Projet :ANR-09-CORD-026
Titre : VideoSense - Reconnaissance multimodale de concepts enrichis (statiques, dynamiques, émotionnels) dans des vidéos multilingues au travers de langages pivots.
Equipe : GETALP (LIG)

Ensemble de descripteurs textuels (Version provisoire du 7 juin 2011)

Livrables :	L3.4 et L3.5
Auteurs :	Francis Brunet-Manquat, Jérôme Goulian, Alexandre Labadié, Didier Schwab, Gilles Sérasset
Affiliation :	UPMF-UJF-CNRS, LIG-GETALP
Version :	0
Date :	09/06/2011
Type de document :	interne

Table des matières

1 Introduction

Afin de compléter, améliorer ou influencer les descripteurs extraits des séquences vidéo (et/ou audio), le projet vidéosense prend le parti d'exploiter les méta-données textuelles attachées au séquences vidéo formant le corpus d'application. Pour prendre en compte ces méta-données textuelles, nous avons choisi d'utiliser des descripteurs originaux, de nature sémantique et indépendants des langues des méta-données. Ces descripteurs (les vecteurs conceptuels) sont des vecteurs de dimension fixée, chaque dimension représentant une composante sémantique (identifiable ou non, selon l'expérience abordée). Ces descripteurs sont donc de même nature que les descripteurs video/audio et seront donc intégrés au processus de classification au même titre que les autres descripteurs. Dans un premier temps nous évoquons quelques unes des possibilités existant dans l'état de l'art avant de nous focaliser sur les vecteurs conceptuels puis présentons quelques unes des expériences actuellement menées.

2 Les descripteurs textuels : état de l'art

2.1 Introduction générale, définitions et notations

2.1.1 Mot, item lexical, terme

Une définition communément admise considère qu'un mot est une suite de caractères graphiques formant une unité sémantique et pouvant être distinguée par un séparateur (généralement un blanc typographique ou un signe de ponctuation [1]) [Larousse, 2004] [Robert, 2000]. Toutefois cette définition reste relativement floue et souvent inadéquate aux problèmes qui nous sont posés. [2]

Dans ce document, nous utiliserons donc une définition plus "lexicale" et, préférerons en toute rigueur, *item lexical* plutôt que *mot*. Nous définissons un item lexical comme « *une suite de caractères formant une unité sémantique et pouvant constituer une entrée de dictionnaire* ». Ainsi, ‹*voiture*›, ‹*clair*›, ‹*être*› tout comme ‹*pomme de terre*›, ‹*moulin à vent*›, des locutions verbales comme ‹*tirer le diable par la queue*›, des mots ayant un sens différent au singulier et au pluriel comme ‹*orgue*› et ‹*orgues*› [3] et même des termes techniques comme ‹*pompe bivalve à échappement central*› sont des items lexicaux. Nous réserverons *item lexical*, *lemme* ou *terme* pour la forme canonique (cf. morphologie 2.1.3) tandis que nous utiliserons plutôt *mot* pour les formes fléchies. Habituellement, *terme* est, comme son nom l'indique, plutôt utilisé dans un cadre terminologique ; toutefois nous le considérerons ici dans une acception plus large comme désignant une entrée d'un lexique général. Ainsi, nous considérerons *terme* et *lemme* comme parfaitement synonymes d'*item lexical*. Nous insistons sur *item lexical* pour bien mettre l'accent sur l'idée que ce qui nous importe ici est que ce soit une entrée de dictionnaire.

En ce qui concerne les notations, les items lexicaux sont notés en petits caractères, en italique et entre apostrophes (‹*vie*›), les mots en petits caractères et entre guillemets (‹‹vie››).

2.1.2 Niveaux de traitement linguistique

Compréhension et production de textes peuvent se décomposer en une suite de traitements symétriques. La compréhension, comme le présente la figure 1, est constituée d'un *traitement morphologique* qui consiste à identifier les items lexicaux possibles à partir d'une forme fléchie donnée, puis d'un *traitement syntaxique* qui cherche à identifier les relations qui existent entre les mots d'une phrase (sujets, verbes, compléments, ...), d'un *traitement sémantique* qui cherche à capturer le "sens" des phrases et enfin d'un *traitement pragmatique* qui consiste à trouver la signification complète d'un texte liée en grande partie à la présupposition. Une phase de production est décomposable en une séquence inverse. Cette décomposition reste théorique et un certain nombre de cas d'ambiguïtés nécessite des informations issues du niveau suivant.

1. Nous nous restreignons volontairement ici aux langues indo-européennes écrites.
2. Elle ne tient pas compte, par exemple, des locutions non-connexes (où un mot peut s'intercaler dans la locution) comme ‹*mettre les pieds dans le plat*›.
3. Un ‹*orgue*› est un instrument de musique tandis que des ‹*orgues*› (au féminin) sont « *en géologie, des prismes d'une grande régularité formés lors du refroidissement d'une coulée de lave, basaltique le plus souvent, perpendiculaire à la surface.* » [Larousse, 2004]

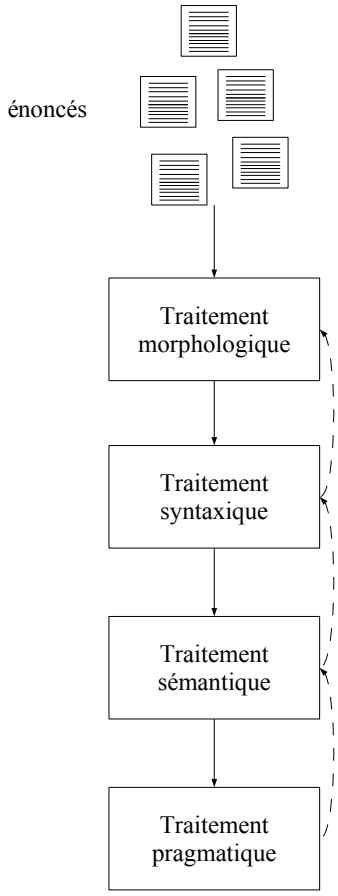

FIGURE 1 – Schéma général d'analyse de textes.

Il serait illusoire de penser que l'analyse d'un énoncé ne consiste qu'à *filtrer* des informations en se limitant, pour chaque niveau, à faire un choix parmi les possibilités proposées au niveau précédent. En effet, la complexité de la langue tant au niveau de la construction des mots, des phrases, des idées qu'au niveau de son évolution (parfois par l'apparition de nouvelles tournures, plus couramment de nouveaux termes souvent de façon éphémère) laisse difficile voire impossible une couverture totale des possibilités de chaque niveau. Ce modèle par couches permet éventuellement à un niveau supérieur de compléter les informations d'un niveau inférieur. Ainsi, dans le cas où le niveau morphologique ignorerait l'existence de ‹*orgues*› (au pluriel) le niveau sémantique pourra, s'il connaît cette forme, la lui indiquer.

Les traitements morphologiques et syntaxiques sont relativement bien connus et étudiés en linguistique. Nous rappelons brièvement dans cette partie les principes sur lesquels ils reposent. Les deux niveaux suivants, en revanche, qui nous intéressent plus particulièrement dans le cadre de VidéoSense, sont encore très méconnus. Ils font l'objet de bien des théories et sont détaillés en 2.2.

2.1.3 Niveau morphologique

Niveau morphologique en linguistique En linguistique, la *morphologie* est l'étude de la façon dont sont formés les mots. On appelle *morphèmes* les unités minimales significatives qui constituent les mots. Par exemple, le mot ‹‹fleurs›› est constitué de deux morphèmes : le radical (ou base) correspondant à l'item ‹*fleur*› et le suffixe marquant le pluriel *s*. Il existe deux catégories de morphèmes :

– les *morphèmes lexicaux* qui correspondent aux items lexicaux ou à une légère variante ;

– les *morphèmes grammaticaux*, autrement appelés *affixes*. Situé avant le radical, un affixe est dit *préfixe*, après le radical, il est dit *suffixe* et dans le radical, *infixe*.

On peut distinguer deux types de formations morphologiques :

– *flexion* : Les mots dits *fléchis* comportent un radical et une ou plusieurs désinences. Les désinences sont les morphèmes porteurs des indications de nombre et de genre pour les noms, adjectifs et déterminants, de temps, de personnes et de mode pour les verbes. Ainsi, ‹‹lisions›› est constitué du radical *lis-* issu de l'item ‹*lire*›, de la désinence temporelle *-i-* et de la désinence personnelle *-ons* ([Lehmann & Martin-Berthet, 1998], p. 132) tandis que ‹‹rattes›› est lui formé par *rat* (radical) + *te* (féminin) + s (pluriel). En aucun cas, la flexion ne modifie la catégorie syntaxique ;
– *dérivation* : On parle de dérivation lorsqu'un mot est formé à partir d'un autre en y adjoignant un ou plusieurs affixes porteurs de sens. Ainsi le sens du radical se trouve modifié et contrairement aux flexions, les dérivations peuvent amener à une nouvelle catégorie syntaxique. Si on prend pour exemple l'adjectif ‹‹inacceptable›› , il est formé de *in* (affixe de contraire) + *accept* (le radical, le verbe ‹*accepter*›) + *able* (affixe de possibilité).

Dans la suite, lorsque nous parlerons de la *morphologie* pour un item lexical ou un mot, il s'agira d'un abus de langage qui désigne les informations que nous pouvons déduire de la morphologie de cet item ou de ce mot. Ainsi, nous aurons les *catégories grammaticales* (*nom, pronom, adjectif, adverbe, etc.*), le *genre* (*masculin, féminin, neutre*), le nombre (*singulier, pluriel*), le mode (*transitif, intransitif*), ...

La *forme canonique* d'un mot est la forme de ce mot telle qu'on peut la trouver comme entrée d'un dictionnaire par opposition à la forme fléchie. Par définition, un item lexical est donc toujours dans une forme canonique. Traditionnellement, suivant la nature de l'item, une forme particulière est choisie :

– *verbe* : à l'infinitif ;
– *nom* : au singulier (s'il existe) ;
– *adjectif* : au masculin singulier

On peut remarquer que, pour les mots invariables, formes fléchie et canonique sont identiques.

Niveau morphologique en TALN Habituellement, dans la phase d'analyse, il s'agit de reconnaître les items lexicaux des textes. Il s'agira ainsi de retrouver à partir d'un mot les possibilités de radical et d'affixe ainsi que leurs caractéristiques grammaticales. Par exemple, ‹*charges*› peut être le nom féminin ‹*charge*› au pluriel comme le verbe ‹*charger*› à l'indicatif ou au subjonctif présent. Cette opération est aussi appelée *lemmatisation*. Une partie des ambiguïtés peut être levée au niveau syntaxique du processus d'analyse.

2.1.4 Niveau syntaxique

Niveau syntaxique en linguistique La syntaxe étudie la manière dont les mots se combinent pour former des phrases. La *structure syntaxique* des phrases, un arbre syntaxique résultant d'une analyse, peut être représentée au moyen de structures de dépendances ou de structures syntagmatiques (appelées également structures de *constituants*).

Dans une représentation syntagmatique, les syntagmes (groupes de mot) ont, comme les mots, leurs propres catégories grammaticales (syntagme verbal, syntagme nominal, syntagme prépositionnel, ...). Les règles syntaxiques permettent de décrire chacun des constituants de la phrase :

– leur *nature* : morphologie pour les mots, verbal, nominal, adjectival pour les syntagmes ;
– leur *structure hiérarchique* : la manière dont les mots se regroupent pour former des syntagmes et la manière dont les syntagmes se regroupent pour former la phrase ;
– leur *fonction syntaxique* : le rôle qu'ils tiennent à leur niveau de la hiérarchie : sujet, verbe, complément, ...

Dans une représentation en dépendances, la structure de la phrase est organisée autour d'un mot (par exemple le verbe) appelé la tête, auquel sont attachés des modificateurs (les noms sujets et compléments) ; chaque modificateur pouvant à son tour posséder des modificateurs. Le concept fondamental associé aux structures de dépendances est celui de *relation* entre les mots. Etant donné deux mots du langage, on établit entre eux une relation de dominé (dépendant) à dominant (gouverneur) et on peut représenter cette relation par un arc entre deux nœuds, chaque nœud étant étiqueté par un mot.

Niveau syntaxique en TALN Après la phase d'analyse morphologique, un certain nombre de solutions sont envisageables pour les mots d'une phrase. Une analyse syntaxique permet grâce aux règles de ne conserver que les solutions qui sont possibles. Par exemple, prenons la phrase « *Des charges supplémentaires seront retenues contre l'accusé.* ». Le mot ‹‹charges››, comme nous l'avons vu dans la partie consacrée à la morphologie, peut être le *nom féminin* ‹*charge*› au pluriel comme le verbe ‹*charger*› à l'indicatif ou au subjonctif présent. La morphologie possible des mots constituant le groupe nominal sujet « *des charges supplémentaires* » (pour ‹‹des››, *déterminant pluriel*, pour ‹‹supplémentaires››, *adjectif masculin ou féminin pluriel*) rendent ici la seule solution possible pour ‹‹charges›› *nom féminin pluriel*.

La tâche pour mettre au point des analyseurs fiables est complexe puisqu'il n'existe pas à l'heure actuelle de règles grammaticales pouvant couvrir l'ensemble des phrases correctes dans aucune des langues existantes. Ainsi, deux grandes familles d'analyseurs coexistent ([Bangalore, 1997], p. 5) :

- *approche symbolique* : ces analyseurs se basent sur des règles grammaticales et nécessitent donc une recherche et une implémentation de ces règles. On peut citer dans cette catégorie, ARIANE [Boitet, 2000], l'analyseur du GREYC [Vergne & Giguet, 1998], IPF [Wehrli, 1992], SYGMART [Chauché, 1984] ;
- *approche statistique* : ces analyseurs se basent sur des méthodes d'apprentissage à partir de corpus annotés manuellement ou automatiquement pour produire des règles pondérées [Church, 1988, Collins, 1997, Muñoz *et al.*, 2000].

Quelle que soit la technique utilisée, les analyseurs syntaxiques ne renvoient pas tous un arbre syntaxique complet.

De même, une analyse syntaxique ne peut pas toujours lever toutes les ambiguïtés. Ainsi, certaines phrases comme « *La petite brise la glace.* » ne peuvent être totalement désambiguïsées à ce niveau de traitement. Deux interprétations syntaxiques sont ici possibles. Dans la première, ‹‹petite›› et ‹‹glace›› sont des noms, ‹‹brise›› est la troisième personne du présent de l'indicatif du verbe ‹*briser*› (ie. une petite fille casse un miroir) tandis que dans la deuxième, ‹‹petite›› correspond à l'adjectif ‹*petit*›, ‹‹glace›› au verbe ‹*glacer*› et ‹‹brise›› à l'item lexical ‹*brise*› (ie. un léger vent donne froid à quelqu'un ou quelque chose de féminin). Si syntaxiquement, il est absolument impossible de lever l'ambiguïté, des informations de nature sémantique et pragmatique sur cette phrase peuvent permettre d'émettre des préférences.

2.2 Qu'est-ce que le sens ? Comment le représenter ?

La sémantique est l'étude du sens des énoncés. Bien que fort ancienne, puisque déjà étudiée par les philosophes de l'Antiquité, cette science fait encore l'objet de bien des recherches notamment car aucun moyen de décrire complètement le sens ne fait aujourd'hui l'unanimité.

Peu d'ouvrages traitant de sémantique se risquent à donner ne serait ce qu'une esquisse de définition du terme ‹*sens*›. En effet, le sens est quelque chose de difficile à décrire, souvent considéré dans ces livres comme déjà acquis par le lecteur. ([Polguère, 2003], p. 98) déroge à cette règle en reconnaissant explicitement le caractère intuitif du sens et en présente une approche dont nous nous inspirons largement ici. Alain Polguère propose comme définition du sens :

Le sens d'une expression linguistique est la propriété qu'elle partage avec toutes ses paraphrases.

Cette définition repose sur la notion d'équivalence entre phrases, les paraphrases. Ces équivalences sont loin d'être rares en langue, c'est même une des caractéristiques essentielles des langues naturelles par rapport aux langages artificiels. Ainsi, pour Polguère, la notion de paraphrase est reconnue comme un concept primitif possédé par un locuteur qui permet de définir la notion de sens.

Le sens d'un énoncé est régi par le *principe de compositionnalité sémantique* pour lequel « *le tout est calculable à partir du sens de ses parties* ». Ainsi, un énoncé est directement calculable (dans sa composition lexicale et sa structure syntaxique) à partir de la combinaison du sens de chacun de ses constituants ([Polguère, 2003], p. 134). Par exemple, le sens d'une phrase comme « *L'enfant voit la mer.* » est calculable à partir :

- des items lexicaux ‹*le*›, ‹*enfant*›, ‹*voir*›, ‹*la*›, ‹*mer*› ;
- des règles syntaxiques et morphologiques du français utilisées dans la phrase.

Il est souvent spécifié dans la littérature que les locutions transgressent, au moins en partie, le principe de compositionnalité sémantique. Dans notre approche où un mot est défini comme une des formes fléchies d'un *item lexical* (notion qui englobe les locutions, cf. 2.1.1), le problème ne se pose pas.

De nombreuses théories sur la sémantique ont été élaborées comme la *sémantique du prototype* [Kleiber, 1990], la *sémantique distributionnelle* ou la *sémantique structurale*. En traitement automatique du langage naturel, il s'agira de trouver le sens d'un texte et pour cela, de désambiguïser le sens des mots qui le composent. Prenons l'exemple de la phrase « *La souris est reliée à l'ordinateur.* ». Les traitements morphologique et syntaxique permettront de savoir que le mot ‹‹souris›› correspond à l'item ‹*souris*›. Considérons (pour simplifier) que ce terme a deux sens : le premier correspondant à l'animal ; le deuxième à la souris d'ordinateur. Le traitement sémantique permettra de trouver un *sens préférentiel* (dans notre exemple, la souris d'ordinateur). Le traitement pragmatique, lui, choisira le "bon sens" en fonction du contexte général. On peut imaginer que nous sommes dans un texte où l'on parle d'une petite souris (l'animal) qui se promène et qui se coince la queue dans le tiroir du lecteur de DVD ; alors le sens préférentiel du traitement sémantique ne sera pas celui choisi au cours du traitement pragmatique.

Dans le cadre du projet VidéoSense, nous sommes confrontés aux trois questions principales posées habituellement par ces problèmes :

- *Comment représenter informatiquement le sens ?*
- *Comment alors désambiguïser les mots d'un texte ?*
- *Comment calculer le sens d'un texte ?*

Plusieurs approches existent. Nous détaillons dans ce document principalement celles qui sont proches des méthodes employées en recherche d'information.

2.3 Représentations d'origine distributionnaliste [4]

2.3.1 Approche distributionnelle

La linguistique distributionnelle [Harris *et al.*, 1989] est le nom donné aux recherches menées aux États-Unis par Zelig Sabbatai Harris (*1909 - 1992*) à partir des années 1950 et qui poursuivaient celles de Léonard Bloomfield (*1887 - 1949*). L'analyse distributionnelle cherche à décrire les objets linguistiques en fonction du pouvoir d'associativité qu'ils possèdent ou ne possèdent pas entre eux. Ainsi, l'objectif premier de cette branche de la linguistique est d'examiner les distributions des unités linguistiques (phonèmes, morphèmes, mots) dans un corpus donné.

La linguistique distributionnelle considère que le sens d'un mot peut être défini à partir de l'ensemble des contextes dans lequel il apparaît, en d'autres termes, par l'ensemble des termes qui lui sont cooccurrents dans un corpus. Par exemple, considérons ces quelques phrases extraites du Web :

- « *Seuls les chatons et pas les chats peuvent boire du lait de vache.* »
- « *Le pédiatre a diagnostiqué une allergie au lait de vache.* »
- « *Dis papa, c'est quoi cette bouteille de lait ?* »
- « *À partir du lait, le fermier fait des fromages et des yaourts.* »

Selon la linguistique distributionnelle, la sémantique de l'item ‹*lait*› peut ainsi être décrite grâce aux termes ‹*vache*›, ‹*bouteille*›, ‹*fromage*›, ‹*yaourt*›, ‹*allergie*›, ‹*chat*›, ‹*chaton*›, ...

On pourra dire que deux mots ont un sens proche s'ils sont employés dans des contextes très voisins. Ce sont ces idées qui ont permis la mise au point des vecteurs saltoniens et de leurs dérivés.

En informatique, le sens d'un texte est donné par un vecteur dont les composantes correspondent directement (modèle vectoriel standard) ou indirectement (LSA) aux items lexicaux constituant le texte.

2.3.2 Représentations saltoniennes

À partir de la fin des années 1960, Gerard Salton [5] (*1927 - 1995*) professeur à la *Cornell University* [6] met au point ce que l'on appelle aujourd'hui le *modèle vectoriel standard* (VSM pour *Vector Space Model*). Son application la plus connue est le système de recherche documentaire SMART [7] [Salton, 1971, Salton & McGill, 1983, Salton, 1991]. Suivant des idées issues de la linguistique distributionnelle, les dimensions de l'espace vectoriel sont associées à des *termes d'indexation*, c'est-à-dire aux termes considérés comme les plus discriminants dans le corpus de recherche.

4. Partie fortement inspirée de [Schwab, 2005]
5. http://www.cs.cornell.edu/Info/Department/Annual95/Faculty/Salton.html
6. Ithica, État de New York, États-Unis d'Amérique.
7. Une version gratuite est accessible gratuitement pour la recherche à l'adresse ftp://ftp.cs.cornell.edu/pub/smart/

Indexation des documents : fabrication des vecteurs Si t est le nombre de termes d'indexation, chaque document (et chaque requête) est représenté par un vecteur à t dimensions tel que :

$$D_i = (p_{i_1}, p_{i_2}, \ldots, p_{i_t}) \tag{1}$$

où p_{i_k} est la k-ième composante de D_i et a pour valeur le poids du terme T_k dans le document D_i. Le poids est souvent calculé par une formule de type *tf* $*$ *idf* (*term frequency* $*$ *inverse document frequency*). Les critères pris en compte sont :

- *l'importance du terme dans le document* : on appelle fréquence d'un terme (*term frequency*) le nombre de fois où ce terme apparaît, on parle aussi du *nombre d'occurrences* ou de la *fréquence d'occurrence*. Ce critère doit permettre de prendre en compte le fait que, plus le terme est présent, plus il a une importance dans le texte ;
- *le pouvoir discriminant du terme* : les mots fréquents dans un texte ne sont pas forcément les plus discriminants par rapport au corpus entier. Par exemple, identifier un grand nombre d'occurrences du terme ‹*lait*› dans un corpus dont le sujet central est justement le lait ne va pas permettre de différencier les divers documents. C'est pour contrebalancer ces cas que la prise en compte de la fréquence inverse en document est nécessaire. Il s'agit d'une évaluation de l'importance du terme dans l'ensemble du corpus. Plus le terme est présent, moindre sera l'idf.

Pour ces deux critères, plusieurs heuristiques peuvent être choisies. Ces dernières sont généralement basées sur la fréquence du terme t dans le document d, notée $f(t,d)$ ainsi que sur le nombre d'occurrences du terme le plus fréquent de d *Max*$(f(t,d))$. Par exemple, pour *tf* on peut trouver :

- *tf* $= f(t,d)$ si on considère que l'importance de l'item n'est donnée que par le nombre d'occurrences dans le texte ;
- *tf* $= log(f(t,d)+1)$: si on considère que l'on doit distinguer de façon moindre deux items ayant un nombre d'occurrences proches si leur fréquence dans le texte est importante et de façon plus importante dans le cas contraire ;
- *tf* $= \frac{f(t,d)}{Max(f(t,d))}$ si on considère que l'importance d'un terme est relative à celle du terme le plus présent dans le document. Notons que cette formule offre aussi l'avantage d'effectuer une certaine normalisation sur les vecteurs produits puisque le poids des composantes n'est pas influencé par la taille du document.

Pour *idf*, les heuristiques sont moins nombreuses, on utilise en général $log(\frac{N}{n})$ où N est le nombre total de documents du corpus et n le nombre de documents du corpus où le terme apparaît au moins une fois.

tf $*$ *idf* est donc la multiplication des valeurs de ces deux critères. Ainsi on pourra choisir comme formule :

$$\frac{f(t,d)}{Max(f(t,d))} \times log(\frac{N}{n}) \tag{2}$$

Exploitation des vecteurs La similarité entre deux documents D_a et D_b (ou entre un document et une requête dans le cas de SMART) est donnée par la formule (parfois dite *du cosinus*) :

$$\mathbb{R}^{\mathrm{t}} \times \mathbb{R}^{\mathrm{t}} \rightarrow [0,1] : \quad \mathrm{sim}(\mathrm{D_a}, \mathrm{D_b}) = \frac{\sum_{\mathrm{k}=1}^{\mathrm{t}} \mathrm{p_{a_k}} \times \mathrm{p_{b_k}}}{\sum_{\mathrm{k}=1}^{\mathrm{t}} \mathrm{p_a}2 \times \sum_{\mathrm{k}=1}^{\mathrm{t}} \mathrm{p_b}2} \tag{3}$$

Les document les plus proches du document (ou de la requête) sont ceux qui maximisent la similarité (l'angle entre les vecteurs est alors le plus petit). On peut ainsi obtenir une liste ordonnée des documents les plus proches d'un autre document ou, dans un cas de recherche d'informations, de la requête.

Problèmes posés par la méthode Le premier problème du modèle vectoriel standard est aussi posé à l'ensemble des représentations vectorielles : la mise à jour de la base ne peut pas se faire de façon incrémentale. En effet, l'utilisation de méthodes basées sur le critère *idf* entraîne obligatoirement le recalcul de l'ensemble des vecteurs lors de l'ajout du moindre document au corpus.

Le second problème concerne le choix des termes d'indexation qui entraîne trois conséquences notables :

- plus le nombre de termes retenus est important, plus fines sont les représentations ;
- plus le nombre de termes retenus est important, plus longues sont les opérations à réaliser (tant en indexation qu'en exploitation) et plus grande est la taille des données à stocker ;

– plus le nombre de termes retenus est important, moins la différence entre les documents les plus proches et les documents les plus éloignés d'un document donné est faible.

Suivant les corpus, le nombre d'item lexicaux différents peut être relativement important. De la sorte, si la méthode utilisée pour choisir les termes d'indexation ne fait que sélectionner ces items, les vecteurs obtenus seront de très grande taille. L'approche doit donc être menée d'une manière plus fine. Elle peut être basée sur un antidictionnaire pour éliminer certains termes inadéquats, sur une stemmatisation pour extraire la racine des termes (tous les mots ayant la même racine seront alors considérés par la même composante), ou sur une lemmatisation.

2.3.3 Une approche psycholinguistique : LSA

Le modèle LSA (*Latent Semantic analysis*), appelé souvent aussi LSI pour *Latent Semantic indexing*, a été créé en psycholinguistique pour simuler l'acquisition de connaissances d'un être humain à partir de grands corpus de textes. Techniquement, LSA est une variante du modèle vectoriel standard qui cherche à la fois à réduire le nombre de dimensions des vecteurs et à améliorer la représentation en rajoutant des informations sur la structure sémantique implicite des unités linguistiques représentées par leurs dépendances cachées [Deerwester *et al.*, 1990].

En effet, les auteurs considèrent que le co-texte d'un item I n'apporte pas suffisamment d'informations sur le sens puisqu'on ne sait rien des liens sémantiques qu'entretiennent les mots de ce co-texte avec les items qui n'apparaissent pas conjointement à I. Par exemple, le co-texte de ‹*chaise*› peut être donné par {‹*s'asseoir*›, ‹*repos*›, ‹*bureau*›, ‹*siège*›, ‹*cuisine*›, ...} mais si un item comme ‹*fauteuil*› n'apparaît pas dans les co-textes de ‹*chaise*›, aucune information sur les rapports sémantiques entre les deux termes ne sera disponible. L'idée est donc de croiser les informations de cooccurrence de chaque item, c'est-à-dire ce que l'on appelle les *affinités de second ordre* [Grefenstette, 1994]. Dans LSA, le sens des termes est donc engendré par les enchaînements de cooccurrences, à savoir les liens implicites. Pour résumer, dans LSA, deux items sont similaires si leurs co-textes sont similaires. Deux co-textes sont similaires s'ils comportent des termes similaires [Lemaire & Dessus, 2003].

Dans un premier temps, la technique LSA consiste à construire à la manière du modèle vectoriel standard des vecteurs correspondant aux mots (dans ce cas l'unité du co-texte utilisée est généralement le paragraphe) ou aux documents. Dans un deuxième temps, il s'agit de regrouper les vecteurs dans une matrice et d'effectuer une décomposition en valeurs propres. Seuls les k premiers vecteurs propres sont pris en compte, l'espace de représentation est donc réduit à k dimensions. Une composante ne correspond pas à un terme particulier, ce qui empêche toute interprétation directe et ne rend possible que les comparaisons entre les vecteurs. La valeur de k ne doit pas être trop importante pour éviter le bruit et doit être suffisamment faible pour éviter les trop grandes pertes d'information. La valeur optimale de k a été estimée empiriquement pour l'anglais autour de 300 [Deerwester *et al.*, 1990].

LSA utilise deux mesures. La première, identique à celle utilisée dans le modèle vectoriel standard, permet d'estimer la similarité entre deux mots ou deux groupes de mots, à partir du cosinus entre les angles des vecteurs correspondants. La seconde mesure caractérise la connaissance que LSA a sur un mot ou sur un groupe de mots, à partir de la longueur du vecteur associé. Cette mesure, beaucoup moins utilisée dans la littérature, dépend de la fréquence des mots et de la diversité des contextes dans lesquels ils apparaissent.

Outre la recherche documentaire, la technique LSA a été utilisée dans plusieurs applications comme l'extraction de métaphores [Kintsch, 2000] ou pour la segmentation automatique des textes [Bestgen, 2004].

Dernièrement, [Gamallo Otero & Bordag, 2011] montre que, comparativement à des modèles plus simples à mettre en œuvre, LSA n'est efficace ni en temps de calcul ni en précision sur une tâche d'extraction de traductions.

2.4 Représentations symboliques connexionnistes

Ces représentations forment des graphes dont les sommets correspondent à des objets lexicaux (item, acceptions) et les arêtes à des relations sémantiques. Parmi ces relations, on peut distinguer les *relations de hiérarchie* (*hyperonymie/hyponymie*, *holonymie/méronymie*) des *relations symétriques* (*synonymie/antonymie*). Les relations sémantiques sont décrites au moyen d'outils formels conçus sur le modèle des fonctions mathématiques : les fonctions lexicales ([Polguère, 2003], p. 131), ([Mel'čuk *et al.*, 1995], p. 127). Il existe autant de fonctions lexicales qu'il existe de liens lexicaux et chaque fonction lexicale est identifiée par un nom particulier. Deux classes de fonctions lexicales sont identifiées : les *fonctions lexicales paradigmatiques* et les *fonctions lexicales syntagmatiques*.

2.4.1 Réseaux sémantiques

Les réseaux sémantiques tirent leur origine de la psychologie expérimentale et plus particulièrement des travaux concernant l'organisation mentale des concepts, menés à la fin des années 1960 [Quillian, 1968].

Un réseau sémantique est une représentation des connaissances sous forme de graphe orienté étiqueté. Comme le dit François Rastier, « *La valeur de connaissance d'un réseau ne réside ni dans ses nœuds, ni dans ses liens, mais dans l'interrelation de ses constituants.* » [Rastier, 2004]. Les nœuds correspondent ainsi aux concepts et les arcs, aux relations entre ces concepts. La relation typique, celle de taxonomie, est la relation *Sorte-de*. Par exemple, pour représenter l'existence d'un *étalon* nommé *Tornado*, on ajoute simplement un nœud au réseau (cf. 2).

(a) Ce réseau représente « *un étalon est un cheval.* ».

(b) On rajoute un nœud au réseau pour indiquer que Tornado est un étalon

FIGURE 2 – Réseaux sémantiques élémentaires.

On constate sur l'exemple 2 que la représentation des deux premières informations (« *Un étalon est un cheval* » et « *Tornado est un étalon* ») permet de déduire facilement par transitivité que « *Tornado est un cheval* ».

L'exemple précédent présente une composition entre deux relations. Il est possible de retrouver les informations contenues dans le graphe par simple *héritage des propriétés* en suivant les arcs *Sorte-de*. Cette composition des relations permet de réaliser une économie d'espace mémoire pour la représentation du réseau mais ralentit le temps de traitement. Le réseau de la figure 3 permet de retrouver que les *étalons* en général et *Tornado* en particulier possèdent des *sabots*.

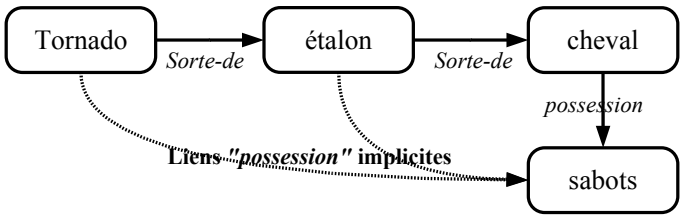

FIGURE 3 – Héritage de propriétés

Tous les raisonnements sur le réseau peuvent être modélisés par une *table de composition des relations* qui contient l'ensemble des compositions autorisées dans le réseau et leurs relations résultantes respectives. Par exemple, le réseau de la figure 3 contient la propriété $\textit{Sorte-de} \circ \textit{possession} = \textit{possession}$.

Dans la famille des réseaux sémantiques, il convient de citer les *graphes conceptuels* (GC) [Sowa, 1984] [Sowa, 2000] [8]. Un GC est un graphe biparti étiqueté dont les deux classes de sommets correspondent à des *concepts* et des *relations conceptuelles* entre ces concepts. La figure 4 présente un exemple de graphe conceptuel avec la phrase « *John va à Boston en bus.* ». L'avantage principal qu'offrent les GC est que le modèle est muni d'une sémantique en logique du premier ordre qui est adéquate et complète par rapport à la déduction. Les applications des graphes conceptuels concernent, entre autres, la génération automatique de langage [Nogier, 1991] ou la recherche d'informations [Genest, 2000].

8. Une pré-version de cet ouvrage est disponible en ligne à l'adresse http://www.jfsowa.com/krbook/index.htm

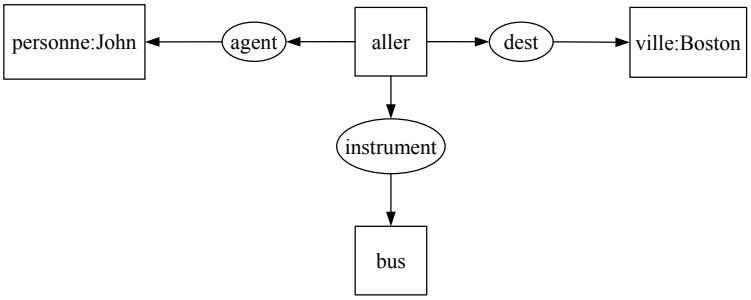

FIGURE 4 – Exemple de graphe conceptuel : « *John va à Boston en bus.* » [Sowa, 2000]

2.4.2 Les réseaux d'aujourd'hui : WordNet

WordNet est une base de données lexicale pour l'anglais développée sous la direction de George Armitage Miller (*né en 1920*) par le *Cognitive Science Laboratory* de l'université de Princeton (États-Unis d'Amérique). Il se veut représentatif du fonctionnement de l'accès au lexique mental humain.

WordNet est organisé en ensembles de synonymes appelés synsets. À chaque synset correspond un concept. Le sens des termes est décrit dans WordNet par trois moyens :

- leur *définition*
- le *synset* auquel ce sens est rattaché.
- les *relations lexicales* qui unissent entre eux les synsets. Ces relations sont ici l'hyperonymie, la méronymie ainsi que l'antonymie.

La version 3 de WordNet [9] compte 155287 termes ce qui constitue une couverture relativement large de la langue anglaise. Les relations lexicales présentes dans WordNet ne connectent que les termes de même morphologie. Il y a donc une hiérarchie pour les noms, une pour les adjectifs, une pour les verbes et enfin une dernière pour les adverbes. Un extrait de la hiérarchie des noms est présenté dans la figure 5.

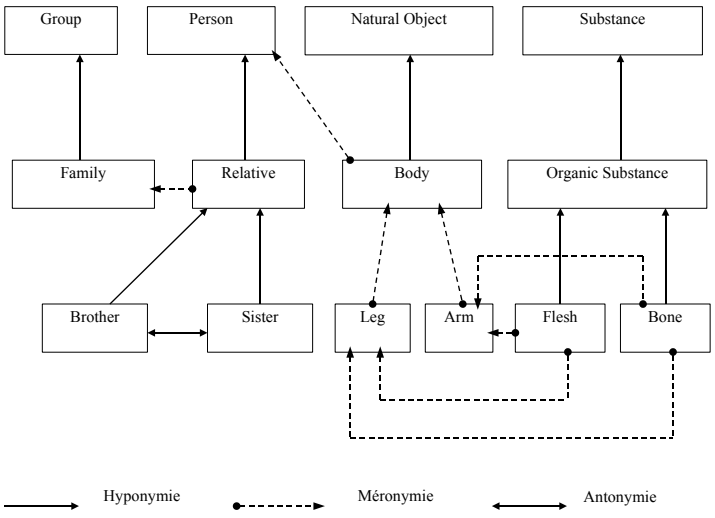

FIGURE 5 – Extrait de la hiérarchie des noms

Dans [Harabagiu *et al.*, 1999], les auteurs de *WordNet* (alors à sa version 1.6) relèvent six faiblesses dans la construction de leur réseau : (1) le manque de liens entre les hiérarchies ; (2) le nombre limité de relations entre

9. Voir http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html.

	sur terre	sur rail	deux roues	individuel	payant	4 à 6 personnes	intra-urbain	transport d'objets	transport de personnes
voiture	+	-	-	+	-	+	~	~	+
taxi	+	-	-	~	+	+	~	~	+
autobus	+	-	-	-	+	-	+	~	+
autocar	+	-	-	-	+	-	-	~	+
métro	+	+	-	-	+	-	+	~	+
train	+	+	-	-	+	-	-	~	+
avion	-	-	-	~	+	~	-	~	+
moto	+	-	+	+	-	-	~	~	+
bicyclette	+	-	+	+	-	-	~	~	+

FIGURE 6 – Analyse sémique des véhicules selon Pottier

termes traitant du même sujet ; (3) le manque de relations morphologiques ; (4) l'absence de relations thématiques ; (5) l'absence de certains sens de mots ; (6) le manque d'uniformisation et de cohérence dans les définitions. Si les points 3, 5 et 6 ne nous intéressent pas dans cet article, nous allons montrer l'apport des vecteurs conceptuels pour la résolution des autres, tous trois formant le problème du tennis.

2.5 Approche componentielle (ou sémique)

2.5.1 Le sens vu comme la composition de primitives

La linguistique componentielle postule que le sens d'un terme peut être défini par un ensemble d'éléments de sens plus petits, appelés, suivant les diverses écoles, sèmes, noèmes, traits sémantiques, atomes de sens, primitives de base... La linguistique componentielle tire son origine des années 1940 et des travaux de Hjelmslev [Hjelmlev, 1968] sur l'analyse en composants sémantiques (comparaison des termes en fonction des sèmes qui les composent).

L'analyse sémique s'attache à identifier pour un certain nombre de termes l'ensemble des sèmes qu'ils comportent. Même si ces sèmes ne sont pas des atomes de sens, mais des traits distinctifs, ils supposent l'existence dans l'esprit humain de primitives de sens, les sèmes n'en étant alors que des compositions s'opposant entre elles. Ainsi, contrairement à la distributionnalité présentée en 2.3.1 qui est une théorie purement linguistique et qui donc ne repose pas sur un postulat cognitif, il s'agit ici de comprendre comment les mots coexistent dans notre esprit ([Nyckees, 1998], p. 216).

Parmi les analyses sémiques les plus connues figurent celles effectuées par Bernard Pottier [Pottier, 1964]. La figure 6 présente l'analyse sémique de certains véhicules. Les signes + et - marquent la présence ou l'absence du trait tandis que ~ spécifie que le trait est indifférent.

Par ailleurs, les primitives de sens doivent permettre d'exprimer la signification de tout énoncé quelle que soit la langue dans laquelle il est exprimé. Ainsi, dans la théorie atomiste, les primitives sémantiques sont nécessairement universelles et surtout elles sont à la fois *indépendantes* et *antérieures* au langage. Ainsi, elles devraient nécessairement se retrouver présentes dans toutes les langues.

Limites de l'approche componentielle

Chez les linguistes À la suite d'études approfondies sur les langues les plus diverses durant presque trente ans, une liste de 35 primitifs universaux a été présentée [Wierzbicka, 1993]. Le principal problème posé par cette liste est que, là où on simplifie les choses en présentant un nombre fort restreint de primitives, on augmente singulièrement la difficulté de représenter le sens d'un terme. Un deuxième problème concernant cette vision est son caractère utopique. En effet, la notion ne rentre dans les primitifs que si elle est universelle donc si elle est présente dans l'ensemble des langues du monde. Or du fait de leur grand nombre (environ 6000), cette étude semble particulièrement difficile à réaliser. Pour cette raison purement pratique, on ne pourra donc jamais être totalement certain de l'universalité d'un concept.

Chez les informaticiens Les créateurs des systèmes informatiques des années 1970 comme Schank [Schank, 1972] et Wilks sont directement héritiers de l'analyse sémique et à ce titre, ils cherchent principalement quelles primitives permettraient de représenter l'ensemble des sens en langue. Ainsi, Yorick Wilks énonce quelques critères très généraux pour fabriquer un ensemble de primitives [Wilks, 1977] :

1. *finitude* : l'ensemble de primitives doit être fini et de relativement faible dimension. En particulier, cette dernière doit être très largement inférieure au nombre de sens à décrire ;
2. *étendue* : les primitives doivent couvrir l'ensemble de l'intervalle des sens à exprimer ;
3. *complétude* : toutes les informations sur le sens d'une entité doivent pouvoir être décrites grâce à l'ensemble de primitives ;
4. *canonicité* : la description d'une entité doit être unique et non-ambiguë ;
5. *indépendance* : aucune primitive ne doit pouvoir être décomposable en un ensemble d'autres ;
6. *non-réductibilité* : l'ensemble de primitives ne peut être remplacé par un ensemble plus petit.

Ces recherches ont fait l'objet de nombreuses discussions tout au long des années 1970 [Winograd, 1978] et jusqu'aux années 1980. Les systèmes basés sur ces primitives étaient lourds et les résultats loin d'être satisfaisants. Les critères proposés pour construire ces listes sont souvent jugés trop généraux pour être utiles mais comme le note [Sabah, 1996], « *les tentatives de réfutation n'ont pas apporté d'idées beaucoup plus constructives en tout cas pour les mises en oeuvre informatiques* ».

Le problème de l'antériorité et de l'indépendance au langage Les tenants de l'approche componentielle considèrent que tous les locuteurs humains partagent un ensemble d'atomes de sens et donc que ceux-ci sont alors forcément antérieurs au langage. Pottier présente plusieurs arguments qui sont, selon lui, favorables à ces idées et qui recoupent ceux que nous avons en partie déjà constatés dans les parties précédentes :

- *traductions* : il semble possible d'effectuer des traductions entre tout couple de langues, du français au chinois, du chinois à l'égyptien,... Il doit ainsi exister un espace conceptuel hors langage commun à l'ensemble de l'humanité qui permet le passage d'une langue à une autre ;
- *acquisition des informations* : en France pour la plupart des gens l'année *1515* évoque la bataille de Marignan tout comme *1789* évoque le début de la Révolution française. Pourtant, qui se souvient exactement où, quand et comment il a appris ces dates ? Les a-t-on lues, entendues ? Au mieux, on croit se souvenir l'avoir appris à l'école mais rien n'est vraiment sûr. Pourtant, on a retenu ces notions. Il semble donc exister un niveau conceptuel indépendant du langage.

L'argument de la traduction est toutefois très contestable. En effet, la traduction est en grande partie le fruit de compromis des traducteurs. Certains concepts issus de la culture et de l'environnement des locuteurs sont présents dans des langues et ne le sont pas dans d'autres. Il s'agit donc pour un traducteur de chercher dans l'autre langue comment exprimer le mieux possible les idées d'un énoncé.

L'antériorité et surtout l'indépendance sont aussi largement contestables. L'évolution culturelle de l'Homme s'est fortement accélérée lorsque celui-ci a acquis le langage. Il a été plus à même de transmettre aux générations suivantes comment couper la viande, ce qui était bon, ce qui était dangereux. Les peuples ont acquis des savoirs, acquis des croyances. Ainsi, la plupart des concepts humains se sont trouvés à la fois qualitativement et quantitativement modifiés par l'apparition des langues, et considérablement réorganisés par les échanges entre les êtres humains ([Nyckees, 1998], p. 220).

2.5.2 Les proto-vecteurs d'idées : une première expérience utilisant des listes préétablies.

Au début des années 1990, Jacques Chauché, dans le but de réaliser un système de Traduction Automatique, propose de représenter le "sens" [10] des items lexicaux grâce à un espace vectoriel dont les axes seraient associés à un ensemble de concepts définis *a priori*. Dans une telle expérience, le choix de cet ensemble définit l'espace vectoriel et est donc, par conséquent, très important. Jacques Chauché considère que ce choix doit être « *assez général pour*

10. Dans [Chauché, 1990] Jacques Chauché met lui-même sens entre guillemets.

permettre le codage d'un mot quelconque et ne doit pas être construit pour l'expérience afin d'éviter une prédétermination des sens. ». Il préfère ainsi utiliser une liste de 416 concepts déjà définie par les rédacteurs de l'encyclopédie Universalis pour leur *organum* [Universalis, 1968].

Le sens d'un item est défini comme un vecteur de cet espace. Pour construire un vecteur, il associe au sens à définir un ensemble de concepts proches sémantiquement. Quatre types d'associations sont définies : associations fortes, associations faibles et leurs contraires. Ces derniers ont été introduits en prévision d'un traitement futur de l'antonymie, mais finalement ne semblent jamais avoir été employés. Ainsi, si le concept A est opposé au concept B et si la liste des associations fortes positives contient A, celle des associations fortes négatives contiendra B. Les poids choisis sont de 1 pour les associations fortes et de $0,5$ pour les associations faibles. Par exemple, l'item ‹*valeur*› dans son sens de **prix (sens commercial)** noté *valeur*/**prix**, est associé aux concepts suivants :

association forte	:	*prix, commerce*
association faible	:	*monnaie*

Le vecteur de ‹*valeur*› calculé à partir de ces associations est donc $(1; 1; 0,5)$ dans l'espace vectoriel à trois dimensions qui a pour axes *(prix, commerce, monnaie)*. On voit ici une différence majeure avec la théorie componentielle classique. Alors que celle-ci considère les concepts comme des primitives, des atomes et les utilise donc de manière booléenne (le concept est présent ou non), les proto-vecteurs d'idées ne considèrent pas les concepts comme des atomes et donc permettent de quantifier l'importance du concept, de l'idée, dans le terme.

Dans l'expérience présentée [Chauché, 1990], les associations ont été définies par 6 personnes différentes ce qui a amené à des différences notables.

Ainsi, pour le terme ‹*bilan*›, un premier codeur a choisi :

– *Association forte* : *ACCUMULATION*, *CAPITAL*, *CONVERGENCE*, *DÉNOMBREMENT*, *GESTION* ;
– *Association faible* : *ASSOCIATION*, *CONNAISSANCE*, *HISTOIRE*, *INDUCTION*, *INFORMATION*, *INTÉGRATION-DES-SENS-DATA*.

Tandis qu'un second associe lui :

– *Association forte* : *AVOIR*, *CONNAISSANCE*, *BIEN*, *DESCRIPTION*, *INFORMATION*, *QUANTIFICATION*, *REPRÉSENTATION* ;
– *Associations faible* : *CAPITAL*, *ACCUMULATION*, *ACQUIS*, *APPROXIMATION*, *CRÉDIT*, *MESSAGE*, *OBSERVATION*, *PROPRIÉTÉ*, *POSSESSION*, *PREUVE*, *REFLET*, *SIGNAL*, *SOURCE*.

Pour comparer les sens entre eux, la distance utilisée est la distance euclidienne. Lors d'une désambiguïsation, le sens choisi sera celui dont le vecteur sera le plus proche des termes de référence. Ainsi, si on compare le sens de ‹*bilan*› présenté ci-dessus avec les différents sens possibles de l'item ‹*cours*›, on obtient :

1. *cours*/**monnaie** : $94,0$
2. *cours*/**durée** : $104,5$
3. *cours*/**déplacement** : $105,5$
4. *cours*/**polycopié** : $106,25$
5. *cours*/**niveau** : $107,5$
6. *cours*/**enseignement** : $108,25$
7. *cours*/**rue** : $108,5$

Le sens choisi dans ce cas est le premier, *cours*/**monnaie**.

Ce modèle vectoriel est le précurseur de celui que nous utilisons dans le projet VidéoSense. Nos descripteurs textuels (vecteurs conceptuels basés sur le thésaurus Larousse) sont introduits ci-après et détaillés dans la partie 3.

2.6 Les vecteurs d'idées ; une variante : les vecteurs conceptuels

Comme nous l'avons vu précédemment (cf. 2.5.1), les tenants de l'approche componentielle, considèrent que tous les locuteurs humains partagent un ensemble d'atomes de sens et donc qu'ils sont forcément antérieurs au langage. Toutefois, deux objections peuvent être soulevées. La première, d'ordre cognitif, considère que l'évolution de l'homme et sa différenciation avec les autres espèces animales s'est réellement accélérée du fait de l'invention du langage. La seconde, d'ordre plus pragmatique, concerne la difficulté à trouver une combinaison de primitives de base permettant de représenter le sens d'un terme lorsqu'elles sont trop peu nombreuses et ainsi trop abstraites.

Il ne s'agit pas, pour nous, de formuler des hypothèses sur l'organisation des concepts chez l'humain mais plutôt de chercher à représenter le sens par une méthode à la fois calculable et efficace. Nous préférons ainsi considérer un ensemble de concepts qui ne seraient pas forcément indépendants les uns des autres mais grâce auxquels il serait relativement aisé de définir les sens des termes. Les travaux de Jacques Chauché (cf. 2.5.2) ont montré la faisabilité d'une telle approche.

Ces concepts ne sont pas alors à envisager comme des concepts correspondant à un être humain en particulier mais plutôt comme les concepts fondamentaux d'une société humaine particulière dont les membres partagent un certain nombre de faits culturels. Ils évoluent au cours de l'histoire et de l'acquisition de nouvelles techniques. Des concepts comme ceux concernant le feu ou les outils sont apparus durant la préhistoire, ceux qui concernent les téléphones portables ou les ordinateurs peuvent aujourd'hui être considérés alors qu'ils ne l'auraient pas été il y a cent ou cinquante ans. C'est pour cette raison que nous considérerons qu'un ensemble de primitives de sens ne devrait et ne pourrait être choisi que pour une certaine société humaine à une certaine époque. Il s'agit de considérer un ensemble de concepts permettant de représenter l'ensemble des idées exprimables pour une langue à une époque donnée.

Le modèle des vecteurs d'idées est la projection de la notion linguistique de champ sémantique dans le modèle mathématique d'espace vectoriel. Cette approche vectorielle est fondée sur des propriétés mathématiques bien connues sur lesquelles il est possible d'effectuer des manipulations formellement pertinentes auxquelles sont attachées des interprétations linguistiques raisonnables. Chaque segment textuel peut ainsi se voir attribuer un vecteur représentant les idées qu'il véhicule et une distance basée sur l'angle entre deux vecteurs peut alors être introduite pour pouvoir estimer la proximité thématique entre deux segments. À partir d'une base de données contenant l'indexation des termes d'une langue, une opération d'*analyse sémantique* permet de calculer un vecteur pour un texte donné. Cette opération est à la base des applications principales des vecteurs d'idées : traduction, recherche d'informations, résumé automatique, ...

Nous ne détaillons ici qu'une des deux variantes des vecteurs d'idées : celle des *vecteurs conceptuels*, construits grâce à un apprentissage à partir de dictionnaires à usage humain présentés sous forme électronique.

2.6.1 Modèle des vecteurs d'idées

Le modèle des vecteurs d'idées est basé sur la projection de la notion linguistique de champ sémantique dans le modèle mathématique d'espace vectoriel. À partir d'un ensemble de notions élémentaires, les concepts, représentés sous forme de vecteurs (dits *vecteurs génératifs*), on peut construire de nouveaux vecteurs d'idées et les associer à tout segment textuel (items lexicaux, phrases, textes, ...). Il est ainsi possible d'effectuer des manipulations formellement bien fondées auxquelles nous pouvons attacher des interprétations linguistiques raisonnables. L'hypothèse principale sur laquelle repose ce modèle est que les vecteurs génératifs constituent un espace générateur pour l'ensemble des mots de la langue.

Depuis 1998, plusieurs expérimentations sur les vecteurs d'idées ont été expérimentées. Toutes se sont faites parallèlement et s'influençaient largement les unes les autres. Nous nous intéressons plus particulièrement ici à trois expériences basées sur les vecteurs conceptuels. La première a été implémentée pendant environ six ans (1999-2005) par Mathieu Lafourcade, la seconde par Didier Schwab pour sa thèse (2001-2006) enrichie d'une expérience menée conjointement par Lim Lian Tze à l'*Universiti Sains Malaysia* sur l'anglais (2006-2007). La troisième est celle menée dans le cadre du projet Videosense.

Historiquement, les vecteurs d'idées sont le prolongement des travaux effectués par Jacques Chauché au début des années 1990 que nous avons présentés au chapitre précédent (cf. proto-vecteurs d'idées section 2.5.2).

Vecteurs génératifs et espace des vecteurs d'idées Le principe de base des vecteurs d'idées est semblable à celui de la linguistique componentielle. Il suppose l'existence d'une atomisation de la signification, c'est-à-dire que le sens d'un terme peut être décomposé en éléments de sens plus petits (cf. 2.5). Dans notre modèle, nous considérons que ces éléments de sens plus petits, que nous appelons *concepts*, peuvent être représentés par des vecteurs de $\mathbb{R}^{+n}$ et qu'ils sont susceptibles de générer l'ensemble des vecteurs d'idées. Les vecteurs des concepts sont appelés *vecteurs génératifs*.

Nous notons l'ensemble des vecteurs d'idées ϑ, celui correspondant aux items lexicaux est noté ω. Les concepts sont notés (à la Mel'čuk) en majuscules (VIE), les items en italique et entre guillemets (‹*vie*›). Enfin, $V(x)$ correspond au vecteur d'idées d'un élément x quel que soit cet élément, item lexical, concept ou segment textuel.

Vecteurs d'idées, première approximation Nous pouvons considérer, en première approximation, que les vecteurs d'idées sont construits grâce à des combinaisons linéaires de vecteurs génératifs. Soit $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ un ensemble fini de n concepts, soit $\mathcal{F} = \{V(c_1), V(c_2), \ldots, V(c_n)\}$ la famille de vecteurs correspondants, soit l un item lexical et soit $\alpha_i \in \mathbb{R}^+$ l'intensité de c_i dans l, alors nous posons que :

$$V(l) = \frac{V}{\| V \|} \text{ où } V = \sum_{i=1}^{n} \alpha_i V(c_i) \tag{4}$$

Un vecteur d'idées est le vecteur normé d'une combinaison linéaire des vecteurs génératifs. Par exemple, si nous considérons que ‹*Ferrari*› peut être construit à partir des idées *VOITURE*, *ROUGE*, *RAPIDE*, le vecteur d'idée associé est :

$$V(\text{‹}\textit{Ferrari}\text{›}) = \frac{\alpha_{\textit{VOITURE}} V(\textit{VOITURE}) + \alpha_{\textit{RAPIDITÉ}} V(\textit{RAPIDITÉ}) + \alpha_{\textit{ROUGE}} V(\textit{ROUGE})}{\| \alpha_{\textit{VOITURE}} V(\textit{VOITURE}) + \alpha_{\textit{RAPIDITÉ}} V(\textit{RAPIDITÉ}) + \alpha_{\textit{ROUGE}} V(\textit{ROUGE}) \|} \tag{5}$$

La valeur de $\alpha_{\textit{VOITURE}}$ (respectivement $\alpha_{\textit{RAPIDITÉ}}$, $\alpha_{\textit{ROUGE}}$) est alors fonction de l'importance de l'idée dans l'item lexical ‹*Ferrari*›.

Espace des vecteurs d'idées et interprétation linguistique Les vecteurs d'idées sont des vecteurs de $\mathbb{R}^n$ où n correspond au nombre de vecteurs génératifs. lorsqu'ils sont associés à des segments textuels, ils sont normés à 1 et forment donc une hyper-sphère de rayon 1. En pratique, plus n est grand, plus fines sont les descriptions de sens offertes par les vecteurs, mais plus leur manipulation informatique est lourde. Le choix de n doit donc être un compromis entre une meilleure finesse de la représentation et des contraintes matérielles.

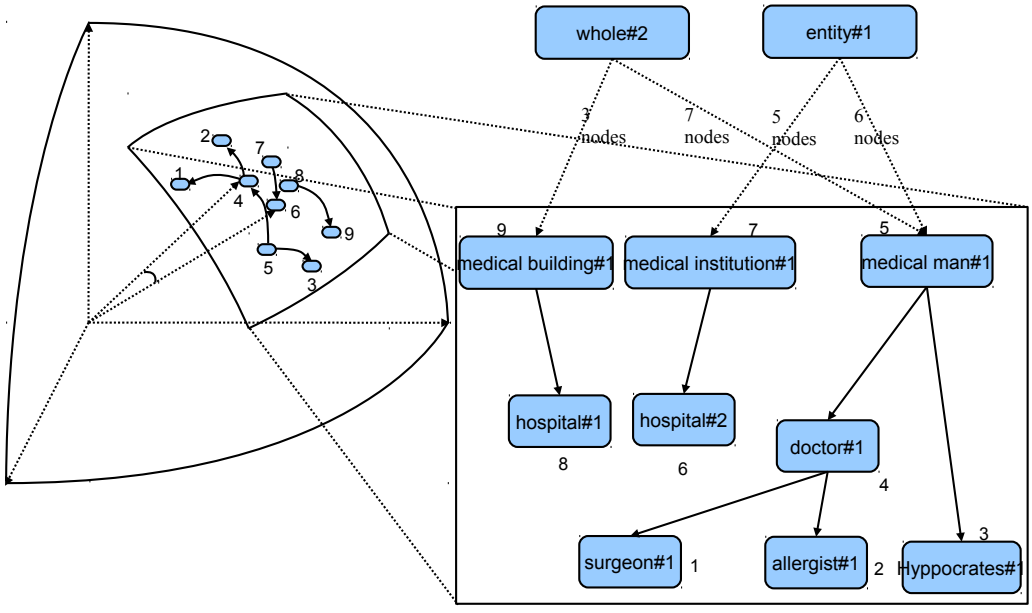

FIGURE 7 – Projection d'un réseau (ici WordNet) sur une hyper-sphère de rayon 1

En pratique, par construction, les composantes des vecteurs d'idées sont positives. On peut donc dire que ce sont des vecteurs de $\mathbb{R}^{+n}$. En toute généralité, et pour respecter les axiomes des espaces vectoriels normés, nous n'oublierons pas que les vecteurs d'idées sont des vecteurs de $\mathbb{R}^n$. Toutefois, par souci de simplification, lorsqu'il s'agira de définir les opérations sur les vecteurs, nous prendrons en compte la positivité des composantes des vecteurs, en particulier en ce qui concerne les domaines de définition.

Précisons toutefois que l'espace des vecteurs d'idées n'est pas un sous-espace vectoriel dont les vecteurs génératifs seraient une base ou une famille génératrice.

Formellement, nous avons $\vartheta \equiv \mathbb{R}^n$. Toutes les opérations vectorielles qu'il est possible d'effectuer dans un espace vectoriel normé sont donc réalisables dans ϑ. La différence est que nous cherchons à donner à ces vecteurs une interprétation linguistique voire psycholinguistique. Ils représentent des idées qui peuvent faire référence éventuellement à un item lexical, un terme de la langue. Les opérations réalisables sur les vecteurs d'idées peuvent donc elles aussi avoir des interprétations linguistiques qu'il convient d'analyser.

Vecteurs normés Dans notre modèle, la norme n'est pas considérée comme une information qualitative. En effet, nous considérons que les idées ne prennent sens que si elles sont appréciées les unes par rapport aux autres. Cette affirmation est vraie tant à l'intérieur des vecteurs qu'entre les vecteurs. Ainsi, il est plus pertinent de comparer les proportions des différentes idées à l'intérieur d'un terme, d'un vecteur, plutôt que de les analyser de façon absolue.

Il en est de même entre deux vecteurs. Que signifierait la comparaison de deux vecteurs qui n'ont pas la même norme ? Si on prend l'exemple du calcul du vecteur d'idées d'un texte quelconque par une méthode d'analyse sémantique (cf. 2.6.4), le principe est, en schématisant beaucoup, de faire une somme pondérée des vecteurs. Si les vecteurs n'étaient pas normés, la norme d'un vecteur serait un simple indicateur de la longueur du texte à partir duquel il a été construit.

Si, au cours des opérations que nous présenterons par la suite (cf. 2.6.3), la norme d'un vecteur n'est pas toujours égale à l'unité, en particulier dans le cas de l'utilisation du produit terme à terme, il n'en est jamais de même lorsqu'il s'agit d'un vecteur correspondant à un segment textuel. Par exemple, les vecteurs d'idées stockés dans une base de données vectorielle ont tous une norme égale à 1.

2.6.2 Distance et voisinage thématique

Un des premiers outils utilisés pour vérifier la cohérence d'une base de vecteurs d'idées a été la fonction de voisinage thématique. Elle permet de connaître les items dont le vecteur est le plus proche du vecteur d'un item donné. Cette fonction de voisinage est basée sur la notion de distance thématique.

Distance thématique Il est souvent souhaitable de pouvoir mesurer la proximité entre deux items, c'est-à-dire une distance (ou au moins une mesure) entre leurs vecteurs d'idées. Ces opérations peuvent permettre non seulement d'estimer la cohérence d'une base de vecteurs mais aussi s'avèrent souvent déterminantes pour connaître le sens d'un terme. Il existe de nombreuses distances possibles entre vecteurs, la thèse de Romaric Besançon ([Besancon, 2001], section 2.2.5) en fait une bonne synthèse. Dans le cadre des vecteurs d'idées, la distance angulaire est utilisée pour les opérations de base. En effet, elle permet une meilleure discrimination pour les faibles angles et offre d'intéressantes interprétations géométriques.

Par abus de langage nous parlerons parfois de distance entre deux segments textuels au lieu de parler de distance entre les vecteurs associés à ces deux segments.

Similarité et distance angulaire Soit $Sim(X, Y)$ une des mesures de *similarité* entre deux vecteurs X et Y, utilisée habituellement en recherche d'informations ([Salton & McGill, 1983], p. 121). Cette valeur est le cosinus de l'angle entre les deux vecteurs.

$$\vartheta^2 \rightarrow [0, 1] : \quad Sim(X, Y) = \cos(\widehat{X, Y}) = \frac{X \cdot Y}{\|X\| \times \|Y\|} \tag{6}$$

Nous définissons la fonction de *distance thématique* D_A entre deux vecteurs X et Y comme la distance angulaire entre les deux vecteurs.

$$\vartheta^2 \rightarrow [0, \frac{\pi}{2}] : \quad D_A(X, Y) = \arccos(Sim(X, Y)) \tag{7}$$

Par définition, nous posons :

$$D_A(\vec{\mathbf{0}}, \vec{\mathbf{0}}) = 0 \quad et \quad D_A(X, \vec{\mathbf{0}}) = \frac{\pi}{2} \text{ si } X \neq \vec{\mathbf{0}} \tag{8}$$

avec $\vec{\mathbf{0}}$ dénotant le vecteur nul. Précisons que ce vecteur n'a sans doute pas de représentation en langue. Il s'agirait d'un mot qui n'active aucun concept, l'idée vide.

La distance angulaire est une vraie distance puisqu'elle vérifie les propriétés de réflexivité, symétrie et inégalité triangulaire.

$$séparation \ : D_A(X, Y) = 0 \Leftrightarrow X = Y \tag{9}$$

$$symétrie \ : D_A(X, Y) = D_A(Y, X) \tag{10}$$

$$inégalité\ triangulaire \ : D_A(A, B) + D_A(B, C) \geq D_A(A, C) \tag{11}$$

Pourquoi choisir l'angle entre les deux vecteurs ? Observons la figure 8. L'arc cosinus est une fonction décroissante par rapport à la similarité. Plus cette dernière est grande, plus l'angle entre les deux vecteurs est important. Cette fonction est intéressante dans notre problématique car elle est fortement non linéaire par rapport à la similarité pour les faibles valeurs de l'angle (en dessous de $\frac{\pi}{4}$, là où les comparaisons à effectuer sont les plus fines) tandis qu'elle est pratiquement linéaire au-delà de $\frac{\pi}{4}$.

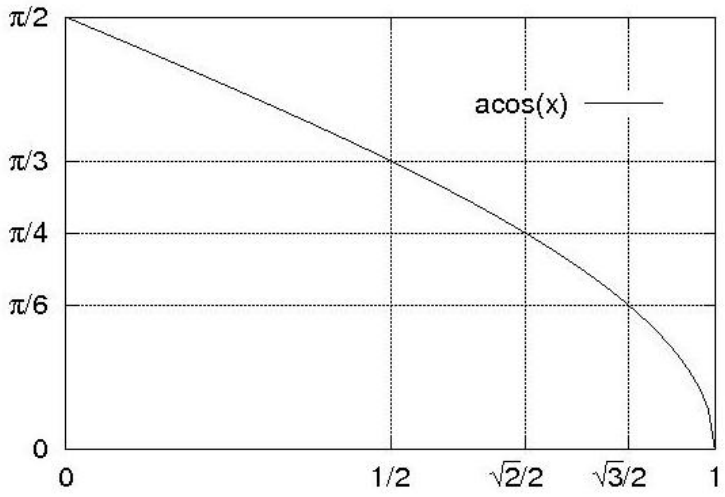

FIGURE 8 – fonction arc cosinus

En effet, un des objectifs des vecteurs d'idées est de participer à des tâches de désambiguïsation sémantique. Il peut arriver que les différences entre les sens soient relativement faibles. Par exemple, l'éloignement des sens *agneau*/**viande** et *agneau*/**animal** semble assez peu important. C'est une des propriétés qui a entraîné le choix de l'angle entre les vecteurs comme l'un de nos principaux outils. La deuxième concerne les relativement bonnes interprétations géométriques qu'il est possible d'effectuer.

Exemples Le tableau 9 présente les distances thématiques (en radians) entre les vecteurs de plusieurs termes. Ces exemples sont réalisés à partir de la base de vecteurs conceptuels dont l'architecture est présentée en 3.

Le tableau est symétrique (symétrie de D_A(X, Y)) et la diagonale est toujours égale à 0 (réflexivité de D_A(X, Y)). On remarquera qu'une valeur prend toute sa signification relativement à une autre. En particulier, il est satisfaisant d'avoir :

- D_A(ᶜ*destin*ᵉ, ᶜ*destinée*ᵉ) $\leq D_A$(ᶜ*destinée*ᵉ, ᶜ*vie*ᵉ) et D_A(ᶜ*existence*ᵉ, ᶜ*vie*ᵉ) $\leq D_A$(ᶜ*destinée*ᵉ, ᶜ*vie*ᵉ) ce qui correspond bien au fait que ᶜ*destin*ᵉ et ᶜ*destinée*ᵉ d'une part, et ᶜ*existence*ᵉ et ᶜ*vie*ᵉ sont plus proches que ᶜ*destinée*ᵉ et ᶜ*vie*ᵉ.
- D_A(ᶜ*vie*ᵉ, ᶜ*mort*ᵉ) $> \frac{\pi}{4}$ (0,78) ce qui dénote un certain éloignement des idées.

D_A	destinée	destin	vie	existence	mort	automobile	train	action	inaction	réaction
destinée	0	0,51	0,82	0,7	0,99	1,29	1,38	1,31	1,14	1,2
destin		0	0,83	0,75	0,99	1,3	1,38	1,25	1,07	1,16
vie			0	0,61	0,89	1,28	1,35	1,3	1,1	1,2
existence				0	0,98	1,37	1,43	1,37	1,25	1,3
mort					0	1,33	1,4	1,32	1,15	1,26
automobile						0	0,88	1,4	1,22	1,29
train							0	1,43	1,3	1,39
action								0	1,01	0,67
inaction									0	0,9
réaction										0

FIGURE 9 – Exemples de résultats de la distance thématique $D_A(\mathrm{X}, \mathrm{Y})$.

- $D_A(\text{‹}vie\text{›}, \text{‹}automobile\text{›}) > \frac{\pi}{3}$ (1,04) ce qui relève d'un éloignement important.
- $D_A(\text{‹}action\text{›}, \text{‹}réaction\text{›})$ est la plus petite valeur de $D_A(\text{‹}action\text{›}, \mathrm{Y})$ car les deux concepts *ACTION* et *RÉACTION* sont relativement proches et que ‹*action*› partage beaucoup moins d'idées avec les autres termes.

Interprétations On remarquera qu'habituellement, les comparaisons entre les valeurs sont plus significatives que les valeurs elles-mêmes, toutefois, on estime empiriquement que :

- si $D_A(X, Y) \leq \frac{\pi}{4}$, X et Y partagent des concepts et sont considérés comme sémantiquement proche.
- si $D_A(X, Y) \geq \frac{\pi}{4}$, la proximité sémantique de A et B est considérée comme faible.
- Aux alentours de $\frac{\pi}{2}$, les sens sont sans rapport.

Intuitivement, cette fonction constitue une évaluation possible de la *proximité thématique*. La métaphore de la nuit étoilée peut aider à appréhender cette idée de distance angulaire pour calculer la proximité thématique. Nous pouvons nous représenter l'espace des sens comme un ciel rempli d'étoiles. Les étoiles sont les items lexicaux. Les mots, tout comme les étoiles, forment des constellations. Certaines parties de l'espace sont très densément peuplées tandis que d'autres sont quasi-désertes. Un sens est une direction de l'espace et non un point. Un observateur ne peut connaître exactement la distance entre une étoile et lui-même mais il connaît la direction de l'astre. Dans le ciel, la distance entre deux étoiles est la distance apparente, l'angle entre les deux. Il en est de même, dans notre espace, avec les items lexicaux.

Voisinage thématique La fonction de voisinage thématique permet de connaître les items lexicaux voisins d'un item lexical donné. On définit $\mathcal{V}$ la fonction de proximité thématique qui renvoie les k items les plus proches en termes de distance angulaire d'un texte Z dans une base vectorielle. Soit :

$$|\mathcal{V}(Z)| = k \qquad \forall X \in \mathcal{V}(Z), \quad \forall Y \notin \mathcal{V}(Z), \quad D_A(\mathrm{X}, \mathrm{Y}) \leq D_A(\mathrm{Y}, \mathrm{Z}) \tag{12}$$

Par exemple, les termes proches et ordonnés par distance thématique croissante des mots ‹*vie*›, ‹*ranger*› et ‹*couper*› pourraient être :

$\mathcal{V}(\text{‹}vie\text{›})$ = ‹*vie quotidienne*›, *VIE*, ‹*s'animer*›, ‹*demi-vie*›, ‹*survivant*›, ‹*avoir la vie devant soi*›, ‹*naissance*›, ‹*viabilité*›, ‹*vital*›, ‹*naître*›, ‹*vivant*›, ‹*assurance-vie*›,...

$\mathcal{V}(\text{‹}ranger\text{›})$ = ‹*trier*›, ‹*cataloguer*›, ‹*sélectionner*›, ‹*classer*›, ‹*distribuer*›, ‹*grouper*›, ‹*ordonner*›, ‹*répartir*›, ‹*aligner*›, ‹*caser*›, ‹*arranger*›, ‹*nettoyer*›, ‹*distribuer*›, ‹*démêler*›, ‹*ajuster*›, ...

$\mathcal{V}(\text{‹}couper\text{›})$ = ‹*cisailler*›, ‹*émincer*›, ‹*scier*›, ‹*tronçonner*› , ‹*ébarber*›, ‹*entrecouper*›, ‹*baptiser*›, ‹*recouper*›, ‹*sectionner*›, ‹*bêcher*›, ‹*hongrer*›, ‹*essoriller*›, ‹*rogner*›, ‹*égorger*›, ‹*écimer*›, ...

2.6.3 Opérations classiques

Nous présentons ici les opérations définies pour les vecteurs d'idées que nous utilisons dans ce document.

Somme vectorielle

Définition Soient X et Y deux vecteurs, leur *somme vectorielle* V est définie par :

$$\vartheta^2 \rightarrow \vartheta \ : V = X + Y \quad | \quad V_i = X_i + Y_i \tag{13}$$

où V_i (resp X_i, Y_i) représente la i-ème composante du vecteur V (resp. X, Y).

Soient X et Y deux vecteurs, leur *somme vectorielle normée* V est définie par :

$$\vartheta^2 \rightarrow \vartheta \ : V = X \oplus Y \quad | \quad V_i = \frac{X_i + Y_i}{\|X + Y\|} \tag{14}$$

L'opérateur $\oplus$ 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, on pose,

$$\vec{\mathbf{0}} \oplus \vec{\mathbf{0}} = \vec{\mathbf{0}}. \tag{15}$$

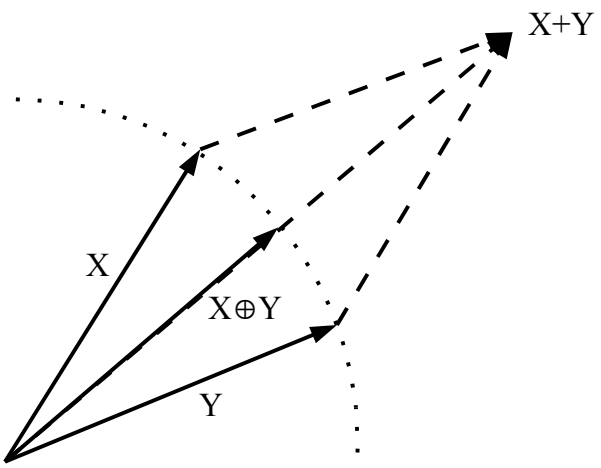

FIGURE 10 – Somme vectorielle normée

De ce qui précède, on peut facilement déduire les propriétés de rapprochement (local et généralisé) :

$$D_A(X \oplus X, Y \oplus X) = D_A(X, Y \oplus X) \leq D_A(X, Y) \tag{16}$$

$$D_A(X \oplus Z, Y \oplus Z) \leq D_A(X, Y) \tag{17}$$

La somme vectorielle est généralisée à n'importe quel nombre de vecteurs par :

$$\vartheta^n \rightarrow \vartheta \ : V = \sum_{i=1}^{n} V(x_i) \quad | \quad V_j = \sum_{j=1}^{n} V(x_i)_j \tag{18}$$

où $V(x_i)$ représente le vecteur d'idée de l'objet x_i, V_j et $V(x_i)_j$ la j-ème composante des vecteurs V et $V(x_i)$. La somme vectorielle normée est généralisée à n'importe quel nombre de vecteurs par :

$$\vartheta^n \rightarrow \vartheta \ : V = \bigoplus_{i=1}^{n} V(x_i) \quad | \quad V_j = \frac{\sum_{j=1}^{n} V(x_i)_j}{\| \sum_{i=1}^{n} V(x_i) \|} \tag{19}$$

Précisons que la somme vectorielle normée binaire n'est pas associative. Toutefois, pour pouvoir simplifier, nous écrirons $\bigoplus_{i=1}^{n} V(x_i)$ au lieu de $V(x_1) \oplus V(x_2) \oplus \ldots \oplus V(x_n)$.

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

Produit terme à terme

Définition Soient X et Y deux vecteurs, leur *produit terme à terme* V est défini par :

$$\vartheta^2 \to \vartheta \ : V = X \odot Y \quad | \quad v_i = x_i y_i \tag{20}$$

Soient X et Y deux vecteurs, leur *produit terme à terme normalisé* V est défini par :

$$\vartheta^2 \to \vartheta \ : V = X \otimes Y \quad | \quad v_i = \sqrt{x_i y_i} \tag{21}$$

Cet opérateur est idempotent ($X \otimes X = X$) et $\vec{\mathbf{0}}$ est absorbant ($X \otimes \vec{\mathbf{0}} = \vec{\mathbf{0}}$). Il peut être généralisé à n'importe quel nombre de vecteurs par :

$$\vartheta^n \to \vartheta \ : V = \bigotimes_{i=1}^{n} V(x_i) \quad | \quad V_j = \sqrt[n]{\prod_{j=1}^{n} V(x_i)_j} \tag{22}$$

L'opérateur $\otimes$ peut être interprété comme un opérateur d'intersection entre vecteurs. Si l'intersection entre deux vecteurs est le vecteur nul, alors ils n'ont rien en commun. On a, de plus, la propriété suivante :

$$\forall X \neq \vec{\mathbf{0}} \quad \forall Y \neq \vec{\mathbf{0}} \quad X \otimes Y = \vec{\mathbf{0}} \Leftrightarrow D_A(X, Y) = \frac{\pi}{2} \tag{23}$$

Interprétation Comme nous venons de le dire, l'opérateur $\otimes$ peut être vu comme une intersection des vecteurs. Du point de vue des vecteurs d'idées, cette opération permet donc de sélectionner les idées communes à un ensemble de termes. Il est utilisé en particulier dans l'opération de contextualisation faible.

Contextualisation faible Lorsque que deux termes sont en présence, pour chacun d'eux, certaines idées se trouvent sélectionnées par le contexte que constitue l'autre terme. Ce phénomène de *contextualisation* consiste à augmenter chaque vecteur de ce qu'il a de commun avec l'autre. Comme nous venons de le voir, les idées communes à deux termes sont données par le produit terme à terme. Ainsi, nous pouvons définir la contextualisation faible $\gamma(X, Y)$ des vecteurs X par Y par :

$$\vartheta^2 \to \vartheta \ : \gamma(X, Y) = X \oplus (X \odot Y) \tag{24}$$

Cette fonction n'est pas symétrique. L'opérateur γ est idempotent ($\gamma(X, X) = X$) et le vecteur nul est un élément neutre ($\gamma(X, \vec{\mathbf{0}}) = X \oplus \vec{\mathbf{0}} = X$).

La propriété de *rapprochement* suivante peut être tirée :

$$D_A(\gamma(X, Y), \gamma(Y, X)) \leq D_A(\gamma(X, Y), Y) \leq D_A(X, Y) \tag{25}$$

$$D_A(\gamma(X, Y), \gamma(Y, X)) \leq D_A(X, \gamma(Y, X)) \leq D_A(X, Y) \tag{26}$$

La contextualisation $\gamma(X, Y)$ rapproche les vecteurs X de Y proportionnellement à leur intersection.

2.6.4 Analyse sémantique de textes : l'algorithme de remontée-redescente

L'objectif des vecteurs d'idées est d'améliorer la plupart des applications du TALN où la sémantique peut jouer un rôle. En *recherche documentaire*, on peut dans une phase de préparation des données affecter un vecteur à chaque texte et dans une phase d'exploitation renvoyer les plus proches du vecteur d'une requête ; en *traduction automatique* il peut s'agir de trouver le vecteur correspondant à l'équivalent le plus proche dans une langue cible ; en *résumé automatique de textes* on peut choisir de privilégier une partie du texte qui représente mieux les idées principales du discours général plutôt qu'une autre ; en *catégorisation* on peut regrouper les textes les plus proches suivant une méthode basée sur la distance angulaire, ... L'idée sous-jacente est donc de pouvoir affecter un vecteur d'idées à tout segment textuel et c'est dans cette perspective qu'a été définie l'analyse sémantique de textes. Cette analyse est différente suivant le type de vecteurs utilisés (sémantique [11] ou conceptuel) cependant le principe général est le même.

Principe L'analyse sémantique de textes permet de calculer le vecteur d'idées d'un texte quelconque. Son principe général est de se baser sur l'hypothèse de compositionnalité de la sémantique linguistique c'est-à-dire que « *le tout est calculable à partir du sens de ses parties* » (cf. 2.2). Le vecteur d'un texte est donc calculé de façon générale par une fonction ayant pour paramètres l'ensemble des vecteurs d'idées des items du texte. En pratique, il s'agit d'une somme pondérée des vecteurs contextualisés des termes de ce texte.

L'idée originale de cette opération est de se baser sur une analyse morpho-syntaxique préliminaire telle que celle présentée en 2.1.4. En effet, il a été montré que l'apport d'informations syntaxiques pour la construction de vecteurs de type saltonien donne de meilleures performances, entre autres, dans le domaine de la recherche d'informations [Besancon, 2001]. Ce constat peut être renouvelé avec les vecteurs d'idées. L'analyse morpho-syntaxique réalisée permet de pondérer par un scalaire le vecteur de chaque mot ou groupe de mots en fonction de son rôle syntaxique [Chauché *et al.*, 2003]. Ainsi, dans le segment « *voile de bateau* », ‹*voile*› est gouverneur syntaxique, son vecteur aura donc un poids plus important que ‹*bateau*›. En revanche, l'inverse sera appliqué pour ‹*bateau à voile*›.

La méthode utilisée se différencie sur la manière d'utiliser le contexte ainsi que sur l'affectation des vecteurs aux feuilles.

Préformatage de textes L'analyse sémantique des textes s'effectue donc en deux parties. Dans une première, on extrait la structure morpho-syntaxique que l'on utilise dans une deuxième partie pour calculer un vecteur d'idées. Il est clair que la première partie influence grandement la seconde. En effet, une mauvaise analyse morpho-syntaxique peut faire insister l'analyse sémantique sur des aspects du texte moins pertinents, par exemple si un gouverneur est mal identifié, voire erroné si l'arbre indique des morphologies ou des fonctions syntaxiques incorrectes. Ce type d'erreur est généré lorsque l'analyseur syntaxique est confronté à des termes qui lui sont inconnus. Un préformatage à partir des informations contenues dans la base de données vectorielle peut influencer bénéfiquement l'analyseur morpho-syntaxique en lui indiquant quelles morphologies sont possibles pour tel ou tel terme. Cette étape préliminaire peut aussi permettre de faciliter l'analyse sémantique si les textes contiennent des constructions de phrases particulières ou de préparer les textes en fonction d'une tâche spécifique. Le préformatage des textes vise donc à améliorer soit l'analyse morpho-syntaxique soit la partie analyse sémantique proprement dite.

Préformatage pour améliorer l'analyse morpho-syntaxique Deux principaux préformatages peuvent être effectués pour améliorer l'analyse morpho-syntaxique des textes : la *gestion des mots inconnus* et la *reconnaissance des locutions*.

- *Gestion des mots inconnus* Les mots inconnus pour l'analyseur morpho-syntaxique entraînent souvent des erreurs dans l'arbre renvoyé : les constituants ne sont pas ou sont mal identifiés. Ce genre de problème est extrêmement fréquent. Bien qu'il ne soit pas rare de trouver des fautes d'orthographe quelle que soit la provenance des textes (extraits de journaux, dictionnaires, dépêches d'agences, etc.), la plupart des mots inconnus sont plutôt le fait de néologismes ou de noms propres que l'analyseur ne peut pas forcément posséder. Des méthodes automatiques de correction existent.
- *Reconnaissance des locutions* Un autre moyen d'aider l'analyseur morpho-syntaxique est de lui indiquer les *locutions semi-figées connexes*, c'est-à-dire les items dont la sémantique ne peut pas être calculée uniquement à

11. non développée dans ce document.

partir de la somme de leurs parties. C'est le cas par exemple de termes comme ‹*moulin à vent*›, ‹*avion de chasse*› ou ‹*lampe de chevet*›. Il est nécessaire de les indiquer à l'analyseur morpho-syntaxique afin que celui-ci ne renvoie pas un sous-arbre pour ces mots.

Préformatage pour améliorer l'analyse sémantique Le préformatage permet aussi de préparer le texte en vue de l'analyse sémantique proprement dite. Il s'agit alors de préparer les textes en vue d'une analyse spécifique comme c'est le cas avec le métalangage des définitions pour les vecteurs conceptuels (cf. 3.5.1).

3 Descripteurs textuels dans VideoSense : les vecteurs conceptuels

Dans le cadre du projet VidéoSense, nous utilisons ce modèle, proche dans sa conception de la théorie componentielle, qui est l'héritier direct de celui des proto-vecteurs d'idées présenté en 2.5.2. Ainsi, il postule que le sens des termes peut être calculé à partir d'un ensemble de concepts. Dans le cadre de videosense, nous étudions 2 modes de construction : un premier basé sur un ensemble de concepts connus *a priori*, un second basé sur un ensemble de concepts émergents.

Pour notre première expérience, nous nous reposons sur des thésaurus généraux. Ces thésaurus ont pour but d'organiser les termes du lexique en fonction des idées qu'ils véhiculent. Ainsi, pour un certain nombre de langues, il existe des thésaurus qui décrivent le lexique en fonction d'une classification mise au point pour cet exercice.

3.1 Les thésaurus

3.1.1 Généralités

Un thésaurus comporte un ensemble de concepts censés permettre de décrire l'ensemble des idées exprimables en langue (*hypothèse du thésaurus*). Un thésaurus n'est pas à proprement parlé d'inspiration componentielle, puisque les premiers sont antérieurs de près d'un siècle à cette théorie, mais s'en rapprochent fortement. Le but d'un thésaurus est, selon les auteurs de [Larousse, 1992], « *d'explorer à partir d'une idée l'univers des mots qui s'y rattachent et de trouver des idées à partir des mots liées à une notion* ».

Un thésaurus est le résultat d'un long processus de tri des items lexicaux d'une langue donnée. Ce tri conduit à la constitution d'une hiérarchie qui diffère donc suivant les idées importantes dans le vocabulaire de telle ou telle société (donc les idées importantes dans telle ou telle société). Ainsi, le thésaurus du français se différencie de celui de l'anglais beaucoup plus raffiné, par exemple, sur des notions comme celle qui touchent au fait religieux.

De même, les thésaurus s'adaptent aux évolutions de la société. Ainsi, comme les rédacteurs du thésaurus Roget le notent dans la préface de la version datant de 1987 [Kirkpatrick, 1987] « *Cette version a été rendue nécessaire par l'extension sans précédent du vocabulaire de l'anglais durant les années 1980, qui reflète les principaux changements d'ordre scientifique, culturel ou social. Les découvertes et les inventions dans le monde des sciences, de la médecine et des technologies ont fait apparaître des termes comme acid rain, AIDS, genetic fingerprinting, nuclear winter, . . .* [12] ».

Le thésaurus Larousse, que nous allons présenter plus particulièrement ici, est inspiré du thésaurus de Roget paru en Grande-Bretagne au milieu du XIXe siècle [Roget, 1852]. Il est constitué de trois parties : (1) *organisation des idées* qui constitue la hiérarchie du thésaurus ; (2) la partie *thésaurus* proprement dite qui permet à partir d'idées de trouver des mots dans le même thème et (3) la partie *index* qui permet de trouver les idées associées aux mots.

3.1.2 Le thésaurus Larousse

La partie *Organisation des idées* : la hiérarchie Larousse Ce thésaurus est basé sur une classification organisée selon une structure hiérarchique d'arbre qui comporte 5 niveaux :

- niveau 0 : 1 concept (*C0 :OMEGA*), la racine de l'arbre. Il faut noter que ce concept n'existe pas explicitement dans la hiérarchie, nous l'avons rajouté pour disposer d'un véritable arbre hiérarchique.
- niveau 1 : 3 concepts (*C1 :LE MONDE*, *C1 :L'HOMME*, *C1 :LA SOCIÉTÉ*)
- niveau 2 : 26 concepts

12. pluies acides, SIDA, empreintes génétiques, hiver nucléaire, . . .

– niveau 3 : 95 concepts
– niveau 4 : 873 concepts, les feuilles de l'arbre.

Afin de les distinguer suivant leur niveau de hiérarchie, nous notons ici les concepts par un *c* concaténé au numéro de niveau du concept, à deux points puis enfin au nom du concept. Par exemple, le concept de niveau 0 oméga est noté *C0 :OMEGA*, le concept de niveau 4 existence est noté *C4 :EXISTENCE*. Pour des raisons de clarté, nous omettrons souvent cette convention d'écriture en ce qui concerne les niveau 4 de la hiérarchie (*C4 :EXISTENCE* sera alors noté *EXISTENCE*).

Les concepts de niveau 4 se succèdent, quand cela s'y prête, en fonction des domaines auxquels ils appartiennent par paires de notions proches, corrélatives ou opposées. Nous avons donc des contraires comme *EXISTENCE* (1) et *INEXISTENCE* (2), *HONNEUR* (641) et *DISCRÉDIT* (642), qui se suivent ainsi que des termes proches thématiquement comme *AMOUR* (600) ou *CARESSE* (601). La figure 11 présente un extrait de cette hiérarchie.

Selon les auteurs, la hiérarchie du thésaurus « *(couvre) méthodiquement l'ensemble des champs notionnels possibles (de la langue)* ». Ainsi, l'ensemble des concepts de niveau 4 permettrait de définir la globalité des termes du lexique. C'est sur cette hypothèse, que nous appelons *hypothèse du thésaurus*, que reposent nos expérimentations.

0 OMEGA
- *1 MONDE*
  - ...
  - *2 ESPACE*
  - *2 TEMPS*
    - *2 TEMPS ET DURÉE*
    - *3 DATE ET CHRONOLOGIE*
      - *4 PASSÉ*
      - *4 PRÉSENT*
      - *4 FUTUR*
      - ...
    - *3 ÉVOLUTION ET HISTOIRE*
    - ...
  - *2 MATIÈRE*
  - *2 VIE*
  - ...
- *1 HOMME*
  - *2 ÊTRE HUMAIN*
  - *2 CORPS ET VIE*
    - *3 CORPS*
      - *4 TÊTE*
      - *4 MEMBRES*
      - *4 MAIN*
      - *4 PIED*
    - *3 FONCTIONS VITALES*
    - ...
  - *2 CORPS ET PERCEPTIONS*
  - *2 ESPRIT*
  - ...
- *1 SOCIÉTÉ*
  - ...

FIGURE 11 – Extrait de la hiérarchie du thésaurus Larousse [Larousse, 1992]

La partie *Thésaurus* : des idées aux mots Cette section est constituée pour permettre au lecteur de trouver des mots en fonctions d'idées. Cette partie du thésaurus comporte 873 articles qui correspondent à chacun des concepts de niveau 4. Les notions traitées sont elles-mêmes divisées en paragraphes ordonnés selon les catégories grammaticales. Chacun de ces paragraphes regroupe des mots proches sémantiquement qu'il est possible de parcourir grâce à des renvois vers des notions communes permettant ainsi de faciliter les associations d'idées ou la recherche d'expressions les plus pertinentes possible.

fable (*nom fem*) : MORALE, MENSONGE, REPRÉSENTATION, RÉCIT
million (*nom masc*) : MULTITUDE, MILLE
échelle (*nom fem*) : MUSIQUE, MONTÉE et MESURE
pêcher (*verbe*) : PÊCHE

FIGURE 12 – Exemple de quelques termes extraits du thésaurus Larousse [Larousse, 1992]

La partie *Index* : des mots aux idées Cette dernière partie est sans nul doute la plus utilisée dans le cadre des vecteurs d'idées puisqu'elle permet de retrouver à partir d'un mot les idées, les thèmes qui lui sont associés. On y trouve, par exemple, que ‹*échelle*› est un *nom commun* et que ses concepts associés sont MUSIQUE, MONTÉE et MESURE. On voit donc que cette partie du thésaurus permet facilement de construire une base de données de vecteurs, une fois les concepts eux-mêmes munis d'un vecteur (vecteurs génératifs cf. 2.6.1). On peut toutefois regretter que les distinctions entre les sens ne soit pas bien marquées. Si on reprend l'exemple d'‹*échelle*› les sens *échelle*/**escalier** et *échelle*/**musique** sont, par exemple, clairement fusionnés.

Les vecteurs conceptuels visent à une représentation ainsi qu'à des traitements plus fins du sens des segments textuels. Ainsi, nous ne considérons pas, comme dans les vecteurs sémantiques, un seul objet lexical pour représenter le sens d'un terme mais plusieurs : les LEXIES. L'objet lexical ITEM LEXICAL correspond au terme et fusionne les informations des différentes lexies. Ces lexies sont constituées à partir d'un apprentissage permanent utilisant des dictionnaires à usage humain. Ce mode de construction de la base de données vectorielles permet d'assurer une certaine couverture lexicale.

3.2 Vecteurs génératifs : origine et interdépendance des concepts

Les vecteurs génératifs sont les seuls vecteurs conceptuels à être construits manuellement. Leur construction est basée sur l'hypothèse forte que les concepts ne sont pas indépendants les uns des autres.

Pour la construction de nos vecteurs génératifs, nous avons choisi un ensemble de concepts issus du thésaurus Larousse [Larousse, 1992] que nous avons présenté en 3.1. L'ensemble de concepts considéré correspond à celui des 873 concepts de niveau 4 de cette hiérarchie. L'interdépendance des concepts se situe à deux niveaux dans la construction des vecteurs génératifs : au niveau hiérarchique, c'est-à-dire en tenant compte de la place que le concept a dans la hiérarchie (*vecteurs génératifs hiérarchiquement augmentés*) et à un niveau transversal à cette hiérarchie (*vecteurs génératifs transversalement augmentés*).

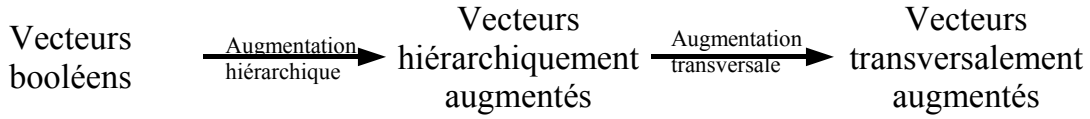

FIGURE 13 – Séquence d'opérations pour la construction des vecteurs génératifs

3.2.1 Interdépendance hiérarchique : vecteurs génératifs hiérarchiquement augmentés

Le point de départ de cette construction est un vecteur booléen. Le vecteur booléen du concept i est le vecteur dont tous les éléments sont à 0 sauf la composante i qui elle est à 1. Cette construction est simple et elle obtient souvent de bons résultats mais semble inadéquate dans plusieurs cas. Il paraît curieux que deux concepts proches comme le sont GUERRE et PAIX partagent quantitativement autant d'idées que PAIX et CHAMPIGNON. Nous l'avons déjà dit, les concepts ne sont clairement pas indépendants et leurs vecteurs respectifs doivent en tenir compte. L'ensemble des concepts défini selon [Larousse, 1992] est hiérarchiquement ordonné selon un arbre (cf. 3.1). La construction des vecteurs génératifs

est basée sur cette structure et plus particulièrement sur la distance ultramétrique entre deux concepts. Il s'agit de la longueur du chemin minimal à parcourir dans l'arbre des concepts pour aller d'un concept à l'autre. Cette distance est définie par :

$$D_u(C, C) = 0 \tag{27}$$

$$D_u(C_1, C_2) = \min \begin{bmatrix} D_u(Sup(C_1), C_2) + 1 \\ D_u(C_1, Sup(C_2)) + 1 \end{bmatrix} \tag{28}$$

où $Sup(X)$ est le père du concept X. Par définition, on a $Sup(racine(arbre)) = racine(arbre)$. Si nous nous référons à la figure 11, nous avons $D_u(tête, membre) = 2$. Tous les concepts frères de tête sont à une distance ultramétrique égale à 2. Nous avons également $D_u(tête, corps) = 1$,
$D_u(tête, fonctionsvitales) = 3$ et $D_u(tête, présent) = 8$. La valeur 8 est d'ailleurs la plus grande possible entre deux concepts de cette hiérarchie puisque leur ancêtre commun le plus éloigné est la racine, située au maximum à quatre niveaux au-dessus.

Grâce à cette distance ultramétrique, nous construisons les *vecteurs génératifs hiérarchiquement augmentés*. Appelons X_i le vecteur booléen correspondant au i-ème concept de la base $\mathcal{C}$. Y_i est le vecteur conceptuel hiérarchiquement augmenté défini par :

$$Y_i = X_i \oplus \bigoplus_{j=0}^{\dim(\mathcal{C})} \frac{1}{2^{D_u(C_i, C_j)}} \times X_j \tag{29}$$

Le vecteur original X_i est ajouté afin que, de tous les vecteurs Y_j, le plus proche de X_i soit toujours Y_i. La figure 14 montre, pour *PAIX*, le vecteur booléen et le vecteur hiérarchiquement augmenté.

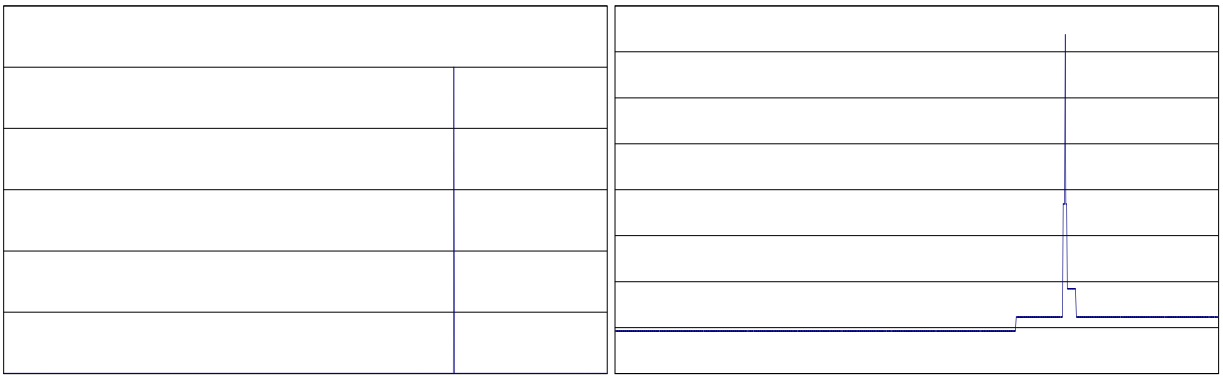

FIGURE 14 – Vecteur du concept *PAIX* et vecteur hiérarchiquement augmenté du concept *PAIX*

3.2.2 Interdépendance transversales : vecteurs génératifs transversalement augmentés

Bien qu'augmenter un vecteur par ses voisins améliore sa qualité, il faut admettre que la hiérarchie des concepts n'est qu'une vue particulière de la façon selon laquelle ils peuvent être organisés. D'autres liens spécifiques peuvent être exhibés. C'est le cas entre *CHAMPIGNON* et *TOXICITÉ* ou *GASTRONOMIE* par exemple. L'augmentation transversale d'un concept C est une opération manuelle réalisée une seule fois, à des ajustements près, qui consiste à énumérer les concepts relatifs à C qui ne sont pas représentés dans la hiérarchie. Le nouveau vecteur est appelé *vecteur transversalement augmenté*.

Par exemple, le concept *PAIX* a comme concepts transversalement associés les concepts *CONCORDE*, *GUERRE*, *CALME*, *SÉCURITÉ*, *REPOS*, *ÉQUILIBRE*. Ces concepts transversaux sont sélectionnés manuellement et peuvent être trouvés dans la partie index de [Larousse, 1992] (cf. 3.1.2).

Si Y_i est le vecteur hiérarchiquement augmenté du concept i défini sur C, nous pouvons calculer le i-ème vecteur augmenté transversalement Z_i en faisant la somme pondérée de tous les vecteurs Y_j avec Y_i. Cette construction assure que le vecteur Z_j le plus proche de Y_i demeure Z_i.

$$Z_i = \bigoplus_{j=0}^{dim(\mathcal{C})} \alpha_{ij}(Y_j \oplus Y_i) \tag{30}$$

où α_{ij} est la pondération du concept transversal j pour le concept i.

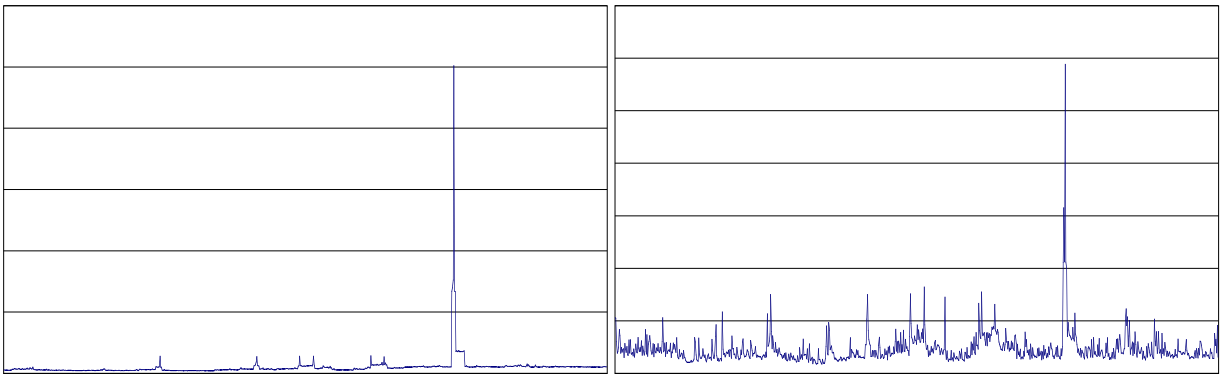

FIGURE 15 – Vecteurs transversalement augmenté du concept *PAIX* et pour l'item ‹*paix*›

3.3 Pourquoi nos vecteurs sont-ils dits "conceptuels" ?

On pourrait penser que le terme ‹*conceptuel*› qui qualifie nos vecteurs provient du fait que leur construction est basée sur ces objets que nous appelons concepts. En pratique, ce nom a été introduit pour marquer le fait que ces vecteurs représentent des idées plus ou moins abstraites, plus ou moins générales.

3.4 Architecture et construction de la base

Nous présentons succinctement ici l'architecture de la base lexicale sémantique. Cette architecture est centrée sur le stockage et l'exploitation de deux objets lexicaux l'objet ITEM LEXICAL et l'objet LEXIE.

3.4.1 Structure des objets lexicaux

Les objets lexicaux sont composés d'un certain nombre d'*informations linguistiques* :
- un **identifiant** ;
- la **morphologie** composé des *catégories grammaticales* (*nom, pronom, adjectif, adverbe, etc.*), du *genre* (*masculin, féminin, neutre*) et du nombre (*singulier, pluriel*) ;
- la **fréquence en usage** c'est-à-dire le nombre de fois (ou au moins une estimation) où l'objet a été rencontré ;
- un **vecteur conceptuel**.

3.4.2 Objets lexicaux

L'architecture est composée de deux sortes d'objets lexicaux, les ITEMS LEXICAUX et les LEXIES. Cette architecture provient essentiellement du mode de fabrication des vecteurs conceptuels c'est-à-dire à partir de dictionnaires à usage humain.

Lexies Les LEXIES constituent le socle de la base lexicale sémantique. À partir de définitions de dictionnaires à usage humain, il est possible d'extraire un certain nombre d'informations linguistiques et de calculer un vecteur conceptuel. Examinons, par exemple, les entrées de [Larousse, 2004] pour l'item ‹*botte*›.

1.botte : #nf# (néerl. bote, touffe de lin) Assemblage de végétaux de même nature liés ensemble : (Botte de paille. Botte de radis.).
2.botte : #nf# (#ethym-it# botta, coup) . Coup de pointe donné avec le fleuret ou l'épée.
3.botte : #nf# (p.-ê. de bot) . Chaussure à tige montante qui enferme le pied et la jambe généralement jusqu'au genou : (Bottes de cuir).

L'idée est pour chacune de ces définitions de fabriquer un objet LEXIE qui comportera :
- un **identifiant** : habituellement pour les lexies, il est constitué du nom du terme et d'un numéro (ex : *botte.1*, *botte.2*, *chat.1*, ...) ;
- la **morphologie** généralement facile à récupérer puisqu'elle est souvent assez bien délimitée ;
- la **fréquence en usage** qui est estimée par des heuristiques, que nous ne détaillerons pas ici, à partir de la fréquence de l'ITEM LEXICAL ;
- un **vecteur conceptuel** calculé à partir du texte de la définition grâce à une analyse sémantique telle que celle présentée en 3.7.

Items Lexicaux Les objets ITEMS LEXICAUX correspondant à un terme sont fabriqués à partir des LEXIES de ce terme. Leur structure est la suivante :
- un **identifiant** qui est généralement le nom du terme (ex : *botte*, *chat*) ;
- la **morphologie** qui rassemble l'ensemble des morphologies des lexies ;
- la **fréquence en usage** qui correspond au nombre de fois où le terme a été repéré dans les textes étudiés par l'analyse sémantique ;
- un **vecteur conceptuel** qui est la somme vectorielle normée des vecteurs conceptuels des lexies.

3.5 Apprentissage des objets lexicaux

La fabrication des objets lexicaux se fait à partir de définitions extraites de dictionnaires à usage humain sous forme électronique. On crée ainsi une LEXIE par définition puis les objets ITEM LEXICAUX à partir de ces LEXIES.

3.5.1 Lexies : apprentissage à partir de définitions issues de dictionnaires classiques

L'apprentissage à partir de définitions n'est pas sans poser un certain nombre de problèmes. Il s'agit d'extraire d'un dictionnaire la ou les entrées correspondant à un terme. À cause de la polysémie ainsi que de l'homonymie, nous pouvons avoir plusieurs définitions pour une même entrée. Nous avons divisé en trois étapes successives le traitement d'une entrée d'un dictionnaire comme nous le présentons sur la figure 16 : (1) un *prétraitement* dont l'objectif est de préparer les données avant les 2 étapes suivantes. Il consiste à séparer les différentes définitions et à les unifier dans un même format prédéfini puis à préparer la définition en vue d'une analyse sémantique. (2) Une *extraction des informations lexicales* et en particulier de la morphologie et enfin (3) le *calcul d'un vecteur conceptuel* à partir de la définition formatée.

Prétraitement des données La première étape consiste à prétraiter les données. En effet, les définitions sont extraites de dictionnaires hétérogènes dont le formatage et les informations disponibles peuvent fortement différer à la fois pour des raisons purement techniques (codage utilisé, format des données : XML, HTML, ...) ainsi que pour des raisons formelles (séparation des sens, des homonymes, ...).

Cette opération d'unification du formatage est fondamentale car elle consiste à convertir le format des données vers un format conçu pour faciliter l'extraction des informations lexicales ainsi que le calcul du vecteur conceptuel des définitions. Cette partie de l'apprentissage est totalement *ad hoc*. Elle doit être conçue pour chaque dictionnaire afin que les deuxième et troisième étapes soient les mêmes, quelle que soit la source. Le prétraitement des données

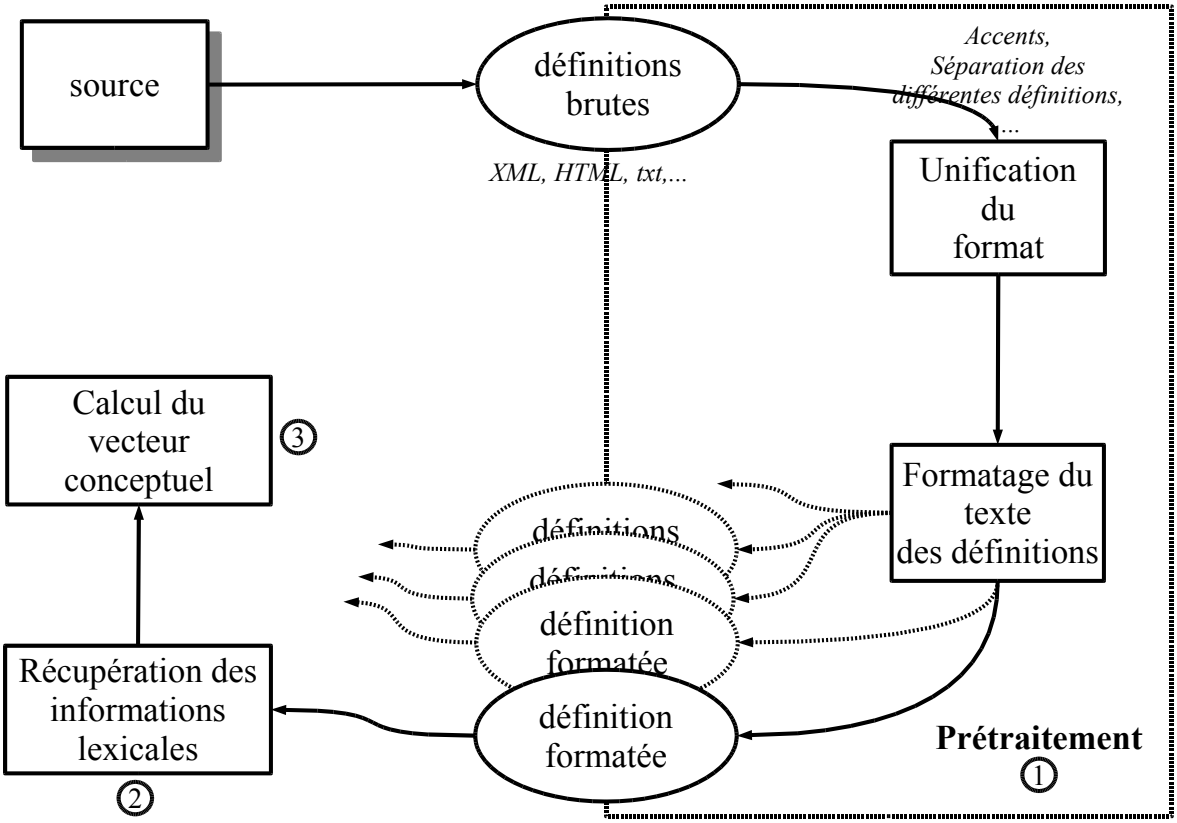

FIGURE 16 – Séquence d'opérations pour l'apprentissage de vecteurs conceptuels à partir d'une source

effectue deux catégories de tâches : une *unification du format* et un *formatage du texte des définitions*. On appelle *définitions non formatées* les textes bruts récupérés depuis les sources et *définitions formatées* les textes obtenus après ce prétraitement.

Unification du format Les dictionnaires utilisés sont sous forme électronique. Les formats des données sont très hétérogènes. Nous pouvons avoir de simples textes comme ceux des dictionnaires papiers, du format HTML pour les dictionnaires disponibles en ligne ou bien du XML. Il est clair que suivant le format, il est plus ou moins simple de repérer les informations pertinentes : pour le XML, langage balisé généralement de façon assez claire, beaucoup plus simplement que pour le simple texte dont il faut chercher à repérer de façon souvent beaucoup plus heuristique les schémas permettant de détecter la morphologie et les séparations entre les différentes définitions. En revanche, un certain nombre de caractères spéciaux (lettres avec diacritiques, symboles, ...) devront être convertis dans le cas du HTML.

Formatage du texte des définitions Cette opération vise à préparer l'apprentissage des vecteurs conceptuels grâce aux textes des définitions. Le formatage est employé à plusieurs niveaux :

– *Gestion du métalangage* L'apprentissage va s'effectuer grâce à une méthode d'analyse sémantique telle que celle présentée en 2.6.4. Comme nous l'avons vu, il est parfois nécessaire de préformater les textes afin de simplifier l'apprentissage. L'étude des définitions pose en particulier des problèmes essentiellement à cause du *métalangage*, c'est-à-dire le langage utilisé pour structurer le discours. Un certain nombre de tournures caractérisent ce métalangage et ne sont pas porteuses de sens : ‹*se dit de*›, ‹*relatif à*›, ‹*action de*›, ‹*nom usuel de*›, ‹*Abr. de*›, ...

Ce prétraitement consiste à rechercher ces tournures dans les définitions et à les remplacer par un symbole. L'analyseur sémantique lorsqu'il le rencontre ne lui attribue aucune sémantique. Ainsi, ces symboles ne seront absolument pas utilisés dans le calcul des vecteurs car considérés comme non porteurs de sens par l'analyseur sémantique. Nous avons répertorié à ce jour [13] un peu moins de 80 tournures. Ce chiffre est en constante augmentation au fur et à mesure de la découverte de nouveaux cas, phénomène de plus en plus rares toutefois.

- *balisage de la morphologie* Cette opération est assez simple, la morphologie étant généralement bien indiquée.
- *formatage des informations thématiques* Ce prétraitement exploite aussi des informations qui peuvent permettre par la suite d'aider l'analyse sémantique. C'est le cas en particulier du domaine qui est un très bon indice du champ sémantique des items constituant une définition. En pratique ceux-ci sont remplacés par des concepts (par exemple, le domaine *COMM.* sera remplacé par COMMERCE et *BOURSE* par BOURSE). De même, certains dictionnaires fournissent des résumés d'où on peut extraire certaines informations simples. Par exemple, un résumé indiquant *poète français* sera annoté par le concept POÉSIE tandis qu'un résumé indiquant *dramaturge* sera annoté par THÉÂTRE.

Extraction des informations lexicales La deuxième étape de l'apprentissage de lexies consiste à extraire les informations lexicales. Pour l'instant, cette partie de l'apprentissage ne consiste qu'en l'extraction des informations concernant la morphologie. Nous verrons dans les chapitres suivants l'ajout d'autres informations lexicales en particulier des informations concernant les relations lexicales.

Calcul des vecteurs conceptuels L'analyse des définitions se fait grâce à la méthode d'analyse sémantique présentée en 3.7. Les définitions formatées au cours du pré-traitement sont analysées et le vecteur résultant de l'analyse est affecté à la lexie.

3.5.2 Noyau

Afin d'amorcer le système d'apprentissage, une partie des termes est indexée de façon manuelle. Ce noyau est constitué d'un ensemble de termes choisis parmi les plus courants et/ou les plus polysémiques. Il s'agit donc de fabriquer manuellement les LEXIES de ces items lexicaux particuliers. L'identifiant et la morphologie sont remplis de manière triviale, la fréquence est, comme pour tous les objets lexicaux, nulle à la création. Le vecteur conceptuel associé à la lexie est fabriqué par la somme pondérée des vecteurs génératifs correspondant aux concepts présents dans cette lexie. Par exemple, nous avons pour l'item lexical ‹*paix*› les lexies manuellement indexées suivantes :

- **identifiant** : paix.1
- **morphologie** : [NOM]
- **fréquence en usage** : 0
- **vecteur conceptuel** : $V(\text{PAIX}) \oplus \frac{1}{2}V(\text{GUERRE}) \oplus V(\text{SÉCURITÉ}) \oplus \frac{1}{2}V(\text{ACCORD})$

- **identifiant** : paix.2
- **morphologie** : [NOM]
- **fréquence en usage** : 0
- **vecteur conceptuel** : $V(\text{REPOS}) \oplus V(\text{CALME}) \oplus V(\text{SILENCE}) \oplus \frac{1}{2}V(\text{ÉQUILIBRE})$

Deux sens ont été indexés pour ‹*paix*›. Le premier se réfère à une *absence de guerre* et le deuxième à une *situation de calme*. L'énumération des concepts pondérés est une tâche difficile car subjective. Dans notre expérimentation, nous laissons cette tâche aux lexicographes qui ont créé le thésaurus Larousse [Larousse, 1992] (cf. 3.1) pour les concepts ainsi que les dictionnaires [Larousse, 2001] et [Robert, 2000] pour les découpages de sens. Seuls les mots parmi les plus importants sont ainsi décrits dans le noyau. Nous nous reposons sur l'apprentissage automatique pour l'indexation en masse. Toutefois, les séparations délicates de sens nécessitent des ajustements manuels.

Les items lexicaux de ce noyau sont considérés comme pertinents. Cet ensemble constitue la base d'items lexicaux à partir de laquelle démarre l'apprentissage (cf. figure 17). Nous cherchons à mettre au point un apprentissage qui

13. Mars 2005

soit le plus cohérent possible afin d'obtenir une base augmentée pertinente. Dans ce chapitre, cet apprentissage se fait uniquement à partir de dictionnaires mais, dans les suivants, il sera basé, en partie, sur les relations sémantiques, en particulier symétriques, comme la synonymie et l'antonymie.

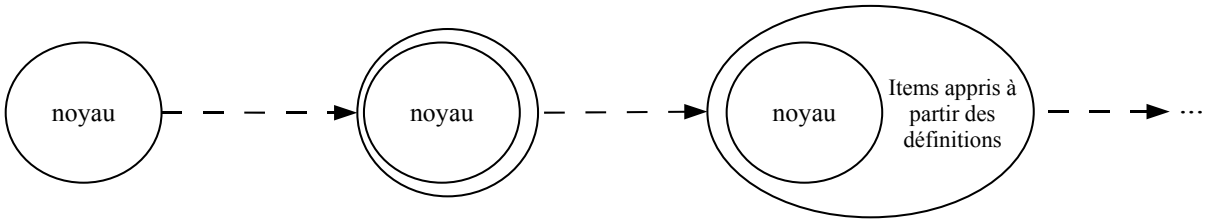

FIGURE 17 – Augmentation de la couverture lexicale grâce à l'apprentissage

3.6 Contextualisation forte

3.6.1 Définition

Tandis que la méthode de contextualisation faible peut être utilisée avec n'importe quel vecteur (cf. 2.6.3), la méthode de contextualisation forte est utilisée, elle, avec les ITEMS LEXICAUX. L'idée est de considérer les vecteurs conceptuels associés à chacune des LEXIES de l'item en fonction des informations contextuelles disponibles. Ces informations peuvent être alors non seulement d'ordre vectoriel mais aussi d'ordre morphologique. Ainsi, lors d'une analyse sémantique, les informations morphologiques des feuilles peuvent être utilisées grâce à cette méthode. La méthode de contextualisation forte ne fait donc pas un choix parmi les divers vecteurs des LEXIES, mais en favorise certains aux détriments d'autres. La fonction de contextualisation forte est ainsi définie par :

$$\omega \times m \times \vartheta \rightarrow \vartheta \quad : \quad V = \Gamma(I, M_c, V_c)$$

où ω est l'ensemble des ITEMS LEXICAUX, m l'ensemble des morphologies et ϑ l'ensemble des vecteurs conceptuels.

Il s'agit ici d'un principe général et diverses méthodes ont été testées. Celle qui suit est actuellement utilisée dans le cadre du projet VidéoSense

$$\Gamma(I, M_c, V_c) = \bigoplus_{L_i}(P_i \odot V(L_i))$$

$$\text{où } P_i = P_{morpho}(L_i, M_c)^l \times P_{ang}(L_i, V_c)^m \times P_{\mathit{freq}}(L_i)^n \tag{31}$$

$$\text{et } l, m, n \in \{0, 1\}$$

où I représente l'item dont on cherche le vecteur contextualisé et qui est composé d'un ensemble de LEXIES $\{L_1, L_2, \ldots, L_i, \ldots, L_n\}$, P_{morpho} le poids morphologique, P_{ang} le poids angulaire et P_{freq} le poids de la fréquence, tous trois définis ci-après.

Le poids P_i est donc la moyenne géométrique pondérée des poids morphologique, angulaire et de fréquence des lexies. Suivant les utilisations de la méthode de contextualisation forte, l'une ou l'autre des informations contextuelles peut ainsi être favorisée ou au contraire être ignorée. Ainsi, dans le cas de l'apprentissage, les exposants l, m, n valent 1, tandis que dans le calcul du vecteur d'un ITEM LEXICAL, aucun des critères n'est considé ($l = m = n = 0$), ce qui correspond à la somme vectorielle normée des vecteurs des LEXIES (cf. 3.4.2).

3.6.2 Poids angulaire

$$\iota \times \vartheta \rightarrow [0, k] \ : P_{ang}(L, X) = Min(k, cotan(D_A(\text{V(L), X}))) \tag{32}$$

où $cotan$ est la fonction cotangente, l'inverse de la fonction tangente ($\frac{1}{tan(x)}$). Expérimentalement, nous avons posé $k = 10$.

3.6.3 Poids de la fréquence

$$\iota \rightarrow [0,1] \ : P_{\textit{freq}} = \frac{\textit{freq}(L)}{\textit{freq}(item(L))} \tag{33}$$

où $item(L)$ renvoi l'ITEM LEXICAL correspondant à la LEXIE L et $\textit{freq}(x)$ renvoie la fréquence de l'objet lexical x.

3.6.4 Poids et distance morphologique

Soit le poids morphologique P_{morpho} défini par :

$$m \times m \rightarrow [0, \frac{\pi}{2}] \ : P_{morpho}(M_1, M_2) = \frac{\pi}{2} - arctan(D_{morpho}(M_1, M_2)) \tag{34}$$

M_1 et M_2 sont des ensembles au sens mathématique du terme. Par exemple, la morphologie de ‹*botte*› peut être vue comme l'ensemble à deux éléments {*nom*, *masculin*} et celle de ‹*orgues*› comme l'ensemble à trois éléments {*nom*, *masculin*, *pluriel*}.

Par définition, on pose :

$$P_{morpho}(M_1, M_2) = \frac{\pi}{2} \text{ si } M_1 = \varnothing \text{ ou } M_2 = \varnothing. \tag{35}$$

La mesure du poids est calculée comme suit :

$$D_{morpho}(M_1, M_2) = \sum_{x \in (M_1 \cup M_2) - (M_1 \cap M_2)} p(X) \text{ où } \begin{cases} p(X) = 1 \text{ si } X \text{ est une catégorie grammaticale} \\ p(X) = 0,5 \text{ sinon} \end{cases} \tag{36}$$

Cette distance est donc uniquement fonction de ce qui sépare les deux morphologies. Par exemple, la distance morphologique entre {*nom*, *masculin*} et {*nom*, *masculin*, *plur*} ne sera que fonction de *plur* et vaudra ainsi $0,5$. De fait, nous avons $D_{morpho}(X, X) = 0$. Voici quelques exemples de poids et de distances morphologiques :

$morpho_1$	$morpho_2$	D_{morpho}	P_{morpho}
{nom}	{nom}	0	$\frac{\pi}{2}$
{nom, masc}	{nom, masc}	0	$\frac{\pi}{2}$
{nom, masc}	{nom}	0,5	1,11
{nom, masc}	{fem}	2	0,47
{nom, masc}	{nom, fem}	1	0,79
{verbe}	{nom}	1,25	0,68
{verbe, intransitif}	{nom masc plur}	2,75	0,35

3.7 Analyse sémantique des textes en remontée-redescente grâce aux vecteurs conceptuels

3.7.1 Algorithmes

L'analyse sémantique des textes en remontée-redescente à l'aide des vecteurs conceptuels est permise par les algorithmes 1 et 2.

Algorithme 1: analyse : algorithme d'analyse sémantique avec les vecteurs conceptuels

Entrée : vecteur conceptuel $V_{contexte}$, A arbre morpho-syntaxique du texte, seuil s
Sortie : vecteur conceptuel du texte
Vecteur V = analyse ($V_{contexte}$, A.racine)
répéter
  Vecteur V_2 = V
  V = analyse(V, A.racine)
jusqu'à $(D_A(V, V_2) < s)$;
retourner V

Algorithme 2: algorithme d'analyse sémantique avec les vecteurs conceptuels : analyse

Entrée : vecteur conceptuel $V_{contexte}$, nœud N
Sortie : vecteur conceptuel du sous-arbre de N
si *N est une feuille* **alors**
  si *N.item.estUnMotVide()* **alors**
    N.vecteur=$V_{\vec{0}}$
  sinon
    N.vecteur=Γ(N.item, N.morpho, $V_{contexte}$)
  si *N.estGouverneur* **alors**
    N.vecteur = $2 \odot$N.vecteur
  retourner *N.vecteur*
sinon
  $V = \vec{0}$
  pour *chacun des fils* f_i *de N* **faire**
    $V = V + analyse(\gamma(N.vecteur, V_{contexte}), f_i)$
    N.vecteur=norm(V)

3.7.2 Principe

Le principe est de faire descendre les informations du vecteur contexte du texte jusqu'aux feuilles de l'arbre en les enrichissant par les informations contenues dans les nœuds de l'arbre. Dans le cas général où nous n'avons aucune information thématique au début de l'analyse, le vecteur contexte utilisé lors de la première descente est le vecteur nul. Dans d'autres cas, l'analyse de définitions par exemple, si des informations du domaine sont spécifiées, le vecteur contexte utilisé sera celui de ce domaine.

Dans un arbre morpho-syntaxique, les feuilles contiennent les informations sur les items ainsi que leur morphologie dans le texte. Ces deux informations sont utilisées pour calculer les vecteurs correspondant aux contextualisations fortes de ces items (cf. 3.6) et les affecter à chacune des feuilles de l'arbre. Les feuilles qui correspondent à des mots vides de sens (déterminants, conjonction de coordination, préposition, ...) se voient affecter un vecteur vide.

La remontée se fait alors de la même manière que celle de l'analyse avec les vecteurs sémantiques. Les vecteurs de chaque nœud sont calculés à partir des vecteurs de leurs fils et de pondérations calculées en fonction de leur rôle syntaxique. Le vecteur de chaque nœud est ainsi calculé récursivement jusqu'au sommet de l'arbre. Ce vecteur possède les idées contenues dans tout mot du texte. À ce moment du calcul, il n'y a eu, dans le cas général, aucune contextualisation. Le vecteur du sommet de l'arbre contient donc les idées pertinentes du texte mais aussi beaucoup de bruit. On effectue à nouveau une descente. On calcule la contextualisation faible du vecteur de chacun des nœuds en fonction de celui de son père. Ainsi, le vecteur contexte n'est pas le même pour tous les nœuds de l'arbre mais est plus directement fonction du sous-arbre dont il est ancêtre. Cette solution améliore largement l'analyse dans le cas d'une phrase comme « *La souris d'ordinateur est posée sur la table du vétérinaire.* ». En effet, si on n'utilise pas un tel mécanisme, le sens de souris serait autant influencé par l'idée d'*INFORMATIQUE* contenue dans ‹*ordinateur*› que par celle

d'*ANIMAL* contenue dans ‹‹vétérinaire››, ce qui empêcherait la désambiguïsation du texte.

Au niveau des feuilles, on effectue une contextualisation forte puis une remontée. Ces opérations sont renouvelées un certain nombre de fois jusqu'à une relative stabilisation du vecteur général c'est-à-dire tant que la distance angulaire entre deux vecteurs du sommet calculés successivement n'est pas inférieure à un certain seuil s.

3.7.3 Exemple

La figure 18 présente l'analyse sémantique de « *La souris d'ordinateur.* » grâce aux vecteurs conceptuels. Considérons le cas général où le vecteur contexte est nul. Nous le faisons redescendre jusqu'aux feuilles de l'arbre en faisant une contextualisation faible aux vecteurs de chaque nœud. Bien entendu, lors de cette première descente, tous ces vecteurs seront nuls. Les feuilles 2 et 5 qui correspondent respectivement à un déterminant et à une préposition et qui sont donc des mots vides de sens se voient affectées d'un vecteur nul. En revanche, le nœud 3 se voit affecté du vecteur conceptuel correspondant à la contextualisation forte de ‹*souris*› avec la morphologie *nom fem* et le vecteur contexte nul. La même opération est réalisée sur le noeud 5 avec ‹*ordinateur*› et *nom masc*.

Lors de la remontée, le nœud 4 est affecté par le vecteur correspondant à la somme vectorielle pondérée entre ses deux nœuds fils 5 et 6. Le vecteur de la feuille est considéré avec un poids de 2 pour cette opération puisqu'il est gouverneur syntaxique du sous-arbre correspondant à « *d'ordinateur* ». Il en est de même pour le vecteur du nœud 3 (‹*souris*›) dans le calcul du vecteur global du texte.

L'opération de redescente-remontée se renouvelle ainsi de suite jusqu'à une stabilisation du vecteur global du texte.

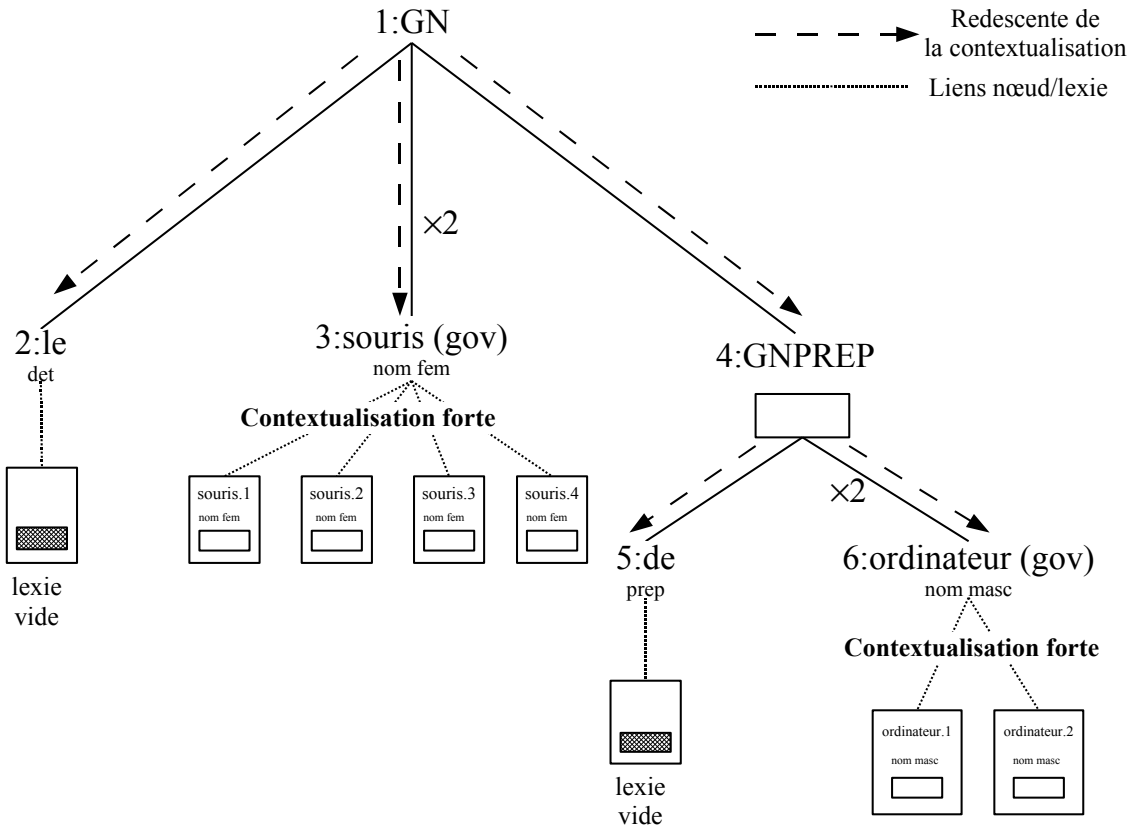

FIGURE 18 – Exemple d'analyse sémantique grâce aux vecteurs conceptuels

3.8 Construction des vecteurs par émergence

Nous utilisons dans le cadre de VideoSense, une deuxième manière de construire les vecteurs conceptuels : par émergence. Cette approche s'affranchit de tout thésaurus et vecteurs de concept comme base de départ. Seule d la taille du vecteur est fixée *a priori*. Le mode de construction des vecteurs est identique au modèle classique à la différence que si un des vecteurs entrant dans la somme est inexistant, car non encore calculé, alors ce vecteur est tiré au hasard. Le processus de calcul est itéré jusqu'à convergence de chaque vecteur.

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

4 Expériences mises en œuvre dans VideoSense

Reprenant les travaux menés depuis quelques années à Montpellier et à Penang en Malaisie par Didier Schwab, nous avons réimplanté l'ensemble des outils permettants la fabrication et l'exploitation de vecteurs conceptuels (Blexisma 3 : Base LEXicale Sémantique Multi-Agent) en Java. Cela se traduit par la création de plus de 300 classes et interfaces disponible sous licence LGPL (Licence publique générale limitée GNU) [14].

Plusieurs expériences sont en cours, toutes utilisent comme ressource un grand graphe lexical construit à partir de Wiktionary [15].

- une première, de type émergence, calcule des vecteurs du français. Un service Web a été mis en place et est exploité par notre partenaire *Ghanni* pour calculer des vecteurs conceptuels décrivant des textes compagnons de leur vidéos. L'anglais et l'allemand sont en phase d'intégration (voir figure 19)
- une seconde, qui aura plusieurs variantes, utilise elle une hiérarchie pour chaque langue. Les noyaux sont en phase de constitution, celui du français, le plus avancé en est environ à 40% (voir figure 20).

Références

[Bangalore, 1997] Srinivas BANGALORE. « *Complexity of lexical descriptions and its relevance to partial parsing.* ». PhD thesis, University of Pennsylvannia., 1997. 2.1.4

[Besancon, 2001] Romaric BESANCON. « *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles de texte* ». Thèse de doctorat, École Polytechnique Fédérale de Lausanne, Laboratoire d'Intelligence Artificielle, 2001. 2.6.2, 2.4.1

[Bestgen, 2004] Yves BESTGEN. « Analyse sémantique latente et segmentation automatique des textes ». Dans les actes de *7èmes Journées internationales d'Analyse statistique des Données Textuelles*, pp 171–181, Louvain-la-Neuve, Mars 2004. 2.3.3

[Boitet, 2000] Christian BOITET. « Handling Texts and Corpuses in Ariane-G5, a complete environment for multilingual MT ». Dans les actes de *ACIDCA'2000, Corpora and Natural Language Processing*, pp 7–11, 2000. 2.1.4

[Chauché, 1984] Jacques CHAUCHÉ. « Un outil multidimensionnel de l'analyse du discours. ». Dans les actes de *COLING'1984 : 10th International Conference on Computational Linguistics*, pp 11–15, Stanford University, California, 1984. 2.1.4

14. http://ligforge.imag.fr/projects/blexisma/
15. http://www.wiktionary.org/

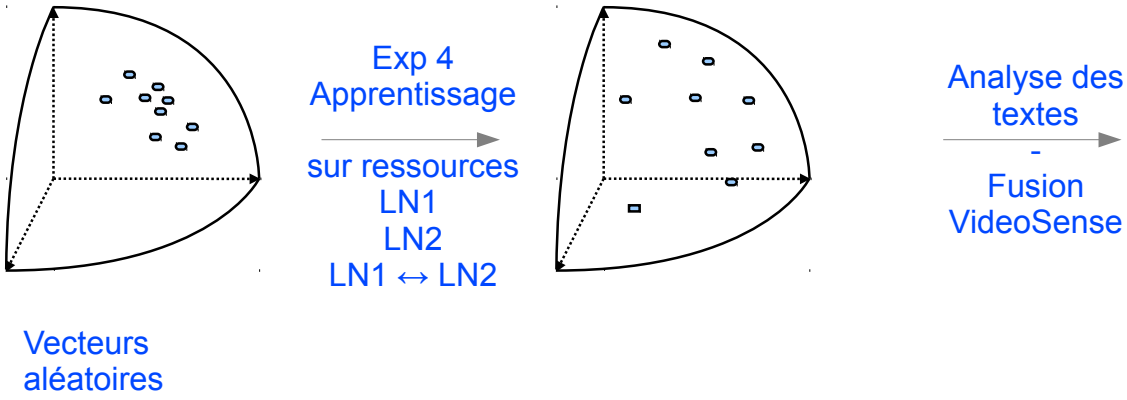

FIGURE 19 – Émergence des vecteurs dans VideoSense. Dans cette expérience, le français, l'anglais et l'allemand sont plongés dans un seul espace vectoriel. Les vecteurs conceptuels sont appris grâce aux informations issus de wiktionnary : définitions et liens de traduction entre les langues

[Chauché, 1990] Jacques CHAUCHÉ. « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance ». *TAL Information*, pp 17–24, 1990. 10, 2.5.2

[Chauché *et al.*, 2003] Jacques CHAUCHÉ, Violaine PRINCE, Simon JAILLET, et Maguelonne TEISSEIRE. « Classification automatique de textes À partir de leur analyse syntaxico-sémantique ». Dans les actes de *TALN 2003*, volume 1, pp 55–64, Batz-Sur-Mer, France, Juin 2003. 2.4.1

[Church, 1988] Kenneth Ward CHURCH. « A stochastic parts program and noun phrase parser for unrestricted text ». Dans les actes de *2nd Conference on Applied Natural Language Processing*, pp 136–143, 1988. 2.1.4

[Collins, 1997] Michael COLLINS. « Three generative, lexicalised models for statistical parsing ». Dans les actes de *the Annual Meeting of the Association of Computational Linguistics*, Madrid, 1997. 2.1.4

[Deerwester *et al.*, 1990] Scott C. DEERWESTER, Susan T. DUMAIS, Thomas K. LANDAUER, George W. FURNAS, et Richard A. HARSHMAN. « Indexing by Latent Semantic Analysis ». *Journal of the American Society of Information Science*, pp 391–407, 1990. 2.3.3

[Genest, 2000] David GENEST. « *Extension du modèle des graphes conceptuels pour la recherche d'informations* ». Thèse de doctorat, Université Montpellier II, 2000. 2.4.1

[Grefenstette, 1994] Gregory GREFENSTETTE. « Corpus-derived first, second and third-order word affinities ». Dans les actes de *6th EURALEX*, Amsterdam, 1994. 2.3.3

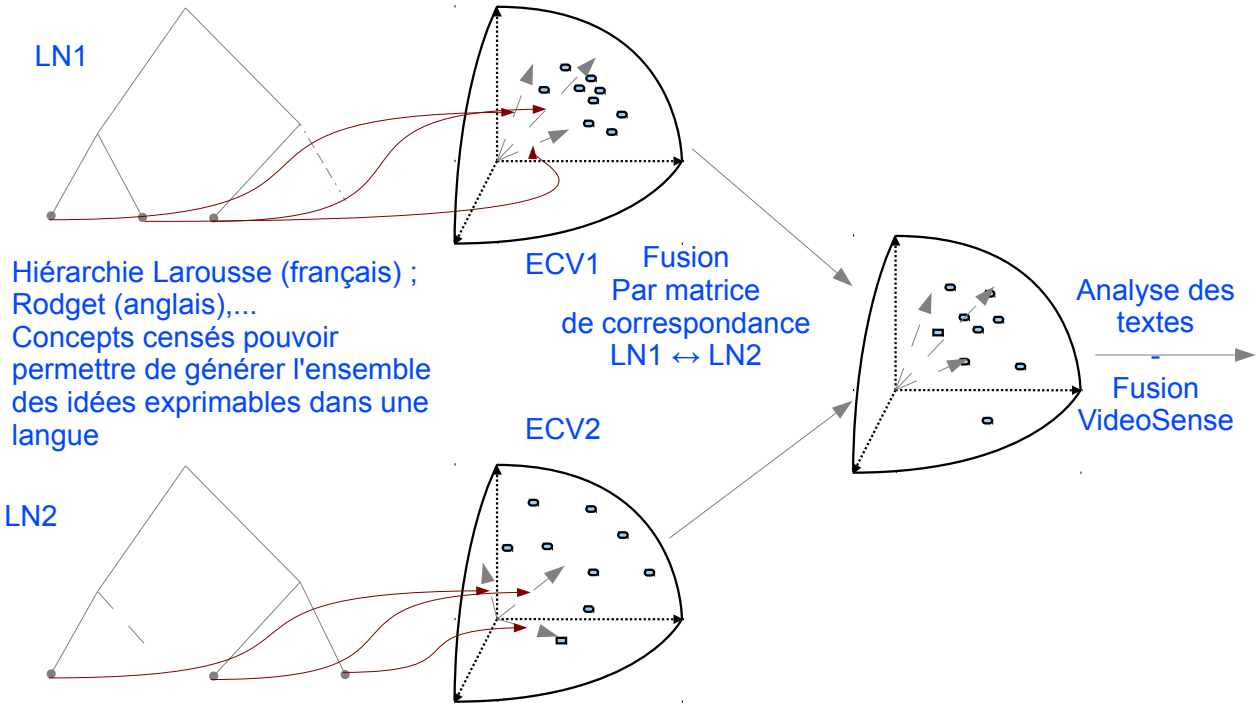

FIGURE 20 – Vecteurs construits à partir de hiérarchies dans VideoSense. Dans cette expérience, une hiérarchie est associée à chaque langue. La projection des langues dans un même espace vectoriel se fait grâce à une matrice de correspondance calculée semi-automatiquement à partir des hiérarchies.

[Harabagiu *et al.*, 1999] Sanda M. HARABAGIU, George Armitage MILLER, et Dan I. MOLDOVAN. « WordNet 2 - A Morphologically and Semantically Enhanced Resource ». Dans les actes de *Workshop SIGLEX'99 : Standardizing Lexical Resources*, pp 1–8, 1999. 2.4.2

[Harris *et al.*, 1989] Zellig S. HARRIS, Michael GOTTFRIED, Thomas RYCKMAN, Paul MATTICK JR., Anne DALADIER, T.N. HARRIS, et S. HARRIS. *The form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 de *Boston Studies in the Philosophy of Science*. Kluwer Academic Publisher, Dordrecht, 1989. 2.3.1

[Hjelmlev, 1968] Louis HJELMLEV. *Prolégolème à une théorie du langage*. Éditions de minuit, 1968. 2.5.1

[Kintsch, 2000] Walter KINTSCH. « Metaphor comprehension : A computational theory ». *Psychonomic Bulletin and Review*, 2000. 2.3.3

[Kirkpatrick, 1987] Betty KIRKPATRICK, . *Roget's Thesaurus of English Words and Phrases*. Penguin books, London, 1987. 3.1.1

[Kleiber, 1990] Georges KLEIBER. *La sémantique du prototype*. Presses Universitaires de France, 1990. 2.2

[Lafourcade, 2006] Mathieu LAFOURCADE. « Conceptual Vector Learning - Comparing Bootstrapping from a Thesaurus or Induction by Emergence ». Dans les actes de *LREC'2006*, Genoa, Italia, Mai 2006. 3.8

[Larousse, 1992] LAROUSSE, . *ThÉsaurus Larousse - des idÉes aux mots, des mots aux idées*. Larousse, 1992. 3.1.1, 11, 12, 3.2, 3.2.1, 3.2.2, 3.5.2

[Larousse, 2001] LAROUSSE, . *Le Petit Larousse Illustré 2001*. Larousse, 2001. 3.5.2

[Larousse, 2004] LAROUSSE, . *Le Petit Larousse Illustré 2004*. Larousse, 2004. 2.1.1, 3, 3.4.2

[Lehmann & Martin-Berthet, 1998] Alise LEHMANN et Francoise MARTIN-BERTHET. *Introduction À la lexicologie. Sémantique et morphologie.* Dunod, Paris, 1998. 2.1.3

[Lemaire & Dessus, 2003] Benoît LEMAIRE et Philippe DESSUS. « Modèles cognitifs issus de l'analyse sémantique latente ». *Cahiers Romans de sciences cognitives*, pp 55–74, 2003. 2.3.3

[Mel'čuk *et al.*, 1995] Igor MEL'ČUK, André CLAS, et Alain POLGUÈRE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, 1995. 2.4

[Muñoz *et al.*, 2000] Marcia MUÑOZ, Vasin PUNYAKANOK, Dan ROTH, et Dav ZIMAK. « A learning approach to shallow parsing ». Dans les actes de *EMNLP-WVL-99*, pp 168–178, 2000. 2.1.4

[Nogier, 1991] Jean-Francois NOGIER. *Génération automatique de langage et graphes conceptuels*. Hermès, 1991. 2.4.1

[Nyckees, 1998] Vincent NYCKEES. *La sémantique*. Belin, 1998. 2.5.1, 2.5.1

[Polguère, 2003] Alain POLGUÈRE. *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal, 2003. 2.2, 2.4

[Pottier, 1964] Bernard POTTIER. « Vers une sémantique moderne ». *Travaux de sémantique et de littérature*, pp 107–137, 1964. 2.5.1

[Quillian, 1968] Ross QUILLIAN. « *Semantic Informatic processing* », Chapitre Semantic memory, pp 227–270. MIT Press, 1968. 2.4.1

[Rastier, 2004] Francois RASTIER. « Ontologie(s) ». *Revue des sciences et technologies de l'information*, pp 15–40, 2004. 2.4.1

[Gamallo Otero & Bordag, 2011] Pablo GAMALLO OTERO et Stefan BORDAG. « Is singular value decomposition useful for word similarity extraction ? ». *Language Resources and Evaluation*, pp 95–119, 2011, Springer Netherlands. http ://gramatica.usc.es/ gamallo/artigos-web/LRE2010Web.pdf. 2.3.3

[Robert, 2000] Le ROBERT, . *Le Nouveau Petit Robert, dictionnaire alphabétique et analogique de la langue francaise*. Éditions Le Robert, 2000. 2.1.1, 3.5.2

[Roget, 1852] Peter Mark ROGET. *Roget's Thesaurus of English Words and Phrases*. Longman, London, 1852. 3.1.1

[Sabah, 1996] Gérard SABAH. « le sens dans les traitements automatiques des langues - le point après 50 ans de recherches ». Dans les actes de *journée ATALA (un demi-siècle de traitement automatique des langues : état de l'art)*, 1996. 2.5.1

[Salton & McGill, 1983] Gerard SALTON et Michael MCGILL. *Introduction to Modern Information Retrieval*. McGrawHill, New York, 1983. 2.3.2, 2.6.2

[Salton, 1971] Gerard SALTON. « The SMART Retrieval System – Experiments in Automatic Document Processing ». 1971. 2.3.2

[Salton, 1991] Gerard SALTON. « The Smart Document Retrieval Project ». Dans les actes de *Proc. of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp 357–358, Chicago, IL, 1991. 2.3.2

[Schank, 1972] Roger C. SCHANK. « Conceptual Dependency : A Theory of Natural Language Understanding ». *Cognitive Psychology*, pp pages 532–631, 1972. 2.5.1

[Schwab, 2005] Didier SCHWAB. « *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* ». Thèse de doctorat, Université Montpellier 2, 2005. 4

[Sowa, 1984] John SOWA. *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley, Reading, 1984. 2.4.1

[Sowa, 2000] John SOWA. *Knowledge Representation : Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, 2000. 2.4.1, 4

[Universalis, 1968] Encyclopædia UNIVERSALIS, . *Encyclopædia Universalis*, volume 17. Encyclopædia Universalis France, 1968. 2.5.2

[Vergne & Giguet, 1998] Jacques VERGNE et Emmanuel GIGUET. « Regards théoriques sur le "Tagging" ». Dans les actes de *TALN'1998*, Paris, France, Juin 1998. 2.1.4

[Wehrli, 1992] Éric WEHRLI. « The IPS system ». Dans les actes de *COLING'92 : 14th International Conference on Computational Linguistics*, pp 870–875, 1992. 2.1.4

[Wierzbicka, 1993] Anna WIERZBICKA. « La quête des primitifs sémantiques : 1965-1992 ». *Langue francaise*, Mai 1993. 2.5.1

[Wilks, 1977] Yorick WILKS. « Good and Bad Arguments About Semantic Primitives. ». *Communication and Cognition*, pp 181–221, 1977. 2.5.1

[Winograd, 1978] Terry WINOGRAD. « On primitives, prototypes, and other semantic anomalies ». Dans les actes de *conference on Theoretical Issues in Natural Language Processing*, pp 25–32, University of Illinois, 1978. 2.5.1