

Modélisation de l'Hyperonymie via la combinaison de réseaux sémantiques et de vecteurs conceptuels

Mathieu Lafourcade, Violaine Prince
{lafourcade,prince}@lirmm.fr

LIRMM - Laboratoire d'informatique, de Robotique
et de Microélectronique de Montpellier
MONTPELLIER - FRANCE.

<http://www.lirmm.fr/~{lafourcade,prince}>

Résumé - Abstract

La sémantique lexicale est un élément clé du Traitement Automatique du Langage Naturel (TALN) en ce qu'elle représente le point de convergence entre les représentations des connaissances (KR) et les ontologies. Parmi les tendances durables en représentation de sémantique lexicale, deux d'entre-elles semblent être conflictuelles ; l'approche WordNet issue des réseaux sémantiques et teintée de KR, et l'approche vectorielle issue des représentations saltoniennes et de la recherche d'information qui a su trouver un large spectre d'applications en TALN. Quand les représentations ontologiques sont nécessaires, l'hyperonymie, l'approximation la plus proche de la relation is-a, est en jeu. Dans cet article, nous montrons comment gérer l'hyperonymie dans le cadre d'une approche vectorielle pour la sémantique, en combinant les réseaux sémantiques et les vecteurs conceptuels, et comment cette modélisation peut être appliquée à de nouvelles fonctions telles que la substitution de termes et l'approximation sémantique qui appartiennent au domaine de la similarité sémantique.

Lexical semantics are a key issue in Natural Language Processing (NLP) since they represent the point of convergence with conceptual Knowledge Representation (KR) and ontologies. Among the well established trends in lexical semantics representations, two trends seem to be conflictual: the WordNet approach, born from semantic networks, and KR-oriented, and the Vector approach, originated from the Saltonian representation in Information Retrieval, which has found a set of applications in NLP. When ontological representation is needed, hyperonymy, the closest approximation to the is-a relation, is at stake. In this paper, we show how to account for hyperonymy within the vector-based frame for semantics, relying on a cooperation between semantic networks and conceptual vectors, and how this can be applied to new functions such as word substitution, and semantic approximation, that belong to the field of semantic similarity.

Mots Clef

Traitement Automatique du Langage Naturel, Apprentissage lexical, Vecteurs conceptuels, Hyperonymie, Réseaux sémantiques

Keywords

Natural Language Processing, Lexical Learning Conceptual Vectors, Hyperonymy, Semantic Networks

1 Introduction

La sémantique lexicale est un aspect important du Traitement Automatique du Langage Naturel (TALN) en ce qu'elle constitue le point de convergence entre les représentations des connaissances (KR) conceptuelles et les ontologies. Elle s'étend largement au domaine du traitement des ressources lexicales, et de nombreux travaux à la fois en TALN et en IA ont été dédiés aux fonctions sémantiques lexicales comme un moyen d'approcher la représentation des sens des mots et leur discrimination. Parmi les tendances classiques en représentation lexicale et sémantique, deux approches semblent conflictuelles : celle de Wordnet [Miller et Fellbaum 1991], [Fellbaum 1998] issue des réseaux sémantiques et teintée de KR, et les approches vectorielles issues des représentations saltoniennes en recherche d'information [Salton 1968], qui ont su trouver de nombreuses applications en TALN. La première est globalement basée sur des logiques et la seconde sur des algèbres d'espaces vectoriels. La première est très efficace pour les relations du type *is-a* (considéré comme une relation conceptuelle souvent incluse dans l'hyponymie), mais demeure pratiquement muette à propos d'autres fonctions lexicales intéressantes comme l'antonymie ou les associations thématiques. La synonymie a été largement traitée [Sparck Jones 1986], [Resnik 1995], mais la discrimination entre synonymie et hyperonymie a souvent amené les chercheurs à se pencher sur des notions plus flexibles comme la similarité sémantique [Resnik 1999]. L'approche vectorielle se situe complètement à l'opposé. Permettant naturellement les associations thématiques, elle s'offre à l'implémentation de nombreuses fonctions de synonymies à grain fin [Lafourcade et Prince 2001] mais aussi de fonctions d'antonymie [Schwab et al. 2002] tout en restant incapable de différencier ou de valider l'existence de relations d'hyperonymie.

Dans cet article, nous montrons (1) comment modéliser l'hyperonymie dans le cadre vectoriel pour la sémantique, en se basant sur une coopération entre les réseaux sémantiques et les vecteurs conceptuels, (2) comment elle peut être appliquée à d'autres nouvelles fonctions comme la substitution de termes et l'approximation sémantique qui toutes les deux font partie du champ de la similarité sémantique. Nous utilisons un réseau sémantique afin d'améliorer l'apprentissage de vecteurs, et de façon symétrique, nous construisons un sous-réseau spécifique de relations d'hyperonymie entre vecteurs sémantiques.

2 Hyperonymie et relations *is-a*

L'**hyperonymie** est une fonction lexicale qui, à partir d'un terme t , associe un ou plusieurs autres termes plus généraux, comme ceux utilisés pour définir t en *genre* (définition aristotélicienne). Les définitions de dictionnaire sont en général en *genre* et en *différences*. Sa fonction symétrique est l'**hyponymie**.

L'hyperonymie, dans la plupart des articles en KR, est assimilée à la relation *is-a* (voir par exemple [Brachman et Schmolze 1985]). Rappelons ici que la relation *is-a* est telle que si X est une classe d'objets, et X' une sous-classe de X , alors $is-a(X', X)$ est vrai. Le terme X est l'argument général tandis que X' est l'argument spécifique. Le problème que nous rencontrons est que l'hyperonymie linguistique n'est pas une relation *is-a* "pure" (héritage simple). Nous trouvons, par exemple, la définition suivante du terme *cheval* : "un animal herbivore, quadrupède, etc." Un bon hyperonyme pour cette définition de *cheval* est un *animal herbivore* ou encore *quadrupède herbivore*. Le terme *animal* est un autre hyperonyme, tandis que la relation *mammifère herbivore is-a mammifère* est vrai. Cependant, thématiquement,

un *cheval* est très proche d'un *herbivore*, alors que *herbivores* est un ensemble d'individus qui peuvent appartenir à plusieurs sous-hiérarchies d'une taxonomie (certains oiseaux, insectes et reptiles sont herbivores, mais aussi métaphoriquement plein d'autres choses). Donc, même si linguistiquement on peut toujours écrire qu'*un cheval est un herbivore*, il demeure que la relation *cheval is-a herbivore* pose problème. Plusieurs solutions sont alors envisageables, dont en premier lieu considérer que l'héritage multiple de *is-a* est possible. Si tel n'est pas le cas, alors si on choisit *cheval is-a mammifère*, il semble nécessaire de représenter les autres relations (*cheval is-a herbivore*, *cheval is-a quadrupède*, etc.) comme des propriétés (ou des rôles ou des vues selon la terminologie utilisée dans la modélisation concernée). En ce qui nous concerne, l'héritage multiple de classes, motivées à la fois par leur utilité (factorisation) et leur interprétation naturelle, nous semble la plus adaptée.

Wordnet et l'hyperonymie. Wordnet a pour relation fondamentale la synonymie. Cependant, en ce qui concerne l'hyperonymie, Wordnet s'apparente à une taxonomie de termes, et en tant que telle, elle ne capture que les relations *is-a*. Un hyperonyme est un super-ordonné linguistique, généralement utilisé dans des définitions, qui capture également certaines propriétés qui ne peuvent pas jouer le rôle de classes par elles-mêmes. Les termes polysémiques ont, en général, de nombreuses définitions et donc souvent beaucoup d'hyperonymes : un *cheval* est également un *véhicule*, ce qui explique pourquoi WordNet est organisé selon un réseau et non pas un arbre. La seule contrainte langagière est que l'hyperonyme doit être plus général (et donc *herbivore* peut être un hyperonyme pour *cheval*) alors que dans un réseau sémantique, chaque étape de la chaînes de classes et sous-classes doit vérifier la relation d'ordre.

Hyperonymie et définition de termes. Comme évoqué précédemment, les hyperonymes peuvent être extraits, quand ils sont inconnus, de définitions dictionnaires. Seuls les concepts généraux, qui ont tendance à jouer le rôle d'hyperonymes et de superclasses (*is-a*) de nombreux termes, ne sont pas définis via une approche aristotélicienne, mais sont appréhendés à travers leurs hyponymes. C'est pourquoi, dans notre modèle de vecteurs conceptuels (présenté dans la section suivante), nous considérons l'existence d'un *horizon d'hyperonymie* au-delà duquel les définitions doivent être inversées : les hyperonymes sont plus difficiles à trouver et sont moins explicatifs que les hyponymes. Le terme *action* est pratiquement au sommet de la taxonomie de WordNet et les définitions de dictionnaires tendent à l'expliquer avec des termes plus spécifiques.

3 Vecteurs conceptuels

Les vecteurs ont été utilisés depuis longtemps en Recherche d'Informations [Salton et MacGill 1983] ainsi que pour la représentation du sens dans le modèle LSI [Deerwester et al. 1990] issu des études d'analyse sémantique latente (LSA) en psycholinguistique. En TALN, [Chauché 1990] a proposé un formalisme pour la projection de la notion linguistique de champ sémantique dans un espace vectoriel, sur lequel notre modèle se fonde. À partir d'un ensemble de notions élémentaires, *les concepts*, il est possible de construire des vecteurs (dits conceptuels) et de les associer à des items lexicaux.¹ L'hypothèse qui considère qu'un ensemble de concepts peut être un générateur pour le langage a depuis longtemps été exposée dans [Rodget 1852] (hypothèse du thésaurus) et a été appliqué par les chercheurs en TALN (par exemple [Yarowsky 1992]). Les termes polysémiques combinent les différents vecteurs correspondant aux différents sens. Cette approche vectorielle est basée sur des propriétés mathématiques connues, et il est donc

¹ Un item lexical est un terme simple ou composé, ou encore une locution. Par exemple *voiture*, *pomme de terre* et *tirer le diable par la queue* sont des items lexicaux.

possible d'effectuer des manipulations formelles auxquelles sont attachées des interprétations linguistiques (ou psycho-linguistiques) raisonnables. Les concepts sont définis par un thésaurus (dans notre prototype appliqué au français, nous avons choisi [Larousse 2001] où 873 concepts sont identifiés - on peut comparer avec le millier de concepts définis dans [Rodget 1852]). Afin d'être cohérent avec l'hypothèse du thésaurus, on considère que cet ensemble constitue un espace générateur pour les mots et leurs sens. Cet espace n'est pas libre (pas de base propre) et n'importe quel terme peut y projeter ses sens selon le principe suivant.

Soit \mathcal{C} un ensemble fini de n concepts, un vecteur conceptuel V est une combinaison linéaire des éléments c_i de \mathcal{C} . Pour un sens A , son vecteur $V(A)$ est la description en extension des activations de tous les concepts de \mathcal{C} . Par exemple, les différents sens de ‘*porte*’ peuvent être projetés sur les concepts suivants (avec leur valeur d'activation entre crochets) :

$$V(\text{‘porte’}) = (\text{OUVERTURE}[0.7], \text{BARRIÈRE}[0.41], \text{LIMITE}[0.35], \text{PROXIMITÉ}[0.31], \text{EXTERIEUR}[0.30], \text{INTERIEUR}[0.29], \dots)$$

En pratique, plus \mathcal{C} est grand, plus fine est la description du sens. En retour, la manipulation informatique est moins aisée. Comme les vecteurs sont globalement denses (très peu de composantes valent 0), l'énumération des concepts activés est longue et passablement difficile à évaluer. On préfère, en général, sélectionner les termes thématiquement les plus proches, c'est-à-dire son *voisinage*. Par exemple, les termes les plus proches de ‘*porte*’ ordonnés par distance thématique croissante sont :

$$\mathcal{V}(\text{‘porte’}) = \text{‘portillon’}, \text{‘portail’}, \text{‘portière’}, \text{‘issue’}, \text{‘ouverture’}, \text{‘vantail’}, \dots$$

Afin de représenter et manipuler certains aspects de sémantique lexicale dans ce cadre vectoriel, nous utilisons des opérations élémentaires sur les vecteurs. Une mesure intéressante est la distance angulaire qui se base sur une mesure classique de similarité.

3.1 Opération sur les vecteurs conceptuels

Distance angulaire. Définissons $Sim(A, B)$ comme une mesure de *similarité* entre deux vecteurs A et B , souvent utilisée en recherche d'information. On suppose ici que les composantes des vecteurs sont positives ou nulles. On définit la *distance angulaire* D_A entre deux vecteurs A et B de la façon suivante :

$$D_A(A, B) = \arccos(Sim(A, B)) \quad \text{et} \quad Sim(A, B) = \cos(\widehat{A, B}) = \frac{A \cdot B}{\|A\| \times \|B\|} \quad (1)$$

avec “ \cdot ” pour le produit scalaire. Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et est une mesure de l'angle formé par les deux vecteurs. On considèrera, en général, que pour une distance $D_A(A, B) \leq \frac{\pi}{4}$, (i.e. moins de 45 degrés) A et B sont thématiquement proches et partagent beaucoup de concepts. Pour $D_A(A, B) \geq \frac{\pi}{4}$, la proximité thématique entre A et B sera considéré comme faible. Aux alentours de $\frac{\pi}{2}$, ils n'ont pas de relation. D_A est une vraie distance et vérifie les propriétés de réflexivité, de symétrie et d'inégalité triangulaire.

Somme et produit terme à terme. *Somme.* Soit X et Y deux vecteurs, on définit V comme leur somme normée :

$$V = X \oplus Y \quad | \quad v_i = (x_i + y_i) / \|V\| \quad (2)$$

Intuitivement, le vecteur somme de X et Y correspond à l'union des propriétés sémantiques de X et Y . Cet opérateur est idempotent (on a $X \oplus X = X$). Le vecteur nul $\vec{0}$ est l'élément neutre et par définition nous avons $\vec{0} \oplus \vec{0} = \vec{0}$.

Produit terme à terme. Soit X et Y deux vecteurs, on définit V comme leur produit terme à terme :

$$V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i} \quad (3)$$

Cet opérateur est idempotent et $\vec{0}$ est absorbant : $V = X \otimes X = X$ et $V = X \otimes \vec{0} = \vec{0}$. Intuitivement, le produit terme à terme de X et de Y correspond à l'intersection de leurs propriétés sémantiques. Cet opérateur est crucial pour l'hyperonymie car un hyperonyme et son hyponyme peuvent être étudiés à l'aide d'une relation d'inclusion de propriétés. Mais, cette intersection est également importante dans le cas de la synonymie et peut fournir certains indices à propos des propriétés de polysémie de certains vecteurs (intersection de nombreux vecteurs différents). Une fonction plus adaptée concernant la mise en évidence de propriétés commune avec conservation relative de celles en différence est donnée avec la contextualisation.

Contextualisation. Quand deux termes sont en présence, certains des sens de chacun d'eux sont mutuellement sélectionnés par le contexte que constitue l'autre terme. Ce phénomène est appelé *contextualisation*. Il consiste à mettre l'accent sur les propriétés communes de chaque sens. Soient X et Y deux vecteurs, on définit $\gamma(X, Y)$ la contextualisation de X par Y comme :

$$\gamma(X, Y) = X \oplus (X \otimes Y) \quad (4)$$

Cet opérateur n'est pas symétrique, il est idempotent ($\gamma(X, X) = X$) et le vecteur nul est l'élément neutre à droite ($\gamma(X, \vec{0}) = X \oplus \vec{0} = X$). La fonction $\gamma(X, Y)$ produit un nouveau vecteur qui est une copie de X rapprochée de Y proportionnellement à leur intersection. La contextualisation est un moyen à faible coût d'amplifier les propriétés saillantes dans un contexte donné. Pour un vecteur d'un terme polysémique, si le vecteur contexte est pertinent, un des sens possibles est *activé* par la contextualisation. Par exemple, le terme *frégate* est ambigu et son vecteur global se trouve entre ceux de *oiseau* et *navire*. Si le vecteur de *frégate* est contextualisé par celui de *mouette* alors le sens d'*oiseau* de frégate émergera relativement à celui de *navire*.

3.2 Fonctions lexicales modélisées : synonymie et antonymie

Synonymie. Deux items lexicaux sont en relation de synonymie s'il y a une équivalence sémantique entre eux. La synonymie est une relation centrale en TALN mais demeure problématique car l'équivalence sémantique n'est pas formalisable en une relation d'équivalence. Il n'est pas nécessaire de vérifier la transitivité [Lewis 1952], et elle peut, au moins partiellement, être confondue avec l'hyperonymie, quand l'équivalence est réduite à une proximité sémantique [Resnik 1999]. Une solution possible dans une approche vectorielle est de définir une synonymie contextuelle (également proposée dans [Gwei et al. 1987]) représentée par une relation entre trois arguments. Cette relation dispose alors des propriétés d'une relation d'équivalence

quand le troisième terme (dit pivot) est fixé. La solution suggérée est nommée *synonymie relative* [Lafourcade et Prince 2001]. La représentation fonctionnelle est la suivante. Nous définissons la fonction de synonymie relative Syn_R entre trois vecteurs A , B et C , le dernier jouant le rôle de pivot comme suit :

$$Syn_R(A, B, C) = D_A(\gamma(A, C), \gamma(B, C)) = D_A(A \oplus (A \otimes C), B \oplus (B \otimes C)) \quad (5)$$

L'interprétation de cette fonction correspond à un test de proximité thématique de deux sens (A et B), chacun augmenté de ce qu'il a en commun avec le troisième (C). L'intérêt d'une telle solution est qu'elle contourne l'effet de la polysémie et restituant une forme de transitivité et une symétrie. Cependant, elle ne fournit pas de distinction réelle, pour un même mot, entre (1) un hyperonyme d'un sens et (2) un vrai synonyme du mot. Ce problème est discuté dans la section suivante, par l'introduction de notions plus flexibles comme la substitution de termes.

Antonymie. *Deux items lexicaux sont en relation d'antonymie s'il existe une symétrie entre leur composants relativement à un axe.* Trois types de symétries ont été définies, inspirées des recherches en linguistique [Palmer 1976]. A titre d'exemple, on n'expose ici que l'*antonymie complémentaire* proposé par [Schwab et al. 2002] (la même méthode étant appliquée aux autres types). Les antonymes complémentaires sont des couples comme *existence/inexistence*, *présence/absence*. L'axe de symétrie, ou encore la référence, correspond aux propriétés sur lesquelles se fondent la projection comme dans les couples père/mère ou père/fils. La représentation fonctionnelle est la suivante : la fonction $AntiLex_S$ retourne les n plus proches antonymes d'un terme A dans le contexte défini par un terme C et la référence R . La fonction partielle $AntiLex_R$ a été définie pour tenir compte du fait que, dans la plupart des cas, le contexte suffit également à définir l'axe de symétrie. $AntiLex_B$ est définie pour déterminer l'axe de symétrie plutôt que le contexte. En pratique, nous avons $AntiLex_B = AntiLex_R$. La dernière fonction est la *fonction d'antonymie absolue*. Les équations sont donnée ci-dessous :

$$\begin{aligned} A, C, R, n &\rightarrow AntiLex_S(A, C, R, n) \\ A, X, n &\rightarrow AntiLex_R(A, X, n) = AntiLex_S(A, X, X, n) \quad \text{avec } X = (C|R) \\ A, n &\rightarrow AntiLex_A(A, n) = AntiLex_S(A, A, A, n) \end{aligned} \quad (6)$$

Une implémentation de ces fonctions dans le modèle des vecteurs conceptuels est détaillée et commentée dans [Schwab et al. 2002]. Contrairement, à la synonymie, ces fonctions sont réalisées partiellement sous forme de réseaux sémantiques et pour partie sous forme de vecteurs conceptuels. En effet, certaines oppositions sont de nature d'abord lexicale, et peuvent éventuellement s'étendre continuellement dans l'espace des sens. Par recherche sur les réseaux sémantiques et les voisinage de termes, on essaye de trouver s'il existe un terme proche sur vecteur antonyme calculé.

3.3 Construction des vecteurs conceptuels

La construction des vecteurs conceptuels est basée sur des définitions de termes provenant de différentes sources (dictionnaires, listes de synonymes, indexation manuelle, extraction diverses depuis des corpus, etc.). Les vecteurs des termes sont construits à partir des définitions. Les définitions sont donc analysées morphologiquement et syntaxiquement (outil SYGMART de [Chauché 1984]) et le vecteur conceptuel correspondant est calculé selon la procédure décrite ci-dessous.

Après filtrage en fonction des attributs morphosyntaxiques, on attache à chaque feuille de l'arbre d'analyse un vecteur conceptuel qui est calculé à partir des vecteurs des k définitions du terme correspondant. La méthode la plus simple (pas nécessairement la meilleure) est de calculer le vecteur moyen : $V(w) = V(w.1) \oplus \dots \oplus V(w.k)$. Si le terme est inconnu (absent du dictionnaire), c'est le vecteur nul qui est proposé.

Les vecteurs sont ensuite propagés vers la racine de l'arbre. Considérons un nœud N de l'arbre ayant p dépendants $N_i (1 \leq ip)$. Le nouveau vecteur de N est la somme pondérée de tous les vecteurs de N_i : $V(N) = \alpha_1 N_1 \oplus \dots \oplus \alpha_p N_p$. Les poids α sont relatifs aux fonctions syntaxiques du nœud. Par exemple, un gouverneur disposera d'un poids augmenté ($\alpha = 2$) par rapport à un nœud standard ($\alpha = 1$), de façon à différencier les vecteurs calculés pour *'bateau à voile'* et *'voile de bateau'*.

Une fois que le vecteur de la racine de l'arbre est calculé, on effectue une propagation vers le bas jusqu'aux feuilles. Le vecteur d'un nœud est contextualisé par ses parents : $V'(N_i) = V(N_i) \oplus \gamma(N_i, N)$. Cette succession de propagations descendantes et montantes est itérée jusqu'à convergence du vecteur conceptuel de la racine ou à défaut jusqu'à un nombre maximum de cycles. En effet, des phénomènes d'oscillations peuvent se produire pour des phrases particulièrement ambiguës. Ces phases descendante et montante permettent de propager, pour chaque terme, le contexte que constituent les autres termes.

Cette méthode d'analyse permet de former, à partir de (1) vecteurs conceptuels déjà existants et (2) des définitions, de nouveaux vecteurs. Elle demande un amorçage à l'aide d'un noyau constitué de vecteurs pré-calculés, en général manuellement indexés sur quelques (environ 2000) termes particulièrement fréquents ou difficiles. Construire un système d'apprentissage cohérent, sans doute nécessite de modéliser les nombreuses relations sémantiques entre items lexicaux, et parmi ces relations la synonymie, l'antonymie et l'hyponymie semblent particulièrement importantes. Une base de vecteurs conceptuels adéquate est obtenue après quelques itérations du système d'apprentissage. On a constaté expérimentalement que le système avait globalement convergé au bout d'environ 20 cycles, mais des variations peuvent survenir selon la nature des définitions (la polysémie et la taille étant des facteurs aggravants). Au moment de l'écriture de cet article plus de 130000 items lexicaux ont été appris pour le français correspondant à plus de 490000 vecteurs. Environ 2000 vecteurs sont directement concernés par l'antonymie, alors que la plupart sont concernés par la synonymie et l'hyponymie. La modélisation et l'implémentation de nos fonctions lexicales a permis de considérablement améliorer la représentation de l'ensemble des vecteurs.

3.4 Importance de l'hyponymie pour le modèle de vecteurs conceptuels

Une approche incluant l'hyponymie s'avère particulièrement pertinente pour améliorer la construction des vecteurs, car la plupart de ceux-ci est construite via l'analyse des définitions hyponymiques (disponibles sur le Web ou dans des dictionnaires en ligne). En pratique, nous avons extrait environ 250000 liens pondérés d'hyponymie pour environ 40000 substantifs (soit 120000 acceptions). En fait, l'ensemble des fonctions lexicales semble être particulièrement utile pour une telle tâche. De façon symétrique, les relations entre les vecteurs sont cruciales pour une telle approche guidée par les données : essayer d'extraire les relations sémantiques de corpus ([Yarowsky 1992]) et ainsi construire une ontologie d'un domaine, ou essayer d'organiser l'information des corpus en se basant sur des hiérarchies *is-a* ([Lee et al. 1993], [Resnik 1999]).

4 Le calcul de l'hyponymie

Notre approche étant à la fois guidée par les données et basée sur des hiérarchies, nous définissons deux mesures qui sont utilisées conjointement. Le modèle de co-occurrence permet de formaliser la substitution de termes et l'approximation sémantique (avec un aspect taxonomique). Le modèle d'inclusion (de propriétés) directement basé sur les vecteurs conceptuels : une sous-classe contient les propriétés de sa superclasse. La combinaison de ces deux modèles permet de construire des réseaux sémantiques partiels qui constitue une *mémoire incrémentale* qui instancie pour des termes les mesures de substitution et d'approximation sémantique. Conjointement à leur construction, ces réseaux sémantiques sont aussi utilisés comme support pour la propagation et l'affinage de vecteurs conceptuels.

4.1 Modèle de co-occurrence

On définit deux mesures entre un terme t et un **candidat hyperonyme** h :

$$M_S(t, h) = \frac{|H \cap T|}{|T|} \quad \text{et} \quad M_T(t, h) = \frac{|H \cap T|}{|H|} \quad (7)$$

T (resp. H) représente l'ensemble de documents dans un corpus donné, qui contiennent le terme t (resp. h). $H \cap T$ représente l'ensemble de documents contenant les deux termes h et t . M_S se rapproche de la notion de *rappel* et M_T de celle de *précision* en Recherche d'Information.

Substitution de termes et approximation sémantique : vers une hiérarchie *is-a* locale. Si on ajoute l'hypothèse que h est un hyperonyme possible, alors M_S évalue dans quelle mesure t peut être substitué par h et constitue ce que nous appellerons une mesure de substitution. De façon similaire, M_T est une mesure d'évaluation de taxonomie, et correspond à une *approximation sémantique* (de la même façon que l'on peut approximer *cheval* par *mammifère*).

Nous avons conduit une expérimentation en utilisant Google (www.google.com) et le nombre de documents trouvés pour chaque requête. Statistiquement, cette approximation semble largement compensée par la grande taille de corpus que représente le Web. Par exemple, nous avons les résultats suivants pour le terme *avion* :

$$\begin{array}{l|l} \text{aéronef} / M_T = 0.26, M_S = 0.025 & \text{appareil volant} / M_T = 0.53, M_S = 0.0007 \\ \text{appareil} / M_T = 0.12, M_S = 0.17 & \text{aéronef plus lourd que l'air} / M_T = 0.52, M_S = 0.00004 \end{array}$$

Le meilleur terme de substitution est *appareil* (plus élevée M_S), mais c'est également la pire des approximations sémantiques. Le terme *appareil volant* est le plus précis (M_T la plus forte), mais ne peut pas raisonnablement être utilisé à la place d'*avion*. Pour un exemple plus détaillé avec *cheval* se référer à l'annexe.

Dans le cas de *cheval*, nous créons un nouveau sens (*cheval/moyen de transport* et *cheval/viande*) et les lions à leur hyperonymes respectifs. Le problème est qu'en commençant à partir de définitions vectorisées, il n'y a pas moyen d'appréhender ces nouveaux sens car ils ne sont pas (encore) identifiés. Pour surmonter cette difficulté, nous lions chacun de ces nouveaux sens au sens le plus proche (selon la distance thématique) déjà existant. Dans l'exemple précédent, nous avons :

- *cheval/moyen de transport* est plus proche de *cheval/mammifère* que de *cheval/unité de puissance*. Cette relation peut être vérifiée sur leurs vecteurs respectifs, et (par-

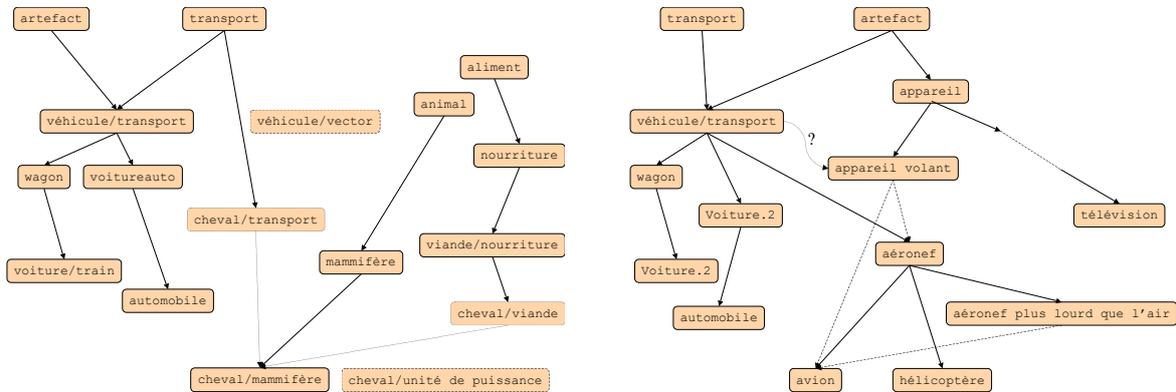


Figure 1: Construction des réseaux sémantiques partiels par insertion d'hyperonymes. L'ajout des hyperonymes trouvés amène à l'identification (1) des propriétés saillantes dans les sens déjà présents ou (2) de nouveaux sens. La distance thématique est utilisée comme un sélecteur de sens. L'approximation sémantique est préférée à la mesure de substitution qui est utilisée de façon inversée lors de la recherche des sens possibles pendant d'une analyse thématique.

fois) par la recherche de patrons lexicaux caractéristiques sur des définitions encyclopédiques.

- *cheval/viande* est plus proche de *cheval/mammifère* que de *cheval/unité de puissance*.

Ces deux mesures sont particulièrement utiles lors d'un processus d'analyse sémantique. En effet, la constitution d'un réseau lexical sur la base des deux mesures M_S et M_T permet de reconnaître les hyperonymes lointains de substitution (faible M_T et fort M_S). Par exemple, on peut lors de l'analyse s'apercevoir que la cohérence thématique du texte est bien plus forte lorsque l'on (re)substitue *avion* à *appareil*. Les candidats à substituer sont déterminés par la structure du réseau et la distance angulaire entre le candidat et le contexte global. Il s'agit d'un processus itératif globalement convergent. Ainsi, en vue de l'analyse, on procède à l'inverse de l'auteur, qui avait, directement ou non, remplacé les termes précis par des hyperonymes plus vagues à des fins stylistiques (par exemple, suppression des répétitions).

4.2 Modèle d'inclusion

On note $V(A)$ le vecteur conceptuel associé au terme A . Si A est un hyperonyme de B , alors les propriétés de A sont incluses dans celles de B . Cela peut-être mesuré via l'intersection de vecteurs et la distance angulaire :

$$H(A, B) \Rightarrow D_A(V(A), \gamma(V(A) V(B))) \leq D_A(V(B), \gamma(V(A), V(B))) \quad (8)$$

Par exemple, nous avons les mesures suivantes entre *cheval/mammifère* et *mammifère* :

$$D_A(V(\text{cheval}), \gamma(V(\text{cheval}) V(\text{mammifère}))) = 0.41$$

$$D_A(V(\text{mammifère}), \gamma(V(\text{cheval}) V(\text{mammifère}))) = 0.25$$

De ces résultats, nous déduisons que les propriétés de ‘*mammifère*’ sont incluses dans celles de ‘*cheval*’. De plus, si par ailleurs on sait que ‘*cheval*’ et ‘*mammifère*’ sont en relation d’hyperonymie, on en déduit que ‘*mammifère*’ est l’hyperonyme. Bien sûr, pour connaître l’existence de cette relation entre deux termes, l’extraction de schémas caractéristiques (par exemple *X est un type de Y*) combiné au modèle de co-occurrence fournissent des indices forts.

4.3 Limites des modèles

Le modèle fonctionne correctement pour des vecteurs qui ont été calculés à partir de définitions hyperonymiques. L’évaluation se fait manuellement, ponctuellement pour les mesures sur certains termes, et globalement sur les réseaux sémantiques construits. Cependant, pour des termes très généraux, où les définitions ont tendance à être hyponymiques (une collection d’exemples), l’inclusion de vecteur est renversée. Plus précisément, la position limite où a lieu le renversement est appelée l’*horizon*. L’horizon est globalement constituée des concepts feuilles de la taxonomie sur laquelle est construit l’espace vectoriel des sens. Quand une définition conduit à un nouveau vecteur (ou à sa révision), les vecteurs des termes présents dans la définition sont mélangés. En conséquence, le vecteur résultat est relativement plat en comparaison des principaux concepts impliqués. Nous avons une mesure formelle de *profil de vecteur* qui est le *coefficient de variation CV* :

$$CV(X) = \frac{s(X)}{\mu(X)} \quad \text{avec} \quad s^2(X) = \frac{\sum_i (x_i - \mu(X))^2}{n} \quad (9)$$

Le terme CV est l’écart-type des composants x_i des vecteurs divisé par leur moyenne μ . C’est une valeur sans unité. Par définition, CV n’est défini que pour les vecteurs non nuls. Si $CV(A) = 0$ alors A fait un angle de $\pi/4$ avec chacun des axes, et à sa valeur maximum (environ 29 quand $n = 873$), nous avons un vecteur booléen (une seule composante à 1 et toutes les autres à 0).

Au-delà de l’horizon, nous avons :

$$H(A, B) \Rightarrow D_A(V(A), \gamma(V(A), V(B))) \geq D_A(V(B), \gamma(V(A), V(B))) \quad (10)$$

Comment évaluer de quel côté de l’horizon un vecteur donné se situe ? En lui-même, le coefficient de variation évalue uniquement la forme générale du vecteur et son taux de *conceptualité* relativement au jeu de concepts initial. Nous avons deux moyens de résoudre ce problème :

1. De se concentrer sur une approche lexicale combinant des fonctions et informations lexicales aux vecteurs. Le modèle de co-occurrence constitue une réponse possible mais on peut également penser aux graphes conceptuels à *la Sowa*. C’est une approche multi-représentation dont la généralisation abusive risque toutefois d’être difficile à gérer.
2. D’inclure comme dimensions supplémentaires dans l’espace vectoriel, chaque concept de la hiérarchie et non pas seulement les concepts terminaux (les feuilles). Cette solution est clairement partielle, car elle ne peut pas être une réponse à celui de la polysémie si on se place au niveau du terme et non celui de l’acception.

5 Discussion et Conclusion

Nous donnons quelques résultats numériques en annexes. Les expériences que nous avons conduites sur une collection de substantifs (et termes composés), ont mis en lumière le problème

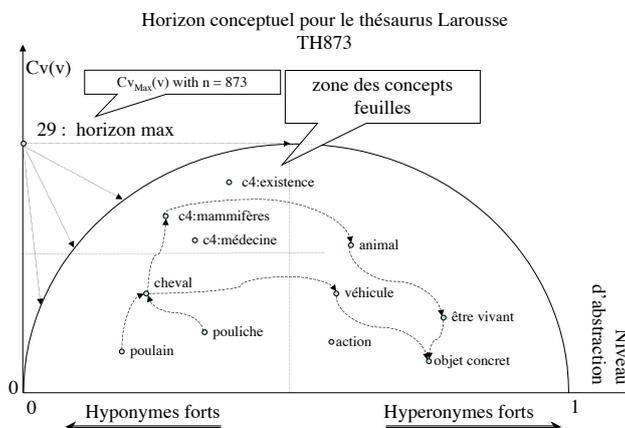


Figure 2: Représentation graphique en 2D de l'horizon conceptuel. L'horizon se situe au plus haut niveau du coefficient de variation (des vecteurs de la base) qui est le plus bas niveau de la hiérarchie du thesaurus. Sur le côté gauche, nous avons une généralisation des concepts, où de façon similaire, le mélange des vecteurs tend à faire baisser le coefficient de variation alors que le niveau d'abstraction augmente. Les termes préfixés par *c4:* correspondent aux concepts de niveau 4 définis dans [Larousse 2001].

posé par l'horizon conceptuel. Cet horizon se situe au niveau le plus bas de la hiérarchie de concepts (nous avons utilisé [Larousse 2001] pour le français). De par la nature de la composition de vecteurs, le modèle d'inclusion doit être inversé quand les termes se situent au-delà de cet horizon.

La détection du passage de part et d'autre de l'horizon est réalisé par des modèles lexicaux. De façon plus précise, elle peut se baser sur le modèle de co-occurrences mais aussi sur l'identification des hyperonymes. La présentation détaillée de la découverte des hyponymes n'est pas l'objet de cet article (et les méthodes sont très classiques, voir [Thelen et Riloff 2002] pour une approche récente d'extraction de catégories taxonomiques), cependant il est intéressant de noter que les termes les plus abstraits (qui correspondent à des classes taxonomiques de grande taille) conti-

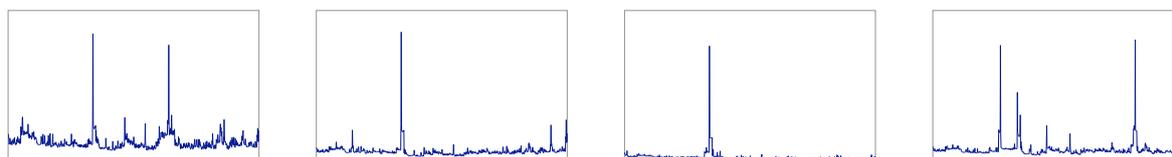


Figure 3: Représentations graphiques des vecteurs des termes *«poulain»*, *«cheval»*, *«Mammifères»* et *«animal»*. Le coefficient de variation augmente de gauche à droite jusqu'au troisième vecteur, celui de *«Mammifères»*, et diminue ensuite pour *«animal»*. En abscisse se trouvent les concepts, en ordonnée la valeur d'activation. Un vecteur (non nul) est plat, si tous les concepts ont même valeur. Son coefficient de variation CV est alors nul.

ennent un grand nombre d'hyperonymes. Selon notre modèle, les fonctions d'hyperonymie et d'hyponymie ne sont pas strictement symétriques (à la fois dans leur usage et leur comportement dans les corpus) et peuvent être utilisées conjointement pour renforcer la construction du réseau sémantique.

Mis à part la construction d'un réseau sémantique servant de base à une expertise de désambiguïsation lexicale et de support à la révision des vecteurs conceptuels, une application de notre modèle est la création d'un outil de paraphrasage. À partir d'un texte donné, un nouveau document sera généré où tout ou partie des termes sont substitué à leurs hyperonymes (ou quasi synonymes). Les résultats initiaux montrent que les paraphrases les plus naturelles sont celles qui maximisent la mesure de substitution et non la précision taxonomique. Un tel outil peut être utilisé pour évaluer globalement la pertinence de notre approche, mais également comme un prétraitement à de la traduction automatique.

Dans cet article, nous avons essayé de montrer comment il est possible de modéliser l'hyperonymie dans le cadre d'une sémantique lexicale vectorielle, en se basant sur une coopération entre les réseaux sémantiques et le modèle des vecteurs conceptuels. Après avoir évalué l'importance, pour la sélection lexicale, de fonctions lexicales comme la synonymie et l'antonymie ainsi que la construction et l'utilisation des vecteurs conceptuels, nous nous sommes concentrés sur l'hyperonymie qui semble plus difficile à aborder sous un angle purement numérique. Notre approche étant à la fois guidée par les données et par la hiérarchie, nous avons d'abord chercher à définir l'impact de l'hyperonymie par des mesures dans des corpus. On peut ainsi définir formellement la substitution de termes et l'approximation sémantique (incluant l'aspect taxonomique).

À partir du modèle théorique, qui combine réseaux sémantiques et vecteurs conceptuels, il a été possible de l'implémenter et de montrer comment l'inclusion a été traitée et quels résultats nous avons obtenus. Bien que satisfaisants, ces résultats tendent à refléter la nature polymorphe de l'hyperonymie : étant plus complexe qu'une simple relation *is-a*, l'hyperonymie doit être contrainte au sein de la tâche à entreprendre. Si l'enjeu est la correction ou l'explication de texte, alors la substitution de termes est une bonne utilisation des propriétés d'hyperonymie. Si la construction d'une taxonomie est l'objectif, alors l'approximation sémantique est un meilleur candidat.

Donc, de façon similaire aux fonctions comme la synonymie et l'antonymie, qui ont été restreintes en ajoutant la notion de *relativité* lorsqu'elles sont confrontées à des bases textuelles, l'hyperonymie n'apparaît pas comme absolue contrairement à la relation *is-a*. Il semble plus productif de décliner l'hyperonymie selon ses fonctions (psycholinguistique, entre autres) et de la définir selon l'utilisation que l'on peut en faire.

Références

- [Brachman et Schmolze 1985] Brachman Ronald J. and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, April–June 1985.
- [Chauché 1984] Chauché J. Un outil multidimensionnel de l'analyse du discours *COLING'84*, vol 1/1, pp. 11-15, 1984.
- [Chauché 1990] Chauché J. Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance. *TA Information*, 31(1): 17–24.1990
- [Deerwester et al. 1990] Deerwester S. and S. Dumais, T. Landauer, G. Furnas, R. Harshman, Indexing

- by latent semantic analysis. *Journal of the American Society of Information science*, 416(6): 391–407, 1990.
- [Fellbaum 1998] C. Fellbaum (ed). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts, 1998.
- [Gwei et al. 1987] Gwei G. M. and E. Foxley. A Flexible Synonym Interface with Application examples in CAL and Help environments. *The Computer Journal* 30 (6): 551–557, 1987.
- [Hearst 1998] Hearst M. A. *Automated discovery of Wordnet relations*, In C. Fellbaum ed. *WordNet : An Electronic Lexical Database* MIT Press, Cambridge, MA, 131–151, 1998.
- [Lafourcade et Prince 2001] Lafourcade M. and V. Prince. *Relative Synonymy and Conceptual Vectors NLPRS01*, pp. 127-134, 2001.
- [Larousse 2001] Larousse, *Thésaurus Larousse - des idées aux mots - des mots aux idées*. Larousse, 1re édition 1992, 2e édition 2001.
- [Lee et al. 1993] Lee J.H., M.H Kim and Y.J. Lee. Information Retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2), 188–207, 1993.
- [Lewis 1952] Lewis, C. I. *The modes of meaning*. in Linsky ed, "Semantics and the philosophy of language". Urbana. NY, 1952.
- [Miller et Fellbaum 1991] Miller G. A. and C. Fellbaum. Semantic Networks in English. in Beth Levin and Steven Pinker (eds.) *Lexical and Conceptual Semantics*, 197–229. Elsevier, Amsterdam, 1991.
- [Palmer 1976] Palmer F. R. *Semantics: A New Introduction*. Cambridge University Press, 1976.
- [Resnik 1995] Resnik P. *Using Information Contents to Evaluate Semantic Similarity in a Taxonomy*, *IJCAI-95*, 1995.
- [Resnik 1999] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, 95–130, 1999.
- [Resnik 1999] Resnik P. Disambiguating noun groupings with respect to WordNet senses. in S. armstrong, K. Church, P. Isabelle, E. Tzoukermann, S. Manzi and D. Yarowsky (eds.) *Natural Language Processing using Large Corpora*, Kluwer Academic, Dordrecht, 1999.
- [Rosch 1978] E. Rosch and B.B. Lloyd (Eds.) *Principles of Categorization* Cognition and Categorization, 27-48. Hillsdale, NJ: Erlbaum; 1978
- [Rodget 1852] Rodget P. *Thesaurus of English Words and Phrases* Longman, London, 1852.
- [Salton 1968] Salton G., *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York, 1968.
- [Salton et MacGill 1983] G. Salton and M. J. MacGill, *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
- [Sparck Jones 1986] Sparck Jones K. *Synonymy and Semantic Classification*. Edinburgh Information Technology Serie, 1986.
- [Schwab et al. 2002] Schwab D., M. Lafourcade and V. Prince *Antonymy and Conceptual Vectors*. In Proc of *COLING'02*, vol. 2/2, 2002.
- [Thelen et Riloff 2002] Thelen M., E. Riloff *A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts* In Proc of *EMNLP 2002*, vol1/1, 2002.
- [Yarowsky 1992] Yarowsky D. *Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora*, *COLING'92*, 454–460, 1992.

6 Annexe

Nous avons les résultats suivants pour le terme *cheval* :

mammifère / $M_T = 0.81$	$M_S = 0.0005$	(a)
animal / $M_T = 0.0986$	$M_S = 0.1523$	(a)
animal domestique / $M_T = 0.133$	$M_S = 0.0035$	(a)
type de mammifère / $M_T = 0.0481$	$M_S = 0.00002$	(a)
espèce / $M_T = 0.1376$	$M_S = 0.0857$	(b)
chevaux / $M_T = 0.4673$	$M_S = 0.2954$	(b)
équitation / $M_T = 0.3498$	$M_S = 0.0991$	(c)
représentation / $M_T = 0.0399$	$M_S = 0.0505$	(d)
jouet / $M_T = 0.1363$	$M_S = 0.0184$	(e)
jouet d'enfant / $M_T = 0.2387$	$M_S = 0.0004$	(e)
cheval de bois / $M_T = 0.2025$	$M_S = 0.0012$	(e)
femme / $M_T = 0.0363$	$M_S = 0.4012$	(f)
femme masculine / $M_T = 0.5692$	$M_S = 0.00003$	(f)
unité / $M_T = 0.033$	$M_S = 0.0647$	(g)
unité arbitraire / $M_T = 0.067$	$M_S = 0.00004$	(g)
unité de puissance / $M_T = 0.1042$	$M_S = 0.0003$	(g)

Nous avons ici plusieurs significations pour *cheval* : (a) l'animal, (b) la classe des chevaux ou l'espèce, (c) l'équitation, (d) la représentation d'un cheval, (e) le cheval de bois, (f) une femme masculine, (g) l'unité de puissance. Le terme *mammifère* est le plus précis pour une taxonomie mais *animal* est un meilleur terme de substitution. Le terme *espèce* est trop vague comparé à *chevaux*. *jouet d'enfant* est précis, mais *jouet* est un meilleur substitut. En général, les termes courts sont de meilleurs substituts (c'est en partie une conséquence du principe d'économie en linguistique) mais la plupart du temps ils sont, d'un point de vue taxonomique, relativement vagues ou ambigus. Nous avons les résultats suivants pour le terme *peinture* :

art / $M_T = 0.133$	$M_S = 0.6913$	(a)
art de peindre / $M_T = 0.649$	$M_S = 0.0016$	(a)
ouvrage / $M_T = 0.2248$	$M_S = 0.0955$	(b)
ouvrage d'un artiste / $M_T = 1.0$	$M_S = 0.00001$	(b)
matière / $M_T = 0.2543$	$M_S = 0.1644$	(c)
produit / $M_T = 0.2301$	$M_S = 0.1755$	(c)
produit à base de pigments / $M_T = 1.0$	$M_S = 0.00004$	(c)
produit à base de pigments en suspension / $M_T = 1.0$	$M_S = 0.00004$	(c)
produit à base de pigments en suspension dans un liquide / $M_T = 1.0$	$M_S = 0.00004$	(c)
couche / $M_T = 0.1443$	$M_S = 0.0876$	(d)
couche de couleur / $M_T = 0.4939$	$M_S = 0.0004$	(d)
description / $M_T = 0.2049$	$M_S = 0.1216$	(e)

Le terme *peinture* peut être : (a) l'*art*, (b) l'*objet peint*, (c) la *matière colorante*, (d) la *couche de peinture*, et (e) la *description*.