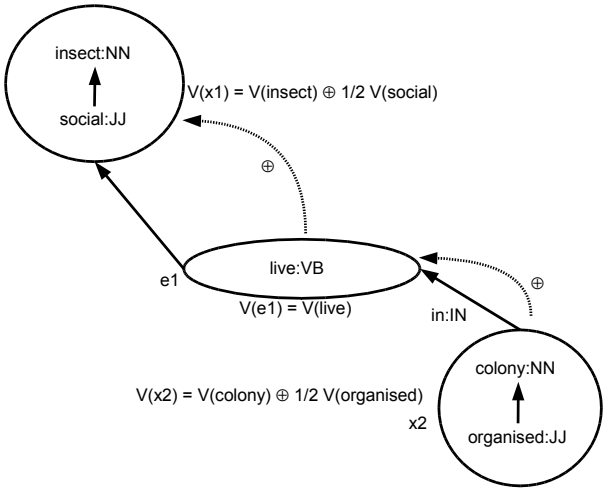


FIG. 1 – Construction du vecteur conceptuel de la définition de fourmi

### 5.2.1 Construction des vecteurs

La construction d'un vecteur conceptuel est effectuée pour chaque nœud du réseau par simple somme pondérée des vecteurs des nœuds reliés. Soit un nœud $N$ relié a $k$ nœuds $N_1 \dots N_k$, le vecteur de $N$, $V(N)$ sera égal à $p_1V(N_1) + p_2V(N_2) + \dots + p_kV(N_k)$ où $p_i$ est le poids du i-ème nœud. Le vecteur somme est ensuite normalisé.

Cette approche entraîne naturellement une agglomération des vecteurs. Il est donc nécessaire d'augmenter le contraste d'un vecteur à la suite de son calcul. Pour ce faire, on calcule le coefficient de variation[4] de V. Si ce dernier ne se situe pas a 10% du CV moyen alors le vecteur subit une opération non linéaire d'amplification (la mise à une puissance $n$ de chaque composante puis normalisation), et ce de façon itérée jusqu'à l'obtention d'un coefficient de variation dans la fourchette acceptable. Cette dernière a été estimée à partir des valeurs obtenues dans les expériences avec concepts prédéfinis.

### 5.2.2 Problème du regroupement d'ensembles distincts

Un dernier problème potentiel est que les vecteurs de deux ensembles distincts (à la fois au sens du réseau lexical et de la thématique) de termes peuvent occuper la même région de l'espace. L'approche du calcul se faisant par activation et les vecteurs étant tirés au hasard à l'initialisation rien n'empêche que cela se produise par accident. Il est donc nécessaire de "séparer" les vecteurs proches mais correspondant pourtant à des parties très différentes du réseau lexical et de la thématique.

La détection de ce phénomène se fait par scrutation du voisinage d'un vecteur conceptuel. Si parmi ses $n$ premiers voisins, la densité de mots n'ayant rien à voir avec le mot étudié est importante alors une action de séparation doit être entreprise.

Cette action de séparation consiste à plonger l'ensemble du réseau dans un champs où les nœuds ont tendance à se repousser. En s'inspirant directement de la physique, une force de répulsion en $1/d^2$ est calculée itérativement entre les nœuds. Pour un nœud donné, on peut ainsi calculer

[4]Le coefficient de variation CV est donné par la formule $\frac{EC(V)}{\mu(V)}$ avec EC(V) l'écart type du vecteur V et $\mu(V)$ la moyenne arithmétique des composantes de V.

un vecteur déplacement qui va l'éloigner des nœuds dont il se trouve trop près. Les nœuds ne se rapprochant pas par voisinage thématique (lors de la première phase du calcul) mais se trouvant proches "par accident" finissent ainsi naturellement par se séparer.

# 6 Conclusion

Dans cet article, nous avons présenté les vecteurs conceptuels construits par émergence. Nous avons montré en quoi ils peuvent aider à résoudre le « *problème du tennis* » de par leur caractère complémentaire aux réseaux lexico-sémantiques dont l'exemple le plus courant dans les recherches actuelles est *WordNet*. En effet, le rappel des réseaux est faible, ils ne permettent pas facilement de représenter le champs sémantique contrairement aux vecteurs tandis que ces derniers ne sont pas suffisants pour représenter des relations comme l'hyperonymie ou la méronymie.

Notre proposition est de tirer profit de cette complémentarité en ajoutant à *WordNet* des vecteurs conceptuels construits à partir des définitions et des relations contenues dans cette base. La méthode proposée ici tient du domaine du continu contrairement à l'ensemble des méthodes que nous avons étudiées dans la littérature qui, elles, font partie du domaine du discret (ajout d'arcs pour les relations, de symboles sur le domaine, etc.).

Nous avons conscience que cette méthode ne permet seulement que de résoudre une partie du « *problème du tennis* ». En effet, les vecteurs conceptuels ne permettent pas d'exhiber les rapports collocationnels non-thématiques entre items. Il s'agit essentiellement des relations qu'Igor Mel'čuk modélise avec ses fonctions lexicales syntagmatiques (Mel'čuk *et al.*, 1995) comme l'intensification (« *peur bleue* » ; *Magn* (‹*peur*›) = ‹*bleue*›)), la dégradation (« *lait tourne* » ; *Degrad* (‹*lait*›) = ‹*tourner*›) ou bien encore le confirmateur (« *argument valable* » ; *Ver* (‹*argument*›) = ‹*valable*›). Comme le remarque (Ferret & Zock, 2006), ces relations font partie de celles qu'il faudrait vraisemblablement avoir dans une base lexicale. Nous partageons ce point de vue, certaines pistes ont été explorées dans (Schwab, 2005) et continuent à l'être actuellement.

# Références

E. AGIRRE, O. ANSA, D. MARTINEZ, et E. HOVY. « Enriching WordNet concepts with topic signatures ». Dans les actes de *NAACL worshop on WordNet and Other Lexical Resources : Applications, Extensions and Customizations*, Pittsburg, USA, 2001.

Romaric BESANÇON. « *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de texte* ». Thèse de doctorat, École Polytechnique Fédérale de Lausanne, Laboratoire d'Intelligence Artificielle, 2001.

Christiane FELLBAUM, . *WordNet : An Electronic Lexical Database*. The MIT Press, 1988.

Olivier FERRET et Michael ZOCK. « Enhancing Electronic Dictionaries with an Index Based on Associations ». Dans les actes de *Proceedings of the 21st International Conference on Computational Linguistics*, pp 281–288, 2006. Association for Computational Linguistics.

Sanda HARABAGIU et Joyce Yue CHAI, . *Usage of WordNet in Natural Language Processing Systems*, Université de Montréal, Montréal, Canada, 1998.

Sanda M. HARABAGIU, George Armitage MILLER, et Dan I. MOLDOVAN. « WordNet 2 - A Morphologically and Semantically Enhanced Resource ». Dans les actes de *Workshop SIGLEX'99 : Standardizing Lexical Resources*, pp 1–8, 1999.

Kevin KNIGHT et Steeve LUK. « Building a Large-Scale Knowledge Base for Machine Translation ». Dans les actes de *AAAI'1994 : National Conference on Artificial Intelligence*, 1994.

Mathieu LAFOURCADE et Frédéric GUINAND. « Ants for Natural Language Processing ». *International Journal of Computational Intelligence Research*, 2006. À paraître.

Mathieu LAFOURCADE et Violaine PRINCE. « Mixing Semantic Networks and Conceptual Vectors : the Case of Hyperonymy ». Dans les actes de *ICCI-2003 (2nd IEEE International Conference on Cognitive Informatics)*, pp 121–128, 2003.

Mathieu LAFOURCADE. « Conceptual Vector Learning - Comparing Bootstrapping from a Thesaurus or Induction by Emergence ». Dans les actes de *LREC'2006*, 2006.

LAROUSSE, . *Thésaurus Larousse - des idées aux mots, des mots aux idées*. Larousse, 1992.

Mathieu MANGEOT-LEREBOURS, Gilles SÉRASSET, et Mathieu LAFOURCADE. « Construction collaborative d'une base lexicale multilingue : Le projet Papillon ». *TAL (Traitement Automatique des langues) : Les dictionnaires électroniques*, pp 151–176, 2003.

Igor MEL'ČUK, André CLAS, et Alain POLGUÈRE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, 1995.

Rada MIHALCEA et Dan MOLDOVAN. « eXtended Wordnet : progress report ». Dans les actes de *NAACL 2001 - Workshop on WordNet and Other Lexical Resources*, Pittsburgh, USA, 2001.

Rada MIHALCEA, Paul TARAU, et Elizabeth FIGA. « PageRank on Semantic Networks, with Application toWord Sense Disambiguation ». Dans les actes de *COLING'2004 : 20th International Conference on Computational Linguistics*, pp 1126–1132, 2004.

Ross QUILLIAN. « *Semantic Informatic processing* », Chapitre Semantic memory, pp 227–270. MIT Press, 1968.

François RASTIER. « *L'isotopie sémantique, du mot au texte* ». Thèse de doctorat d'État, Université de Paris-Sorbonne, 1985.

Didier SCHWAB. « *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* ». Thèse de doctorat, Université Montpellier 2, 2005.

Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Vers l'apprentissage automatique, pour et par les vecteurs conceptuels, de fonctions lexicales. L'exemple de l'antonymie ». Dans les actes de *TALN 2002*, volume 1, pp 125–134, 2002.

Mark STEVENSON. « Augmenting Noun Taxonomies by Combining Lexical Similarity Metrics ». Dans les actes de *COLING'2002 : 19th International Conference on Computational Linguistics*, volume 2/2, pp 953–959, 2002.

Michael ZOCK. « Sorry, What Was Your Name Again, Or How to Overcome The Tip-Of-The Tongue with the help of a computer ? ». Dans les actes de *SemaNet'02 : Building and Using Semantic Networks*, Taipei, Taiwan, 2002.