VIDEOSENSE

Agence Nationale de la Recherche ANR

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 16 mai 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

Ceci est la future intro...

2 État de l'art

2.1 Descripteurs de texte

2.1.1 Qu'est-ce que le sens ? [1]

La sémantique est l'étude du sens des énoncés. Cette science qui, bien que fort ancienne puisque déjà étudiée par les philosophes de l'Antiquité, fait encore l'objet de bien des recherches car non seulement le sens est indispensable dans une phase de compréhension de textes mais aussi car aucun moyen de le décrire complètement ne fait aujourd'hui l'unanimité.

Nombreux sont les ouvrages traitant de sémantique, mais fort rares sont ceux qui 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 car intuitif et 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.

Il n'est pas rare, même pour un locuteur du français, de rencontrer des mots qui lui sont inconnus. Ainsi, nous avons tous vécu un dialogue tel que celui que nous pouvons imaginer entre un maître et son élève :

– Que veut dire ≪ *prendre la poudre d'escampette* ≫ ?
– Cela signifie ≪ *s'enfuir* ≫ ou ≪ *se sauver à toutes jambes* ≫.

Un bon moyen de faire comprendre ce que signifie un mot est donc d'utiliser une expression équivalente, une paraphrase. Sur cette idée, Alain Polguère propose comme définition du sens :

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

On le voit, 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. **??**), 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

1. Partie fortement inspirée de [Schwab, 2005]

(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.

Nous allons alors être confronté 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 ?*

Ce sont, en partie, ces questions que nous étudions dans cette thèse. Lorsque nous envisagerons une désambiguïsation, la sémantique, dont nous parlerons alors, considérera toujours le niveau pragmatique même si pour simplifier le discours nous ne le préciserons pas.

2.2 Représentations d'origine distributionnaliste [2]

2.2.1 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 son maître, 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.

2.2.2 Représentations saltoniennes

À partir de la fin des années 1960, Gerard Salton [3] (*1927 - 1995*) professeur à la *Cornell University* [4] 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 [5] [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.

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 :

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

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

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*). Par cette formule, il s'agit de prendre en compte deux critères :

- *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* $= log(f(t,d)+1)$: la fonction logarithme augmente fortement dans les petites valeurs (¡100) et puis augmente de moins en moins vite. Cette formule du *tf* est donc à choisir 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* ; cf. **??**) :

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

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.

2.2.3 Une approche psycholinguistique : LSA

Le modèle LSA (*Latent Semantic analysis*), appelé souvent aussi LSI pour *Latent Semantic indexing*, a été créé dans un objectif de 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].

Limites de LSA [Otero & Bordag, 2010] montre que LSA n'est efficace ni en temps de calcul ni en résultat (**Partie à augmenter**)

2.3 Représentations symboliques connexionnistes

Ces représentations peuvent être dessinées grâce à des graphes dont les sommets correspondent à des objets lexicaux (item, acceptions) et les arêtes à des relations sémantiques.

2.3.1 Relations sémantiques et fonctions lexicales

Relations sémantiques lexicales (ou relations sémantiques externes) Nous avons déjà vu l'importance des paraphrases puisque ce sont elles qui nous ont permis de présenter ce que nous appelons le sens (cf. **??**). À travers ce petit dialogue, nous pouvons voir que communiquer c'est donc aussi pouvoir comprendre l'équivalence. Mais c'est également pouvoir comprendre les différences entre les phrases, savoir reformuler, développer, condenser ou améliorer

son expression. Il semblerait donc que les relations sémantiques lexicales soient nécessaires à la compréhension linguistique des individus. Elles nous permettent une certaine maîtrise du lexique.

Les relations sémantiques structurent le lexique sur le plan paradigmatique [6]. Nous présentons succinctement ici les six principales pour en donner une première idée nécessaire avant d'appréhender les fonctions lexicales et les réseaux sémantiques. Ces relations seront approfondies dans cette thèse lorsqu'il s'agira pour nous de les formaliser et de les modéliser à l'aide des vecteurs d'idées. Ces six relations sont de deux types, les *relations de hiérarchie* (*hyperonymie/hyponymie*, *holonymie/méronymie*) et les *relations symétriques* (*synonymie/antonymie*).

Relations de hiérarchie Ces relations sont unidirectionnelles et transitives. Elles structurent ainsi le lexique de façon hiérarchique. Il s'agit des deux paires hyponymie/hyperonymie et méronymie/holonymie. Si $\mathcal{R}$ est une relation hiérarchique entre deux items A et B, alors il existe une relation symétrique $\overline{\mathcal{R}}$ telle que :

$$\mathcal{R}(A, B) \equiv \overline{\mathcal{R}}(B, A)$$

Il existe deux paires de relations hiérarchiques : *hyponymie/hyperonymie* et *méronymie/holonymie*. La relation d'hyponymie est la relation hiérarchique qui lie un hyponyme à un item plus général, l'hyperonyme. La relation d'hyperonymie est la relation inverse. Parmi les exemples d'hyponymie, on peut trouver : ‹*chat*›\‹*animal*›, ‹*voilier*›\‹*bateau*›, ‹*bateau*›\‹*véhicule*›, ‹*rose*›\‹*fleur*›.

Relation tout-partie La relation de méronymie est la relation hiérarchique qui lie la partie au tout. Un des éléments de la relation est une partie de l'autre élément. Les deux relations sont symétriques c'est-à-dire que le tout est l'holonyme de la partie tandis que la partie est le méronyme du tout. Parmi les exemples de méronymie, on peut trouver : ‹*voile*›\‹*bateau*› (‹*voile*› est méronyme de ‹*bateau*›), ‹*page*›\‹*livre*›, ‹*mur*›\‹*maison*›, ‹*jour*›\‹*mois*›.

Relations symétriques Les relations symétriques sont la synonymie et l'antonymie. Comme leur nom l'indique, ces relations vérifient la propriété de symétrie. Ainsi, si $\mathcal{R}$ est une relation symétrique entre deux items A et B, alors :

$$\mathcal{R}(A, B) \equiv \mathcal{R}(B, A) \tag{4}$$

En d'autres termes, si A est synonyme (respectivement antonyme) de B, alors B est synonyme (respectivement antonyme) de A.

Synonymie La synonymie est la relation sémantique qu'il existe entre deux items lexicaux qui diffèrent par leur forme mais expriment le même sens ou un sens très proche.
‹*avion*›\‹*aéroplane*›, ‹*écrivain*›\‹*auteur*›, ‹*livre*›\‹*bouquin*›, ‹*chat*›\‹*matou*›.

Antonymie L'antonymie est la relation sémantique qui existe entre deux items lexicaux dont les sens s'opposent. Il existe trois types d'antonymie qui sont caractérisés par l'application ou non d'une propriété [Schwab, 2001, Schwab *et al.*, 2002, Schwab *et al.*, 2005] (‹*vie*›\‹*mort*›, ‹*certitude*›\‹*incertitude*›), une propriété étalonnable (‹*riche*›\‹*pauvre*›, ‹*chaud*›\‹*froid*›) ou un usage (‹*père*›\‹*fils*›, ‹*lune*›\‹*soleil*›).

Fonctions lexicales de production Les fonctions lexicales ont été créées dans le cadre de la théorie linguistique *sens-texte* (*TST*) élaborée par Igor Mel'čuk à Moscou avec des applications au russe jusqu'à la fin des années 1970 puis ensuite à Montréal pour le français. L'un des outils mis au point est le *Dictionnaire explicatif et combinatoire du français contemporain* (*DEC*) dont l'objectif est de décrire, de façon systématique et rigoureuse, les informations permettant à un locuteur de construire toutes les expressions linguistiques correctes de n'importe quelle pensée et ce, dans n'importe quel contexte : c'est un dictionnaire de production. Cette lexicographie est très détaillée et très organisée, ce qui lui permet d'être exploitable pour la fois à un usage humain et pour un usage "machinal" en TALN. Les items décrits dans le DEC le sont donc sous toutes leurs facettes : sémantique, syntaxique, lexico-combinatoire,

acception	autoriser.1		
propriétés grammaticales	*verbe*		
formule sémantique	donner le droit de : X autorise Y à Z.		
régime 1	1=X	2=Y	3=Z
	1. N	2. N *obligatoire*	3. *à* N
	« *Le docteur autorise l'avion à Pierre.* »		
régime 2	1=X	2=Z	3=Y
	1. N	2. N *obligatoire*	3. *à* V *obligatoire*
	« *Le docteur autorise Pierre à embarquer.* »		
Fonctions lexicales	$Syn_{\cap}$	permettre.I laisser.2, tolérer	
	$Anti$	interdire I, ne pas autoriser I.1	
	$Anti_{\subset}$	proscrire	
	$Anti_{\cap}$	défendre, s'opposer, empêcher,...	
	S_0	autorisation	
	$Magn$	formellement, expressément	

FIGURE 1 – Entrée résumée du DiCo pour l'item ‹*autoriser*› [Mel'čuk, 1988]

morphologique, ... Le tableau de la figure 1 présente l'entrée résumée du DiCo pour un des sens de l'item ‹*autoriser*› [Mel'čuk, 1988].

Les fonctions lexicales ont été définies dans le cadre de la TST pour décrire les relations sémantiques lexicales au moyen d'un outil formel conçu sur le modèle des fonctions mathématiques ([Polguère, 2003], p. 131), ([Mel'čuk *et al.*, 1995], p. 127).

Une fonction lexicale f décrit une relation existant entre un item lexical I (l'*argument* de f) et un ensemble d'items lexicaux $\{I_1, I_2, \ldots, I_n\}$ appelé la *valeur* de l'application de f à l'item I. La fonction lexicale f est telle que :

1. l'expression $f(I)$ représente l'application de f à l'item I : $f(I) = \{I_1, I_2, \ldots, I_n\}$
2. chaque élément de la valeur de $f(I)$ est lié à I de la même façon et remplit (à peu près) le même rôle : $\frac{f(I_1)}{I_1} \approx \frac{f(I_2)}{I_2}$

Ainsi, pour illustrer, la fonction lexicale de synonymie Syn modélise le rapport qu'entretiennent entre eux les termes ‹*livre*› et ‹*bouquin*› d'une part et ‹*chat*› et ‹*matou*› d'autre part. Mel'čuk note :

$$\frac{\text{‹}livre\text{›}}{\text{‹}bouquin\text{›}} \approx \frac{\text{‹}chat\text{›}}{\text{‹}matou\text{›}} \approx \frac{\text{‹}matou\text{›}}{\text{‹}chat\text{›}}$$

De même, la fonction lexicale d'antonymie $Anti$ modélise le rapport qu'entretiennent entre eux les termes ‹*certitude*› et ‹*incertitude*› d'une part et ‹*vie*› et ‹*mort*› d'autre part.

$$\frac{\text{‹}certitude\text{›}}{\text{‹}incertitude\text{›}} \approx \frac{\text{‹}vie\text{›}}{\text{‹}mort\text{›}} \approx \frac{\text{‹}mort\text{›}}{\text{‹}vie\text{›}}$$

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*.

Fonctions lexicales paradigmatiques Comme leur nom l'indique, les fonctions lexicales paradigmatiques formalisent les relations sémantiques. On distingue par exemple :

6. Le plan paradigmatique est le plan dans lequel les termes sont unis par leur sens à l'intérieur du lexique.

- *Synonymie* (Syn) :
  Syn(ᶜ*avion*ᵉ) = ᶜ*aéronef*ᵉ, ᶜ*aéroplane*ᵉ, ᶜ*coucou*ᵉ,...
- *Antonymie* ($Anti$) :
  $Anti$(certitude) =ᶜ*incertitude*ᵉ, ᶜ*doute*ᵉ, ᶜ*scepticisme*ᵉ, ...
- *Générique* ($Gener$) : Les génériques sont les équivalents des hyperonymes.
  $Gener$(ᶜ*rose*ᵉ) = ᶜ*fleur*ᵉ ; $Gener$(ᶜ*chat*ᵉ) =ᶜ*animal*ᵉ, ...
- *Dérivés syntaxiques* : Ces fonctions associent à un item sa contrepartie nominale (Substantification S_0), verbale (verbalisation V_0), adjectivale (adjectivisation A_0) ou adverbiale (Adv_0) :
  S_0 (ᶜ*tuer*ᵉ) = ᶜ*meurtre*ᵉ, S_0 (ᶜ*vivre*ᵉ) = ᶜ*vie*ᵉ, V_0(ᶜ*serment*ᵉ) = ᶜ*jurer*ᵉ, S_0(ᶜ*jurer*ᵉ)=ᶜ*serment*ᵉ

Les fonctions lexicales peuvent être indicées par des opérateurs ensemblistes pour indiquer des nuances de sens. Ainsi, on trouvera :

- $\subset$ pour indiquer une inclusion du sens de l'argument dans la valeur de la fonction. On trouve ce cas avec les rapports d'hyperonymie : $Syn_{\subset}$(ᶜ*pigeon*ᵉ) =ᶜ*oiseau*ᵉ ;
- $\supset$ pour indiquer une inclusion du sens de la valeur dans l'argument de la fonction : $Syn_{\supset}$(ᶜ*oiseau*ᵉ) =ᶜ*pigeon*ᵉ ;
- $\cap$ pour indiquer une intersection de sens. Ainsi $Syn_{\cap}$(ᶜ*jouer*ᵉ) =ᶜ*s'amuser*ᵉ puisqu'on peut jouer sans s'amuser et s'amuser sans jouer.

Fonctions lexicales syntagmatiques : collocations Dans toutes les langues, certaines combinaisons d'items lexicaux prévalent sur d'autres sans qu'il ne semble n'y avoir de motif logique. Par exemple, on parle de ≪ *dormir profondément* ≫ plutôt que de *≪ *dormir intensément* ≫ ou *≪ *dormir totalement* ≫ pourtant aucune raison (du moins en synchronie) ne semble expliquer cette préférence. On parle, dans ces cas, de phénomène de *collocation*.

L'énoncé AB (ou BA) formé des items lexicaux A et B est une collocation si, pour produire cette expression, le locuteur sélectionne A librement d'après son sens alors qu'il sélectionne B pour exprimer un autre sens en fonction de A ([Polguère, 2003], p. 134). On appelle A *base de la collocation* et B *collocatif*. On peut citer comme exemples de collocations en français : ᶜ*tir*ᵉ$_{[=A]}$ ᶜ*nourri*ᵉ$_{[=B]}$, ᶜ*peur*ᵉ$_{[=A]}$ ᶜ*bleue*ᵉ$_{[=B]}$, ᶜ*forte*ᵉ$_{[=B]}$ ᶜ*fièvre*ᵉ$_{[=A]}$,ᶜ*dormir*ᵉ$_{[=A]}$ ᶜ*profondément*ᵉ$_{[=B]}$.

Les fonctions lexicales paradigmatiques ont été créées pour rendre compte des collocations non seulement dans le rôle syntaxique que joue le collocatif auprès de la base mais aussi par le sens qu'il exprime. Parmi les fonctions lexicales syntagmatiques, on peut citer :

- Bon qui marque une évaluation positive : Bon(ᶜ*choix*ᵉ) = ᶜ*bon*ᵉ ;
- $Magn$ qui marque l'intensification : $Magn$(ᶜ*majorité*ᵉ) =ᶜ*forte*ᵉ, ᶜ*écrasante*ᵉ.

ou leurs opposés :

- $AntiBon$ qui marque une évaluation négative : $AntiBon$(ᶜ*choix*ᵉ) = ᶜ*mauvais*ᵉ ;
- $AntiMagn$ qui marque une modification inverse à l'intensification : $AntiMagn$(ᶜ*majorité*ᵉ) =ᶜ*courte*ᵉ, ᶜ*faible*ᵉ.

2.3.2 Réseaux sémantiques

Les réseaux sémantiques tirent leur origine de la psychologie expérimentale et plus particulièrement des travaux concernant l'organisation mentale des concepts.

Origines À la fin des années 1960, Alan Collins et Ross Quillian observent que, pour un être humain, le temps d'estimation de la validité d'une phrase comme ≪ *un chien est un mammifère* ≫ est plus long que celui d'une phrase comme ≪ *un chien est un animal* ≫ [Collins & Quillian, 1969] [Collins & Quillian, 1970]. Cette différence dans le délai de réaction semble liée au nombre d'individus de la classe [Juola & Atkinson, 1971] [Landauer & Freedman, 1968] comme le montre l'expérience présentée dans la figure 2.

Ces expériences semblent montrer que les informations associées à certains concepts sont transmissibles aux concepts hyponymes et ne sont pas mémorisées directement avec ceux-ci. En d'autres termes, les concepts spécialisés héritent des propriétés et des attributs des concepts plus généraux auxquels ils sont associés et y adjoignent leurs propres propriétés et attributs. Ainsi, il s'agit d'un *principe d'économie* reposant sur la mise en facteur des connaissances communes à plusieurs sortes d'objets [Bernard, 2000] [Sabah, 1988]. Par exemple, ᶜ*mammifère*ᵉ met en facteur les propriétés et les attributs communs aux différentes espèces qui le composent (ᶜ*Homme*ᵉ, ᶜ*chien*ᵉ, ᶜ*lapin*ᵉ,...) et qui

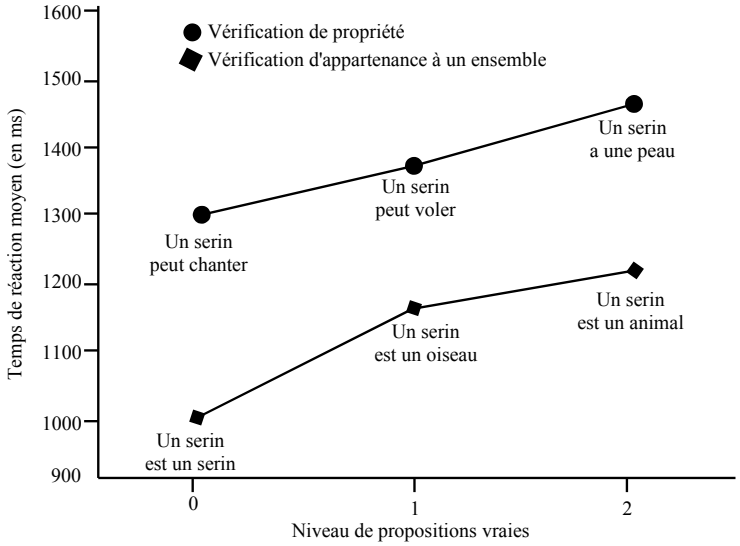


FIGURE 2 – Vérification d'énoncés sémantiques : résultats expérimentaux [Collins & Quillian, 1969]

en retour en héritent. Le figure 3 présente l'automate de Quillian qui est un exemple de hiérarchie centrée autour d'‹*animal*›.

Plusieurs expériences dont celle d'Harold Goodglass (*1920 - 2002*) et Errol Baker concernant l'aphasie [7] corroborent cette thèse [Goodglass & Baker, 1976]. Pour les sujets de l'expérience, il s'agit d'associer ou non à un terme cible des termes cités oralement. Goodglass et Baker constatent que suivant les relations existant entre les termes proposés (similarité, hyponymie, pragmatique) et plus particulièrement le nombre et la nature des attributs sémantiques que partagent les termes, les temps de réponse diffèrent.

Les travaux de Ross Quillian [Quillian, 1968] proposent les réseaux sémantiques comme un modèle de cette mémoire associative.

Modèle

Notions de base 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. 4).

On constate sur l'exemple 4 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* ».

Composition de relations 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 est la transposition dans les réseaux sémantiques du principe d'économie abordé à la section 2.3.2, il permet de réaliser une économie d'espace mémoire pour la représentation du réseau mais ralentit le temps de traitement. Pour Quillian, cognitivement, la longueur du chemin reliant deux nœuds doit refléter le temps

7. L'aphasie est une perte totale ou partielle du langage consécutive à une atteinte cérébrale. Les personnes aphasiques ne peuvent plus (ou avec difficulté) parler, comprendre, lire et écrire. http://membres.lycos.fr/aphasie/

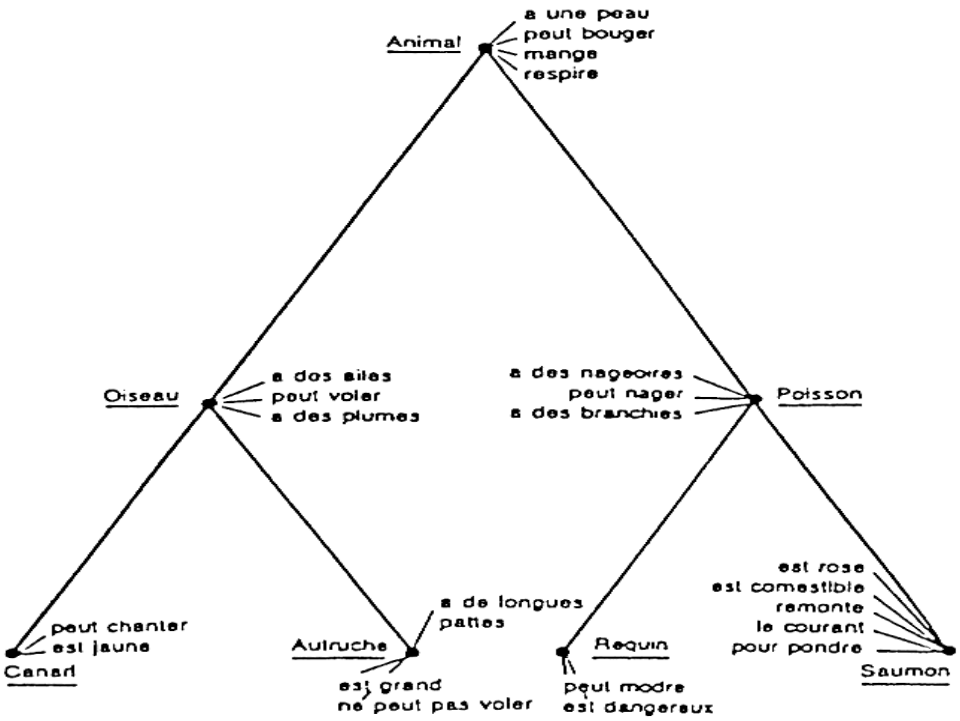
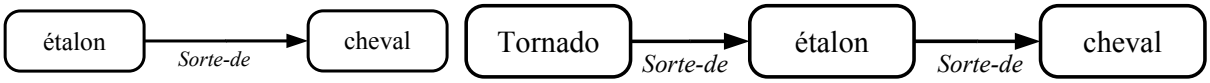


FIGURE 3 – L'automate de Quillian : hiérarchie centrée autour d'‹*animal*›.



(a) Ce réseau représente ≪ *un étalon est un cheval.* ≫. (b) On rajoute un nœud au réseau pour indiquer que Tornado est un étalon

FIGURE 4 – Réseaux sémantiques élémentaires.

mis par des humains pour associer les concepts correspondant à ces nœuds et leurs propriétés. Le réseau de la figure 5 permet de retrouver que les *étalons* en général et *Tornado* en particulier possèdent des *sabots*.

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 5 contient la propriété *Sorte-de* $\circ$ *possession* $=$ *possession*.

Types de nœud Représenter une information comme ≪ *Diego possède un étalon* ≫ peut poser le problème de l'arc à introduire dans le réseau. Il ne faut pas comme sur le réseau de la figure 6(a) mettre un arc de *Possession* entre les nœuds *Diego* et *cheval* puisque cela signifierait que l'ensemble des éléments de la catégorie *cheval* appartient à *Diego*. La représentation correcte verrait un arc entre un propriétaire et une instance de la classe *cheval*, dans l'exemple 6(b) pour *Diego*, *Tornado* et pour *Alec*, *Black*. Il est donc nécessaire de différencier les nœuds d'instance des nœuds de classes.

On veut aussi pouvoir représenter d'autres informations comme des informations temporelles (≪ *Alec possède Black d'octobre 1941 à août 1950* ≫) auquel cas l'état n'est plus simplement codé par un lien mais aussi par un nœud. Ce nœud sera alors une spécialisation de la notion d'*appartenance* comme le montre la figure 7

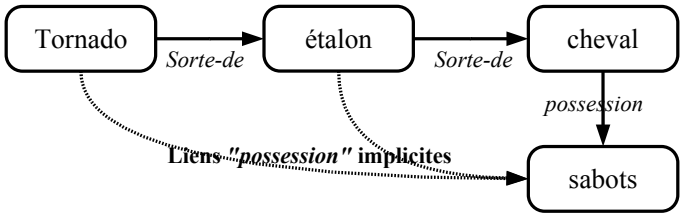
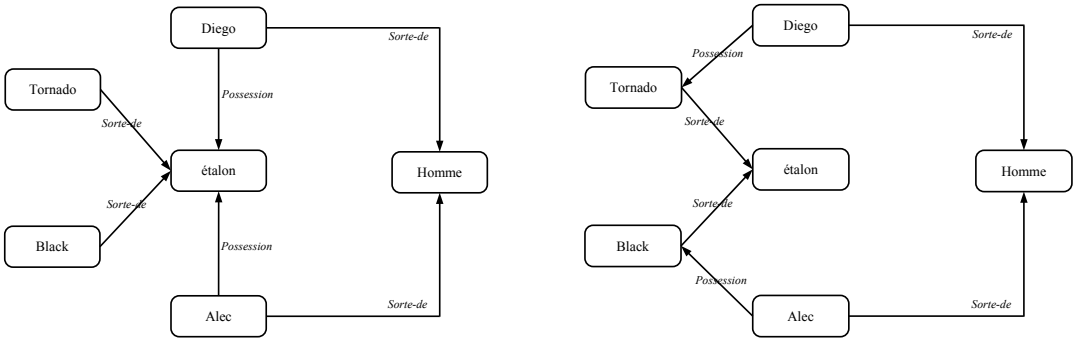


FIGURE 5 – Héritage de propriétés



(a) Représentation erronée des informations dans un réseau

(b) Représentation correcte des informations de la figure

FIGURE 6 – Importance de la différenciation des types de nœuds dans un réseau sémantique

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 8 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].

Les réseaux d'aujourd'hui : WordNet Contrairement aux réseaux sémantiques classiques qui cherchent à la fois à représenter des phrases et à ordonner les connaissances sur le monde (le côté ontologique des réseaux), le projet WordNet est uniquement concentré sur cette deuxième tâche. Libre ensuite aux développeurs de l'utiliser pour d'autres applications. 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 2 de WordNet compte 152059 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 n'y a pas de relations comme la substantification S_0 cf. 2.3.1). Il y a donc une hiérarchie pour les noms, une pour les

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

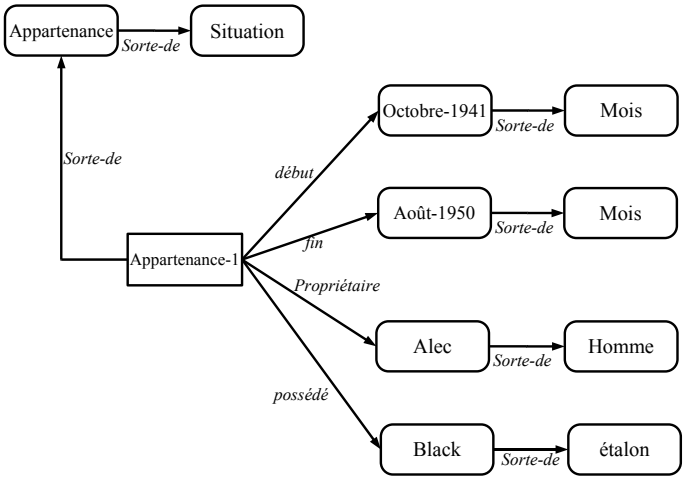


FIGURE 7 – Représentation de l'information ≪ *Alec possède Black d'octobre 1941 à août 1950* ≫

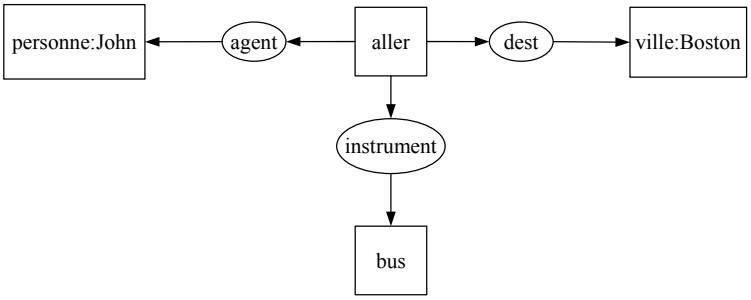


FIGURE 8 – Exemple de graphe conceptuel : ≪ *John va à Boston en bus.* ≫ [Sowa, 2000]

adjectifs, une pour les verbes et enfin une dernière pour les adverbes. Un extrait de la hiérarchie des noms est présenté dans la figure 9.

Limites des réseaux sémantiques L'une des principales critiques adressées aux réseaux sémantiques concerne le modèle de la mémoire associative dont ils sont issus. En effet, diverses études ont montré que certains hyponymes sont plus caractéristiques de leur catégorie que d'autres. Par exemple, une ‹*pomme*› ou une ‹*orange*› sont plus considérées comme appartenant au genre ‹*fruit*› qu'une ‹*noix*› ou une ‹*olive*› si on se réfère aux temps de réaction des sujets. Par ailleurs, Carol Conrad [Conrad, 1972] a montré que les temps de réaction des sujets à des énoncés n'étaient pas seulement fonction du nombre d'individus de la classe et du parcours d'un réseau d'hyponymes, mais aussi de la fréquence des énoncés (cf. figure 10).

Ces remarques sont à la base de la *sémantique du prototype* dont les tenants considèrent que chaque classe contient un prototype, c'est-à-dire un élément plus "typique" que les autres (la ‹*pomme*› pour les fruits par exemple) ; en d'autres termes, qu'il présente dans sa catégorie ≪ *à la fois un maximum de points communs avec les autres éléments de la catégorie et un minimum de points communs avec les éléments de catégories opposées* ≫ [Nyckees, 1998].

Comme nous l'avons dit dans la partie **??**, l'adéquation avec le modèle cognitif ne suffit pas à justifier que ce modèle informatique est insuffisant. Celui-ci est aussi l'objet de critiques de fond, en particulier de la part de François Rastier [Rastier, 2004]. Il lui reproche de n'être qu'une vision du monde. ≪ *Ce qu'on appelle le mobilier ontologique du monde, ce qui est présenté comme naturel et dit par toutes les langues appartient en réalité à un certain type de civilisation* ≫. Certaines langues distinguent ‹*pied*› et ‹*jambe*›, ce que les langues slaves ou malaises ne font pas.

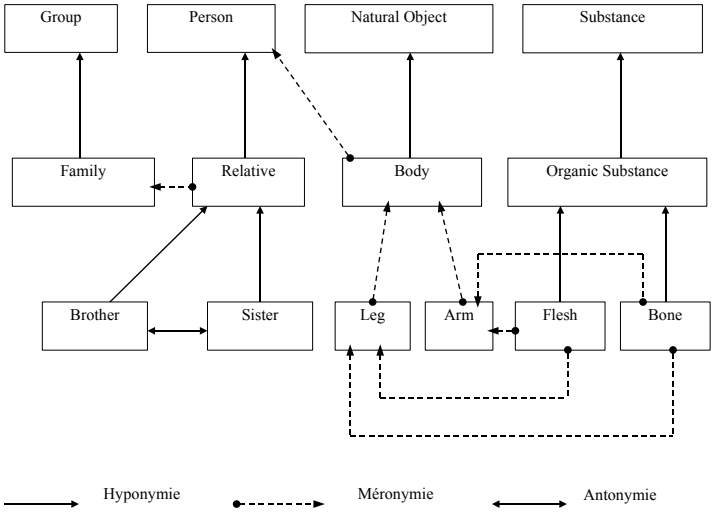


FIGURE 9 – Extrait de la hiérarchie des noms

niveau	fréquent	rare
1	≪ *Un requin peut bouger* ≫	≪ *un saumon a une bouche* ≫
2	≪ *Un oiseau peut bouger* ≫	≪ *Un poisson a des yeux* ≫
3	≪ *Un animal peut bouger* ≫	≪ *Un animal a une peau* ≫

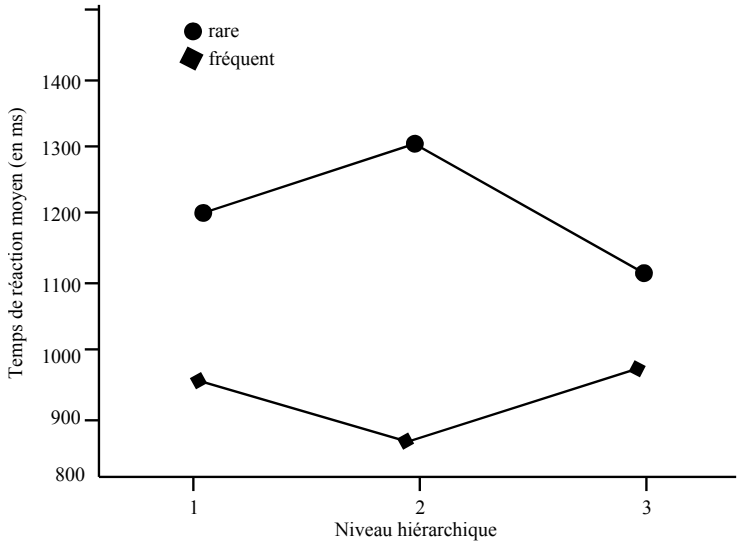


FIGURE 10 – Expérience de Conrad [Conrad, 1972]

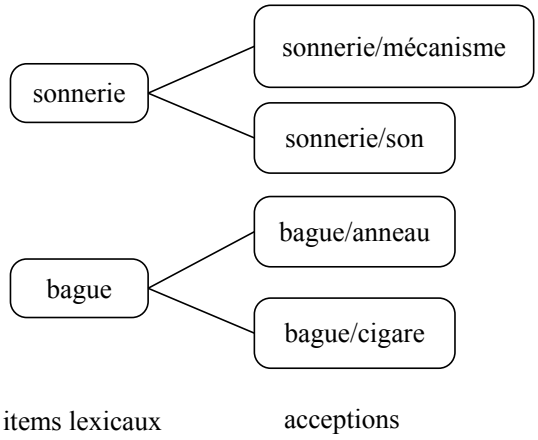


FIGURE 11 – Items lexicaux et acceptions de ‹*bague*› et ‹*sonnerie*›

Les réseaux sont utilisés comme si les sens étaient prédéterminés or ce n'est pas le cas. Il y a des coutumes d'usage mais les langues évoluent et ses coutumes aussi. Certains termes sont créés (néologismes), d'autres prennent de nouvelles acceptions, certaines de ces formations devenant plus fréquentes se verront lexicaliser et finalement figurer dans un dictionnaire. Dans le cas des réseaux, ≪ *on fixe (la langue), on la bloque dans un moment, dans une forme de civilisation ou du moins dans une forme de système économico-technique et on dit voilà ce que c'est que l'esprit humain. Ça s'appelle une prise de pouvoir.* ≫ [9]

2.3.3 Bases d'acceptions

Le modèle de base d'acceptions a été développé à Grenoble depuis le début des années 1990 par Gilles Sérasset et Christian Boitet . Aujourd'hui, sa principale implémentation concerne le projet Papillon dont elle constitue la macro-structure. Ce projet, mené depuis 2000, vise à la constitution d'une base lexicale multilingue linguistiquement riche. Du côté organisationnel, son principal atout est de se baser sur un principe collaboratif qui permet à n'importe quelle personne qui le souhaite de l'enrichir [Mangeot-Lerebours *et al.*, 2003]. La base comprend entre autres l'anglais, le français, le japonais, le malais, le lao, le thaï, le vietnamien et le chinois.

Acceptions Une acception est un sens particulier d'un mot, admis et reconnu par l'usage. Il s'agit d'une unité sémantique propre à une langue donnée [Sérasset & Mangeot, 2001]. Par exemple, en français, l'item lexical ‹*bague*› a, au moins, deux acceptions, l'**anneau**, ou la collerette de papier entourant le cigare (annotée **cigare**) tandis que l'item lexical ‹*sonnerie*› en a deux, le **son** et le **mécanisme** qui l'émet. Une acception est en fait ce qu'on appelle communément ≪ *un sens d'un mot* ≫. Ainsi, pour revenir à l'une de nos préoccupations principales, désambiguïser, c'est trouver quelle est l'acception d'un mot qui semble concorder le mieux avec les autres mots de la phrase.

Base d'acceptions Une base d'acceptions monolingue contient des entités ITEM LEXICAL et des entités ACCEPTION qui regroupent les informations sur les différents sens que peut prendre l'item. Dans le cas du dictionnaire Papillon, par exemple, cette microstructure est basée sur la Lexicographie Explicative et Combinatoire, partie de la Théorie Sens-Texte (cf 2.3.1).

La figure 11 présente l'architecture d'une base par acception avec ‹*bague*› et ‹*sonnerie*› dans un cadre monolingue. Les acceptions ont été annotées pour faciliter la compréhension.

Dans un cadre multilingue [Mangeot-Lerebours *et al.*, 2003], une base contient plusieurs bases d'acceptions monolingues dont les acceptions sont reliées par des acceptions interlingues appelées *axies*. Un exemple d'une telle

9. Extrait de l'émission *Tire ta langue* du 18 Février 2003 sur *France Culture* http://www.radiofrance.fr/chaines/france-culture2/emissions/tire_langue/fiche.php?diffusion_id=11608

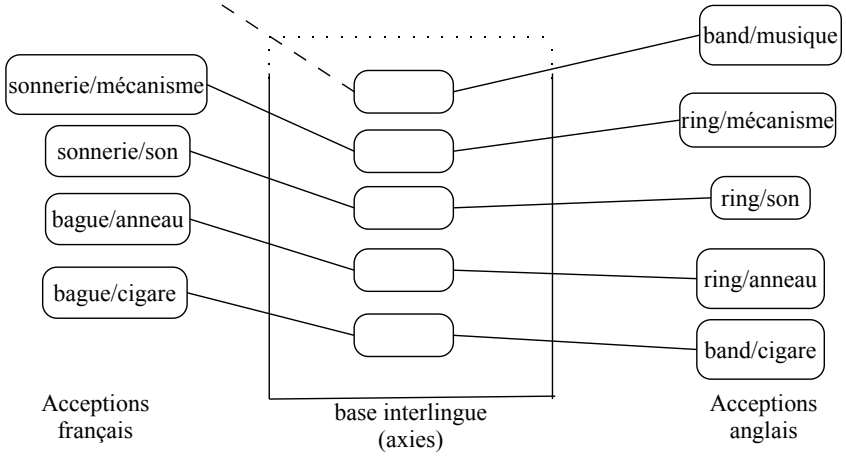


FIGURE 12 – Acceptions et axies en multilingue

architecture est présenté en 12. D'un côté les acceptions monolingues du français, de l'autre celles de l'anglais reliées entre elles par des axies.

Ce système permet de représenter les raffinements de sens de certaines langues. L'exemple de ‹*fleuve*› et ‹*rivière*› est caractéristique. Le français différencie, en effet, « *un cours d'eau qui se jette dans la mer ou l'océan.* » (‹*fleuve*›) et « *cours d'eau qui se jette dans un autre cours d'eau* » (‹*rivière*›) tandis que ni l'anglais ni l'espagnol ne font cette distinction (cf.figure 13).

2.4 Approche componentielle (ou sémique)

2.4.1 Le sens vu comme la composition de primitives

La linguistique componentielle postule que le sens d'un terme peut être défini par un ensemble de primitives de base. Cette idée est le prolongement des réflexions de Leibniz (*1646 - 1716*) qui a passé une partie de sa vie à la recherche d'un *alphabet des pensées*. Si on pense qu'il peut exister un tel alphabet, il doit en exister nécessairement un qui permettrait de représenter les mots qui ne sont, après tout, que des étiquettes accolées à certaines pensées. Les structuralistes, en particulier les linguistes héritiers de Leibniz comme Hjelmslev, Pottier, Greimas ou Rastier, s'inspirent à la fois de ces idées et des théories de la phonologie pour mettre au point l'analyse sémique et la théorie des primitives sémantiques qui en est une conséquence directe. Prenant la suite de Leibniz, la linguiste Wierbicka a étudié de nombreuses langues à la recherche de ces primitives, les informaticiens Wilks et Schank ont essayé de construire de manière moins universelle un ensemble d'atomes de sens qui permettrait à un système informatique de représenter les sens exprimables en langue.

Origine de l'approche componentielle : l'analyse sémique Au début des années 1940, l'étude de la sémantique a pris énormément de retard sur les autres branches de la linguistique ([Nyckees, 1998], p. 206). Les significations des mots sont encore expliquées uniquement par le rapport au monde que chacune d'entre elles entretient. La phonologie, en revanche, est descendue bien en dessous du mot en le décomposant en unités phoniques plus petites, les *phonèmes* eux-mêmes descriptibles suivant la zone de la bouche ou le mode d'articulation qu'il faut utiliser pour les produire. L'étude de la structure des phonèmes a permis de montrer qu'ils pouvaient être considérés comme un faisceau de *traits distinctifs*. Ainsi, si on compare /p/ /b/ et /m/ , il se rejoignent sur le trait de la bilabialité (utilisation des deux lèvres) mais se distinguent sur le trait sonore (vibration des cordes vocales), et le trait nasal (un + dénote l'existence du trait, un - sa non-existence) ([Nyckees, 1998], p. 207) :

	sonore	nasal
/p/	-	-
/b/	+	-
/m/	+	+

Certains sémanticiens se demandent alors si on ne peut pas extraire des structures similaires pour décrire le sens des items lexicaux. En d'autre termes, peut-on faire une description fine du sens des items lexicaux grâce à une décomposition en unités sémantiques minimales ?

La linguistique componentielle suppose donc l'existence d'une atomisation de la signification ([Le Ny, 1979], p. 122), c'est-à-dire que le sens d'un terme n'est plus considéré comme primitif, mais peut être décomposé en éléments de sens plus petits appelés suivant les diverses écoles : sèmes [10], noèmes, traits sémantiques, atomes de sens, primitives, ...

Contrairement aux phonèmes qui ne sont que quelques dizaines [11], il y a des centaines de milliers voire des millions d'items lexicaux dans une langue. Les sémanticiens n'ont donc pas cherché à décrire l'ensemble du lexique, mais se sont restreints à décrire des termes qui ont un trait sémantique commun fort.

La linguistique componentielle tire son origine des années 1940 et des travaux de Hjelmslev sur l'analyse en composants sémantiques.

Hjelmslev et l'analyse en composants sémantiques Le danois Louis Hjelmslev (*1899 - 1965*) prône dès 1943 la comparaison des termes en fonction des sèmes qui les composent (analyse en composants sémantiques) [Hjelmlev, 1968]. Selon cette théorie, les termes ‹*garçon*›, ‹*fille*›, ‹*homme*› et ‹*femme*› peuvent êtres analysés grâce aux traits sémantiques *ANIMÉ*, *HUMAIN*, *MÂLE*, *FEMELLE*, *ADULTE* ([Éco, 1988], p. 136).

‹*garçon*› : *ANIMÉ* + *HUMAIN* + *MÂLE* - *ADULTE*
‹*fille*› : *ANIMÉ* + *HUMAIN* + *FEMELLE* - *ADULTE*
‹*homme*› : *ANIMÉ* + *HUMAIN* + *MÂLE* + *ADULTE*
‹*femme*› : *ANIMÉ* + *HUMAIN* + *FEMELLE* + *ADULTE*

Certains traits sémantiques, comme ici *ANIMÉ*, sont considérés afin de justifier la concordance des termes avec certains verbes. On estime que les phrases ≪ *L'homme respire.* ≫ ou ≪ *Le chien respire.* ≫ sont grammaticalement justes parce que ‹*respirer*› s'adapte positivement au trait *ANIMÉ* et indifféremment à *HUMAIN* ou *ANIMAL*. Par contre, on considère qu'il est incorrect de dire *≪ *Le rocher respire* ≫ (si on exclut les sens métaphoriques) parce que ‹*respirer*› ne peut concerner que des *ANIMÉS*. Ces valences combinatoires du verbe sont dites *restrictions sélectives*. Bien qu'elle ait donné des résultats intéressants, l'analyse en traits sémantiques sert davantage à expliquer les concordances grammaticales qu'à expliquer les concordances sémantiques.

Éco (*né en 1932*) formule deux objections à cette démarche ([Éco, 1988], p. 137).

1. Tandis qu'il est possible d'organiser en système les catégories grammaticales du fait de leur nombre restreint, il est difficile d'organiser les catégories sémantiques qui sont bien plus nombreuses.
2. Le grand nombre de ces catégories fait que s'il est aisé de définir ‹*homme*› par rapport à ‹*femme*›, il l'est moins pour ‹*vache*› en fonction de ‹*brebis*›. Il s'agit dans les deux cas d'animés, d'animaux et de femelles et pourtant il ne s'agit pas de la même chose.

L'analyse sémique de Pottier Parmi les analyses sémiques les plus connues figurent celles effectuées par Bernard Pottier [Pottier, 1964]. La figure 14 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.

Dans cet exemple, chaque ligne représente un sémème, c'est-à-dire l'ensemble des sèmes que le mot comporte. Le seul sème qui appartienne à tous les mots est celui de la dernière colonne, *TRANSPORT DE PERSONNES*. Il constitue aussi, à

10. Greimas appelle sèmes ces éléments. Comme bien souvent dans d'autres disciplines, la terminologie utilisée par les auteurs est loin d'être unifiée. Nous le constaterons à plusieurs reprises dans cette thèse. Le terme ‹*sème*› en est un bon exemple et le lecteur, sous peine de confusion dans cette partie, doit en avoir bien conscience. En effet, un sème est un atome de sens chez Greimas mais chez Pottier, et c'est sa définition qui est le plus souvent utilisée, un ‹*sème*› est plutôt une composition de ‹*noèmes*›, le nom des primitives de sens pour Pottier.
11. En français, on compte, par exemple, 37 phonèmes.

lui seul, le sémème de ‹*véhicule*›, item lexical qui peut s'appliquer à tous les objets dénommés par les autres termes de la liste et qui est donc par rapport à eux, l'hyperonyme le plus proche. On l'appelle *archisémène*.

Tout comme pour une analyse phonologique, il s'agit ici de chercher des traits distinctifs entre les items. Ceux sont uniquement eux qui sont notés dans le tableau (à l'exception de l'archisème). Les sèmes constituent donc l'ensemble minimal permettant de différencier les items étudiés entre eux.

Par la suite, l'analyse sémique a été reprise, en psycholinguistique, par Jean-François Le Ny [Le Ny, 1979] et, en linguistique par François Rastier [Rastier, 1989] et Algirdas Julien Greimas [Greimas, 1986] mais, pour ce dernier, avec une définition du terme ‹*sème*› plus proche d'‹*atome de sens*› que de ‹*trait distinctif*›.

2.4.2 Les primitives de sens

L'analyse sémique s'attache à identifier pour un certain nombre de termes leur sémène, c'est-à-dire l'ensemble des sèmes qu'ils comportent. Même si ces sèmes ne sont pas des atomes de sens [12], 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 2.2.1 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).

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.

La recherche des primitives

Chez les linguistes À la suite d'études approfondies sur les langues les plus diverses durant presque trente ans, Anna Wierbicka a proposé en 1992 une liste de 35 primitifs universaux présentée par la figure 15.

Ces primitifs, selon elle, « *(. . .) sont censés être des universaux d'ordre lexical ; il est présumé qu'ils sont lexicalisés dans toutes les langues du monde. Les recherches en sémantique transculturelle semblent indiquer qu'il y a un nombre plutôt restreint de concepts qui sont bel et bien universels (et probablement innés). L'hypothèse qu'il n'y en aurait qu'une douzaine s'est révélée incorrecte ; il faut multiplier ce chiffre par trois. De toute façon, le nombre de primitifs ne s'élève pas dans les milliers, ni même dans les centaines.* »

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. En effet, comment représenter avec ces primitifs ‹*voiture*›, ‹*lit*› ou ‹*dormir*› ? Dans l'hypothèse où ces primitifs seraient réellement à la base de toute la pensée humaine, il est difficile (impossible ?) de retracer le cheminement parcouru pour arriver au sens de ces termes.

Un deuxième problème concernant cette vision de Wierzbicka 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, rappelons-le, plus de 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 ;

12. Toujours en adoptant la terminologie de Pottier mais pas celle de Greimas.

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 donc 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).

2.4.3 Le Dictionnaire Intégral

Le *dictionnaire intégral* est un réseau sémantique développé par la société Memodata [13] depuis plus d'une quinzaine d'années. Il allie à la fois des connaissances relatives à l'approche componentielle et des fonctions lexicales à la Mel'čuk ; sa couverture lexicale est très étendue (comparable à celle de WordNet) ; enfin, les applications visées sont hétérogènes et s'étendent du résumé automatique au filtrage d'information en passant par la comparaison textuelle et la traduction automatique. C'est pour ces caractéristiques, très proches de celles visées par notre équipe, qu'il a paru intéressant de le présenter ici.Dans sa thèse [Dutoit, 2000], Dominique Dutoit co-fondateur de Memodata, présente ce dictionnaire intégral (DI), sa construction et les diverses applications mises en œuvre.

Architecture du dictionnaire intégral Dans ce modèle, la description de chaque sens de mots (appelé *mot-sens*) se fait selon trois points de vue considérés comme complémentaires : la sémantique componentielle, les fonctions lexicales sémantiques et les propositions courantes.

13. http://www.memodata.com/

Sémantique componentielle La brique de base du dictionnaire intégral est le concept. Chaque concept est désigné par un identifiant qui commence par un '\' et une majuscule comme, par exemple, \FLEUR ou \VENDRE. Les concepts sont organisés en hiérarchie dont les feuilles sont les mots-sens. Selon ses concepteurs [Dutoit, 2000], un concept du dictionnaire intégral est artificiel et peut être conçu selon de très nombreux besoins différents. Ainsi, un concept sert à renseigner de façon non-ambiguë, d'un côté un utilisateur humain sur son contenu (les concepts qui en découlent) et sur son contenant (les concepts dont il découle) d'un autre, la machine sur la façon de l'utiliser (possibilité ou impossibilité de passer d'un concept à un autre).

Il existe deux sortes de concepts :

- les *classes* sont situées dans la partie basse de la hiérarchie et correspondent des sous-hiérarchies de concepts de même morphologie (verbes [/V], noms [/N], adjectifs [/A], adverbes) ;
- les *thèmes*, notés ([T]), qui sont situés dans la partie haute de la hiérarchie générale et correspondent au champ sémantique commun à tous les thème et classes successeurs dans la hiérarchie.

Propositions courantes Les propositions courantes correspondent aux relations syntaxiques qu'il peut exister entre mots-sens. Ils reprennent les idées adoptées par le DEC en ce qui concerne leur régime (cf. 2.3.1) qu'ils adaptent à leur architecture pour éviter en particulier la redondance d'informations. Ainsi alors qu'une relation est notée dans tous les articles du DEC, elle sera mentionnée à part dans le DI.

Fonctions lexicales sémantiques Le DI relie les mots-sens entre eux grâce à des fonctions lexicales sémantiques. Il y en a 66 différentes pour 96 000 liens décrits [Dutoit & Nugues, 2002]. La liste de ces fonctions a été établie à partir de celles de Mel'čuk (cf. 2.3.1) mais chacune a été réétudié en fonction des spécificités du dictionnaire intégral :

- *spécificités syntaxiques* : Certaines fonctions lexicales rendent compte de phénomènes sémantiques décrits en partie par la syntaxe. C'est le cas par exemple des compléments circonstanciels. Ainsi, les dérivés sémantiques nominaux circonstanciels (S_{instr}, S_{loc}, S_{med}, S_{mod}, S_{res}) ne sont pas introduits dans le DI comme des relations de nature sémantique mais comme des relations syntaxiques. Ainsi, la fonction S_{instr} n'existe pas dans le DI mais peut être trouvée grâce au complément circonstanciel de moyen ;
- *spécificités sémantiques* : les concepteurs du DI considère que le sens des mots-sens est donné par les sèmes qu'il met en jeu. Sur cette idée, ils n'ont pas de fonction lexicale sémantique de synonymie telle que celle de Mel'čuk qui sont remplacées par le réseau de concepts.

Construction Cette base est construite manuellement depuis une quinzaine d'années par une équipe de trois personnes [14]. Il contenait en 2002, 16 000 thèmes, 25 000 classes et pour le français 190 000 mots-sens. Il continue actuellement à être construit en particulier vers d'autres langues.

On peut s'interroger sur cette construction manuelle qui certes a le mérite d'offrir une précision manifeste mais est coûteuse en temps. En effet, cette tâche est en grande partie similaire à celle effectuée par les lexicographes pour réaliser un dictionnaire papier. Or, cette tâche leur a pris des dizaines d'années et doit constamment être renouvelée puisque des termes et des sens apparaissent tandis que d'autres disparaissent. Ainsi n'y avait-il pas quelques possibilité pour fabriquer (au moins en partie) le DI à partir de tels dictionnaires ? Nous verrons dans la suite de la thèse que c'est pour ces raisons que nous avons adopté pour les vecteurs conceptuels un apprentissage permanent à partir de dictionnaires à usage humain préexistants à nos expériences.

Applications La société Memodata développe à partir de ce réseau de nombreuses applications commerciales. L'une des plus anciennes est le dicologique qui est la version du dictionnaire intégral de 1992 [15]. Il permet de trouver définitions, synonymes, analogies, rimes, mots-croisés, anagrammes, conjugaisons, féminins et pluriels [Dutoit, 1992]. Ils développent aussi des outils directs (puisque ces données sont stockées telles quelles dans le DI) tel qu'un dictionnaire de synonymes ou un conjugueur ou des outils qui nécessitent des opérations à l'aide du DI comme :

- le *résumé automatique* qui est ici un outil qui relève les mots importants d'un texte c'est-à-dire les mots qui possèdent dans le DI un grand nombre de relations avec les autres ;

14. Communication personnelle.
15. http://www.memodata.com/2004/fr/dicologique/index.shtml

– la *classification automatique de documents* et en particulier le *routage de courriels* qui consiste à acheminer les courriers électroniques vers des boîtes prédéfinies en fonction de critères sémantiques ;
– la *comparaison textuelle* qui consiste à comparer les sèmes et les mots-sens contenus dans deux documents ;
– la *traduction automatique* grâce à des liens "*se traduit par*" entre mots-sens issus de langues différentes.

L'ensemble de ces applications est basé sur deux opérations ensemblistes réalisées sur les sèmes. La première, l'*activation*, permet d'évaluer la ressemblance de deux mots-sens c'est-à-dire les sèmes qu'ils ont en commun tandis que la seconde, le *calcul de proximité* permet d'évaluer, elle, à la fois les ressemblances et les différences.

On peut regretter que les informations relationnelles ne soient pas utilisées dans les tâches de désambiguïsation. Nous le verrons plus loin dans ce mémoire, l'utilisation d'informations de nature relationnelle permet de résoudre des ambiguïtés que l'usage d'informations mutuelles telles que le sont l'activation ou la proximité sémantique (ou dans le cas des vecteurs d'idées que nous aborderons dans le chapitre suivant, la distance thématique) soit ne peuvent résoudre soit peuvent largement aider à résoudre. Il s'agit ici d'un des résultats les plus importants de nos travaux.

2.4.4 Une première expérience utilisant des listes préétablies : les proto-vecteurs d'idées.

Au début des années 1990, Jacques Chauché, dans le but de réaliser un système de Traduction Automatique, propose de représenter le "sens" [16] 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 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 :

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

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 des vecteurs d'idées que nous utilisons dans cette thèse (d'où le nom de *proto-vecteurs d'idées*). Ces vecteurs d'idées sont basés, eux, sur le thésaurus Larousse, présenté dans la partie 2.4.6, qui offre le double avantage de présenter une liste de concepts présentés comme pouvant décrire l'ensemble des idées contenues en langue ainsi qu'une description des idées contenues dans quelques milliers d'items lexicaux.

2.4.5 Notre vision

Comme nous l'avons vu précédemment (cf. 2.4.2), 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. 2.4.4) 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.

Dans le cadre de ses recherches, l'équipe TALN du LIRMM utilise le formalisme des vecteurs d'idées. Ce modèle, proche dans sa conception de la théorie componentielle, est l'héritier direct de celui des proto-vecteurs d'idées présenté en 2.4.4. Ainsi, il postule que le sens des termes peut être calculé à partir d'un ensemble de concepts. Il va de soi que le problème posé par la recherche d'un ensemble de primitives sémantiques n'est pas l'objectif des recherches menées actuellement par l'équipe au sein du LIRMM. Nous préférons nous reposer 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.

2.4.6 Les thésaurus : un exemple, le Larousse

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, . . .* [17] ».

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.

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
- 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 18 présente un extrait de cette hiérarchie. La hiérarchie complète de [Larousse, 1992] se trouve en annexe **??**.

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.

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.

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. **??**). 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.

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

Références

[Bernard, 2000] Gilles BERNARD. « *Intelligence artificielle, linguistique expérimentale, cognition : Principes et mécanismes d'économie des représentations dans la modélisation de systèmes linguistiques* ». Mémoire d'Habilitation à Diriger des Recherches, Université Paris VIII, Paris, France, 2000. 2.3.2

[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. 2.2.3

[Chauché, 1990] Jacques CHAUCHÉ. « Détermination sémantique en analyse structurelle : une expérience basée sur une définition de distance ». *TAL Information*, pp 17–24, 1990. 2.4.4, 16

[Collins & Quillian, 1969] Alan COLLINS et Ross QUILLIAN. « Retrivial time from semantic memory ». *Verbal learning and verbal behaviour*, pp 240–247, 1969. 2.3.2, 2

[Collins & Quillian, 1970] Alan COLLINS et Ross QUILLIAN. « Does category size affect categorization time ? ». *Verbal learning and verbal behaviour*, pp 432–438, 1970. 2.3.2

[Conrad, 1972] Carol CONRAD. « Cognitive economy in semantic memory ». *Journal of Experimental Psychology*, pp 149–154, 1972. 2.3.2, 10

[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. 2.2.3

[Dutoit & Nugues, 2002] Dominique DUTOIT et Pierre NUGUES. « A Lexical Network and an Algorithm to Find Words from Definitions ». Dans les actes de Frank van HARMELEN, , *ECAI2002, Proceedings of the 15th European Conference on Artificial Intelligence*, pp 450–454, Lyon, July 21-26 2002. IOS Press, Amsterdam. 2.4.3, 17(a)

[Dutoit, 1992] Dominique DUTOIT. « A set-theoric approach to Lexical Semantics ». Dans les actes de *COLING'1992 : 14th International Conference on Computational Linguistics*, pp 539–545, Nantes, France, 1992. 2.4.3

[Dutoit, 2000] Dominique DUTOIT. « *Quelques opérations Sens -¿ texte et texte -¿ Sens utilisant une sémantique linguistique univerliste a priori* ». Thèse de doctorat, Université de Caen, Novembre 2000. 2.4.3, 2.4.3

[Éco, 1988] Umberto ÉCO. *Le signe : histoire et analyse d'un concept*. Livre de poche. Labor, 1988. 2.4.1

[Genest, 2000] David GENEST. « *Extension du modèle des graphes conceptuels pour la recherche d'informations* ». Thèse de doctorat, Université Montpellier II, 2000. 2.3.2

[Goodglass & Baker, 1976] Harold GOODGLASS et Errol BAKER. « Semantic field, naming and auditory comprehension in aphasy ». *Brain and Language*, pp 359–374, 1976. 2.3.2

[Grefenstette, 1994] Gregory GREFENSTETTE. « Corpus-derived first, second and third-order word affinities ». Dans les actes de *6th EURALEX*, Amsterdam, 1994. 2.2.3

[Greimas, 1986] Algirdas Julien GREIMAS. *Sémantique Structurale*. PUF, 1986. 2.4.1

[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. 2.2.1

[Hjelmlev, 1968] Louis HJELMLEV. *Prolégolème à une théorie du langage*. Éditions de minuit, 1968. 2.4.1

[Juola & Atkinson, 1971] James F. JUOLA et Richard C. ATKINSON. « Memory scanning for word versus categories ». *Journal of verbal learning and verbal behaviour*, pp 449–452, 1971. 2.3.2

[Kintsch, 2000] Walter KINTSCH. « Metaphor comprehension : A computational theory ». *Psychonomic Bulletin and Review*, 2000. 2.2.3

[Kirkpatrick, 1987] Betty KIRKPATRICK, . *Roget's Thesaurus of English Words and Phrases*. Penguin books, London, 1987. 2.4.6

[Kleiber, 1990] Georges KLEIBER. *La sémantique du prototype*. Presses Universitaires de France, 1990. 2.1.1

[Landauer & Freedman, 1968] Thomas LANDAUER et Jonathan FREEDMAN. « Information Retrivial from long term memory : category size and recognition time ». *Journal of verbal learning and verbal behaviour*, pp 291–331, 1968. 2.3.2

[Larousse, 1992] LAROUSSE, . *ThÉsaurus Larousse - des idÉes aux mots, des mots aux idées*. Larousse, 1992. 2.4.6, 2.4.6, 18, 19

[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. 2.2.3

[Le Ny, 1979] Jean-Francois LE NY. *La sémantique psychologique*. PUF, Paris, 1979. 2.4.1, 2.4.1

[Mangeot-Lerebours *et al.*, 2003] Mathieu MANGEOT-LEREBOURS, Gilles SÉRASSET, et Mathieu LAFOURCADE. « Construction collaborative d'une base lexicale multilingue : Le projet Papillon ». *TAL (Traitement Automatique des langues) : Les dictionnaires électroniques*, pp 151–176, 2003. 2.3.3, 2.3.3

[Mel'čuk, 1988] Igor MEL'ČUK. *Dictionnaire explicatif et combinatoire du francais contemporain*, volume 2. Les presses de L'université de Montréal, Montréal, 1988. 1, 2.3.1

[Mel'čuk *et al.*, 1995] Igor MEL'ČUK, André CLAS, et Alain POLGUÈRE. *Introduction à la lexicologie explicative et combinatoire*. Duculot, 1995. 2.3.1

[Nogier, 1991] Jean-Francois NOGIER. *Génération automatique de langage et graphes conceptuels*. Hermès, 1991. 2.3.2

[Nyckees, 1998] Vincent NYCKEES. *La sémantique*. Belin, 1998. 2.3.2, 2.4.1, 2.4.2, 2.4.2

[Otero & Bordag, 2010] Pablo Gamallo OTERO et Stefan BORDAG. « Is Singular Value Decomposition Useful for Word Similarity Extraction ? ». *Language Resources and Evaluation*, 2010. 2.2.3

[Polguère, 2003] Alain POLGUÈRE. *Lexicologie et sémantique lexicale*. Les Presses de l'Université de Montréal, 2003. 2.1.1, 2.3.1, 2.3.1

[Pottier, 1964] Bernard POTTIER. « Vers une sémantique moderne ». *Travaux de sémantique et de littérature*, pp 107–137, 1964. 2.4.1

[Quillian, 1968] Ross QUILLIAN. « *Semantic Informatic processing* », Chapitre Semantic memory, pp 227–270. MIT Press, 1968. 2.3.2

[Rastier, 1989] Francois RASTIER. *Sémantique et Recherche Cognitive*. Presses Universitaires de France, 1989. 2.4.1

[Rastier, 2004] Francois RASTIER. « Ontologie(s) ». *Revue des sciences et technologies de l'information*, pp 15–40, 2004. 2.3.2, 2.3.2

[Roget, 1852] Peter Mark ROGET. *Roget's Thesaurus of English Words and Phrases*. Longman, London, 1852. 2.4.6

[Sabah, 1988] Gérard SABAH. *L'intelligence artificielle et le langage*. Hermès, Paris, 1988. 2.3.2

[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. 2.4.2

[Salton & McGill, 1983] Gerard SALTON et Michael MCGILL. *Introduction to Modern Information Retrieval*. McGrawHill, New York, 1983. 2.2.2

[Salton, 1971] Gerard SALTON. « The SMART Retrieval System – Experiments in Automatic Document Processing ». 1971. 2.2.2

[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. 2.2.2

[Schank, 1972] Roger C. SCHANK. « Conceptual Dependency : A Theory of Natural Language Understanding ». *Cognitive Psychology*, pp pages 532–631, 1972. 2.4.2

[Schwab, 2001] Didier SCHWAB. « Vecteurs conceptuels et fonctions lexicales : application À l'antonymie ». Mémoire de dea, Université Montpellier II, LIRMM, Juillet 2001. 2.3.1

[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. 1, 2

[Schwab *et al.*, 2002] Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Antonymy and Conceptual Vectors ». Dans les actes de *COLING'2002 : 19th International Conference on Computational Linguistics*, volume 2/2, pp 904–910, Taipei, Taiwan, Août 2002. 2.3.1

[Schwab *et al.*, 2005] Didier SCHWAB, Mathieu LAFOURCADE, et Violaine PRINCE. « Extraction semi-supervisée de couples d'antonymes grâce À leur morphologie ». Dans les actes de *TALN'2005*, pp 73–82, Dourdan, Juin 2005. 2.3.1

[Sérasset & Mangeot, 2001] Gilles SÉRASSET et Mathieu MANGEOT. « Papillon lexical databases project : monolingual dictionaries and interlingual links ». Dans les actes de *NLPRS 2001*, pp 119–125, 2001. 2.3.3

[Sowa, 1984] John SOWA. *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley, Reading, 1984. 2.3.2

[Sowa, 2000] John SOWA. *Knowledge Representation : Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA, 2000. 2.3.2, 8

[Universalis, 1968] Encyclopædia UNIVERSALIS, . *Encyclopædia Universalis*, volume 17. Encyclopædia Universalis France, 1968. 2.4.4

[Wierzbicka, 1993] Anna WIERZBICKA. ≪ La quête des primitifs sémantiques : 1965-1992 ≫. *Langue francaise*, Mai 1993. 15

[Wilks, 1977] Yorick WILKS. ≪ Good and Bad Arguments About Semantic Primitives. ≫. *Communication and Cognition*, pp 181–221, 1977. 2.4.2

[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. 2.4.2

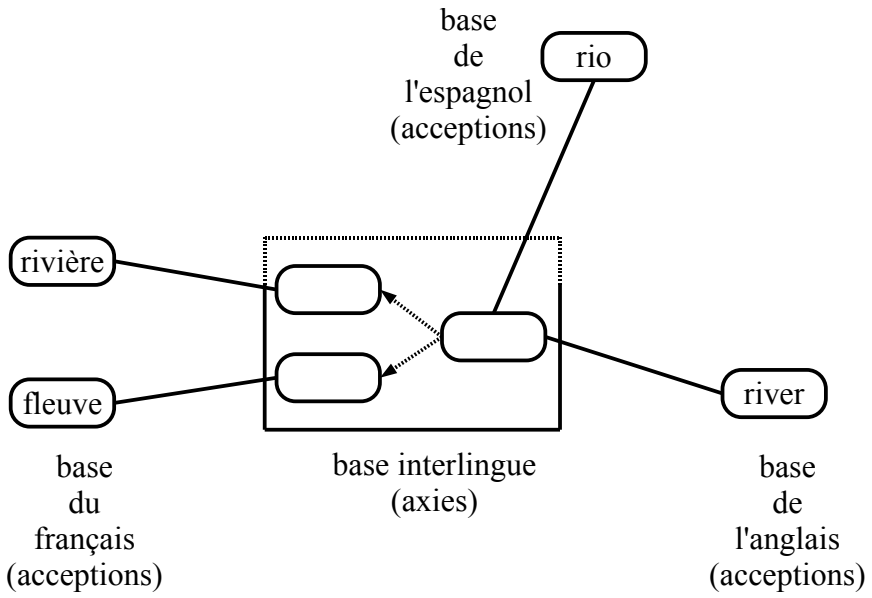


FIGURE 13 – Exemple de raffinement de sens.

	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	+	-	+	+	-	-	~	~	+

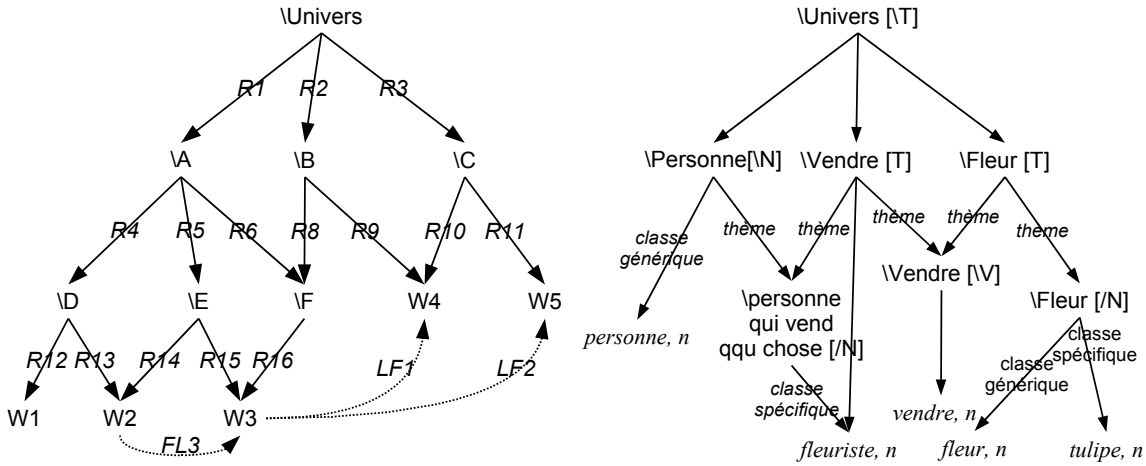
FIGURE 14 – Analyse sémique des véhicules selon Pottier

ˁ*je*ˀ, ˁ*tu*ˀ, ˁ*quelqu'un*ˀ, ˁ*quelque chose*ˀ, ˁ*on*ˀ (ou ˁ*les gens*ˀ),
ˁ*penser*ˀ, ˁ*savoir*ˀ, ˁ*dire*ˀ, ˁ*éprouver*ˀ, ˁ*vouloir*ˀ,
ˁ*ceci*ˀ, ˁ*le même*ˀ, ˁ*autre*ˀ, ˁ*un*ˀ, ˁ*deux*ˀ, ˁ*tous*ˀ, ˁ*beaucoup*ˀ,
ˁ*faire*ˀ, ˁ*arriver à/dans*ˀ,
ˁ*ne pas vouloir*ˀ (ou ˁ*non !*ˀ), ˁ*si*ˀ, ˁ*pouvoir*ˀ (ou ˁ*peut-être*ˀ), ˁ*comme*ˀ, ˁ*à cause de*ˀ, ˁ*près*ˀ (temps et lieu), ˁ*quand*ˀ (ou ˁ*temps*ˀ),
ˁ*où*ˀ (ou ˁ*endroit*ˀ), ˁ*après*ˀ (ou ˁ*avant*ˀ), ˁ*au-dessous de*ˀ (ou ˁ*au-dessus de*ˀ),
ˁ*avoir*ˀ (des parties), ˁ*différentes espèces*ˀ,
ˁ*bon*ˀ, ˁ*mauvais*ˀ, ˁ*grand*ˀ, ˁ*petit*ˀ

FIGURE 15 – Les 35 primitifs sémantiques d'Anna Wierzbicka [Wierzbicka, 1993]

classe	nombre	exemples
entités	19	*HUMANITÉ*, *SUBSTANCE*, *OBJET PHYSIQUE*
actions	34	*CAUSER*, *COULER*, *FRAPPER*, *ÊTRE*
cas	19	*VERS*, *DANS*, *AGENT*, *LIEU*
qualificatifs	16	*BON*, *CONTENANT*
indicateurs de type	2	*QUALITÉ*, *MANIÈRE*

FIGURE 16 – Quelques-uns des éléments primitifs de Wilks



(a) Le graphe de concepts, mots, relations et fonctions lexicales du dictionnaire intégral (extrait de [Dutoit & Nugues, 2002])

(b) Une partie du dictionnaire intégral

FIGURE 17 – Architecture du dictionnaire integral.

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

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

FIGURE 19 – Exemple de quelques termes extraits du thésaurus Larousse [Larousse, 1992]