Synonymies et vecteurs conceptuels

Mathieu Lafourcade, Violaine Prince
LIRMM : Dép. ARC, grp. TAL, Université Montpellier 2
161, rue Ada, 35392 Montpellier Cedex 5, France
www.lirmm.fr/˜lafourca - www.lirmm.fr/˜prince

Résumé - Abstract

La synonymie est une relation importante en TAL mais qui reste problématique. La distinction entre synonymie relative et synonymie subjective permet de contourner certaines difficultés. Dans le cadre des vecteurs conceptuels, il est alors possible de définir formellement des fonctions de test de synonymie et d'en expérimenter l'usage. **Mots-clés.** Synonymie relative, synonymie subjective, approches statistiques, distances thématiques.
Synonymy is a pivot relation in NLP but remains problematic. Putting forward, the distinction between relative and subjective synonymy, allows us to circunvent some difficulties. In the framework of conceptual vectors, it is then possible to formalize test functions for synonymy and to experiment their use. **Keywords.** Relative synonymy, subjective synonymy, statistical approaches, thematic distances.

1 Introduction

La synonymie est, avec l'hyperonymie, une des fonctions lexicales qui ont été les plus étudiées en traitement automatique des langues, sans parler de la linguistique. Parmi les travaux les plus récents, dans la communauté française, fortement dédiés à la synonymie, on peut citer [Hamon et al. 1999] qui traite de l'extraction de synonymes, et s'appuie essentiellement sur des liens de synonymie pour faire émerger des structures de connaissances dans des textes techniques ; ou encore, [Selva 1999], qui fonde en grande partie l'apprentissage du français langue seconde sur les fonctions lexicales et en particulier, la synonymie. On pensera également aux travaux multiples sur l'extraction des relations sémantiques en général, à partir de corpus : citons, dans le paysage francophone, les actions incitatives (au sein de l'ARC A3 de l'Aupelf-Uref), les thèses (par exemple [Morin 1999]), et dans la communauté anglo-saxonne, les travaux réalisés à partir de Wordnet [Hearst 1998].
Si la synonymie est une relation étudiée en TAL c'est parce qu'elle permet, entre autres : a) d'aider à la constitution de dictionnaires ; b) de réaliser une recherche d'information plus fine que le simple appariement d'une chaîne de caractères ; c) de ne pas multiplier les concepts dans les bases de connaissances (un même concept sera associé à une liste de termes *synonymes*) ; d) de gérer une qualité stylistique en génération.
Quelques propriétés de la synonymie sont cependant problématiques. Telle qu'elle est entendue, la synonymie devrait être une relation d'équivalence sémantique entre termes. Or, pour que cette équivalence soit exploitable sur le plan formel, la relation devrait être réflexive, symétrique et transitive. Malheureusement, ces propriétés ne sont pas toujours vérifiées comme nous le montrons plus loin.
Dans cet article, nous abordons d'abord les problèmes liés à la synonymie, ce qui nous conduit à définir plusieurs types de synonymies et à préciser le concept de synonymie relative. Nous présentons ensuite de façon générale, les notions et les traitements liés aux vecteurs conceptuels

qui sont au cœur du formalisme que nous avons choisi d'adopter. Nous précisons enfin les différentes fonctions de test liées aux synonymies. Ces fonctions sont basées sur les vecteurs conceptuels.

2 Synonymies

Un des premiers *défauts* connus de la synonymie, en tant que relation entre termes c'est qu'elle n'est pas nécessairement transitive, lorsque l'on prend les termes deux à deux, sans plus de précautions [Lewis 1952]. Ainsi par exemple, ‹*trier*› et ‹*ordonner*›, ‹*trier*› et ‹*choisir*›, qui sont synonymes deux à deux, sont tels que ‹*ordonner*› et ‹*choisir*› ne sont pas synonymes. En pratique, au moins trois concepts sont désignés par ‹*trier*› : *ORDONNER*⌈*trier une liste de cinquante éléments*⌋ les éléments de la liste sont mis dans un ordre donné mais aucun d'eux n'est soustrait ; *CHOISIR*⌈*c'est un personnel trié*⌋ ou les personnes sélectionnées constituent un sous-ensemble d'un ensemble possible de personnes ; et enfin, *RÉPARTIR*⌈*trier le courrier*⌋. [Fischer 1973] montre que la synonymie est au mieux une relation de tolérance[1].
Le deuxième défaut de la synonymie est qu'elle peut se confondre au moins partiellement avec l'hyperonymie. Par exemple, ‹*morceler*› a pour synonyme ‹*couper*› alors qu'il en est hyponyme. On a, en effet, *MORCELER*⌈*couper en plusieurs morceaux*⌋, par opposition à l'idée de couper en deux, ou couper dans le sens de soustraire une partie. Ce défaut fait apparaître une fragilité dans la symétrie de la relation, ce qui remet même en cause son statut de relation de tolérance au sens fort. En effet, si un hyperonyme apparaît comme synonyme parce que le terme partage avec lui toutes ses propriétés, en revanche sur un plan sémantique, l'hyponyme n'est pas un synonyme. Ainsi ‹*cisailler*› n'apparaît pas comme synonyme de ‹*couper*›, alors que l'inverse semble plus admissible.
Enfin, il y a une petite "déconvenue" prévisible qui veut que deux hyponymes d'un même terme, tout en étant parents, ne sont pas forcément synonymes. Si ‹*poignarder*› et ‹*abattre*› sont hyponymes d'‹*assassiner*›, ils ne sont pas en mesure de présenter des qualités de synonymie. Ce qui amène à définir la synonymie comme la qualité, pour deux termes, de partager le plus grand nombre de contenus sémantiques[2], ou avoir la plus grande base commune possible (lorsqu'il s'agit d'une représentation plus numérique ou topologique). À l'inverse, on remarquera également, que la polysémie d'un terme peut faire qu'il ait plusieurs hyperonymes.
Par conséquent, si l'on souhaite exploiter les liens de synonymie entre termes pour faire de l'indexation, de la recherche d'information dans un corpus, ou pour générer du texte, il est important de définir des relations de synonymie qui pourraient avoir de meilleures propriétés que celles de la synonymie vue *in abstracto*.

2.1 Notion de Synonymie relative

Pour pallier le premier défaut, nous avions proposé dès 1991 [Prince 1991] une notion de *synonymie relative* qui part du principe que deux termes peuvent être synonymes par rapport à l'idée centrale développée par un troisième, ou par un des deux. Cette notion avait déjà été appréhendée par [Sabah 1984] sous la forme de *synonymie approchée contextuelle* dans un modèle lexical de type *réseau sémantique*. Ainsi, ‹*trier*› et ‹*choisir*› sont synonymes par rapport au concept discriminant de *CHOISIR*, alors que ‹*trier*› et ‹*ranger*› sont synonymes par rapport à *RANGER*.

[1] Une relation de tolérance peut-être symétrique et réflexive mais n'est pas transitive. Il existe plusieurs niveaux de tolérance, selon le nombre de propriétés vérifiées.

[2] La sémantique componentielle dirait : le plus grand nombre de sèmes communs.

Avec un troisième terme, cela peut fonctionner de la même manière. Le concept *ORDONNER*, permet de lier ‹*trier*› et ‹*ranger*›, ‹*trier*› et ‹*ordonner*›, ‹*trier*› et ‹*ventiler*›. L'idée est que tous les synonymes, par rapport à un même tiers (qui peut être en l'occurrence, l'un des termes), sont synonymes entre eux deux à deux ; dès lors, la synonymie relative au tiers est transitive. Par exemple, ‹*choisir*› et ‹*sélectionner*› sont synonymes par rapport à *CHOISIR*, par conséquent, ‹*trier*› et ‹*sélectionner*› le sont aussi.
L'intérêt d'une telle relation est qu'elle devient alors une relation d'équivalence (une démonstration formelle en a été faite dans [Prince op. cit]), ce qui rend toute sa valeur au lien de synonymie.

2.2 Notion de Synonymie subjective

Si on veut utiliser plus largement la synonymie dans le cadre de l'indexation ou de la recherche d'information, il faudrait tenir compte d'une notion de synonymie "forcée" par un point de vue. Bien que deux hyponymes d'un même terme ne soient pas synonymes si on se place dans le champs sémantique le plus proche de ces termes, il n'en va pas de même lorsque l'on s'éloigne sémantiquement d'eux.
Ainsi, dans des textes consacrés à l'usinage, thème qui fera figure de concept *point de vue*, ‹*cisailler*› et ‹*morceler*› seront forcément différenciés et doivent l'être lors de l'indexation. En revanche, dans un texte consacré au transport, on peut négliger la différence entre ces termes, voire les assimiler à leur hyperonyme ‹*couper*›. C'est cette capacité à "confondre" des termes parce que leur différence sémantique est faible au regard de la thématique générale que nous nommons *synonymie subjective*.
La synonymie subjective reste une notion opératoire définie pour des besoins de recherche d'information, par opposition à a) la synonymie héritée qui est une propriété fonctionelle de l'hyperonymie, et b) la synonymie relative, où le concept pivot est un des sens des termes polysèmes à comparer, et qui est une propriété fonctionnelle de la polysémie.

3 Vecteurs conceptuels

Dans le cadre de recherche sur la représentation du sens en TALN et son application à la recherche d'information, nous nous concentrons sur la représentation de l'aspect thématique (des segments textuels tels que les documents, paragraphes, syntagmes, etc.) sous la forme de vecteurs *conceptuels* [Lafourcade et Sanford 1999]. Cette approche tire son origine de [Chauché 1990] pour l'utilisation d'un jeu de concepts prédéterminé, mais s'inspire aussi du modèle vectoriel [Salton et MacGill 1983] et du modèle LSI [Deerwester et all. 1990] pour la reconnaissance et l'exploitation de l'inter-dépendance des concepts. Par contre, son application à l'indexation et la recherche d'information textuelle se distingue nettement de [Salton 1988], en ce qu'elle se base explicitement pour son calcul sur la géométrie et les variables morphosyntaxiques des arbres d'analyse structurelle issus du texte et non pas sur une analyse de surface par mots-clés. D'une façon générale les documents sont traités indépendamment (ce qui constitue une différence majeure d'avec LSI) et l'accent est mis sur la sélection lexicale en contexte. Les mêmes considérations peut être faite avec [Resnik 1995] à propos de l'usage exclusif de taxonomies.
Pour mémoire, le modèle de vecteurs conceptuels s'appuie paradigmatiquement sur la projection dans un modèle mathématique de la notion linguistique de champ sémantique. Les concepts sont définis selon un thésaurus (en ce qui nous concerne, il s'agit de la langue française [Larousse 1992] où 873 concepts sont répertoriés). L'hypothèse principale est que cet ensemble forme un

espace générateur pour les mots de la langue (espace qui n'est probablement pas libre). Dès lors, tout mot se projette sur cet espace selon le principe énoncé ci-après.

3.1 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 a été exposé dans [Lafourcade 2001]. Il est clair, que pour des vecteurs denses (ayant tès peu de coordonnées nulles), l'énumération des concepts activés est vite fastidieuse et surtout difficile à évaluer. On préferera en général, procéder par sélection de termes thématiquement proches (cf 3.2). 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*›, ‹*tronçonnner*› , ‹*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 est lourde (on rappelle que dans nos expérimentations, $\dim(\mathcal{C}) = 873$, ce qui correspond au niveau 4 des concepts définis dans [Larousse 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 140000 définitions correspondants à 52000 mots vedettes (pour 25000 mots monosémiques et 27000 mots polysémiques - pour ces derniers le nombre moyen de définitions, certaines éventuellement redondantes, étant de 4.54).

3.2 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 "·" 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).

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

$$D_A(X, Y) = \arccos(Sim(X, Y)) \tag{2}$$

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 $D_A(X,Y) \leq \pi/4$, X et Y sont sémantiquement proche 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 $\pi/2$, 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 évidement pas vrai.

Il s'agit d'une vraie distance (contrairement à la mesure de similarité) et elle vérifie les propriétés de réflexivité (3), symétrie (4) et inégalité triangulaire (5) :

$$D_A(X,X) = 0 \tag{3}$$
$$D_A(X,Y) = D_A(Y,X) \tag{4}$$
$$D_A(X,Y) + D_A(Y,Z) \geq D_A(X,Z) \tag{5}$$

Par définition, nous posons : $D_A(\vec{\mathbf{0}},\vec{\mathbf{0}}) = 0$ et $D_A(X,\vec{\mathbf{0}}) = \pi/2$ avec $\vec{\mathbf{0}}$ dénotant le vecteur nul[3]. 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[4] dans le tableau qui suit, nous avons les distances angulaire (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, par exemple : 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*› ; 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	0.517	0.662 d_1	0.611 d_2	0.551	0.441	0.462
ranger		0.0	0.829	0.6	0.523	0.409 d_4	0.444
choisir			0.0	0.848 d_3	0.77	0.796	0.758
ordonner				0.0	0.595	0.523	0.519
ventiler					0.0	0.471	0.391
classer						0.0	0.36
répartir							0.0

3.3 Opérateurs

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

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

[3]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.

[4]Tous les exemples de cet article sont issus de `<http://www.lirmm.fr/~lafourca>`

Cet opérateur est idempotent et nous avons $X \oplus X = X$. Le vecteur nul $\vec{\mathbf{0}}$ est l'élément neutre de la somme vectorielle et, par définition, nous posons que $\vec{\mathbf{0}} \oplus \vec{\mathbf{0}} = \vec{\mathbf{0}}$. 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 ($V = X \otimes \vec{\mathbf{0}} = \vec{\mathbf{0}}$).

Contextualisation et Anti-contextualisation. Lorsque que deux termes sont en présence, pour chacun d'eux certains de leur 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. À des fins opératoires, nous définissons également la fonction *opposée*, l'*anti-contextualisation*. Soit X et Y deux vecteurs, on définit $\Gamma(X, Y)$ (resp. $\overline{\Gamma}(X, Y)$) comme la contextualisation (resp. l'anti-contextualisation) de X par Y comme :

$$\Gamma(X, Y) = X \oplus (X \otimes Y) \qquad \text{et} \qquad \overline{\Gamma}(X, Y) = X \ominus (X \otimes Y) \tag{13}$$

Ces fonctions ne sont pas symétriques. L'opérateur Γ est idempotent ($\Gamma(X, X) = X$) et le vecteur nul est un élément neutre ($\Gamma(X, \vec{\mathbf{0}}) = X \oplus \vec{\mathbf{0}} = X$). L'opérateur $\overline{\Gamma}$ est nulpotent ($\overline{\Gamma}(X, X) = X \ominus X = \vec{\mathbf{0}}$) et $\vec{\mathbf{0}}$ est également un élément neutre. On remarquera (sans les démontrer) que nous avons les propriétés (de *rapprochement* et d'*éloignement*) suivantes :

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

$$D_A(\overline{\Gamma}(X, Y), \overline{\Gamma}(Y, X)) \geq \{D_A(X, \overline{\Gamma}(Y, X)), D_A(\overline{\Gamma}(X, Y), Y)\} \geq D_A(X, Y) \tag{15}$$

La contextualisation $\Gamma(X, Y)$ rapproche le vecteur X de Y proportionnellement à leur intersection. L'anti-contextualisation $\overline{\Gamma}(X, Y)$ procède inversement. Dans la tableau qui suit, nous avons dans la partie supérieure les valeurs de (a) $D_A(\Gamma(X, Y), \Gamma(Y, X))$ et dans la partie inférieure les valeurs de (b) $D_A(\overline{\Gamma}(X, Y), \overline{\Gamma}(Y, X))$.

$b \backslash a$	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.269	0.363	0.322	0.288	0.228	0.239
ranger	2.183	0.0	0.474	0.316	0.273	0.211	0.23
choisir	2.401	2.17	0.0	0.485	0.434	0.451	0.425
ordonner	2.382	2.374	2.314	0.0	0.313	0.272	0.27
ventiler	2.334	2.303	2.282	2.483	0.0	0.244	0.201
classer	2.505	2.481	2.313	2.648	2.535	0.0	0.185
répartir	2.476	2.388	2.364	2.637	2.53	2.761	0.0

4 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 :

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

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).

4.1 Propriétés

Pour rendre compte des trois propriétés théoriques de la relation de synonymie (réflexivité, symétrie et 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 une forme de transivité pour 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, quelque soit le point de vue, la synonymie relative ne peut que rapprocher A et B.

4.2 Exemples

Dans la 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) $Syn_R(X, Y, \textit{\textbf{trier}})$.

On voit bien apparaître ici la mise en lumière de la polysémie. Nous avons, par exemple, $Syn_R($***classer***, ***ranger***, ***trier***$)$ valant $0,283$, ce qui indique une forte synonymie relative de ‹*classer*› et ‹*ranger*› par rapport à ‹*trier*›, chose que la distance angulaire correspondante $(0,409)$ n'indiquait pas aussi fortement. À l'inverse, $Syn_R($***choisir***, ***ordonner***, ***trier***$)$ vaut $0,636$, ce qui montre que ‹*choisir*› et ‹*ordonner*› ne sont pas synonymes entre eux par rapport à ‹*trier*›, alors qu'ils sont deux synonymes possible 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\a	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.517	0.662	0.611	0.551	0.441	0.462
ranger	**0.402**	0.0	0.829	0.6	0.523	0.409	0.444
choisir	**0.5**	**0.623**	0.0	0.848	0.77	0.796	0.758
ordonner	**0.478**	**0.43**	**0.636**	0.0	0.595	0.523	0.519
ventiler	**0.435**	**0.365**	**0.575**	**0.435**	0.0	0.471	0.391
classer	**0.369**	**0.283**	**0.607**	**0.385**	**0.344**	0.0	0.36
répartir	**0.376**	**0.309**	**0.57**	**0.383**	**0.272**	**0.268**	0.0

5 Synonymie Subjective

Nous définissons la fonction de *synonymie subjective* Syn_S entre trois vecteurs A, B et C, ce dernier jouant le rôle de point de vue, comme suit :

$$\begin{aligned} Syn_S(A, B, C) &= D_A(\overline{\Gamma}(A, C), \overline{\Gamma}(B, C)) \\ &= D_A(A \ominus (A \otimes C), B \ominus (B \otimes C)) \end{aligned} \tag{17}$$

L'interprétation naturelle consiste à considérer C comme un point de vue à partir duquel A et B sont comparés. Plus le point de vue C s'éloigne de A et de B, plus ceux-ci semblent se confondre. À l'inverse, plus C est proche de A et B (plus il se trouve *entre* eux) plus il est à même de les distinguer.
Avec $A \neq B \neq C$, nous avons donc : $\|C\| \to \infty \Rightarrow Syn_S(A, B, C) \to 0$.

5.1 Propriétés

Certaines des propriétés de la synonymie subjective sont analogues à celle de la synonymie relative, mis à part la dernière qui est originale.

1. $Syn_S(A, B, C) = Syn_S(B, A, C)$ Nous avons commutativité pour les deux premiers arguments, par simple héritage de la commutativité de la distance angulaire.
2. $Syn_S(A, B, \vec{\mathbf{0}}) = D_A(A \ominus \vec{\mathbf{0}}, B \ominus \vec{\mathbf{0}}) = D_A(A, B)$ Si le point de vue est le vecteur nul on se ramène à la distance angulaire.
3. $Syn_S(A, A, C) = D_A(A \ominus (A \otimes C), A \ominus (A \otimes C)) = 0$ Deux sens identiques sont toujours à une distance angulaire égale à 0 quelque soit le point de vue.
4. $Syn_S(A, B, E) + Syn_S(B, C, E) \geq Syn_S(A, C, E)$
   Héritage de l'inégalité triangulaire.
5. $Syn_S(A, B, C) \geq D_A(A, B)$ Héritage de l'éloignement.
6. $Syn_S(A, B, B) = D_A(A \ominus (A \otimes B), B \ominus (B \otimes B)) = D_A(A \ominus (A \otimes B), \vec{\mathbf{0}}) = \pi/2$
   $Syn_S(A, B, B) = Syn_S(A, B, A)$ Si le point de vue est l'un des sens considérés, la discrimination de sens est maximale.

5.2 Exemples

Dans la 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) $Syn_S(X, Y, \textbf{\textit{trier}})$.

b\a	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.517	0.662	0.611	0.551	0.441	0.462
ranger	**1.571**	0.0	0.829	0.6	0.523	0.409	0.444
choisir	**1.571**	**1.643**	0.0	0.848	0.77	0.796	0.758
ordonner	**1.571**	**1.433**	**1.624**	0.0	0.595	0.523	0.519
ventiler	**1.571**	**1.395**	**1.543**	**1.36**	0.0	0.471	0.391
classer	**1.571**	**1.259**	**1.741**	**1.292**	**1.323**	0.0	0.36
répartir	**1.571**	**1.324**	**1.613**	**1.245**	**1.132**	**1.158**	0.0

On notera, en particulier, que la colonne correspondant à ‹*trier*› n'a que des valeurs égales à $\pi/2$, ce qui est conforme à la propriété 6. On remarque donc que la synonymie subjective agit comme un "objectif". Plus un terme se rapproche de point de vue, plus la discrimination est forte. Par exemple, ‹*répartir*› et ‹*ventiler*› gardent le meilleur score $(1, 132)$ car ils sont très proches entre eux. Par contre, ‹*ordonner*› et ‹*choisir*› ont un score supérieur à $\pi/2$. Dans ce cas la polysémie est bien discriminée. C'est également le cas de ‹*classer*› et ‹*choisir*› $(1, 741)$.

6 Conclusion

Les travaux que nous avons menés sur la synonymie dans des sources de connaissances lexicales ont montré que : 1) Dans une modélisation globaliste comme celle des vecteurs conceptuels, où l'on travaille à partir de mots qui invoquent des idées et non pas sur des concepts qui se combinent en mots, la synonymie a des propriétés pouvant s'exprimer sous formes de mesure. 2) Pour que ces mesures de synonymie nous rapprochent des bonnes propriétés mathématiques (équivalence ou quasi-équivalence) que l'on voudrait leur voir attribuer, nous avons été amenés à définir deux types de synonymies : a) la synonymie relative, qui permet par rapport à un thème donné, de montrer les groupements de termes qui seraient quasi-équivalents entre eux; et b) la synonymie subjective, qui apparaît comme un discriminateur fort, si le thème est sémantiquement proche des termes à examiner, ou au contraire, comme un mécanisme de lissage si le thème est éloigné.
Nous poursuivons nos travaux avec l'aide de ces deux mesures pour réaliser de la détection d'hyperonymie. Si cette dernière apparaît comme évidente quand on travaille dans le sens concept $\rightarrow$ mot, elle est beaucoup plus difficile à asserter dans le sens mot $\rightarrow$ concept. La synonymie relative et la synonymie subjective qui traquent toutes deux à la fois la ressemblance et la différence sémantiques, forment une structure fonctionelle de base à partir de laquelle nous cherchons à reconstruire bon nombre de fonctions lexicales définies en linguistique.

Références

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

Deerwester S. and S. Dumais, T. Landauer, G. Furnas, R. Harshman, *Indexing by latent semantic analysis.* In Journal of the American, Society of Information science, 1990, 416(6), p 391-407.

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

Hamon, T. et D. Garcia, A. Nazarenko, *Détection de liens de synonymie : complémentarité des ressources générales et spécialisées.* Terminologies Nouvelles. 1999, pp 61-69.

Hearst, M. A. *Automated discovery of Wordnet relations.* In C Fellbaum ed. "Wordnet An Electronic Lexical Database". MIT Press, Cambridge, MA, 1998, pp 131-151.

Lafourcade M. et E. Sandford, *Analyse et désambiguïsation lexicale par vecteurs sémantiques.* In Proc. of TALN'99 (Cargèse, July 1999) pp 351-356.

Lafourcade M. *Lexical sorting and lexical transfer by conceptual vectors.* In Proc. of The First International Workshop on MultiMedia Annotation (MMA'2001) (Tokyo, January 2001) 6 p.

Larousse. *Thésaurus Larousse - des idées aux mots - des mots au idées.* Larousse, ISBN 2-03-320-148-1, 1992.

Lewis, C. I. *The modes of meaning.* in Linsky ed, "Semantics and the philosophy of language". Urbana. NY, 1952.

Morin, E. *Extraction de liens sémantiques entre termes à partir de corpus techniques.* Thèse de doctorat de l'Université de Nantes, 1999.

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. 1991, CNRS. 22 p.

Resnik P. *Using Information contents to evaluate semantic similarity in a taxonomy.* In Proceedings of IJCAI-95, 1995.

Riloff E. and J. Shepherd, *A corpus-based bootstrapping algorithm for Semi-Automated semantic lexicon construction.* In Natural Language Engineering, Vol 5, part 2, June 1995, pp 147-156.

Sabah, G. *Différentes notions de synonymies liées à la compréhension du langage.* Actes du colloque de l'Association pour la Recherche Cognitive

1984.

Salton G. *Term-Weighting Approaches in Automatic Text Retrieval.* McGraw-Hill computer science serie. McGraw-Hill, Volume 24, 1988.

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

Selva T. *Ressources et activités pédagogiques dans un environnement informatique d'aide à l'apprentissage lexical du français langue seconde.* Thèse d'Université, Université de Franche-Comté, Besançon, octobre 1999, 210 p.

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

Ploux S. et B. Victorri *Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes.* In TAL, Vol 39/1, 1998 pp 161-182.

Annexes - Résultats de synonymie relative

ranger* \ *choisir	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.392	0.459	0.437	0.385	0.302	0.318
ranger	0.398	0.0	0.598	0.471	0.416	0.336	0.363
choisir	0.519	0.703	0.0	0.601	0.519	0.55	0.512
ordonner	0.461	0.442	0.698	0.0	0.435	0.382	0.381
ventiler	0.421	0.366	0.654	0.439	0.0	0.344	0.273
classer	0.36	0.287	0.689	0.396	0.342	0.0	0.265
répartir	0.368	0.312	0.661	0.389	0.279	0.269	0.0

ordonner* \ *ventiler	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.348	0.464	0.46	0.39	0.306	0.325
ranger	0.344	0.0	0.591	0.45	0.365	0.276	0.308
choisir	0.456	0.568	0.0	0.629	0.541	0.573	0.536
ordonner	0.438	0.428	0.601	0.0	0.455	0.41	0.404
ventiler	0.409	0.385	0.544	0.454	0.0	0.345	0.273
classer	0.302	0.277	0.55	0.384	0.368	0.0	0.266
répartir	0.316	0.302	0.517	0.378	0.312	0.261	0.0

classer* \ *répartir	*trier*	*ranger*	*choisir*	*ordonner*	*ventiler*	*classer*	*répartir*
trier	0.0	0.344	0.455	0.433	0.382	0.299	0.316
ranger	0.345	0.0	0.567	0.426	0.365	0.283	0.307
choisir	0.451	0.563	0.0	0.598	0.515	0.548	0.509
ordonner	0.439	0.428	0.595	0.0	0.432	0.384	0.378
ventiler	0.381	0.358	0.515	0.429	0.0	0.346	0.271
classer	0.303	0.278	0.544	0.383	0.337	0.0	0.268
répartir	0.325	0.312	0.511	0.385	0.279	0.273	0.0