Vecteurs conceptuels et structuration émergente de terminologies

Mathieu Lafourcade — Violaine Prince — Didier Schwab

LIRMM (CNRS - Université Montpellier 2)
161, rue Ada - F-34392 Montpellier Cedex 5
{lafourca,prince,schwab}@lirmm.fr

RÉSUMÉ. *Cet article présente les principaux avantages du modèle vectoriel pour la sémantique lexicale. Outre une représentation robuste, ce modèle permet l'émergence de relations entre termes, comme celles de synonymie et d'antonymie relatives. Nous décrivons le formalisme vectoriel utilisé, ainsi que les fonctions de base qui permettent de déterminer la notion de proximité thématique. L'extension de la méthode d'indexation d'un terme issu d'un document de spécialité se base en particulier sur une notion de pliage et de dépliage de vecteurs entre espaces vectoriels. Tout terme défini par d'autres termes qui n'appartiennent pas forcément à la terminologie, va imposer l'union des bases génératrices, et donner lieu à un dépliage du vecteur dans une base plus grande. Nous montrons comment la distribution lexicale, l'antonymie et la synonymie agissent comme des révélateurs de structure et réalisent une mise en relation transversale (non liée aux liens ontologiques) et instantanée dans l'espace vectoriel étendu. Les apports des expériences effectuées permettent de nous focaliser sur l'intérêt de ce type de méthode pour les terminologies.*

ABSTRACT. *This paper presents some advantages of the conceptual vector model for lexical semantics. Besides being a robust representation, this model allows the emergence of relations between terms, as relative synonymy and antonymy. We describe the underlying model, as well as the basic functions which allow to define the notion of thematic proximity. The extension of the indexation mechanism of a term extracted from a document of some speciality domain is mainly based on the notion of vector folding and unfolding between vector spaces. Any term defined thanks to other terms which may not belong to the specialty terminology, will impose the union of the vector space generative families, and lead to vector unfoldings toward a larger base. We show how the lexical distribution, the antonymy, and the synonymy, play as structure spotlight and transversally brigde terms across extended vector spaces. Some experiments focus us on the interest of this kind of approach for terminology.*

MOTS-CLÉS : *vecteurs conceptuels, structuration lexicale, synonymie relative, antonymie relative, extension ontologique.*

KEYWORDS: *conceptual vectors, lexical structuration, relative synonymy, relative antonymy, ontological expansion.*

1. Introduction

La constitution automatique ou semi-automatique de bases terminologiques pour indexer des documents spécialisés dans un domaine donné est une tâche relativement ardue mais qui a déjà donné des fruits : elle a conduit les chercheurs à considérer la structuration de la terminologie comme une réponse au problème de son exploitation [Hamon et Nazarenko 2001]. Si la structuration permet, dans certains cas, de résoudre des questions relatives à la représentation, elle ne résout pas celui de la double appartenance d'un texte : un texte spécialisé comprend aussi bien des termes techniques que des termes généraux. De fait, cela impose l'usage d'un lexique général aussi bien que des lexiques terminologiques, quand il faut indexer ce texte, le classer, et surtout quand on veut l'utiliser comme base d'acquisition lexicale.

En outre, si la structuration permet de dériver des connaissances sur le lexique, elle n'a pas toujours un rôle de premier plan quand il est question d'**augmenter** ces lexiques terminologiques à partir de l'analyse de textes ou de définitions (parmi les travaux qui proposent une acquisition de ce type, on peut penser par exemple à [Barrière and Copeck 2001]). D'une part, ces textes contiennent forcément des termes généraux, ce qui nous ramène au problème précédent, et d'autre part, on s'aperçoit que la structure est un résultat local émergeant du calcul du sens, et non pas forcément une donnée stable fournie *a priori*. La polysémie due aux usages est en grande partie responsable de ce phénomène.

Notre objectif, en nous interressant à la terminologie a été : (**1**) d'explorer les potentialités du modèle que nous étudions (le modèle vectoriel) en termes de propositions de structuration de la terminologie ; (**2**) d'augmenter nos lexiques avec de l'information terminologique appropriée ; (**3**) d'être capable d'indexer et d'analyser thématiquement des textes spécialisés grâce aux structures émergentes ; (**4**) de proposer une application de (2) et (3) dans un domaine donné, ici l'économie, dans la mesure où nous pouvons bénéficier d'une ontologie spécialisée de type thésaurus, que nous savons vectoriser dans ce domaine. Dans cet article, nous montrons comment nous réalisons ces objectifs grâce à des mécanismes de structuration de la terminologie qui *émergent* de l'application de fonctions lexicales comme la synonymie et l'antonymie.

2. Problématique

L'exploitation, la plus automatique possible, d'un lexique terminologique pousse à se positionner sur deux points : le rôle du lexique général (vs. le lexique de spécialité) et la structuration de la terminologie pour indexer des textes susceptibles d'être spécialisés.

2.1. *Les rôles respectifs des lexiques*

On peut difficilement se passer de l'évocation d'un lexique général : l'économie de moyens que l'on pensait réaliser en constituant des lexiques spécialisés (versus un lexique général) est mise en échec. En réalité, le problème qui se pose est celui de la représentation : si on utilise une ontologie[1] pour rechercher des termes par appariement, alors la terminologie est un sous-ensemble de l'ontologie générale. En revanche, si on s'appuie, comme nous le faisons sur le modèle vectoriel, c'est au contraire la structure génératrice la plus petite connue qui sert de *base* au lexique général. Dans notre cas, les termes du lexique général (environ $65\,000$ entrées à ce jour) peuvent, de manière satisfaisante, se décliner en termes de composantes dans un espace vectoriel défini par une ontologie générale limitée à 873 concepts feuilles[2], et fondée sur le thésaurus Larousse [Larousse 2001]. La définition des vecteurs de concepts, qui constituent les briques élémentaires de la construction des vecteurs des différents sens de chaque terme, se fait à partir de l'organisation de ces même concepts en une ontologie.

Si 873 concepts servent de famille génératrice pour la représentation d'un nombre quelconque de termes (actuellement $65\,000$ mots), c'est que la *finesse* de représentation est ici relativement faible. Autrement dit, les termes[3] de spécialité ne peuvent se contenter d'une ontologie aussi peu précise, au risque d'être tous considérés comme des quasi-synonymes. Par exemple, les termes complexes tels que ‹*droit du travail*›, et ‹*économie de marché*› sont très proches dans leur description générale, puisqu'ils ont une composante très forte sur le concept *ÉCONOMIE ET DROIT* qui appartient à l'ensemble générateur du thésaurus. Le *maillage* du thésaurus général est trop large. Il faut donc proposer pour les termes de spécialité une possibilité de maillage fin. C'est pourquoi nous avons mené une expérience en utilisant une hiérarchie de concepts issue de l'OCDE [OCDE 1991], définissant environ $2\,000$ concepts feuilles sur la thématique économique. L'objectif que nous avons cherché à atteindre est **l'indexation lexicale de textes susceptibles d'être spécialisés**, avec un lexique général et des lexiques de spécialité (qu'on appelera terminologies), sur la base du modèle des vecteurs conceptuels. L'ontologie spécialisée choisie dans le domaine économique, à partir d'une arborescence fournie par des experts, offre un maillage beaucoup plus précis et décrit finement les termes de spécialité, en particulier les termes complexes (groupes nominaux pour la plupart) dont on sait qu'il est nécessaire de les repérer correctement [Bourrigault 1993] lorsque l'on vise une indexation fine.

---

1. Dans le sens communément admis actuellement qui est l'arborescence des notions fondamentales ou concepts d'un domaine. Si tant est que cette arborescence existe, elle est unique, et est acceptée comme un consensus d'expertise.
2. Pour le thésaurus Larousse, un concept est une expression ou un terme qui sert de notion fondamentale et à laquelle on fait référence pour classifier le vocabulaire.
3. Les termes sont des mots comme ‹*capitalisme*› ou des groupes nominaux comme ‹*droit des sociétés*›. Ils servent à exprimer des notions particulières et peuvent servir d'index.

2.2. *La structuration de la terminologie*

De nombreuses recherches proposent d'utiliser des terminologies préstructurées. Compte tenu de l'aspect *associatif* du modèle vectoriel utilisé, nous avons adopté la démarche symétrique. Nous avons plutôt choisi de faire émerger des structures dans la terminologie, qui sont datées et dynamiques. En effet, dans un environnement terminologique très fortement évolutif comme celui que nous avons mis en place, les structures sémantiques sont sujettes à modification. On a alors intérêt à rechercher les *indicateurs de structure* plutôt qu'à réifier des structures prédites (comme des relations *a priori*).

Dans cet article, nous indiquerons les principaux avantages du modèle vectoriel pour la sémantique lexicale (section 3) : outre une représentation robuste, ce modèle permet de faire émerger des relations entre termes, comme les relations de synonymie relative [Lafourcade et Prince 2001] et d'antonymie relative [Schwab 2001]. Nous décrirons ensuite rapidement le formalisme vectoriel utilisé, ainsi que les fonctions de base qui permettent de déterminer la notion de proximité entre termes (section 4). Les relations émergentes sont formalisées dans la section 5. L'extension de la méthode d'indexation de tout terme issu d'un document de spécialité est proposée dans la section 6 : elle se base en particulier sur une notion de pliage et de dépliage de vecteurs entre espaces vectoriels. Tout terme t, défini par d'autres termes qui n'appartiennent pas forcément à la terminologie, va imposer l'union des bases génératrices, et donc donner lieu à un dépliage du vecteur dans une base plus grande. La section 7 montre comment la distribution lexicale, l'antonymie et la synonymie agissent comme des révélateurs de structure et réalisent une mise en relation transversale (non liée aux liens ontologiques) et instantanée dans l'espace vectoriel étendu. Nous conclurons enfin sur les apports de l'expérience effectuée, en nous focalisant sur l'intérêt de ce type de méthode pour enrichir les terminologies.

3. Modèle vectoriel pour la sémantique lexicale

Le modèle vectoriel n'est pas récent, puisqu'il a été au départ introduit par Salton en informatique documentaire [Salton 1968]. Sa réhabilitation dans les recherches en TALN est en revanche relativement récente, car elle a été essentiellement motivée par la mise à disposition des chercheurs de grandes bases de textes grâce au web en particulier, alors que précédemment, ces recherches passaient par des phases ardues de constitution de corpus d'expérience. L'approche que nous avons s'inspire de la version de 1983 du modèle vectoriel de Salton [Salton and MacGill 1983], mais elle en diffère en ce que nous faisons l'hypothèse qu'il existe un jeu de concepts prédéterminé qui peut jouer le rôle d'ensemble générateur et que ce jeu est celui défini par les lexicologues quand ils réalisent un thésaurus [Chauché 1990]. Les concepts de cet ensemble sont par définition interdépendants : la famille considérée n'est pas libre et ne constitue pas une base vectorielle proprement dite. Cette interdépendance est aussi attestée dans un modèle comme LSA [Deerwester *et al.* 1990] qui non seulement la reconnaît mais aussi l'exploite.

Le modèle vectoriel a été appliqué par Salton à l'indexation et à la recherche d'information textuelle en 1988 [Salton 1988]. Si ce dernier utilisait une analyse de surface par mots-clés pour alimenter ses vecteurs, notre démarche s'en distingue nettement : elle se base explicitement pour son calcul sur la géométrie et les variables morpho-syntaxiques des arbres d'analyse structurelle issus du texte. D'une façon générale, les documents sont traités indépendamment les uns des autres, alors que dans LSA, le traitement se fait de manière liée. De plus nous mettons l'accent sur la sélection lexicale en contexte (voir Bourrigault 1993 *op. cit.*) alors que des travaux comme celui de [Resnik 1995] font un usage exclusif de taxonomies.

3.1. *Avantages du modèle vectoriel pour la représentation du sens*

Le modèle de vecteurs conceptuels s'appuie sur la projection dans un modèle mathématique de la notion linguistique de champ sémantique. Tout terme (lexie) et tout concept est projetable sur les vecteurs de la famille génératrice, et est donc représenté par un vecteur *conceptuel*. Mieux encore, on peut calculer le thème de tout segment de texte tel que documents, paragraphes, syntagmes, etc. sous forme de vecteur conceptuel : c'est le *sens* du segment en question [Lafourcade et Sandford 1999]. Cette représentation homogène du sens, quelle que soit la granularité, est très avantageuse pour la classification des textes, l'indexation et la recherche évoluée d'information.

De plus, la représentation vectorielle ne fait aucune hypothèse *a priori* sur les relations conceptuelles. L'ontologie de départ mise à part, on ne se fonde sur aucune relation casuelle pour dériver du sens, et on n'inclut aucune contrainte sémantique. C'est un modèle purement calculatoire qui donne une *image* sémantique instantanée dans un état donné du dictionnaire conceptuel. Ce dernier est en apprentissage permanent, avec augmentation des définitions dès lors qu'une nouvelle source lexicologique électronique est disponible.

3.2. *Méthode d'acquisition lexicale*

Le principe du dictionnaire fondé sur le modèle des vecteurs conceptuels est celui de l'apprentissage de définitions et de concepts à partir de toute source lexicologique. Chaque définition de dictionnaire, expression en langage naturel fournie par des lexicologues, est analysée avec l'analyseur morphosyntaxique SYGMART[4], et un arbre d'analyse est produit. À partir de là ; des pondérations sont calculées et un vecteur conceptuel est produit pour représenter le sens donné par l'analyse de la définition. Ce vecteur entre alors dans le calcul du vecteur conceptuel du terme défini, ce qui fait que tout vecteur conceptuel (sauf ceux correspondant aux concepts de la famille génératrice) est modifié au fur et à mesure de l'analyse des définitions. Ce qui varie, c'est la valeur de la composante, et donc ce que nous appellerons par la suite *intensité*. L'avantage d'un tel système est qu'il peut acquérir non seulement de nouvelles lexies,

4. Développé par Jacques Chauché.

mais aussi de nouveaux sens à des lexies données en fonction d'un réarrangement des intensités que prennent les concepts sollicités.

3.3. *Relations sémantiques induites*

Dans cet espace vectoriel conceptuel, on sait définir une notion de proximité sémantique en calculant une distance angulaire entre vecteurs (section 4.2). Cela signifie que l'on a une représentation de sens *proches*, sans pour autant valoriser correctement cette proximité. Le formalisme développé ci-après amène quelques remarques. On ne sait pas bien encore décliner cette proximité en relation d'hyperonymie ou d'hyponymie, qui sont caractéristiques des ontologies. En revanche, on arrive assez bien à mettre en valeur des relations transverses telles que la synonymie et l'antonymie et qui sont très utiles lorsqu'il s'agit justement de faire émerger une microstructuration. Les paragraphes suivants définissent les propriétés générales de ces fonctions lexicales telles que nous les avons expérimentées dans [Lafourcade et Prince 2001] (*op. cit.*) et [Schwab 2001] (*op. cit.*).

3.3.1. *Synonymie relative*

La synonymie est une *relation d'équivalence* permettant de substituer un terme (ou un segment) à un autre terme (ou segment), sans modifier le sens global de l'énoncé [Sparck Jones 1986]. En tant que relation lexicale, la synonymie n'a malheureusement pas les bonnes propriétés des relations mathématiques d'équivalence, simplement parce qu'à cause de la polysémie, la propriété de transitivité n'est pas vérifiée [Fischer 1973]. Il y a même des cas où la symétrie aussi n'est pas vérifiée (un hyperonyme peut être donné comme synonyme pour son hyponyme, mais pas l'inverse) [Cruse and Togia 1995].

Pour pouvoir néanmoins exploiter les propriétés d'équivalence qui sont fort utiles, nous avons défini une synonymie relative, c'est-à-dire une évaluation de la possibilité de substituer un terme (un segment, un concept...) à un autre, dans le contexte d'un troisième, pratique d'ores et déjà admise dans [Gwei and Foxley 1987]. Nous avons montré [Prince 1991] qu'il s'agissait d'une relation de pseudo-équivalence (en ce qu'elle est pseudo-transitive), et nous en avons proposé une démonstration dans [Lafourcade et Prince 2001] dans le cadre des vecteurs conceptuels, sur le lexique général. Nous en donnons la formalisation dans la prochaine section.

3.3.2. *Antonymie relative*

Habituellement, l'antonymie est définie comme une notion d'incompatibilité entre deux termes. À la lumière de la représentation vectorielle, nous préférons plutôt considérer une notion de *symétrie* qui se définit comme suit : *deux termes sont en relation d'antonymie si on peut exhiber une symétrie de leurs traits sémantiques par rapport à un axe.* La symétrie peut se décliner de différentes manières selon la nature de son support. On distingue, comme supports de symétrie :

– une **propriété** affectant une valeur étalonnable (valeur élevée, valeur faible) : par exemple, *chaud, froid* sont des valeurs symétriques de température, sur une échelle implicite.

– l'**application d'une propriété** (applicable/non applicable, présence/absence) : par exemple, *informe* est antonyme de tout ce qui a une forme, *insipide, incolore, inodore, etc.* de tout ce qui pourrait avoir saveur, couleur, odeur, ... [Justeson and Katz 1991].

– l'**existence d'une propriété** ou d'un **élément considéré comme symétrique par l'usage** (e.g. *soleil/lune*), ou par des **propriétés naturelles ou physiques des objets considérés** (e.g. *mâle/femelle, tête/pied* ...). Les antonymes vont alors par *paires* [Fellbaum 1995].

Notre idée est que les constructions d'antonymes sont dépendantes du type de support de symétrie. Il peut alors exister plusieurs types d'antonymes pour un même terme, comme il peut ne pas en exister d'évidents, si la symétrie n'est pas immédiatement décelable. En tant que fonction lexicale, comparée à la synonymie, on peut dire que si la synonymie est la recherche de la ressemblance avec comme test la substitution (*x est synonyme de y si x peut "remplacer" y*), l'antonymie est la recherche de la symétrie avec comme test la recherche du support de la symétrie (*x est antonyme de y s'il existe un support de symétrie t tel que x symétrique de y par rapport à t*). Par exemple, ‹*chaud*› est antonyme de ‹*froid*› car ‹*température*› offre un support de symétrie.

De même que pour la synonymie, l'antonymie s'apprécie toujours en contexte. Par exemple, ‹*frais*› peut être le contraire de ‹*tiède*›, ‹*chaud*›, ‹*racorni*›, ‹*flétri*›, ‹*maladif*›, ‹*rassis*›, ‹*confit*›, ‹*sec*›, ‹*surgelé*›, ‹*pourri*›, ... La prochaine section montre comment nous avons formalisé la représentation du sens en vecteurs conceptuels, les règles de composition des vecteurs, et les fonctions associées aux relations sémantiques de synonymie et d'antonymie décrites ci-dessus.

4. Le modèle des vecteurs conceptuels

4.1. *Principe*

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, le vecteur V_A est la description (en extension) des activations des concepts de $\mathcal{C}$. Par exemple, les sens de ‹*ranger*› et de ‹*couper*› peuvent être projetés sur les concepts suivant (les CONCEPT[*intensité*] étant ordonnés par intensité décroissante) :

V_{ranger} = (CHANGEMENT[*0.84*], VARIATION[*0.83*], ÉVOLUTION[*0.82*], ORDRE[*0.77*], SITUATION[*0.76*], STRUCTURE[*0.76*], RANG[*0.76*] ...).

V_{couper} = (JEU[*0.8*], LIQUIDE[*0.8*], CROIX[*0.79*], PARTIE[*0.78*] MÉLANGE[*0.78*] FRACTION[*0.75*] SUPPLICE[*0.75*] BLESSURE[*0.75*] BOISSON[*0.74*] ...).

La description du processus d'apprentissage calculant les valeurs respectives des intensités pour chaque coordonnées d'un vecteur est exposé dans [Lafourcade 2001]. Il est clair, que pour des vecteurs denses (ayant très peu de coordonnées nulles), l'énumération des concepts activés est vite fastidieuse et surtout difficile à évaluer. On préférera en général procéder par sélection de termes thématiquement proches. Par exemple, les termes proches (et ordonnés par distance thématique décroissante) des mots ‹*ranger*› et ‹*couper*› sont :

‹***ranger***› : ‹*trier*›, ‹*cataloguer*›, ‹*sélectionner*›, ‹*classer*›, ‹*distribuer*›, ‹*grouper*›, ‹*ordonner*›, ‹*répartir*›, ‹*aligner*›, ‹*caser*›, ‹*arranger*›, ‹*nettoyer*›, ‹*distribuer*›, ‹*démêler*›, ‹*ajuster*› ...	‹***couper***› : ‹*cisailler*›, ‹*émincer*›, ‹*scier*›, ‹*tronçonner*› , ‹*ébarber*›, ‹*entrecouper*›, ‹*baptiser*›, ‹*recouper*›, ‹*sectionner*›, ‹*bêcher*›, ‹*hongrer*›, ‹*essoriller*›, ‹*rogner*›, ‹*égorger*›, ‹*écimer*›, ...

En pratique, plus $\mathcal{C}$ est grand, plus fines seront les descriptions de sens offertes par les vecteurs, mais plus leur manipulation informatique peut être lourde, surtout si l'on traite beaucoup de données. on rappelle que dans nos expérimentations sur le lexique général, $\dim(\mathcal{C}) = 873$, ce qui correspond au niveau 4 des concepts définis dans (Larrousse, *op. cit.*) La construction d'un lexique conceptuel (ensemble de triplets *(mot, variables morphologiques, vecteur)*) est réalisée automatiquement à partir de corpora (de définitions, de thésaurii, etc. (Lafourcade *op. cit.*)). Au moment de l'écriture de cet article, le corpus du français représente environ $210\,000$ définitions correspondants à $65\,000$ mots vedettes (pour $31\,000$ mots monosémiques et $34\,000$ mots polysémiques – pour ces derniers le nombre moyen de définitions, certaines éventuellement redondantes, étant de 4.61).

4.2. *Distance angulaire*

Il est souhaitable de pouvoir mesurer la proximité entre les sens représentés par deux vecteurs (et donc celle de leur mot associé). Soit $Sim(X, Y)$ la mesure de *similarité*, utilisée habituellement en recherche d'informations, entre deux vecteurs définie selon la formule (1) ci-dessous (avec "$\cdot$" étant le produit scalaire). On notera que l'on suppose ici que les composantes des vecteurs sont toujours positives ou nulles (ce qui n'est pas nécessairement le cas). Enfin, nous définissons une fonction de *distance angulaire* D_A entre deux vecteurs X et Y selon la formule (2).

$$Sim(X, Y) = \cos(X, Y) = \frac{X \cdot Y}{\|X\| \times \|Y\|} \quad (1)$$

$$D_A(X, Y) = \arccos(Sim(X, Y)) \quad (2)$$

Intuitivement, cette fonction constitue une évaluation de la *proximité thématique* et est en pratique la mesure de l'angle formé par les deux vecteurs. On considérera, en général, que pour une distance $D_A(X, Y) \leq \pi/4$ (soit environ $0,78$ radians ou encore 45 degrés), X et Y sont sémantiquement proches et partagent des concepts. Pour

$D_A(X,Y) \geq \pi/4$, la proximité sémantique de A et B sera considérée comme faible. Aux alentours de $\pi/2$ (soit environ $1,57$ radians ou 90 degrés), les sens sont sans rapport. La synonymie (dans son acception la plus large) est incluse dans la proximité thématique, cependant elle exige, de plus, la concordance des catégories morphosyntaxiques. L'inverse n'est évidemment pas vrai.

La distance angulaire est une vraie distance (contrairement à la mesure de similarité) et elle vérifie les propriétés de réflexivité (3), symétrie (4) et inégalité triangulaire (5) (qui peut jouer un rôle de pseudo-transitivité) :

$$D_A(X,X) = 0 \tag{3}$$

$$D_A(X,Y) = D_A(Y,X) \tag{4}$$

$$D_A(X,Y) + D_A(Y,Z) \geq D_A(X,Z) \tag{5}$$

Par définition, nous posons : $D_A(\vec{\mathbf{0}},\vec{\mathbf{0}}) = 0$ et $D_A(X,\vec{\mathbf{0}}) = \pi/2$ pour tout X avec $\vec{\mathbf{0}}$ dénotant le vecteur nul[5]. On considérera, en toute généralité, l'extension du domaine image de D_A à $[0,\pi]$ afin de comparer des vecteurs ayant des composantes négatives. Cette généralisation ne change pas les propriétés de D_A. On remarquera, de plus, que la distance angulaire est insensible à la norme des vecteurs (α et β étant des scalaires) :

$$D_A(\alpha X, \beta Y) = D_A(X,Y) \quad \text{avec} \quad \alpha\beta > 0 \tag{6}$$

$$D_A(\alpha X, \beta Y) = \pi - D_A(X,Y) \quad \text{avec} \quad \alpha\beta < 0 \tag{7}$$

Par exemple[6] dans le tableau qui suit, nous avons les distances angulaires (en radian) entre les vecteurs de plusieurs termes. Le tableau est symétrique (à cause de la symétrie de D_A) et la diagonale est toujours égale à 0 (à cause de la réflexivité de D_A). On remarquera qu'une valeur prend toute sa signification relativement à une autre. En particulier, il est satisfaisant d'avoir : (a) $d_1 \leq d_3$ et $d_2 \leq d_3$ ce qui correspond bien au fait que ‹*trier*› et ‹*ordonner*› d'une part, et ‹*trier*› et ‹*choisir*› sont "plus synonymes" que ‹*ordonner*› et ‹*choisir*›. On remarquera aussi que d_3 est supérieure à $\pi/4$, ce qui dénote un éloignement sémantique qui commence ; (b) d_4 est la plus petite valeur de $D_A(\text{ranger}, Y)$ car les concepts *CLASSER* et *RÉPARTIR* sont relativement proches, et de plus ‹*ranger*› est par ailleurs polysémique (*CLASSER*, *RASSEMBLER* et *NETTOYER*) et seul *CLASSER* est présent dans le tableau.

$D_A(X,Y)$	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0	0,517	0,662 d_1	0,611 d_2	0,551	0,441	0,462
ranger		0	0,829	0,6	0,523	0,409 d_4	0,444
choisir			0	0,848 d_3	0,77	0,796	0,758
ordonner				0	0,595	0,523	0,519
ventiler					0	0,471	0,391
classer						0	0,36
répartir							0

5. Le vecteur n'est sans doute pas représenté par un mot de la langue. Il s'agit d'une idée qui n'active... aucun concept ! C'est l'idée vide.
6. Tous les exemples de cet article sont issus de `<http ://www.lirmm.fr/~lafourca>`

L'espace vectoriel conceptuel est muni de deux lois de composition interne : la somme (et son opération symétrique, la soustraction) et du produit terme à terme (on ne définit pas ici son opération symétrique) qui sont détaillées dans le prochain paragraphe.

4.3. *Opérateurs*

Somme vectorielle. Soit X et Y deux vecteurs, on définit V comme leur somme normée :

$$V = X \oplus Y \quad | \quad v_i = (x_i + y_i)/\|V\| \tag{8}$$

Cet opérateur est idempotent, autrement dit nous avons $X \oplus X = X$. Le vecteur nul $\vec{\mathbf{0}}$ est l'élément neutre de la somme vectorielle et nous avons $\vec{\mathbf{0}} \oplus \vec{\mathbf{0}} = \vec{\mathbf{0}}$ (par idempotence). De ce qui précède, nous déduisons (sans le démontrer) les propriétés de rapprochement (local et généralisé) :

$$D_A(X \oplus X, Y \oplus X) = D_A(X, Y \oplus X) \leq D_A(X, Y) \tag{9}$$

$$D_A(X \oplus Z, Y \oplus Z) \leq D_A(X, Y) \tag{10}$$

Soustraction vectorielle. Soit X et Y deux vecteurs distincts, on définit V comme leur soustraction normée :

$$V = X \ominus Y \quad | \quad v_i = (x_i - y_i)/\|V\| \tag{11}$$

Cet opérateur n'est pas idempotent et on aura par définition : $V = X \ominus X = \vec{\mathbf{0}}$. On remarquera que, dans le cas général, les valeurs v_i peuvent être négatives et que la fonction de distance a son image sur $[0, \pi]$.

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

$$V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i} \tag{12}$$

Cet opérateur est idempotent ($V = X \otimes X = X$) et $\vec{\mathbf{0}}$ est absorbant ($V = X \otimes \vec{\mathbf{0}} = \vec{\mathbf{0}}$). Cette opérateur n'est pas défini pour des vecteurs à composantes négatives.

Contextualisation faible. Lorsque deux termes sont en présence, pour chacun d'eux certains de leurs sens se trouvent sélectionnés par le contexte que constitue l'autre terme. Ce phénomène de *contextualisation* consiste à augmenter chaque sens de ce qu'il a de commun avec l'autre. Soit X et Y deux vecteurs, on définit $\Gamma(X, Y)$ comme la contextualisation de X par Y comme :

$$\Gamma(X, Y) = X \oplus (X \otimes Y) \tag{13}$$

Cette fonction n'est pas symétrique. L'opérateur Γ est idempotent ($\Gamma(X,X) = X$) et le vecteur nul est un élément neutre ($\Gamma(X,\vec{\mathbf{0}}) = X \oplus \vec{\mathbf{0}} = X$). On remarquera (sans les démontrer) que nous avons les propriétés dites de *rapprochement* suivantes :

$$D_A(\Gamma(X,Y),\Gamma(Y,X)) \leq \{D_A(X,\Gamma(Y,X)), D_A(\Gamma(X,Y),Y)\} \quad (14)$$

$$\{D_A(X,\Gamma(Y,X)), D_A(\Gamma(X,Y),Y)\} \leq D_A(X,Y) \quad (15)$$

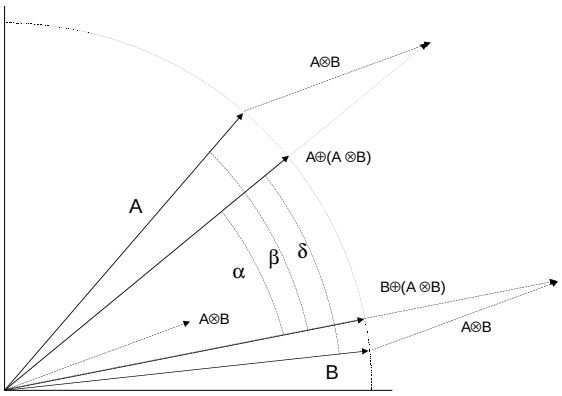

Figure 1. *Représentation géométrique en 2D de la contextualisation faible. L'angle α est la distance $D_A(\Gamma(A,B),\Gamma(B,A))$, β est la distance $D_A(A,\Gamma(B,A))$ et δ est la distance $D_A(\Gamma(A,B),B)$*

La contextualisation $\Gamma(X,Y)$ rapproche le vecteur X de Y proportionnellement à leur intersection. Dans la tableau qui suit, nous avons dans la partie supérieure les valeurs de $D_A(\Gamma(X,Y),\Gamma(Y,X))$.

$b\backslash a$	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0	0,269	0,363	0,322	0,288	0,228	0,239
ranger		0	0,474	0,316	0,273	0,211	0,23
choisir			0	0,485	0,434	0,451	0,425
ordonner				0	0,313	0,272	0,27
ventiler					0	0,244	0,201
classer						0	0,185
répartir							0

5. Synonymie et antonymie relatives

5.1. *Fonction de synonymie relative*

Nous définissons la fonction de *synonymie relative* Syn_R entre trois vecteurs A, B et C, ce dernier jouant le rôle de pivot, comme suit :

$$\begin{aligned} Syn_R(A,B,C) &= D_A(\Gamma(A,C),\Gamma(B,C)) \\ &= D_A(A \oplus (A \otimes C), B \oplus (B \otimes C)) \end{aligned} \quad (16)$$

$$Syn_A(A,B) = Syn_R(A,B,A \oplus B) \quad (17)$$

La synonymie absolue Syn_A n'est qu'un cas particulier de la synonymie relative ou A et B constituent leur propre contexte. L'interprétation correspond bien à celle présentée ci-dessus, à savoir que l'on cherche à tester la proximité thématique de deux sens (A et B), chacun augmenté de ce qu'il a de commun avec un tiers (C).

5.1.1. *Propriétés*

Pour rendre compte des trois propriétés théoriques de la relation de synonymie relative (réflexivité, symétrie et pseudo-transitivité), nous les vérifions comme suit :

1. $Syn_R(A,A,C) = 0$ La réflexivité est héritée de celle de la distance angulaire D_A.
2. $Syn_R(A,B,C) = Syn_R(B,A,C)$ La symétrie pour les deux premiers arguments, provient également de celle de la distance angulaire.
3. $Syn_R(A,B,E) + Syn_R(B,C,E) \geq Syn_R(A,C,E)$ C'est un héritage de l'inégalité triangulaire de D_A. Elle représente la pseudo-transivité de la synonymie relative. Elle est en outre plus précise que la vérification de la propriété de transitivité : elle indique que la distance entre A et C/E est au pire égale à la somme des mesures de synonymie de A et B/E d'une part, et B et C/E d'autre part.
4. $Syn_R(A,B,0) = D_A(A \oplus \vec{\mathbf{0}}, B \oplus \vec{\mathbf{0}}) = D_A(A,B)$ Le vecteur nul $\vec{\mathbf{0}}$ ramène la synonymie relative à la distance angulaire.
5. $Syn_R(A,B,C) \leq D_A(A,B)$ Par héritage du rapprochement de D_A, quel que soit le point de vue, la synonymie relative ne peut que rapprocher A et B.

5.1.2. *Exemples*

Dans le tableau qui suit, nous avons dans la partie supérieure le rappel des valeurs de (a) $D_A(X,Y)$ et dans la partie inférieure les valeurs de (b) $Syn_R(X,Y,\textbf{\textit{trier}})$. On voit bien apparaître ici la mise en lumière de la polysémie. Nous avons, par exemple, $Syn_R(\textbf{\textit{classer}},\textbf{\textit{ranger}},\textbf{\textit{trier}})$ valant $0,283$ (soit environ 16°), ce qui indique une forte synonymie relative de ‹*classer*› et ‹*ranger*› par rapport à ‹*trier*›, chose que la distance angulaire correspondante ($0,409$, ou environ 23°) montrait aussi, mais avec

moins d'acuité. À l'inverse, $Syn_R(\textbf{\textit{choisir}}, \textbf{\textit{ordonner}}, \textbf{\textit{trier}})$ vaut $0,636$ (ou 36°), ce qui montre que ‹*choisir*› et ‹*ordonner*› s'*éloignent* l'un de l'autre par rapport à ‹*trier*›, alors qu'ils sont deux synonymes possibles de ‹*trier*›. La synonymie relative apparaît comme un bon indicateur de polysémie : ‹*choisir*› et ‹*ordonner*› relèvent majoritairement des deux "zones" sémantiques différentes. De plus, ‹*ordonner*› est lui-même polysémique.

$b\backslash a$	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0	0,517	0,662	0,611	0,551	0,441	0,462
ranger	**0,402**	0	0,829	0,6	0,523	0,409	0,444
choisir	**0,5**	**0,623**	0	0,848	0,77	0,796	0,758
ordonner	**0,478**	**0,43**	**0,636**	0	0,595	0,523	0,519
ventiler	**0,435**	**0,365**	**0,575**	**0,435**	0	0,471	0,391
classer	**0,369**	**0,283**	**0,607**	**0,385**	**0,344**	0	0,36
répartir	**0,376**	**0,309**	**0,57**	**0,383**	**0,272**	**0,268**	0

5.2. *Fonctions d'antonymie relative*

L'identification de plusieurs types d'antonymie (voir la section 2), implique l'existence de plusieurs fonctions d'antonymie. Toutefois, ces fonctions sont toutes basées sur une même méthode que nous explicitons ci-dessous.

5.2.1. *Principes et définitions*

La fonction $Anti_R$ de construction d'un vecteur antonyme V d'un vecteur A selon un vecteur contexte V_c, définie en termes linguistiques en 3.3.2, se note comme suit :

$$V = Anti_R(A, V_c)$$

Comme pour la synonymie, les diverses fonctions $Anti$ dépendent du contexte mais, contrairement à la synonymie, elles ne peuvent pas être indépendantes de l'organisation des concepts. Elles nécessitent d'identifier pour chaque concept et pour chaque contexte un vecteur qui sera considéré comme son opposé. Il faut donc construire une liste de triplets $\langle concept, contexte, vecteur\rangle$ appelé *listes d'antonymes*. Cette liste peut comprendre, par exemple, l'antonyme de *EXISTENCE* qui serait le vecteur V(*INEXISTENCE*) quel que soit le contexte. Elle peut contenir aussi l'antonyme d'*AMOUR*, qui serait, lui, constitué des vecteurs *DÉSACCORD*, *AVERSION* et *INIMITIÉ*. On remarquera, que le concept de *HAINE* n'existe pas dans l'ontologie utilisée (Larousse). Nous considérons que l'antonyme d'un terme qui ne possède pas d'antonyme(s) avéré(s) est ce terme lui-même (une discussion sur ce point est proposée dans [Schwab 2001] et [Schwab *et al.* 2002]). Il est important de noter que cette liste est différente pour chaque type d'antonymie. Il suffit donc de dresser autant de listes que de types d'antonymie examinés.

Fonction $AntiC$. La fonction $AntiC$ renvoie en fonction d'un concept c_i de $\mathcal{C}$ et d'un vecteur contexte V_c le vecteur considéré comme le vecteur antonyme dans une

liste d'antonymes. Cette fonction se traduit donc par une simple exploration de la liste d'antonymes. On la note comme suit :

$$V = AntiC(c_i, V_c)$$

Par exemple, nous pouvons avoir :

$AntiC(\textsc{existence}, V_c) = V(\textsc{inexistence}) \quad \forall V_c$
$AntiC(\textsc{amour}, V_c) = V(\textsc{désaccord}) \oplus V(\textsc{aversion}) \oplus V(\textsc{inimitié}) \quad \forall V_c$
$AntiC(\textsc{destruction}, V(\textsc{travaux publics})) = V(\textsc{construction})$
$AntiC(\textsc{destruction}, V(\textsc{écologie})) = V(\textsc{préservation})$

5.2.2. *Construction du vecteur antonyme*

Définitions. Nous définissons les fonctions d'antonymie relative et absolue comme :

$$V = Anti_R(A, V_c)$$
$$V = Anti_A(A) = Anti_R(A, A)$$

Construction du vecteur conceptuel antonyme. Le but est, à partir de deux vecteurs conceptuels, un pour l'item lexical dont nous voulons l'antonyme, l'autre pour le contexte, de construire un vecteur opposé. L'idée est d'insister sur les notions saillantes des vecteurs A et V_c. Si ces notions peuvent être opposées, alors l'antonyme doit posséder les idées inverses dans la même proportion. Une fonction d'antonymie est définie comme suit :

$$Anti_R(A, V_c) = \bigoplus_{i=1}^{N} P_i \times AntiC(c_i, V_c)$$

avec comme définition pour le poids P_i :

$$P_i = A_i^{1+\mathrm{CV}(A)} \times max(A_i, V_{c_i})$$

et A_i la ième composante de A :

$$A = \ < A_0, A_1, \ldots, A_{\dim(\mathcal{C})} >$$

Le poids P a été défini empiriquement à la suite d'expérimentations. Clairement, la fonction ne pouvait pas être symétrique, puisque le rôle de *vecteur à opposer* et celui de *vecteur contexte* ne sont pas interchangeables. Nous ne devons pas avoir $Anti_R$(V(‹*chaud*›),V(‹*température*›)) = $Anti_R$(V(‹*température*›),V(‹*chaud*›)). La puissance $1 + CV(V_{\text{item}})$ a donc été introduite pour insister d'avantage sur les idées présentes dans le vecteur que nous voulions opposer. Nous avons aussi remarqué que plus un vecteur était conceptuel (proche du vecteur d'un concept) plus il était intéressant d'augmenter cette puissance. C'est la raison pour laquelle cette puissance comprend le *coefficient de variation* [7] qui est un bon indice de la "conceptualité". Enfin, nous avons

7. Le coefficient de variation est donnée 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.

introduit la fonction *max* afin de considérer les idées de l'item, même si celles-ci ne sont pas présentes dans le référent. Par exemple, si l'on cherche l'antonyme de ‹*froid*› dans le contexte de ‹*température*›, le poids de ‹*froid*› doit être important même s'il n'est pas présent dans le vecteur représentant ‹*température*›.

Une conséquence importante de notre définition de l'antonymie est que l'antonyme d'un item sans antonyme avéré est l'item lui-même. Celui-ci est alors considéré comme positionné sur l'axe de symétrie ([Schwab 2001] et [Schwab *et al.* 2002] *op. cit.*). Notre formalisation nous a permis de passer d'une fonction d'antonymie discrète (approche *linguistique* classique) à une fonction d'antonymie continuement définie sur l'espace des sens.

5.2.3. *Mesure d'évaluation de l'antonymie*

Il semble pertinent de savoir si deux items lexicaux peuvent être l'antonyme l'un de l'autre afin de posséder un outil comparable à la synonymie relative. Nous avons donc créé une mesure d'évaluation de l'antonymie. Soient les vecteurs A et B. La question est de savoir si on peut dire s'ils sont antonymes dans le contexte C. La distance d'antonymie M_{anti-R} est la mesure de l'angle formé par la somme des vecteurs A et B et la somme de leur opposés $Anti_{c_R}(A, C)$ et $Anti_{c_R}(B, C)$. Soient les mesures d'antonymie relative et absolue :

$$M_{anti-R}(A, B, C) = D_A(A \oplus B, Anti_R(A, C) \oplus Anti_R(B, C)) \quad (18)$$

$$M_{anti-A}(A, B) = D_A(A \oplus B, Anti_A(A) \oplus Anti_A(B)) \quad (19)$$

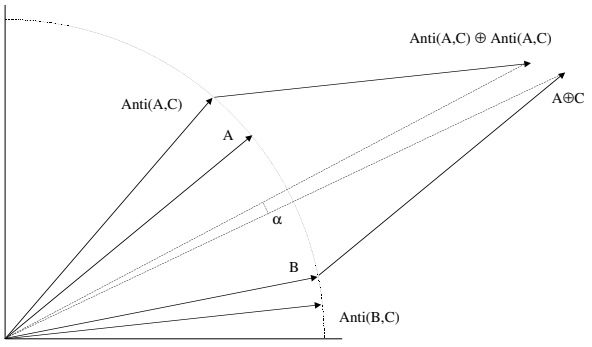

Figure 2. *Représentation géométrique en 2D de la mesure d'évaluation de l'antonymie par l'angle* α

La mesure d'antonymie n'est pas une distance. Ce n'est qu'une pseudo-distance. Elle vérifie les propriétés de réflexivité, symétrie et inégalité triangulaire uniquement dans le sous ensemble des items qui n'ont pas d'antonymes. Dans le cas général, elle

ne vérifie pas la réflexivité. Les composantes des vecteurs conceptuels sont positives et nous avons la propriété : $Dist_{anti} \in [0, \frac{\pi}{2}]$. Plus la mesure est petite, plus les deux items lexicaux sont antonymes dans le contexte. En revanche, ce serait une erreur de considérer que deux antonymes seraient à une distance avoisinant $\pi/2$. Deux items lexicaux à $Mant = \pi/2$ l'un de l'autre n'ont aucune idée en commun[8]. Nous pouvons plutôt voir ici l'illustration que deux antonymes ont certaines idées en commun, celles qui ne sont pas opposables ou celles qui le sont mais dont l'activation est proche. Ils ne s'opposent que par certaines activations de concepts [Cruse and Togia 1995]. Une distance de $\pi/2$ entre deux items lexicaux devrait être plutôt interprétée comme le fait que ces deux items lexicaux n'ont que peu d'idées en commun, une sorte d'anti-synonymie. Ce résultat confirme le fait que l'antonymie n'est pas exactement l'inverse de la synonymie mais lui est très liée. L'antonyme d'un item ‹*m*› n'est pas un mot qui ne partage aucune idée avec ‹*m*› mais un item qui s'oppose à ‹*m*› sur certaines idées !

5.2.4. *Exemples*

Nous avons par exemple :

M_{anti-A}(*EXISTENCE*,*INEXISTENCE*)$= 0,03$
M_{anti-A}(‹*existence*›, ‹*automobile*›) $= 1,06$
M_{anti-A}(‹*existence*›,‹*inexistence*›) $= 0,44$
M_{anti-A}(*AUTOMOBILE*, *AUTOMOBILE*) $= 0,006$
M_{anti-A}(*EXISTENCE*, *AUTOMOBILE*) $= 1,45$
M_{anti-A}(‹*automobile*›, ‹*automobile*›) $= 0,407$

Les exemples ci-dessus illustrent bien ce que nous disions auparavant. Les concepts *EXISTENCE* et *INEXISTENCE* sont très fortement antonymes en antonymie complémentaire. L'effet de la polysémie explique que les items ‹*existence*› et ‹*inexistence*› soient moins antonymes que les concepts. En antonymie complémentaire, *AUTOMOBILE* est son propre antonyme. La mesure de l'antonymie entre *AUTOMOBILE* et *EXISTENCE* est un exemple de notre remarque précédente sur les vecteurs qui ne partagent que peu d'idées. Aux alentours de $\pi/2$, cette mesure se comporte comme la distance angulaire. D'ailleurs, nous avons $D_A(existence, automobile) = 1,464$ (soit un peu moins de $\pi/2$).

5.3. *Vecteurs conceptuels et passage à la terminologie*

Comme on a pu le voir, le modèle des vecteurs conceptuels permet non seulement de travailler sur la composition de sens, mais aussi peut faire émerger des relations sémantiques transverses correspondant aux fonctions lexicales de synonymie et d'antonymie. Ce que nous avons montré a été réalisé sur un lexique général fondé sur une ontologie de même type. Dans la prochaine section, nous allons montrer comment un tel dispositif permet d'exploiter et d'extraire et d'enrichir des terminologies spécifiques et donc de mieux assister le traitement de textes à fort caractère technique.

8. Ce cas de figure est purement théorique, il n'existe dans aucune langue deux items lexicaux qui ne partagent aucune idée.

6. Projection ontologique de vecteurs conceptuels

6.1. *Extensions ontologiques*

6.1.1. *Généralités*

On considérera en toute généralité deux ontologies G (pour générale) et S (pour spécialisée). La première (G) est universelle et est censée engendrer (par définition) tout les mots de la langue et couvre (de façon grossière) tous les champs sémantiques. La seconde (S) ne contient que les termes de sa spécialité et ne couvre (en détail) que les champs sémantiques de son (ou ses) domaines. Parmi les propriétés premières de G et de S : S a de fortes chances d'être *localement* beaucoup plus précise que G, et l'intersection entre G et S ne doit pas être nulle. La première propriété est nécessaire pour rendre S intéressante (on verra dans ce qui suit une formalisation de ces propriétés). Ce qui est présenté ensuite peut s'étendre à n ontologies de spécialités.

Pour demeurer dans le même paradigme que précédemment, on estime que G et S sont des familles génératrices d'espaces vectoriels. Dans la suite, on parlera de G comme de l'espace vectoriel défini par l'ontologie G (idem pour S). Tout vecteur d'un espace E n'est comparable qu'avec un autre vecteur de E : on comparera donc les vecteurs de G (respectivement S) entre eux. Sauf indication contraire, tout vecteur est normé.

Pour traiter un texte technique qui, comme nous l'avons dit, comprend aussi bien des termes techniques que des formulations générales, il importe de considérer l'espace généré par $G \cup S$, que l'on appellera par la suite GS.

6.1.2. *Notion de maillage de la description*

On remarquera que les termes de G sont inclus dans GS, et que plus l'ontologie S est spécialisée, plus le rapport *(nombre de termes $\times$ nombre moyen de sens)/ nombre de concepts* est faible. Cela provient du fait que la description est plus précise et donc que la *maille* descriptive est plus serrée. Pour le moment dans nos expériences pour G (Thésaurus Larousse) nous avons $65\,000$ entrées et environ 5 sens en moyenne par entrée. Compte tenu de la dimension de G (873), cela donne $65\,000 \times 5/873 = 372$. Pour les textes techniques nous avons actuellement repéré environ $10\,000$ lexies concernées (en analysant des définitions) et nos premières constatations font état d'environ 2 sens en moyenne par lexie de spécialité. Comme nous l'avons dit précédemment, les textes techniques sont mieux référencés sur GS que sur S seulement. La dimension de GS de l'ordre de 2873 : somme des dimensions de S, $2\,000$ (nombre de concepts dans l'ontologie de l'OCDE considérée), et G soit 873. Ce qui donne pour GS : $10\,000 \times 2/2873 = 6,96$ soit environ 7. On voit bien que la différence de taille de la maille est assez spectaculaire en G et GS car il y a un facteur 53. Evidemment, à la limite (si l'on dispose d'une ontologie, ou d'une union d'ontologies, aussi spécialisée que possible sans synonymie exacte) ce rapport devrait tendre vers 1 (chaque terme est associé à un concept). Cette limite est tout à fait illusoire quand on traite des textes généraux et n'a de sens que pour des textes spécialisés.

6.1.3. *Quantité d'information*

En revanche, le produit des termes (qui représente la quantité d'information à stoker) reste dans les mêmes ordres de grandeur : pour G, on a : $65\,000 \times 5 \times 873 = 283\,725\,000$ (soit au moins 4 fois plus en taille physique, soit environ 1,2 Go) et pour GS (si on ne traite que les textes techniques) $10\,000 \times 2 \times 2873 = 57\,460\,000$ (soit au moins 4 fois plus en taille physique, soit environ 225 Mo).

On remarque que la taille du lexique sémantique (ensemble des sens) spécialisé est presque cinq fois plus petite que celle du lexique sémantique général, ce qui nous ramène à déplacer le problème de la taille de l'ontologie (plus petite traditionnellement si elle est technique, plus grande pour nous) vers celui de la taille du lexique sémantique, en d'autre termes, la *quantité d'information*.

6.1.4. *Commentaire sur l'exhaustivité d'une couverture*

Pour des raisons opérationnelles, il est clair que nous ne souhaitons pas représenter tous les mots de la langue par des vecteurs de l'espace vectoriel GS (le produit serait déraisonablement égal à $910\,000\,000$ soit environ 4 Go). Ce serait non seulement coûteux, mais de plus n'apporterait rien à la finesse d'analyse, la plupart des mots n'appartement pas aux champs sémantiques décrits par la très grande majorité des concepts (ceux-ci étant en grande partie issus de S en propre).

6.1.5. *Apprentissage*

Nous avons analysé, pour obtenir les vecteurs conceptuels correspondants, les termes de spécialité (ici l'économie) à partir de leurs définitions, dont celles issues du DAFA (Dictionnaire d'Apprentissage du Français des Affaires). Notre objectif était, au départ, de construire les lexies à partir de S, autant que faire se peut, et de ne basculer ensuite sur GS que si cela était nécessaire. Nous nous sommes aperçus très vite que, en phase d'apprentissage, et surtout à partir de dictionnaires, nous avions presque systématiquement accès à des termes généraux (hors de S).

Par exemple, la première définition du terme ‹*marché*› est : *Lieu physique ou virtuel d'échanges*. Au mieux, seul le terme ‹*échange*› pourrait se projeter sur S. Ce qui rendrait les termes ‹*échange*› et ‹*marché*› (dans son sens 1) synonymes ! Il est donc nécessaire de tenir compte des vecteurs de ‹*lieu*›, ‹*physique*› et ‹*virtuel*› qui ne sont pourtant définis que dans G. C'est pourquoi nous traitons essentiellement des vecteurs dans GS et que nous avons défini une opération de ‹*dépliage*› (qui déploie un vecteur d'un sous-espace G ou S dans GS) afin d'obtenir un vecteur $D(v)$ de GS à partir de v de G. Ce vecteur ne porte pas plus d'information que v, mais rend possible le calcul dans GS.

On ne sait jamais si, dans une définition, une occurence d'un terme fait référence à un sens général ou spécifique. C'est pourquoi, souvent en pratique l'apprentissage s'amorce avec une combinaison des deux possibilités. Les sens probables émergent par activation des informations mutuelles des occurences des autres termes de la définition.

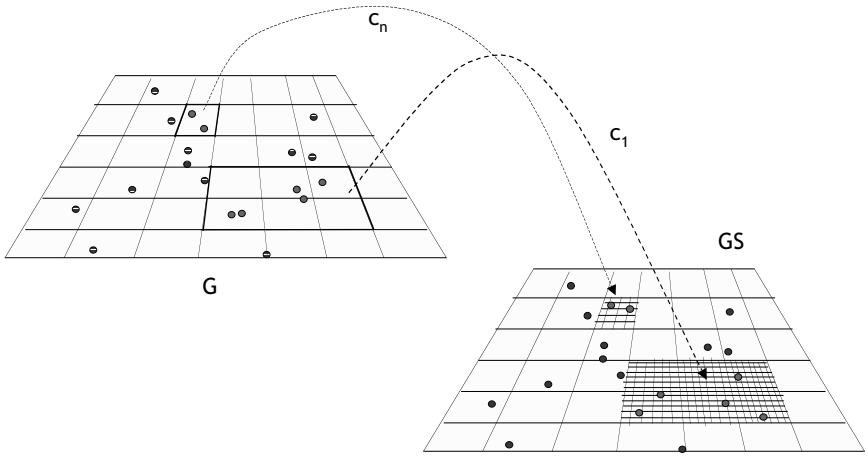

Figure 3. *Affinement du maillage et correspondances entre espaces vectoriels*

6.2. *Dépliage et pliage de vecteurs*

6.2.1. *Correspondances ontologiques*

À un concept c_G de G, on peut associer un ensemble de concepts de S. On appellera une telle association $< C_G, \{C_{S,1}, \ldots, C_{S,n}\} >$ une *correspondance ontologique*. Par exemple, le concept *ÉCONOMIE* de G est associé à toute la sous-arborescence de S contenant ce terme (économie politique, économie de marché, économie dirigé, microéconomie, macroéconomie, etc.)

On se donne, comme contrainte, que l'ensemble des correspondances couvre tout S. C'est-à-dire que l'ensemble des concepts atteint dans S est égal à S. Il s'agit donc d'une surjection. En revanche, ce n'est absolument pas une injection, car il existe des concepts de G qui ne sont pas dans les champs sémantiques de S (qui, on le rappelle sont par définition plongés dans G).

6.2.2. *Dépliage*

La fonction de dépliage D est une projection d'un vecteur v_G de G sur GS : $v_G \rightarrow v_{GS}$, qui permet d'affiner la représentation de ce vecteur si celui-ci est concerné par

les concepts de S. C'est ce que l'on nomme aussi *l'extension ontologique* du vecteur v.

D(v) est un vecteur de $G \cup S$ et se compose comme un vecteur de G suivi d'un vecteur de S, et $dim(D(v)) = dim(G) + dim(S)$. La première partie (nommée Kern) de $D(v)$ est v. La seconde partie (nommée Ext) se calcule comme suit à partir de v (de G) et de la liste des correspondances $\mathcal{C}$:

Procédure déplier (v_G, $\mathcal{C}$) $\rightarrow v_{GS}$

soit $P =< 0, \cdots, 0 >$ % P est un vecteur de taille $\dim(S)$;
pour chaque $< C_G, \{C_{S,1}, \ldots, C_{S,n}\} >$ de $\mathcal{C}$ **faire**
  soit $x = v(C_G)$ % x est la valeur de v à la composante C_G
  chaque composante $\{C_{S,1}, \ldots, C_{S,n}\}$ de P est incrémenté de x
fin pour
$Ext = p_1 * VC_1 + \cdots + p_{\dim(S)} * VC_{\dim(S)}$
v est normé
retourner v

Fin Procédure déplier

Le vecteur $P =< p_1, \ldots, p_n >$ représente les pondérations pour la somme des $\dim(S)$ vecteurs des concepts VC_i de S. C'est à partir de P que l'on construit Ext. On remarquera dans le vecteur obtenu ne contient jamais de zéro si les vecteurs des concepts invoqués ne contiennent pas de zéros. C'est en effet le cas par construction pour les vecteurs de G. Les vecteurs sont donc très denses.

6.2.3. *Pliage*

La fonction de pliage P est une projection d'un vecteur v_{GS} de GS sur G : $v_{GS} \rightarrow v_G$. Pour plier un vecteur de GS sur G, il suffit de *supprimer* Ext. En pratique, on crée un vecteur qui ne contient que Kern. Cela permet de récupérer sur une plus petite base, les termes peu touchés directement par la spécialité ou d'en avoir aussi une acception plus générale.

$$D(v)_{GS} \quad = \quad < x_1, \ldots, x_{dim(G)}, \ldots, x_{dim(G)+dim(S)} >$$

$$\rightarrow \quad < x_1, \ldots, x_{dim(G)} > \quad = \quad v_G$$

Si les procédures de construction (mais également, dans une moindre mesure, celles du noyau et de l'apprentissage) pour les vecteurs des concepts sont *bonnes* alors les concepts de G qui sont *relié* à ceux de S ont bougé si ceux de S ont été modifiés (et réciproquement). Les procédures présentées assurent cette propriété.

Le pliage est une projection qui perd de l'information, en particulier s'il s'agit de termes à la fois généraux et spécialisés, comme ‹*échange*› ou encore ‹*marché*›, cependant l'activation des concepts de la partie G reflète l'activation des concepts de S.

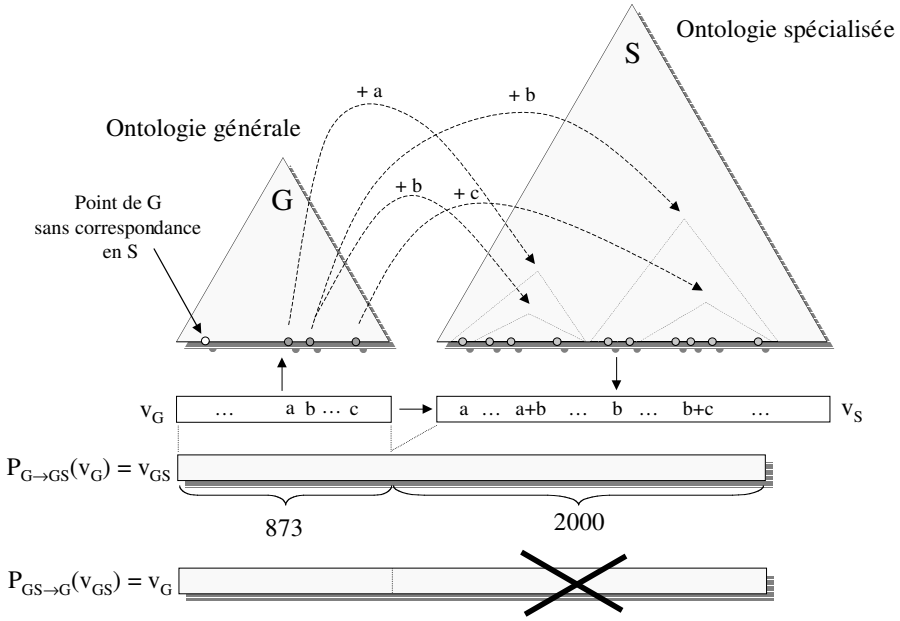

Figure 4. *Correspondances ontologiques. Le pliage est une projection P de GS sur G et le dépliage une projection P de G sur GS*

6.3. *Propriétés*

Une première propriété concerne la composition des fonctions D (de dépliage) et P (de pliage).

$$P(D(v)) = v$$

Déplier puis replier un vecteur équivaut à la fonction identité. Mais dans le cas général, nous n'avons pas l'inverse $D(P(v)) \neq v$ puisqu'il y a perte d'information. Nous avons également une réduction relative de la distance angulaire D_A :

$$D_A(v1, v2) \leq D_A(D(v1), D(v2))$$

Ce phénomène peut se traduire ainsi : *l'extension ontologique augmente la synonymie (hors apprentissage)*. Cela se démontre (et s'expérimente) à partir de la définition de la distance angulaire donnée en section 4.2. Par contre, le *raffinement ontologique* (c'est-à-dire l'analyse d'un terme dans GS au lieu de G) peut soit :

1) Réduire la synonymie (c'est-à-dire augmente la distance sémantique) pour deux termes de spécialité.

Deux termes quasi identiques dans G s'éloignent conceptuellement, ce qui permet de les discriminer davantage. Par exemple : ‹*finances publiques*› et ‹*fiscalité*› sont dans G à $D_A = 0,3$ (environ 17 degrés). Dans S, nous avons $D_A = 1,2$ (environ 69 degrés).

2) Augmenter la synonymie par réduction de la polysémie.

Par exemple, dans G, le terme t1 a deux sens t11 et t12, et t2 a deux sens t21 et t22. On suppose que t11 et t22 sont deux sens synonymes. La distance globale de t1 et t2 peut être (relativement) élevée car t12 et t21 constituent du bruit. En revanche, comme dans S seuls t11 et t22 appartiennent au domaine, nous avons (dans S) t1 et t2 qui sont monosémiques et synonymes.

C'est globalement le cas pour les termes ‹*profit*›, ‹*bénéfice*› et ‹*produit*›.

Dans S, les termes de spécialité sont moins polysémiques, et chacun a une description très fine et séparée des autres. Dans G, ces termes sont souvent fortement polysémiques, les descriptions sont moins fines et moins séparées (elles ont tendance à s'agglutiner en classes d'équivalence lors de l'application de filtres sémantiques basés sur la distance angulaire).

6.4. *Construction de vecteurs ontologiques de* GS

Construction des vecteurs de concepts de S. Les vecteurs de S se construisent comme ceux de G. Pour mémoire, il s'agit de s'appuyer sur l'ontologie pour construire les $\dim(S)$ vecteurs des concepts de S. Cette construction utilise la distance ultramétrique et les activation transverses éventuelles.

Construction des vecteurs de concepts de GS. La question est en fait de savoir comment *ajouter* le vecteurs G (et lequel) à chaque vecteur de concept de S que l'on a produit précédemment. Une solution est d'*inverser* le dépliage (dépliage inverse). On applique la même procédure que *déplier*, mais à partir d'un vecteur de S, on construit un vecteur de G. On peut trivialement inverser une correspondance en une liste de correspondances :

$$\begin{aligned} \mathcal{C} &= < C_G, \{C_{S,1}, \ldots, C_{S,n}\} > \\ \rightarrow \quad (< C_{S,1}, \{C_G\} >, \ldots, < C_{S,n}, \{CG\} >) &= \mathcal{C}' \end{aligned}$$

Le vecteur de S peut être concaténé à gauche de son dépliage inverse qui produit la partie sur G :

$$\text{déplier}(v_S, \mathcal{C}') + v_S \quad \rightarrow \quad v_{GS}$$

Par la suite, la construction des vecteurs du noyau de GS (extensible à tous les termes de S) et l'apprentissage des termes sur GS s'effectue comme pour G.

7. Fonctions de filtres sémantique et changement d'espace ontologique

La densité lexicale permet de mesurer le degré d'appartenance d'un terme (ou d'un de ses sens) à une ontologie donnée. Cette mesure se base sur les variations observées

pour ce terme entre l'espace vectoriel général et l'espace vectoriel spécialisé. Il en est de même pour les deux relations que sont la synonymie et l'antonymie.

7.1. *Distribution et concentration lexicale*

La distribution lexicale $\mathcal{D}_E(\mathrm{t})$ d'un terme t dans l'espace S est la répartition des termes en fonction de leur distance à t. Par exemple la figure 5 représente la distribution lexicale du terme ‹*marché*› de G. Les termes se répartissent en général autour d'un sommet. On observe, systématiquement quel que soit le terme choisi, une *petite bosse* située entre le sommet et $\pi/2$ qui est un point d'accumulation des vecteurs des concepts (qui sont naturellement éloignés des vecteurs des termes).

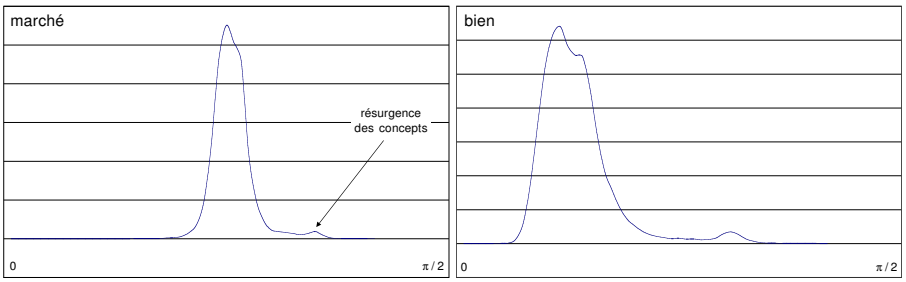

Figure 5. *Distribution lexicale de* ‹marché› *et de* ‹bien› *dans* G

On définit l'*intervalle de proximité thématique*, la fonction $\mathcal{I}_{E,\mathrm{f}}(\mathrm{t})$ avec ($0 \leq f \leq 1$) qui retourne l'intervalle le plus petit dans lequel se trouve la fraction f du lexique de G, *la plus proche* de t (en excluant évidement t lui-même). Par exemple, $\mathcal{I}_{\mathcal{G},0,5}(\mathrm{marché})$ correspond à l'intervalle qui contient la moitiée des mots de G qui sont les plus proche thématiquement du terme ‹*marché*›. Nous avons ici, $\mathcal{I}_{G,0.5}(\mathrm{marché}) = [0,25\,;0,91]$. Il s'agit géométriquement de dire que la moitié des termes de G par rapport à ‹*marché*› se situent entre les deux hypersphères de rayon $0,25$ et $0,91$. Nous avons, $\mathcal{I}_{G,0,5}(\mathrm{bien}) = [0,16\,;0,43]$.

On appelle la *concentration lexicale* $\partial\mathcal{I}$ d'un intervalle $\mathcal{I}$, le pourcentage du lexique couvert divisé par l'écart de cet intervalle. Ici, $\partial\mathcal{I}_{G,0,5}(\mathrm{marché}) = 0,5/(0,91 - 0,25) = 0,76$ et $\partial\mathcal{I}_{G,0,5}(\mathrm{bien}) = 0,5/(0,43 - 0,16) = 1,85$.

Si la concentration lexicale de t est faible alors la courbe est décalée vers $\pi/2$ (il y a peu de termes autour de t). Et inversement, si elle est importante, la courbe est décalée vers 0.

Un terme peut avoir une densité lexicale forte dans plusieurs cas (non exclusifs) :

1) Le terme appartient à un champ sémantique très riche. Par exemple, les noms d'insecte ont une densité lexicale très forte, tout simplement parce qu'ils sont en très grand nombre.

2) Le terme est souvent utilisé dans des définitions comme hyperonyme (il a une forte valeur conceptuelle). C'est le cas de terme comme ‹*insecte*›, ‹*plante*›, ‹*élément*›, ‹*travail*›, ‹*nombre*›... Ces termes sont fortement hyperonymiques, mais des termes très généraux (comme ‹*homme*› ou ‹*former*›) bien qu'ayant une fréquence très élevée ont une concentration lexicale faible. Ils ne sont pas particulièrement porteur de sens, et donc ne participent que peu à la constitution du sens d'un mot.

3) Le terme est très polysémique. Il a tendance à se ramener aux deux cas ci-dessus.

Pour un terme t, nous pouvons nous intéresser, selon l'espace vectoriel considéré, à deux facteurs. D'une part à la variation de la densité lexicale (c'est-à-dire formellement à la taille de l'intervalle), et d'autre part à la variation de positions de cet intervalle.

On remarque que globalement la concentration lexicale *chute* avec l'extension ontologique. C'est-à-dire qu'un terme très polysémique dans G, ne correspond qu'à un nombre de sens très réduit (voire unique) dans S. Il s'agit par exemple du cas de ‹*marché*›, qui prend des sens très spécifiques dans l'ontologie de l'OCDE (et dans les définitions du DAFA). Par exemple, $\partial\mathcal{I}_{S,0,5}(\text{marché}) = 0,68$. De façon, encore plus nette, nous obtenons $\partial\mathcal{I}_{G,0,5}(\text{bien}) = 0,5/(0,43 - 0,16) = 0,8$. Dans S, ‹*bien*› ne correspond qu'au substantif masculin dont la définition est *chose produite pour satisfaire un besoin* et qui correspond directement à un des concepts de S. Ce terme est donc très conceptuel, mais ne constitue pas un terme fortement hyperonymique.

On rappelle que dans nos expériences, tous les termes d'un domaine de spécialité S sont inclus dans l'espace vectoriel général G. Pour un terme t_S issu de S, on peut donc faire la constatation suivante :

$$\partial\mathcal{I}_{G,\mathrm{f}}(\mathrm{t}_S) \geq \partial\mathcal{I}_{S,\mathrm{f}}(\mathrm{t}_S)$$

La densité lexicale dans G est plus forte que dans S. En effet, le maillage étant plus fin dans S, ce terme est mieux discriminé. Le passage du terme t_S de S dans G se fait par pliage.

Si nous avons (dans de rares cas) :

$$\partial\mathcal{I}_{G,\mathrm{f}}(\mathrm{t}_S) \leq \partial\mathcal{I}_{S,\mathrm{f}}(\mathrm{t}_S)$$

cela indique qu'un terme de S dispose d'un certain nombre de termes proches qui s'éloignent dans G. C'est possible, si les termes en question sont fortement polysémiques dans G ou très généraux. Par exemple, dans S le terme de ‹*concentration*› peut être utilisé de façon elliptique pour de nombreux termes associés ayant un sens très précis (non réellement calculables par composition) : *concentration d'entreprise*, *concentration verticale*, *concentration horizontale*, *concentration d'un secteur*, *concentration dans un secteur*... Par contre dans G, le terme de concentration est très général et les plus proches voisins sont globalement plus éloignés que dans S. Ce dernier phénomène est plus rare que le premier. C'est pourquoi globalement la densité lexicale chute.

La distribution et la concentration lexicale fournissent ainsi des filtres permettant de savoir effectivement si un terme *t* peut (ou non) appartenir à une ontologie de specialité S. Il s'agit des mesures fournissant un degré de confiance. En fixant, *a priori*, une valeur seuil, il est ainsi possible d'extraire automatiquement le vocabulaire spécialisé d'un domaine. Ce vocabulaire se compose des termes de spécialité et les sens des mots généraux qui sont pertinents pour cette spécialité. Par exemple, dans la terminologie pétrolière, le terme ‹*poisson*› est bien sélectionné comme étant un *segment de trépan brisé et logé au fond du puits de forage et que l'on doit aller pêcher*.

La synonymie et l'antonymie permettent, elles, d'affiner cette extraction terminologique en établissant entre les termes des relations de *sens proches* et des *sens en opposition*.

7.2. *Utilisation de la synonymie comme un révélateur de structures*

Il s'agit ici de l'étude dystopique de la synonymie relative, c'est-à-dire de la comparaison de son comportement entre les espace G et GS. La densité lexicale consitue une fonction macroscopique à l'échelle du lexique. À l'inverse, la synonymie relative ici est une fonction microscopique à l'échelle du terme. La synonymie relative constitue ainsi une fonction typique de filtrage sur les points des espace vectoriels. Il s'agit ici d'étudier le comportement de la fonction de synonymie relative $Syn_R(A, B, C)_{\mathcal{E}}$ selon que l'on considère pour espace vectoriel $\mathcal{E}$, l'espace général G ou l'espace augmenté GS.

Considérons tout d'abord le cas plus simple de la synonymie absolue (qui n'est qu'un cas particulier de la synonymie relative). Nous cherchons ici à comparer les valeurs $Syn_A(A, B)_{\mathcal{GS}}$ et $Syn_A(A, B)_{\mathcal{G}}$.

On peut distinguer plusieurs cas selon l'appartenance des termes à S :

1) A, B sont dans S. Dans ce cas si :

$$Syn_A(A, B)_{\mathcal{GS}} \leq Syn_A(A, B)_{\mathcal{G}}$$

alors les termes sont discriminés dans S grâce à l'affinement du maillage. C'est le cas de termes comme ‹*commerçant*›, ‹*marchant*›, ‹*détaillant*›, ‹*grossiste*›, ‹*négociant*›, ‹*fournisseur*›, ‹*revendeur*›, ... Tous ces termes sont quasiment synonymes dans G mais sont très différents dans GS. Par contre si :

$$Syn_A(A, B)_{\mathcal{GS}} \geq Syn_A(A, B)_{\mathcal{G}}$$

les termes se sont rapprochés dans S. Il s'agit d'un cas typique où la polysémie dans GS éloigne deux termes qui ont un sens proche en commun. C'est par exemple, le cas pour ‹*travail*› et ‹*emploi*›, ou encore ‹*traitement*› et ‹*salaire*›.

2) Soit A, soit B est dans S. Dans ce cas, on a toujours une diminution de la synonymie. $Syn_A(A, B)_{\mathcal{GS}} \geq Syn_A(A, B)_{\mathcal{G}}$.

3) Ni A ni B ne sont dans S, les deux termes deviennent bien plus synonymes. On a donc bien $Syn_A(A,B)_{\mathcal{GS}} \leq Syn_A(A,B)_{\mathcal{G}}$.

Pour la synonymie relative $Syn_R(A,B,C)$, la question se ramène à évaluer la situation selon que C est ou non un terme de S. Si C est un terme acceptable (au sens de la concentration lexicale) pour S alors les mesures de synonymie sont plus pertinentes. Inversement, si C n'est pas un terme acceptable pour S, les mesures de synonymie sont dégradées. En particulier, si ni A, ni B, ni C ne sont dans S, cela se ramène bien au troisième cas ci-dessus. Cela signifie que la synonymie relative est un bon indice de structure lorsque le pivot de cette structure (C) est pertinent pour l'ontologie.

7.3. *Utilisation de l'antonymie comme un révélateur de structures*

On ne considère ici que la fonction d'antonymie globale (telle qu'elle est décrite dans [Schwab 2001] et [Schwab *et al.* 2002] *op. cit.*). La relation d'antonymie peut émerger, disparaître ou être conservée quand on passe de G vers GS, d'une part, et de GS vers G, d'autre part. La terminologie procédant par métaphore et composition des termes génériques, la relation antonymique est souvent préservée (par exemple, les *médias froids* et les *médias chauds* de McLuhan).

La plupart du temps, pour le vocabulaire fortement terminologique, l'antonymie utilisée sera du troisième type (*duale*, qui concerne les oppositions culturelles) à cause de l'utilisation intensive de la métaphore dans la création terminologique. On ne peut en effet guère déduire par l'analyse le sens strict des termes qui s'opposent essentiellement à travers l'organisation de l'ontologie (et non forcément en tant que tels). Par exemple (OCDE) :

1) ‹*mortalité*› ↔ ‹*fécondité et planification de la famille*›

2) ‹*zone rurale*›↔ ‹*zone urbaine*›

3) ‹*groupes d'ages*› ↔ ‹*groupes éthniques*›

En revanche, les définitions des termes teminologiques (par exemple issus du DAFA) font largement appel à l'opposition. Ce qui peut alimenter la construction incrémentale de fonctions d'antonymie. Ces fonctions peuvent ensuite jouer un rôle de filtre au même titre que la synonymie, afin de faire émerger des structures cachées dans les agglomérations (ou séparations) de vecteurs. Par exemple : ‹*économie de marché*› = ‹*économie libérale*› ↔ ‹*économie dirigée*›.

L'étude de l'antonymie est similaire à celle de la synonymie. Nous cherchons donc à comparer, dans un premier temps, les valeurs de $M_{anti-A}(A,B)_{\mathcal{GS}}$ et de $M_{anti-A}(A,B)_{\mathcal{G}}$.

1) A, B sont dans S. Dans ce cas si :

$$M_{anti-A}(A,B)_{\mathcal{GS}} \leq M_{anti-A}(A,B)(A,B)_{\mathcal{G}}$$

alors les termes sont plus fortement antonymes dans GS que dans G. Il s'agit encore une fois de l'effet de l'affinement du maillage. C'est le cas de termes comme ‹*économie libérale*› (A) et ‹*économie dirigée*› (B) : $M_{anti-A}(A,B)_{\mathcal{G}} = 0,6$ et $M_{anti-A}(A,B)_{\mathcal{GS}} = 0,3$. Nous avons aussi, le cas de ‹*travail*› (A) et ‹*chômage*› (B) à cause de la polysémie de ‹*travail*›. Nous avons : $M_{anti-A}(A,B)_{\mathcal{G}} = 0,35$ et $M_{anti-A}(A,B)_{\mathcal{GS}} = 0,48$. Par contre si :

$$M_{anti-A}(A,B)_{\mathcal{GS}} \geq M_{anti-A}(A,B)(A,B)_{\mathcal{G}}$$

les termes sont moins antonymes dans GS que dans G. Il s'agit d'un cas ou les concepts sur lesquels s'opposent les termes dans G, soit ne s'opposent plus dans S ou ne sont pas pertinents (et donc ne s'opposent plus). C'est le cas avec ‹*boucher*› (A) et ‹*poisonnier*› (B) car dans G *poisson* et *viande* s'opposent dualement. Nous avons : $M_{anti-A}(A,B)_{\mathcal{G}} = 0,57$ et $M_{anti-A}(A,B)_{\mathcal{GS}} = 0,48$. Les concepts liés à *poisson* et *viande* ne sont pas pertinents dans S et donc leur poids dans GS s'en trouve considérablement amoindri.

2) Soit A, soit B est dans S. Dans cas, la variation dépend de leur opposition potentielle dans S.

3) Ni A ni B ne sont dans S, les deux termes deviennent beaucoup moins antonymes.

8. Conclusion

Les expériences que nous avons menées autour de l'intégration d'une ontologie de spécialité (ici, le domaine économique) à une ontologie générale fondée sur les concepts du thésaurus, et munie du dispositif calculatoire du modèle vectoriel, nous ont permis d'aboutir aux conclusions suivantes.

1. Lorsqu'il faut analyser, classer ou indexer des textes de spécialité, la meilleure solution consiste à utiliser une union entre l'ontologie générale et l'ontologie de spécialité parce que les textes de spécialité ne contiennent pas que des termes techniques. Le passage de l'une à l'autre a été décrit dans la section 6 de l'article à l'aide de procédures et d'algorithmes testés et finalisés.

2. Lorsqu'un apprentissage automatique de concepts spécialisés est réalisé à partir de définitions fournies dans des dictionnaires, cette union d'ontologies s'avère indispensable, puisque tous les mots de la définition peuvent alors contribuer à fournir les éléments pour le calcul du sens.

3. Dans le modèle des vecteurs conceptuels, l'ontologie de spécialité est beaucoup plus fournie que l'ontologie générale, contrairement à l'approche d'*arborescence de connaissances* classique. En revanche, ce que nous avons découvert est que la quantité d'information à stocker est plus petite pour une analyse de textes techniques que pour une analyse de textes généraux. Nous avons ainsi déplacé le problème de la taille depuis l'ontologie vers la quantité d'information.

4. L'ontologie de spécialité permet un maillage plus serré de la représentation du sens, donc une meilleure discrimination sémantique entre des termes qui apparaîtraient proches. Inversement, le calcul du sens et des distances sur cette ontologie permettent de rendre très proches, voire *synonymes*, des sens qui, projetés sur l'ontologie générale, ne le seraient pas. La polysémie, caractéristique principale du lexique général est alors circonscrite au profit des sens spécialisés des termes invoqués.

5. Justement, la notion de synonymie calculée, ainsi que celle d'antonymie (qui permet de traiter d'éventuelles négations) est l'un des grands apports du modèle vectoriel tel que nous le pratiquons (section 5). Dans la majorité des cas, les travaux sur la synonymie partent d'une synonymie prédite ou fournie ([Ploux et Victorri 1998] et [Hathout 2001]). Le modèle vectoriel permet de tester la validité d'une proximité supposée et, dans son raffinement, celui de la synonymie relative, il permet d'explorer les relations qu'entretiennent les termes autour d'un terme dit de *contexte*. Emerge alors une microstructuration, ou plus exactement une microtopologie, qui permet de revisiter l'espace vectoriel lexical avec une notion de *densité lexicale* (section 7) au voisinage d'un vecteur. L'antonymie, à laquelle les définitions de dictionnaires font largement appel pour mettre en contraste une notion par rapport à une autre, aussi bien que la synonymie relative, sont des révélateurs de structure émergeante et dynamique.

Ces conclusions partielles vont dans le sens d'une conclusion plus générale : faire appel à une terminologie de spécialité pour traiter des textes techniques est non seulement faisable dans le modèle vectoriel, mais celui-ci permet d'unir les ontologies, de discriminer des sens, de circonscrire la polysémie, et de faire émerger une microstructuration qui pourra être modifiée au gré de l'apprentissage continu que permet le modèle. Les expériences menées ont permis l'intégration d'une terminologie sous forme d'une ontologie de 2 000 concepts feuilles (issue de l'OCDE) et d'analyser des textes définitoires en provenance, entre autres, du dictionnaire des affaires (le DAFA). la construction de l'ontologie terminologique est déjà achevée et les liens émergeants transverses de synonymie et d'antonymie ont été utilisés pour constater des rapprochements entre notions et une amélioration de la discrimination sémantique. Ces tests peuvent être répétés par tout utilisateur qui le souhaite sur un site web[9], où le système est à la disposition de tous.

9. Bibliographie

[Barrière and Copeck 2001] Barrière C., Copeck T., "Building Domain Knowledge from Specialized Texts", *TIA 2001*, Nancy, 2001.

[Bourrigault 1993] Bourrigault D., "Analyse locale pour le repèrage des termes complexes dans les textes", *TAL*, vol. 34, n° 2, p. 105-118, 1993.

[Chauché 1990] Chauché J., "Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance", *TA Information*, vol. 31, n° 1, p. 17-24, 1990.

9. <http ://www.lirmm.fr/~lafourca>

[Cruse and Togia 1995] Cruse D.A., Togia P., “Towards a cognitive model of antonymy”, *Lexicology*, vol. 1, p. 113-141, 1995.

[DAFA 2001] Verlinde S., Selva T., *DAFA - Dictionnaire d’Apprentissage du Français des Affaires*, http ://www.projetdafa.net.

[Deerwester *et al.* 1990] Deerwester S., Dumais S., Landauer T., Furnas G., Harshman R., “Indexing by latent semantic analysis”, *Journal of the American Society of Information Science*, 416(6), p. 391-407, 1990.

[Fellbaum 1995] Fellbaum C., “Co-occurrence and antonymy”, *International Journal of Lexicography*, vol. 8, p. 281-303, 1995.

[Fischer 1973] Fischer W. L., *Äquivalenz und Toleranz Strukturen in der Linguistik zur Theory der Synonyma*, Max Hüber Verlag, München, 1973.

[Gwei and Foxley 1987] Gwei G.M., Foxley E., “A Flexible Synonym Interface with application examples in CAL and Help Environments”, *The Computer Journal*, vol. 30 n˚6, p. 551-557, 1987.

[Hamon et Nazarenko 2001] Hamon T., Nazarenko A., “La structuration de terminologie : une nécessaire coopération”, *TIA 2001*, Nancy, 2001.

[Hathout 2001] Hathout N., “Analogies morpho-synonymiques. Une méthode d’acquisition automatique de liens morphologiques á partir d’un dictionnaire de synonymes”, *TALN 2001*, Tours, vol. 1, p. 223-232, juillet 2001.

[Hearst 1998] Hearst M.A., “Automated discovery of Wordnet relations”, In C. Fellbaum ed. *Wordnet An Electronic Lexical Database*, MIT Press, Cambridge, MA, p. 131-151, 1998.

[Justeson and Katz 1991] Justeson J.S., Katz S., “Co-occurrences of antonymous adjectives and their contexts”, *Computational Linguistics*, vol. 17, p. 1-19, 1991.

[Lafourcade et Sandford 1999] Lafourcade M., Sandford E., “Analyse et désambiguïsation lexicale par vecteurs sémantiques”, *TALN’99*, Cargèse. p. 351-356, juillet 1999.

[Lafourcade 2001] Lafourcade M., “Lexical sorting and lexical transfer by conceptual vectors”, *First International Workshop on MultiMedia Annotation* (MMA’2001), Tokyo, 6 p, January 2001.

[Lafourcade et Prince 2001] Lafourcade M., Prince V., “Synonymies et vecteurs conceptuels”, *TALN 2001*, Tours, p. 233-242, juillet 2001.

[Larousse 2001] Larousse, *Thésaurus Larousse - des idées aux mots - des mots aux idées.* Larousse, 1992.

[OCDE 1991] OCDE, “Macrothesaurus”, http ://info.uibk.ac.at/info/oecd-macroth/, 1991.

[Prince 1991] Prince V., “Notes sur l’évaluation de la réponse dans TEDDI : introduction d’une relation d’équivalence pour la synonymie relative”, *Notes et Documents LIMSI*, 91-20, CNRS, 1991.

[Resnik 1995] Resnik P., “Using Information contents to evaluate semantic similarity in a taxonomy”, *IJCAI-95*, 1995.

[Riloff and Shepherd 1995] Riloff E., Shepherd J., “A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction”, *Natural Language Engineering*, vol. 5, part. 2, p. 147-156, 1995.

[Salton 1968] Salton G., *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York, 1968.

[Salton and MacGill 1983] Salton G., MacGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[Salton 1988] Salton G., *Term-Weighting Approaches in Automatic Text Retrieval*, *McGraw-Hill computer science series*, McGraw-Hill, vol. 24, 1988.

[Schwab 2001] Schwab D., "Vecteurs conceptuels et fonctions lexicales : application à l'antonymie", Mémoire de DEA Informatique, 2001.

[Schwab *et al.* 2002] Schwab D., Lafourcade M., V. Prince V., "Amélioration de la représentation sémantique lexicale par les vecteurs conceptuels : le rôle de l'antonymie", *JATD 2002*, vol. 2, p. 701-712, 2002.

[Sparck Jones 1986] Sparck Jones K., *Synonymy and Semantic Classification*, Edinburgh Information Technology Serie, 1986.

[Ploux et Victorri 1998] Ploux S., Victorri B., "Construction d'espaces sémantiques á l'aide de dictionnaires de synonymes. " *TAL*, vol. 39, n° 1, p. 161-182, 1998.

[Yarowsky1992] Yarowsky D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", *COLING'92*, Nantes, p. 454-460, 1992.