

**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**Université Saad Dahleb de Blida**



**Faculté des sciences**

Département d'informatique

Mémoire présenté par

ALIOUI Youcef

AIT IGRINE Djamel

**En vue d'obtenir le diplôme de Master**

Domaine : Mathématique et Informatique

Filière : Informatique

Option : Ingénierie des logiciels

**Titre :**

**Une nouvelle approche de calcul de contenu informationnel pour la mesure de similarité sémantique utilisant Wordnet**

**Promoteur : M<sup>ER</sup> NEHEL Djilali**

**Soutenue le :**

**devant le jury composé de :**

-M

*Sarfira*

**Président**

-M

*Toubaline*

**Examineur**

-M

*Otteld Aissa*

**Examineur**

MA-004-221-1

## *Remerciements :*

*Tout d'abord, nous tenons à rendre grâce à DIEU tout puissant pour nous avoir donné le courage et la détermination nécessaire pour finaliser ce travail.*

*Nous tenons à remercier avec gratitude notre promoteur Djilali NAHAL qui a endossé son rôle de la meilleure façon qui soit. Nous retiendrons sa disponibilité, son aide indéfectible, ses conseils avisés et ses idées riches ainsi que sa sympathie et ses encouragements.*

*Par ailleurs ; nous rendons un vibrant hommage à l'ensemble du corps professoral du département d'informatique de l'université Saad BAHLAB de Blida qui ont contribué activement à notre formation pendant notre cursus universitaire.*

*Nous ne pouvons pas terminer sans remercier nos parents respectifs qui, par leur amour et leur soutien nous ont permis de mener à terme ce travail.*

*Merci*

## *Dédicaces*

*A ceux qui sont la cause de mon existence.*

*A ceux qui sont toujours là pour moi.*

*A ceux qui sont la lumière de mes yeux.*

*A mes parents, ma très chère maman, que dieux la garde pour mois.*

*A mon père, que dieux l'accueille dans son vaste paradis.*

*A mes très chers frères et sœurs.*

*A toute la famille ALIOUI.*

*A mes amis, et surtout Abdou, Laarbi.*

*A mon binôme Djamel.*

*A mon promoteur M. NAHAL.*

*A toute personne qui va lire ce mémoire.*

*Youcef*

## *Dédicaces*

*Je dédie ce modeste travail :*

*A mes très chères parents en témoignage de ma profonde  
gratitude et mon incontestable reconnaissance, leurs  
sacrifices, la confiance qu'ils m'accordent, leurs soutien  
permanent*

*Et tout l'amour dont ils m'entourent*

*Sans oublier mon grand-père que DIEU le garde pour moi*

*A mes très chers frère et sœurs, Abderrahmane, Sabrina, loubna, Manel, Asma*

*A mes amis Yacine, Nassime, Mohammed, Hmida, Amine, Alaa, Moumouh, Sofiane, Anes,  
Alayachi*

*A mon binôme Youcef*

*A mon promoteur M. NAHAL*

*Djamel*

## **Résumé**

Dans le cadre de la quantification de la similarité sémantique entre deux mots ou deux concepts, nous proposons dans ce travail une nouvelle méthode de calcul du contenu informationnel (CI) d'un concept en utilisant le thesaurus WordNet. Cette méthode n'a pas besoin des ressources externes mais elle exploite seulement la taxonomie « IS A » de Wordnet. Principalement, nous utilisons l'arbre des hyperonymes d'un concept donné avec la profondeur ainsi que l'arbre des hyponymes (précisément les feuilles) et la hauteur pour exprimer le contenu informationnel d'un concept. La nouvelle approche remédie à certaines limites dans d'autres idées de calcul du CI utilisant Wordnet. De plus cette méthode est couplée à des mesures de similarité sémantique entre concepts, elle a montré de bons résultats comparés aux jugements humains.

## **Mots clés**

Wordnet, similarité sémantique, contenu informationnel, concept.

## ABSTRACT

The semantic similarity quantification between words or concepts is an important topic. Therefore, we propose in this article a new method to compute concept informational contents (IC) using the WordNet thesaurus. This approach offers a thorough use of the relation specification/generalization “IS A” or hyperonym/hyponym. The new method does not need external resources but it exploits only taxonomy “IS A”. Mainly, we use subtree of the concept hyperonyms with the depth which is significant in hierarchical structure “IS A”, also, we use the hyponyms (leaves exactly) of a concept and its height in the taxonomy. The new approach cures some limiting in other CI computing idea using WordNet. Moreover, this method coupled to semantic similarity measurement, it showed good performances compared to human judgments.

**KEYWORDS:** informational content, semantic similarity, WordNet, concept.

## Liste des figures

- p.07 –figure 1.1** architecture générale de wordnet [Florentina Vasilescu, 2003]
- p.10- figure1.2** hiérarchies de noms selon wordnet 1.7.1 [Florentina Vasilescu, 2003]
- p.12-figure1.4** représentation des adjectifs descriptifs dans wordnet [[Florentina Vasilescu, 2003]
- p.14-figure1.4** représentation des synsets et des relations dans wordnet [Siddharth Patwardhan, 2003]
- p.15-figure1.5** le réseau « is-a » de wordnet [Siddharth Patwardhan, 2003]
- p.15-figure1.6** interface de wordnet 2.1
- p.21-figure2.1** les relations conceptuelles utilisées par (Rada et al, 1989) [Zargayouna, 2005]
- p.25-figure2.2** extrait N° 01 de wordnet présenté dans (Lin, 1998) avec les probabilités correspondantes aux différents concepts [Zargayouna, 2005]
- p.26-figure2.3** extrait N° 02 de wordnet [Alexander Budanitsky, 1999]
- p33-figure2.4** représentation graphique des résultats obtenus par Rubinstein et Goodenogh [Alexander Budanitsky, 1999]
- p37-figure2.5** représentation graphique des résultats obtenus par Miller et Charles [Alexander Budanitsky, 1999]
- p38-figure2.6** comparaison de la mesure de Resnik avec le jugement humain
- p38-figure2.7** comparaison de la mesure de Jiang et Conrath avec le jugement humain
- p39-figure2.8** comparaison de la mesure de Lin avec le jugement humain
- p.44-figure3.1** exemple de calcul de CI pour quelques concepts par la méthode de Sebti [HADJ Taieb et al. 2012]
- p.46-figure3.2** extrait N° 03 de wordnet 2.1
- p.51-figure4.1** Les sous arbre d’hypernymes de (atropine) et de (obidoxime chloride), extrait de wordnet 2.0, [HADJ Taieb et al. 2012].
- p.52-figure4.2** Le sous arbre hyponymes de (stuff) Extrait N° 04 de wordnet 2.1
- p.53-figure4.3** représentation de l’anti racine dans un extrait de wordnet

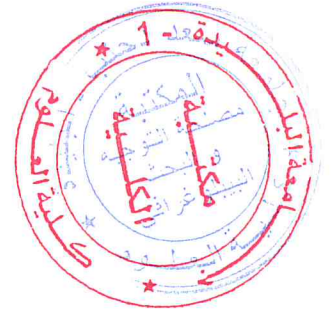
<b>p.60-figure4.4</b>	exemple de calcul de CI.
<b>p.66-figure4.5</b>	extrait N° 05 de wordnet 2.1
<b>p.67-figure3.2</b>	extrait de wordnet 2.1
<b>p.69-figure 4.6</b>	Comparaison de notre méthode avec celle de Lin et le jugement humain



## Liste des tableaux

- p.08-tableau1.1** suffixes et terminaisons par catégorie grammaticale [Florentina Vasilescu, 2003]
- p.08-tableau1.2** exemples d'exceptions par catégorie grammaticale [Florentina Vasilescu, 2003]
- p.13-tableau1.3** les relations sémantiques de wordnet [Siddharth Patwardhan]
- p.16-tableau1.4** nombre de mots, synsets et sens dans wordnet [Florentina Vasilescu, 2003]
- p.16-tableau1.5** répartition des mots dans wordnet en monosémiques et polysémiques [Florentina Vasilescu, 2003]
- p.30-tableau2.1** jugement humain et similarité obtenue par différentes approches : par Rubinstein et Goodenough [Alexander Budanitsky, 1999]
- p.32-tableau2.2** jugement humain et similarité obtenue par différentes approches : par Miller et Charles [Alexander Budanitsky, 1999]
- p.39-tableau2.3** coefficients de corrélation entre le jugement humain et les différentes approches [Alexander Budanitsky, 1999]
- p.56-tableau4.1** comparaison entre les différents modèles de CI
- p.57-tableau4.2** les valeurs de CI calculés par différentes approches de [Ferid et khaddem, 2014].
- p.59-tableau4.3** les valeurs de CI calculés par notre approche
- p.61-tableau4.4** les valeurs des similarités sémantiques pour chaque couple dans la liste des jugements humains de [Miller & Charles, 1991] calculés par notre les différentes approches
- p.64-tableau4.5** les valeurs des similarités sémantiques pour chaque couple dans la liste des jugements humains de [Miller & Charles, 1991] calculés par notre formule de CI couplé avec Lin.
- p.66-tableau4.6** les contenus informationnels calculés avec notre formule
- p.66-tableau4.7** les CI des concepts obidoxime chloride et atropine
- p.67-tableau4.8** les CI des concepts discutés dans la figure 3.2
- p.68-tableau4.9** comparaison des résultats obtenus par notre formule avec les différentes formules de similarité

**p.68-tableau4.10** Les coefficients de corrélation entre les jugements humains de similarité (Miller & Charles et Rubenstein & Goodenough) et les mesures de similarité proposées.



## Table des matières

### Introduction générale

1. Contexte générale.....	1
2. Problématique.....	2
3. Objectifs.....	2
4. Organisation du mémoire.....	3

## Partie I Etat de l'Art

### Chapitre 01 présentation de Wordnet

1. Introduction .....	5
2. C'est quoi wordnet. ....	5
3. Description de wordnet .....	5
4. Architecture de wordnet.....	6
5. Formes des mots .....	7
6. Les noms .....	9
7. Les verbes .....	10
8. Les adjectifs .....	11
9. Les adverbes.....	12
10. Relations .....	13
10.1 relations sémantiques (entre synsets) .....	13
11. Interface de wordnet 2.1 .....	15
12. Quelques données statistiques .....	16
13. Mesure de similarité .....	17
14. Conclusion.....	17

## **Chapitre 02 la similarité sémantique**

<b>1. Introduction.....</b>	<b>19</b>
<b>2. Définitions et concepts de base.....</b>	<b>19</b>
<b>2.1 Ontologie.....</b>	<b>19</b>
<b>2.2 Concept.....</b>	<b>20</b>
<b>3. Similarité sémantique.....</b>	<b>20</b>
<b>4. Mesure de similarité.....</b>	<b>20</b>
<b>5. Classification des approches de mesure de similarité.....</b>	<b>21</b>
<b>5.1 Approches basées sur les arcs.....</b>	<b>21</b>
<b>5.2 Approches basées sur les nœuds.....</b>	<b>24</b>
<b>5.2.1 La similarité de Resnik.....</b>	<b>25</b>
<b>5.2.2 La similarité de Jiang et Conrath.....</b>	<b>26</b>
<b>5.2.3 La similarité de Lin.....</b>	<b>27</b>
<b>6. Comparaison avec le jugement humain.....</b>	<b>28</b>
<b>7. Conclusion.....</b>	<b>40</b>

## **Chapitre 03 Les approches de calcul de contenu informationnel**

<b>1. Introduction.....</b>	<b>42</b>
<b>2. Les approches de calcul de contenu informationnel.....</b>	<b>42</b>
<b>2.1 l'utilisation d'une ressource externe.....</b>	<b>42</b>
<b>2.2 l'utilisation de la taxonomie « is-a » de wordnet.....</b>	<b>43</b>
<b>2.2.1 IC basé sur les hyponymes.....</b>	<b>43</b>
<b>2.2.2 IC de Zhou.....</b>	<b>43</b>
<b>2.2.3 IC de Sanchez.....</b>	<b>43</b>
<b>2.2.4 IC de Sebti.....</b>	<b>44</b>
<b>2.2.5 IC de HADJ Taieb.....</b>	<b>45</b>

3. Les limites des approches citées.....	46
4. Conclusion.....	47
<b>Partie II mise en oeuvre</b>	
<b>Chapitre 04 Nouvelle approche de calcul de CI</b>	
1. Introduction .....	50
2. Notre approche.....	50
2.1. Les principes du calcul du CI .....	50
2.1.1 La relation Hyperonyme/Hyponyme « IS-A » .....	50
2.1.2 La signification de la profondeur dans la taxonomie « IS A » .....	53
2.1.3 La hauteur .....	53
3. Nouvelle méthode de calcul de CI .....	54
4. Comparaison avec d'autres approches de calcul de CI .....	55
5. Expérimentation .....	65
5.1. Balayage des limites .....	65
5.2 Comparaison avec des estimations humaines de similitude .....	67
6. Implémentation .....	59
6.1 Langage utilisé .....	69
6.2 Description des fonctions de notre application.....	70
7. Conclusion .....	71
8. Conclusion générale.....	73

***Introduction  
générale***

## Introduction générale

### 1. contexte générale :

La similarité sémantique, c'est-à-dire l'appréhension de la liaison entre deux concepts, est une capacité de l'homme que les machines ne savent que très mal reproduire. Ainsi, pour un humain, il est évident que les concepts de *crayon* et de *papier* sont liés, beaucoup plus que ceux de *parapluie* et *fer à repasser* en tout cas. Mais il est très difficile de le formaliser car rien, en surface, ne permet de le décider. Pour ce faire, il faut utiliser des ressources sémantiques : les ontologies, c'est-à-dire des bases de connaissances. Elles seules permettent de montrer les liens (hypéronymie, antonymie, etc.) entre des concepts.

Dans les dictionnaires explicatifs, on trouve très souvent des synonymes ou des antonymes pour un mot quelconque. Par exemple si on dit : Est-ce que réacteur relié avec avion ? Est-ce que mouton a une relation avec mammifère ? Est-ce que médecin relié avec hôpital, et s'il est, est-ce que cette relation est plus forte que celle de médecin avec chirurgien ?

La réponse à ces questions implique la notion de *similarité des mots* qui peut se représenter par une valeur scalaire qui définit comment deux mots se relient. Plus concrètement, si la similarité entre le mot *m1* et le mot *m2* est quantifiée par *sim (m1, m2)*, on peut dire que «médecin» est plus proche de «hôpital» que de «chirurgien» si on a :

*sim (hôpital, médecin) > sim (hôpital, chirurgien)* et vice versa.

L'utilisation de la similarité sémantique est une problématique qui ne cesse de prendre de l'essor d'une année à une autre dans le traitement automatique de langages naturels (TALN) et dans la recherche d'information (RI). La recherche d'information (RI) est un champ d'investigation évident pour la similarité sémantique. En effet, les problèmes de polysémie et de synonymie de nos langues génèrent des ambiguïtés dans les recherches. [FURNAS, 1987] par exemple montre les difficultés de consensus dans le choix de termes pour les indexations et pour les recherches.

La probabilité que le même terme soit choisi par deux individus pour décrire une entité quelconque est bien inférieure à 20% [FURNAS, 1987]. Et même si on utilise un thésaurus contraint, avec une liste de mots acceptés (par exemple, pour des formulaires de saisie avec des codes ou des intitulés prédéfinis), la probabilité ne dépasse pas 70% [FURNAS, 1987]. C'est pourquoi il est nécessaire de passer à un niveau sémantique, pour éviter ces problèmes de syntaxe et de comparaison terme-à-terme. Ainsi, en utilisant une ontologie, il doit être possible de savoir que l'« avocat » dont parle ce document est un fruit vert et que celui de cette requête est un défenseur, ou que « chat » et « matou » réfèrent tous les deux au même concept. L'ontologie qui a le plus été utilisée par ces travaux est WordNet; un thésaurus en langue anglaise assez étoffé. Même s'il s'agit d'une ontologie à *minima*, « légère », elle est plus complète que beaucoup d'autres et simple à utiliser.

Les racines de la similarité remontent à [Quillian, 1968] et [Collins, 1975] qui ont pensé à numériser le sens ou la sémantique. Cela permet de dépasser de nombreux problèmes liés aux comparaisons terme à terme. De nombreuses applications comme la désambiguïsation du sens des mots [Resnik, 1999] vu le problème de polysémie, extraction et la recherche d'information, détection et correction des lapsus [Budanitsky, 2001], segmentation du texte [Kozima, 1994], recherche d'images [Smeulders *et al.*, 2000], l'indexation sémantique

[Zargayouna, 2004a et 2004b] et le résumé automatique [Lin, 2003] utilisent désormais la similarité sémantique dans le but d'obtenir de meilleurs résultats.

Un problème récurrent de ces applications est la mesure de proximité entre concepts. Elle a été étudiée par de nombreux auteurs, et deux grandes approches basées sur les arcs [Hirst-St-Onge, 1998], [Leacock et Chodorow, 1998], [Wu et Palmer, 1994] en exploitant la structure de la ressource sémantique utilisée, et les approches utilisant le contenu informationnel [Resnik, 1995], [Jiang et Contrath, 1997] et [Lin, 1998].

Le contenu informationnel est la mesure de spécificité d'un concept ; les grandes valeurs de contenu informationnel (CI) sont associées avec les concepts les plus spécifiés (ex : mouton), tandis que, ceux avec des petites valeurs sont plus générale (ex : mammifère). Le calcul de CI est basé sur la fréquence d'apparition des mots dans un corpus. Dans Wordnet, la fréquence associée à un concept est incrémentée à chaque fois que le concept apparaitre. Ceci est important car à chaque fois qu'un concept générale apparaitre ceci implique l'apparition des concepts spécifique.

## **2. Problématique :**

Nous pouvons voir que l'IC joue un rôle important dans des mesures sémantiques de similarité. Comment acquérir une valeur appropriée d'IC ?

Il existe plusieurs approches pour le calcul de cette valeur et deux grandes approches sont distinguées la première utilise une analyse statistique d'un corpus pour attribuer des probabilités à chaque concept. Tandis la deuxième méthode elle exploite la structure de WordNet sans faire recours à une ressource externe. Elle inclut le modèle de CI basé sur les hyponymes. Le modèles basé les feuilles basée, le modèle basé sur les relations et le modèle qui prend en considération la structure détaillée de l'ontologie Wordnet.

## **3. Objectifs :**

Notre objective c'est de trouver une approche de calcul du contenu informationnel qui donne de meilleur résultat adapté d'IC afin de mesurer la similarité sémantique entre n'importe qu'elle couple de deux concepts.

Nous évaluons cette approche qui calculent cette valeur, notre but est fortement efficace distinguer les différentes concepts et obtenir la valeur d'IC la plus adéquate.

Nous visons que cette approche couplée avec quelle mesure de similarité sémantique et montrer qu'elle a des bonnes corrélations comparés avec des jugements humains.

## **4. organisation de mémoire :**

Ce mémoire est organisé en deux parties :

La première est composée de trois chapitres qui définissent les concepts de base de la similarité sémantique ainsi que les différentes approches de mesures.



Nous présentons dans le chapitre 01 le WordNet une présentation détaillée. Dans le chapitre 02 nous présentons la similarité sémantique et les différentes approches de mesure ; celles qui se basent sur les corpus, et celles qui se basent sur WordNet. Le troisième chapitre est consacré pour la présentation des différentes approches de calcul de contenu informationnel avec présentation de quelques limites de ces approches.

La partie II présente notre travail, elle contient un seul chapitre. Dans le chapitre 04 nous présentons notre nouvelle approche de calcul de contenu informationnel, ainsi que sa performance et son comparaison avec le jugement humain.

Nous terminons avec une conclusion générale qui synthétisera notre travail et par quelques perspectives.

*Partie I Etat de l'Art*

*Chapitre 01 présentation de*

*Wordnet*

## 1. Introduction :

WordNet est une ressource lexicale de large couverture, développée depuis plus de 20 ans pour la langue anglaise. Elle est utilisable librement, y compris pour un usage commercial, ce qui en a favorisé une diffusion très large. Plusieurs autres ressources linguistiques ont été constituées (manuellement ou automatiquement) à partir de, en extension à, ou en complément à WordNet.

Des programmes issus du monde de l'Intelligence Artificielle ont également établi des passerelles avec WordNet. L'ensemble constitue un « écosystème » complet couvrant des aspects lexicaux, syntaxiques et sémantiques. Combinées, ces ressources fournissent un point de départ intéressant pour des développements sémantiques en TAL ou dans le cadre du Web sémantique, tels que la recherche d'information.

## 2. C'est quoi wordnet ? :

WordNet est une base de données lexicale développée depuis 1985 par des linguistes du laboratoire des sciences cognitives de l'université de Princeton. C'est un réseau sémantique de la langue anglaise, qui se fonde sur une théorie psychologique du langage. La première version diffusée remonte à juin 1991. [W01]

Son but est de répertorier, classer et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. Le système se présente sous la forme d'une base de données électronique qu'on peut télécharger sur un système local. Des interfaces de programmation sont disponibles pour de nombreux langages.

## 3. Description de WordNet :

Dans WordNet l'information est structurée autour de groupes de synonymes nommés *synsets*. Un synset comporte la liste des synonymes exprimant un même concept, la définition du concept, éventuellement des exemples d'usage, et les relations de ce concept avec d'autres concepts. Les relations entre les synsets sont de deux types :

- **lexical** - les relations sont exprimées à partir des formes des mots. On trouve les relations suivantes :

- antonymie – deux mots sont antonymes s'ils comportent des sens opposés l'un à l'autre (par exemple, *skilled/unskilled*, *animate/inanimate*, *alignment/nonalignment*, *live\_in*, *sleep\_in/live\_out*, *sleep\_out*). Dans WordNet l'antonymie est considérée plutôt comme une relation entre les formes des mots parce que, dans beaucoup de cas, elle suppose l'ajout d'un préfixe (un-, in-, non-) ou d'un suffixe (-less) ou une préférence pour une certaine forme lexicale (dans l'usage fréquent, *light* est un antonyme de *heavy*, mais pas un antonyme de *ponderous* qui est pourtant un synonyme de *heavy*);
- pertainymie – relation appliquée aux adjectifs relationnels de WordNet pour indiquer le nom de provenance (par exemple, *academic* est relié par ce type de relation au synset *academia*, *academe*);

- participe – un adjectif est en relation de participe avec le verbe d'où il dérive (WordNet relie, par exemple, l'adjectif *applied* au synset *use, utilize, utilise, apply, employ*);
- voir aussi sont des renvois qui apportent des informations supplémentaires à la description d'un synset (par exemple, le synset *drink, imbibe* comporte une relation vers le synset *drain\_the\_cup, drink\_up = drink to the last drop*);
- dérivé d'un adjectif – relation qui relie un adverbe et l'adjectif d'où il dérive (*negatively/negative*);

- **sémantique** – les relations sont établies à partir des sens des mots. On trouve les relations :

- hyperonymie/hyponymie – on entend par hyperonyme un terme dont le sens inclut d'autres termes, qui sont ses hyponymes. Par exemple, le synset *canine, canid* est l'hyperonyme de *dog, domestic\_dog, Canis\_familiaris* qui est lui-même l'hyperonyme de *working\_dog*, hyperonyme de *Eskimo\_dog, husky*;
- méronyme/holonyme – relation de type partie/tout, membre/groupe établie entre deux synsets nominaux, par exemple : *hound, hound\_dog* est un membre méronyme du holonyme *pack*;
- engendrement (entailment) – relation qui suppose l'enchaînement logique entre deux synsets verbaux, comme par exemple, *wear, have\_on* suppose logiquement *dress, get\_dressed*;
- cause – relation qui exprime l'aspect causatif / résultatif entre deux synsets verbaux (*show/see, produce, bring\_on, bring\_out/appear*);
- attribut – relation qui relie les adjectifs descriptifs de WordNet avec les noms qu'ils peuvent déterminer (par exemple, l'adjectif *short* est relié par une relation d'attribut au synset *duration, length*).

WordNet ne traite pas de relations de type syntagmatique, i.e. de relations établies entre les mots appartenant à des catégories syntaxiques différentes dans le cadre de la phrase, les 4 catégories fondamentales (nom, verbe, adjectif et adverbe) étant traitées séparément (sauf les relations de pertainymie, participe, dérivé décrites plus haut).

#### 4. Architecture de WordNet

Le système WordNet comporte quatre parties [Tengi, 1998] : les fichiers sources écrits par les lexicographes, le logiciel (Grinder) pour la conversion de ces fichiers dans la base lexicale proprement-dite, la base de données lexicale et des logiciels d'interface entre l'utilisateur et la base de données lexicale (voir Figure 1.1) (source [Florentina Vasilescu, 2003]):

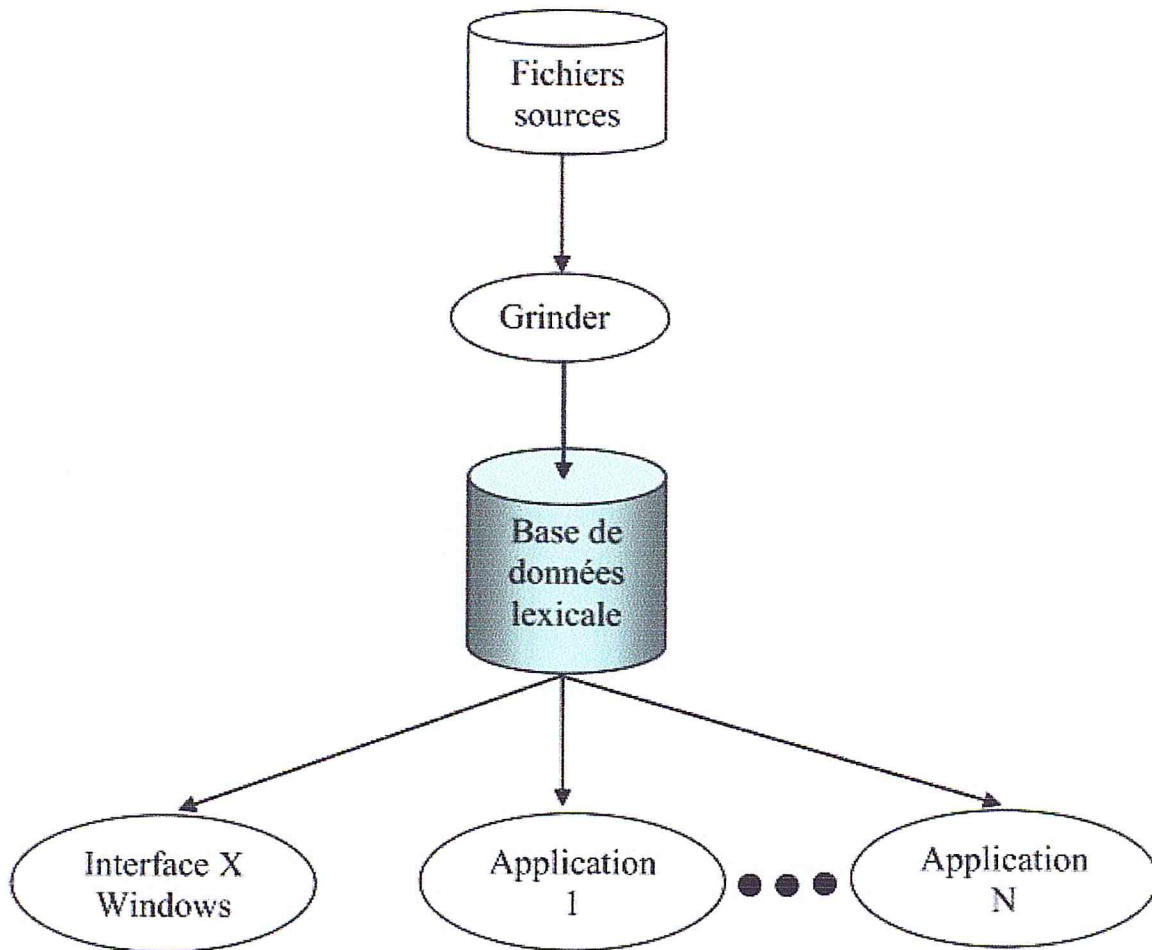


Figure 1.1 Architecture générale de WordNet

Les fichiers sources, créés par les lexicographes, sont le produit d'une analyse minutieuse des relations de type lexical et sémantique entre les mots ainsi que d'une étude sur la fréquence des sens, à partir de corpus sémantiquement annotés.

Le logiciel Grinder a pour but de compiler les fichiers des lexicographes dans un format approprié au traitement automatique, facilitant aussi la détection des erreurs structurales, la construction des pointeurs sémantiques et lexicaux, l'assignation des nombres représentant la fréquence d'usage de chaque sens. La sortie du programme comporte les fichiers de données et d'index, en format ASCII, qui constitue le cœur de WordNet, à savoir la base de données. Le format ASCII de la base permet à un utilisateur d'aller chercher facilement l'information dont il a besoin, par l'intermédiaire de ses propres programmes. C'est de cette façon que nous utilisons WordNet dans ce travail. Il est également possible d'accéder à l'information via une interface dédiée.

##### 5. Forme des mots :

Les mots dans WordNet sont représentés par leur forme canonique [Florentina Vasilescu, 2003] (de base) :

Singulier pour les noms (exemple *book, table*); infinitif court pour les verbes (exemple *be, read*); degré positif pour les adjectifs (exemple *good, lovely*). Les mots composés, faisant référence à un même concept, sont encodés par une succession de mots individuels, reliés par *Under score* (exemple *fontain\_pen, take\_for\_granted*).

Pour un traitement plus facile des textes en langage naturel, WordNet inclut aussi des modules de programmes à fonctions morphologiques et des fichiers d'exceptions, permettant d'obtenir la forme de base à partir de la forme instanciées des mots. Les tableaux ci-dessous indiquent le jeu des suffixes et des terminaisons qui par leur suppression et/ou ajout mènent à la forme de base, ainsi que des exemples extraits des fichiers d'exceptions. (Source [Florentina Vasilescu, 2003]):

Noms		Verbes		Adjectifs	
<i>Suffixe</i>	<i>Terminaison</i>	<i>Suffixe</i>	<i>Terminaison</i>	<i>Suffixe</i>	<i>Terminaison</i>
s		s		er	
ses	s	ies	y	est	
xes	x	es	e	er	e
zes	z	es		est	e
ches	ch	ed	e		
shes	sh	ed			
		ing	e		
		ing			

Tableau 1.1 Suffixes et terminaisons par catégorie grammaticale

Noms		Verbes		Adjectifs		Adverbes	
<i>Forme instanciée</i>	<i>Forme de base</i>	<i>Forme instanciée</i>	<i>Forme de base</i>	<i>Forme instanciée</i>	<i>Forme de base</i>	<i>Forme instanciée</i>	<i>Forme de base</i>
activities	activity	accompanied accompanies accompanying	accompany	angrier angriest	angry	best better	well
halves	half	overrunning overruns	overrun	madder maddest	mad	deeper deeper	deeply
men	man	prying	pry	uglier ugliest	ugly	farther further	far
sports_arenas	sports_arena	shook_hands	shake_hands	wetter wettest	wet	harder hardest	hard

Tableau 1.2 Exemples d'exceptions par catégorie grammaticale

## 6. Les noms

Les noms dans WordNet et [G.A. Miller 1998] sont hiérarchisés en plusieurs niveaux de généralité/spécificité par l'intermédiaire des relations d'hyperonymie et d'hyponymie entre synsets. Par exemple, la séquence hyperonymique {*robin, redbreast*} @-> {*bird*} @-> {*animal, animate\_being*} @-> {*organism, life\_form, living\_thing*} décrit une hiérarchie de termes à partir des plus spécifiques vers les plus généraux qui désigne alors une relation de type IS-A ou IS-A-KIND-OF. Un parcours inverse peut être aussi tracé, en utilisant les relations d'hyponymie ou de spécialisation entre les concepts.

Selon ce principe, les noms de WordNet sont divisés en plusieurs hiérarchies, chaque hiérarchie comportant un élément de départ unique ou une racine dont les traits sont hérités par tous les hyponymes.

Entre ces hiérarchies existent certaines références croisées mais, en général, elles couvrent des domaines conceptuellement et lexicalement distincts, à partir de 11 mots-racines (voir Figure 1.2) (source [Florentina Vasilescu, 2003]) :

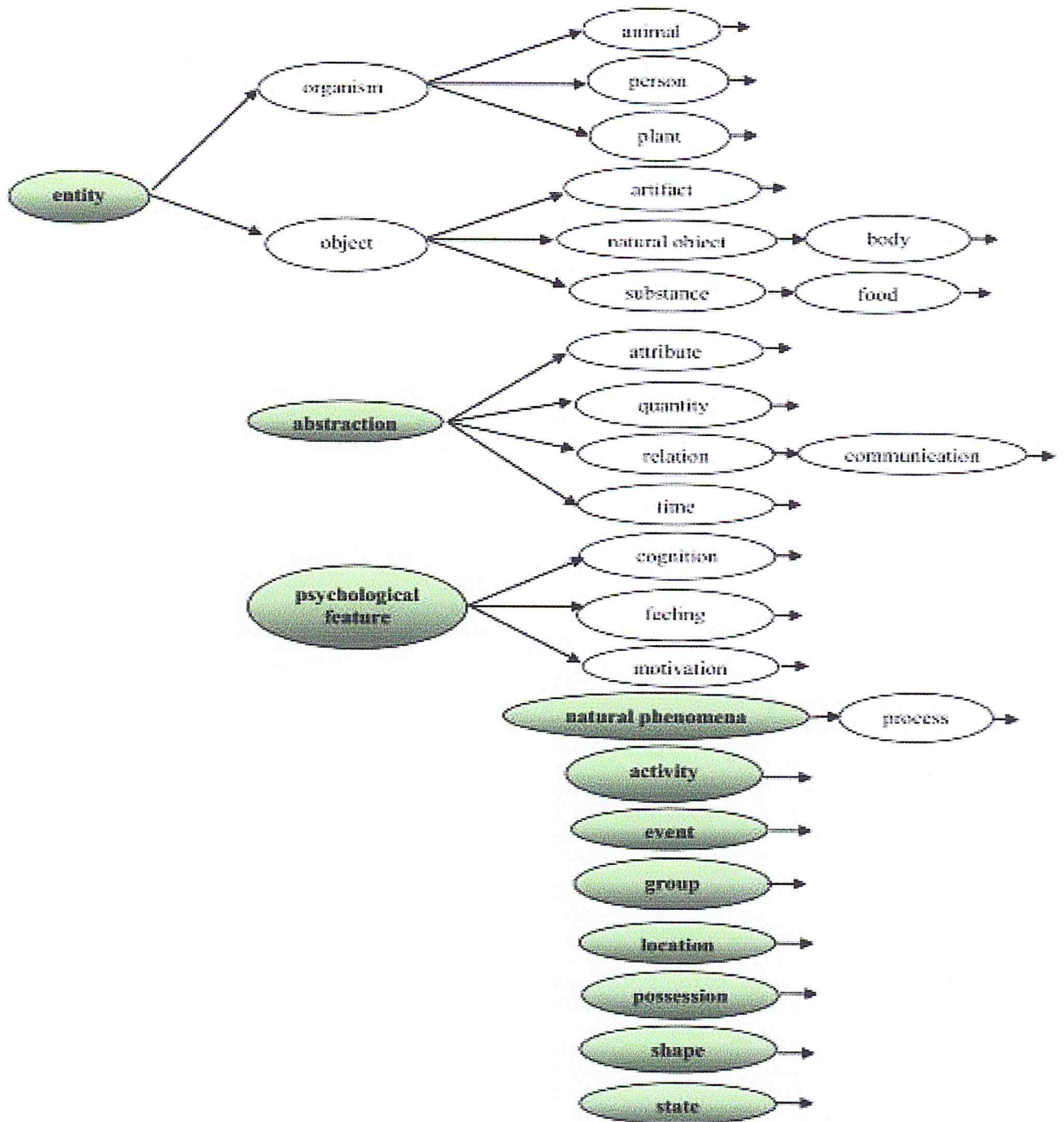


Figure 1.2 Hiérarchies de noms selon WordNet 1.7.1

## 7. Les verbes

Selon le même principe hiérarchique que pour les noms, les verbes dans WordNet [Fellbaum, 1998] sont divisés en plusieurs champs sémantiques regroupant des verbes de mouvement, perception, contact, communication, compétition, changement, cognition, consommation, création, émotion, possession, soin et fonctions du corps, ou encore comportement social et interaction. En plus de ces groupes il y a aussi une catégorie hétérogène englobant les auxiliaires, les verbes de contrôle (*like, want, fail, prevail, succeed*),



l'aspectuel *begin* et des concepts élaborés du verbe *be* de type ressemble, *belong* et *suffice*. Certains champs sémantiques sont représentés par plusieurs arbres indépendants, comme par exemple les verbes de mouvement qui comportent deux racines exprimant deux concepts distincts (*move1* – mouvement de translation, *move2* – mouvement sans déplacement) et les verbes de communication dissociés en deux branches indépendantes, verbes de communication verbale et non verbale.

A la différence des noms où la hiérarchisation est réalisée par l'intermédiaire des relations de type IS-A ou IS-A-KIND-OF, pour les verbes une telle sorte de classification semble inadéquate sans une transformation nominale préalable. La relation utilisée dans WordNet pour la catégorisation des verbes est la troponymie qui peut être exprimée par la phrase :

*V1 est un troponyme de V2 si V1 est V2 d'une certaine manière*

Par exemple, les troponymes du verbe *fight* dénotent l'occasion ou la forme de l'action (*battle, war, tourney, duel, feud*), les troponymes d'un verbe de communication encodent l'intention, la motivation du locuteur (*examine, confess, preach*) ou le medium de communication (*fax, e-mail, phone, telex*), etc. La troponymie inclut aussi l'engendrement (entailment) et la coexistence temporelle, i.e. V1 est un troponyme d'un verbe plus général V2 si V1 suppose implicitement V2 et les actions de V1 et de V2 se déroulent en même temps. Par exemple, *march* est un troponyme de *walk* parce que *marching* suppose aussi *walking* et ils sont nécessairement coexistants du point de vue temporel. Par contre, la paire *snore / sleep* n'exprime pas une relation troponymique parce que *snore* suppose *sleep* mais leurs actions ne sont pas nécessairement temporellement coexistantes.

Comme les noms et les adjectifs, les verbes dans WordNet sont réunis en groupes de synonymes. Pourtant, si on prend en compte la définition exacte de la synonymie qui caractérise des mots interchangeable dans la plupart des contextes, en anglais, le nombre de verbes qui sont de vrais synonymes (*shut / close*) est assez réduit. Par conséquent, dans la majorité des cas, les synsets relient des verbes qui expriment le même concept mais qui ne sont pas substituables dans un contexte ou un registre linguistique donné (*begin / commence, end / terminate, rise / ascend, behead / decapitate etc.*). WordNet fait ces distinctions d'usage par l'intermédiaire des définitions et des exemples attachés à chaque synset.

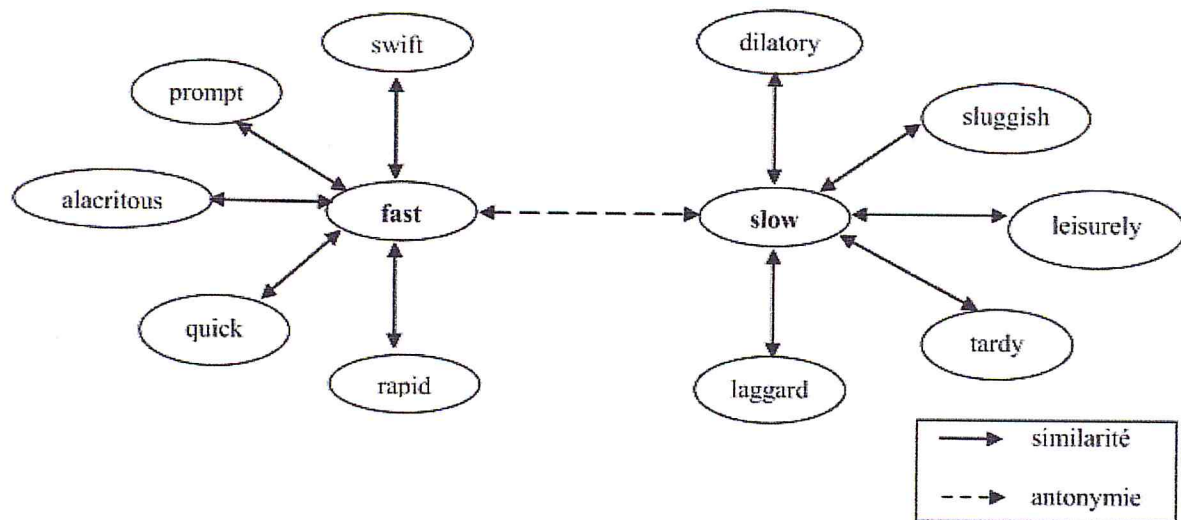
## 8. Les adjectifs

Les adjectifs dans WordNet [K.j Miller, 1998] sont sous-catégorisés en adjectifs descriptifs et relationnels. Selon le type d'adjectif, il y a une représentation différente dans WordNet. A la différence des noms et des verbes, il n'y a pas de hiérarchie dans WordNet pour la représentation des adjectifs.

Les adjectifs descriptifs de type *beautiful, interesting, possible, married* sont implicitement reliés à la notion d'attribut, i.e. dire que x est Adj suppose l'existence d'un attribut A tel que  $A(x) = \text{Adj}$ . Par exemple, la phrase *The package is heavy* implique l'attribut *WEIGHT* tel que :  $\text{WEIGHT}(\text{package}) = \text{heavy}$ . Les antonymes *heavy / light* peuvent être considérés ainsi comme des valeurs possibles de l'attribut *WEIGHT*. Les adjectifs descriptifs sont reliés par des relations de type *attribut* aux noms qu'ils peuvent modifier (par exemple, *heavy* est relié au nom *weight*).

Les adjectifs descriptifs sont généralement organisés en clusters de synonymes par l'intermédiaire de la relation d'antonymie entre des adjectifs dits de têtes (*head synsets*) autour desquels peuvent apparaître des adjectifs satellites reliés aux adjectifs-têtes par une relation de similarité.

La figure 1.3 montre les relations d'antonymie directe entre les adjectifs-têtes *fast* et *slow* ainsi que les relations de similarité entre les adjectifs satellites et têtes (*rapid* est similaire à *fast*, *tardy* à *slow*) et d'antonymie indirecte entre les clusters opposés (*laggard* est un antonyme indirecte de *rapid* ou de *fast*, *quick* de *leisurely* etc.).



Figur1.3 Représentation des adjectifs descriptifs dans WordNet (source [Florentina Vasilescu, 2003])

Une sous-classe des adjectifs descriptifs regroupe les adjectifs qui sont des formes participiales des verbes avec lesquels ils sont reliés par des relations de type participe (*breaking* est relié au synset *break*). Ce type d'adjectif ne comporte pas d'antonymes.

Les adjectifs relationnels sont des adjectifs dérivés de noms, comme par exemple *electrical* est un dérivé du nom *electricity*. Cette relation implique un lien de type sémantique et morphologique avec le nom d'origine. Pourtant le lien morphologique n'est pas toujours direct, comme dans le cas de l'adjectif *dental* relié au synset *tooth* via le mot latin *dens*.

A la différence des adjectifs descriptifs, les adjectifs relationnels ne suppose pas une relation d'attribut avec le nom déterminé et n'acceptent pas de degrés de comparaison (les expressions telles que : \* *the hygiene is dental* ou \* *the very electrical field* ne sont pas acceptables).

## 9. Les adverbes

L'organisation sémantique des adverbes dans WordNet est assez simple [K.j Miller, 1998], car il n'y a pas de hiérarchies ou de clusters comme pour les autres catégories grammaticales. Chaque synset peut comporter un adverbe (éventuellement ses synonymes et/ou antonymes), un pointeur vers l'adjectif à partir duquel l'adverbe est dérivé (s'il y en a un, comme par

*exemple quick-quickly, extreme-extremely*) et la partie de définition et d'exemple d'usage caractérisant le synset en question.

## 10. Relations

### 10.1 Relations sémantique (entre synsets)

Le tableau suivant présente les relations sémantiques de Wordnet par catégorie (Source: [Siddharth Patwardhan, 2003])

Relation	Description	Example
Hypernym	is a generalization of	<i>furniture</i> is a hypernym of <i>chair</i>
Hyponym	is a kind of	<i>chair</i> is a hyponym of <i>furniture</i>
Troponym	is a way to	<i>amble</i> is a troponym of <i>walk</i>
Meronym	is part/substance/member of	<i>wheel</i> is a (part) meronym of a <i>bicycle</i>
Holonym	contains part	<i>bicycle</i> is a holonym of a <i>wheel</i>
Antonym	opposite of	<i>ascend</i> is an antonym of <i>descend</i>
Attribute	attribute of	<i>heavy</i> is an attribute of <i>weight</i>
Entailment	entails	<i>ploughing</i> entails <i>digging</i>
Cause	cause to	to <i>offend</i> causes to <i>resent</i>
Also see	related verb	to <i>lodge</i> is related to <i>reside</i>
Similar to	similar to	<i>dead</i> is similar to <i>assassinated</i>
Participle of	is participle of	<i>stored</i> (adj) is the participle of "to <i>store</i> "
Pertainym of	pertains to	<i>radial</i> pertains to <i>radius</i>

Tableau 1.3. Les relations sémantiques de wordnet

La figure ci-dessus représente une taxonomie de wordnet (Source : [Siddharth Patwardhan, 2003])

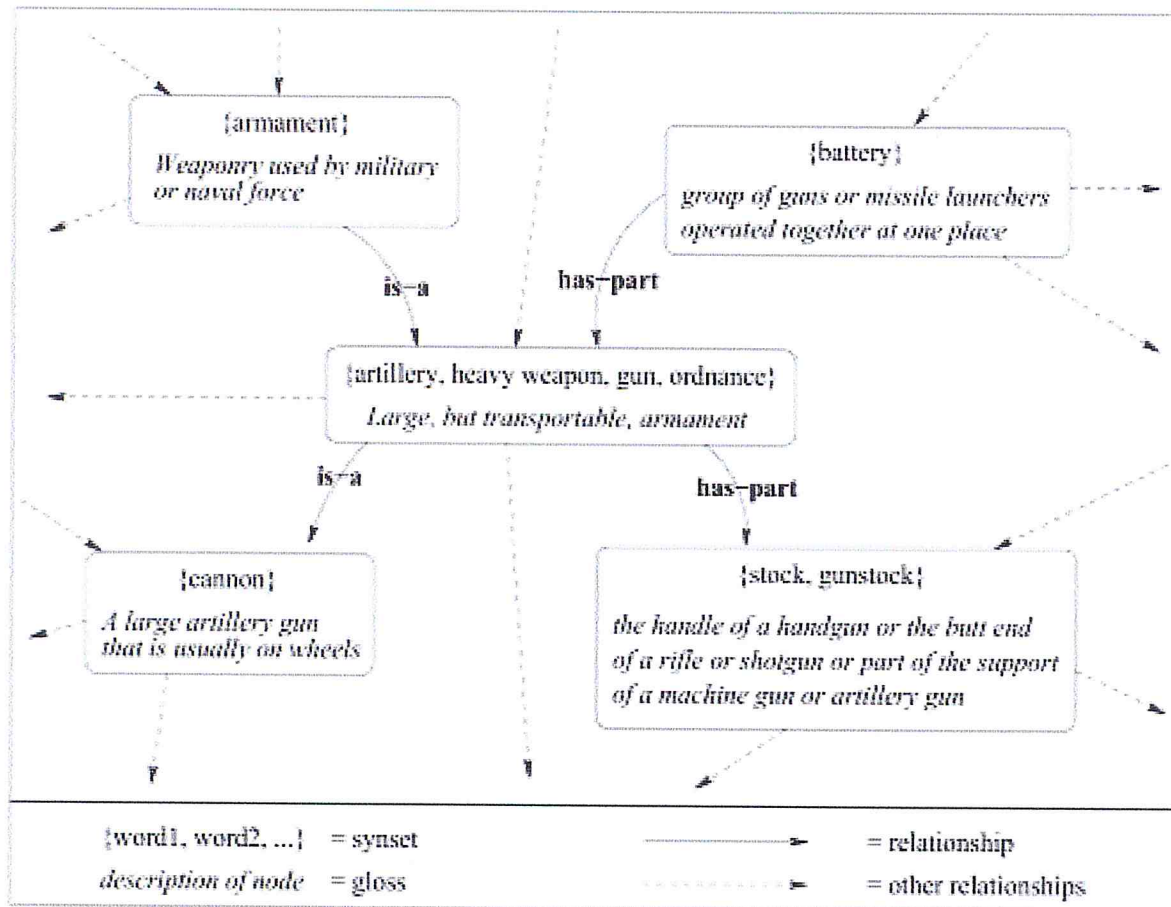


Figure 1.4. Représentation des synsets et des relations dans wordnet

Une des relations de wordnet qui nous intéresse est la relation « *is a kind of* » ou tout simplement « *is a* », cette relation entre synsets est appliquée entre noms et verbes. Elle organise les synsets noms et verbes avec une grande hiérarchie (ou arbre). Chaque arbre a une seule racine. Les concepts les plus générales sont les *hypernymes* des concepts les plus spécifiques. Par exemple, *entity* est le plus général concept dans la hiérarchie des noms et il est la racine de cette hiérarchie. La hiérarchie des verbes contient une petite partie par rapport à celle des noms, ce qui il rend la hiérarchie des verbes plus pratique que celle des noms pour la mesure de similarité.

La figure suivante représente un exemple d' hiérarchie « *is a* » de wordnet: (Source: [Siddharth Patwardhan, 2003])

Les chiffres de 1 à 5 représente les cinq sens du mot *car*.

## 12. Quelques données statistiques :

Voici une présentation quantitative du contenu de WordNet :

Le tableau 1.4 montre la structure de WordNet en nombre de mots, nombre de synsets et nombre de sens, globalement et par catégorie grammaticale. Du nombre total de formes décrites, la plupart sont des noms (74.6%) [Florentina Vasilescu, 2003], le reste étant constitué par des adjectifs (14.6%), des verbes (7.6%) et des adverbes (3.2%). La polysémie (nombre de sens par mot) se manifeste dans Wordnet par le fait qu'il y a des mots qui peuvent appartenir à plusieurs synsets (146350 formes traitées / 111223 synsets).

Tableau 1.4 .Nombre de mots, synsets et sens dans WordNet : source [Florentina Vasilescu, 2003])

Partie de discours	Nombre de mots	Nombre de synsets	Nombre de sens
<i>Noms</i>	109195	75804	134716
<i>Verbes</i>	11088	13214	24169
<i>Adjectifs</i>	21460	18576	31184
<i>Adverbes</i>	4607	3629	5748
<i>Total</i>	146350	111223	195817

Le tableau 1.5 donne des informations sur le caractère monosémique ou polysémique des mots selon leur catégorie grammaticale [Tengi, 1998]. Les verbes présentent le taux de mots polysémiques le plus élevé, par rapport au nombre total des formes verbales décrites dans WordNet (46.6%), tandis que les autres catégories comportent un taux de formes polysémiques plus bas (noms polysémiques – 13.2%, adjectifs polysémiques – 25.5%, adverbes polysémiques - 17%).

Tableau 1.5 Répartition des mots dans WordNet en monosémiques et polysémiques source [Florentina Vasilescu, 2003]

Partie de discours	Mots monosémiques	Mots polysémiques
<i>Noms</i>	94685 (86.8%)	14510 (13.2%)
<i>Verbes</i>	5920 (53.4%)	5168 (46.6%)
<i>Adjectifs</i>	15981 (74.5%)	5479 (25.5%)
<i>Adverbes</i>	3820 (83%)	787 (17%)
<i>Total</i>	120406 (82.3%)	25944 (17.7%)

### 13. Mesures de similarité

Une utilisation possible de l'ontologie fournie par WordNet est la définition de métriques heuristiques de « distance sémantique » entre les synsets. Cette métrique est basée sur la distance à parcourir dans le graphe, combinée ou non avec le Contenu Informationnel. Elle permet de quantifier la similarité de deux concepts. Elle peut également servir dans un cadre de désambiguïsation lexicale.

### 14. Conclusion

Nous avons présenté WordNet, avec ces différentes relations, ainsi que des exemples de ces relations. Nous avons travaillé avec Wordnet par des raisons qui tiennent, d'un côté, de sa complète disponibilité sur Internet (base de données, documentation, fichiers sources etc.). Et d'autre coté, de son utilité comme ontologie et sa facilité pour l'utilisation.

On a préféré a travailler avec la version 2.1 de wordnet par raison que son ontologie est racinée par une seule racine (par contre les autres versions comme 1.7 par exemple contiennent plusieurs racine dans leurs structures) ce qui rendre le travail plus facile, en plus elle est plus riches en nombre *synset* qu'autres version (plus de 74000 *synsets*).

Dans la section suivante on va parler de *la similarité sémantique* et les différentes approches pour la mesurer, ainsi qu'une comparaison de ces approches avec le jugement humain.

## ***Chapitre 02***

### ***La similarité sémantique***

## 1. Introduction

La similarité sémantique, c'est-à-dire l'appréhension de la liaison entre deux concepts, est une capacité de l'homme que les machines ne savent que très mal la reproduire. Ainsi, pour un humain, il est évident que les concepts de *crayon* et de *papier* sont liés, beaucoup plus que ceux de *parapluie* et fer à repasser en tout cas. Mais il est très difficile de le formaliser car rien, en surface, ne permet de le décider. Pour ce faire, il faut utiliser des ressources sémantiques : les ontologies, c'est-à-dire des bases de connaissances. Elles seules permettent de montrer les liens (hypéronymie, antonymie, méronymie, etc.) entre des concepts.

Les recherches sur ce sujet se font sur plusieurs domaines : intelligence artificielle, psychologie, sciences cognitives, et ce depuis de nombreuses années. Les modèles de calcul de la similarité sémantique se retrouvent dans de multiples applications, avec pour but de donner à ces dernières des connaissances supplémentaires pour raisonner sur leurs données. En bio-informatique, les bases de données génomiques et protéiques comportent de très nombreuses annotations textuelles qu'il est possible d'utiliser lors de l'interrogation de ces bases en utilisant une ontologie (Gene Ontology par exemple).

La similarité sémantique a comme principaux champs d'application la recherche d'information et le traitement automatique de la langue.

La recherche d'information (RI) est un champ d'investigation évident pour la similarité sémantique. En effet, les problèmes de polysémie et de synonymie de nos langues génèrent des ambiguïtés dans les recherches. [FURNAS, 1987] par exemple montre les difficultés de consensus dans le choix de termes pour les indexations et pour les recherches. La probabilité que le même terme soit choisi par deux individus pour décrire une entité quelconque est bien inférieure à 20% .

Il est important de noter que dans la littérature, on parle aussi de proximité sémantique (*Semantic relatedness*) qui est une notion plus large que la similarité sémantique. En effet la proximité sémantique prend en considération tout type de relation entre les concepts. Ainsi deux concepts peuvent être proches sémantiquement par leur similarité (ex. *voiture* et *automobile*), mais aussi par d'autres relations comme *partie-de* (*voiture-roue*) ou *contraire* (*guerre-paix*).

## 2. Définitions et concepts de base :

### 2.1 Ontologie :

Le concept d'ontologie peut avoir plusieurs définitions selon le type de l'ontologie et son utilisation. Dans notre cas, une ontologie peut se définir comme un triplé (C, R, I), où C est l'ensemble des concepts de l'ontologie, R est l'ensemble des relations (hyponyme, hypernymes ...) entre les concepts, et I est l'ensemble des instances.

#### Définition

Une ontologie  $O = \langle S, R, \text{type}, \top, \perp \rangle$



- S : l'ensemble des concepts de l'ontologie.
- R : l'ensemble des relations entre les concepts.
- type : une fonction qui associe un type à une relation (ex. is-a(R) est la relation hiérarchique de spécialisation).
- $\top$  : est la racine de l'ontologie.
- $\perp$  : est l'anti-racine de l'ontologie.

Pour notre travail, WordNet peut donc être représenté comme une ontologie, où S est l'ensemble des synsets, R est l'ensemble de relations (hyperonymie, hyponymie, antonymie, etc.).  $\top$  est le concept racine (pour wordnet 2.1 c'est le concept *entity*).  $\perp$  est un concept virtuel représentant l'anti racine.

## 2.2 Concept :

Un concept est l'abstraction d'une réalité. Il est défini par un vecteur d'attributs  $V_i = (T_i, P_1, \dots, P_k, R_1, \dots, R_j)$  tel que  $T_i$  est le terme qui décrit le concept, les  $P_i$  représentent les propriétés du concept et les  $R_i$  sont les relations du concept avec les concepts voisins. Ces différents attributs vont être utilisés pour comparer la sémantique des différents concepts.

## 3. Similarité sémantique :

La similarité sémantique est une notion définie entre deux concepts soit au sein d'une même hiérarchie conceptuelle, soit - dans le cas d'alignement d'ontologies - entre deux concepts appartenant respectivement à deux hiérarchies conceptuelles distinctes. La similarité sémantique indique que ces deux concepts possèdent un grand nombre d'éléments en communs (propriétés, termes, instances).

## 4. Mesure de similarité :

Le calcul de la similarité entre deux concepts permet de déterminer s'ils sont équivalents ou indépendants sémantiquement. Cette mesure est basée sur la terminologie du concept, ses propriétés et ses relations avec son voisinage. En fait, deux concepts qui possèdent la même terminologie et les mêmes propriétés et s'ils ont des relations identiques avec des voisins similaires, il y a une forte chance qu'ils soient identiques.

Une mesure de similarité sémantique fondée sur l'hypothèse dépend fortement à la fois de source de données à partir duquel les données sont constituées et des moyens utilisés pour réaliser la mesure.

Deux approches principales sont utilisées pour la mesure de similarité entre concepts dans une ontologie : (i) en utilisant la structure arborescente ou (ii) en utilisant le contenu informatif des différents concepts en intégrant des mesures statistiques. D'autres approches proposent de combiner les deux. La plupart de ces propositions portent sur WordNet, elles

peuvent cependant être généralisées à une ontologie puisqu'elles exploitent la structure taxonomique. WordNet possède l'avantage d'être une ressource assez fournie, le grand nombre de synsets ainsi que sa facilité d'accès ont fait une ressource très utilisée en RI et en TAL.

## 5. Classification des approches de mesure de similarité

### 5.1 Approches basées sur les arcs :

Nous sommes dans le cadre d'un graphe dont les nœuds sont des concepts. Il paraît donc évident d'utiliser les chemins (suite d'arcs du graphe) pour mesurer la distance entre les concepts. [Rada et al. 1989] ont été, les premiers à suggérer que la similarité dans un réseau sémantique peut être calculée en se fondant sur les liens taxonomiques «is-a». Plus généralement, le calcul de la similarité entre concepts peut être fondé sur les liens hiérarchiques de spécialisation/généralisation. Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est alors de calculer la distance entre les concepts par le chemin le plus court. Les auteurs soulignent que cette proposition est valable pour tous les liens de type hiérarchique (est-un (is-a), sorte-de (kind-of), partie-de (part-of) mais doit être adaptée pour d'autres types de liens (cause, etc.).

Leurs formule de calcul de similarité est comme suit :

$$Sim_{Rada}(s_1, s_2) = d(s_1, s_2) = N_1 + N_2$$

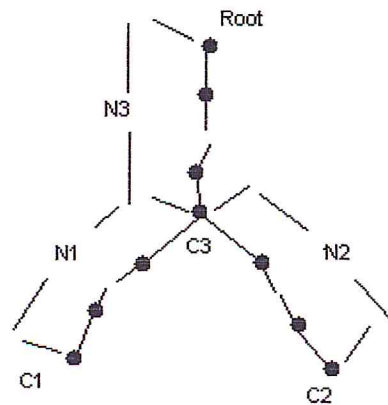


Figure2.1 les relations conceptuelles utilisés par (Rada et al. 1989) (source [Zargayouna, 2005])

D'autres mesures utilisent la notion de plus petit généralisant commun, c'est-à-dire le généralisant commun à c1 et c2 le plus éloigné de la racine. [Wu & Palmer, 1994] ont défini une mesure de similarité entre concepts pour la traduction automatique entre l'anglais et le mandarin chinois. Pour éviter les problèmes d'ambiguïté, leur mesure s'applique à un domaine conceptuel qui correspond à un point de vue donné pour lequel un mot a un seul sens.

Ainsi la mesure de WU et PALMER considère que la similarité entre deux concepts est la longueur du chemin qui les sépare dans une hiérarchie, leurs formule est comme suit :

$$\text{sim}_{\text{wp}}(c1, c2) = \frac{2 * \text{deth}(\text{lso}(c1, c2))}{\text{len}(c1, \text{lso}(c1, c2)) + \text{len}(c2, \text{lso}(c1, c2)) + 2 * \text{deth}(\text{lso}(c1, c2))}$$

Avec :

*Lso* (*c1*, *c2*) est le plus petit généralisant (PPG) de *c1* et *c2*

*Len* est le plus court chemin qui sépare *c1* de *c2* dans wordnet en passant par leurs PPG

*Depth* (*c*) est la profondeur de *c* (de *c* vers la racine) dans le réseau « is a » de wordnet.

Pour la figure ci-dessus Posant :  $N3 = \text{deth}(\text{lso}(c1, c2))$ ,  $N1 = \text{len}(c1, \text{lso}(c1, c2))$ ,  $N2 = \text{len}(c2, \text{lso}(c1, c2))$

Plus formellement cette mesure devient :

$$\text{sim}_{\text{WP}}(C_1, C_2) = \frac{2 * \text{profondeur}(C_3)}{\text{profondeur}_c(C_1) + \text{profondeur}_c(C_2)}$$

Où  $C_3$  est le *ppac* ( $C_1, C_2$ ), *profondeur*( $C_3$ ) est le nombre d'arcs minimal qui séparent  $C$  de la racine et *profondeur*<sub>*c*</sub> ( $C_i$ ) le nombre d'arcs minimal qui séparent  $C_i$  de la racine en passant par  $C$ . La profondeur de  $C$  est une profondeur globale qui permet de normaliser le calcul par rapport à la position des concepts dans la hiérarchie. Deux concepts identiques ont une similarité maximale de 1, plus les concepts sont éloignés, plus la mesure décroît, elle atteint 0 pour deux concepts qui ont  $\top$  (racine) comme *ppac*.

Cette mesure a l'avantage d'être rapide à calculer, en restant aussi expressive que les autres méthodes (elle a d'aussi bonnes performances que les autres mesures de similarité).

[Leacock et Chodorow, 1998] ont proposé une mesure basé sur celle de [Rada *et al.* 1989] mais avec une normalisation par rapport al profondeur de toute la taxonomie avec un logarithme :

$$\text{sim}_{\text{LC}}(c1, c2) = -\log \frac{\text{len}(c1, c2)}{2 * \text{max depth}(c)}$$

Avec : *max depth*(*c*) est la profondeur maximale de *c* dans wordnet

[Hirst & St-Onge 1998] calculent la proximité sémantique qui est une notion plus large que la similarité sémantique. Toutes les relations dans WordNet sont prises en compte. Les liens sont classés comme haut (hyperonymie et meronymie), bas (hyponymie et holonymie) et horizontal (antonymie). il est à noter que cette mesure calcul la similarité entre les mots et pas entre les sens des mots (en effet dans wordnet un mot possède en générale plusieurs sens).

En plus du classement des types de relation, les auteurs définissent un degré de relation:

- Très fort : (*extra-strong*) fondé sur la forme de surface des mots.
- Fort : (*strong*) selon trois cas : (i) quand les mots font partie du même synset, (ii) quand les mots ont une relation horizontale entre leurs synsets, ou (iii) quand un mot fait partie d'un mot-composé.
- Moyen-fort : (*medium-strong*) quand il y a un chemin acceptable entre deux mots.
- Faible : il s'agit de tous les autres mots.

Les mots qui ont une relation extra-forte ou forte ont une similarité constante de  $3 * T$  et  $2 * T$  respectivement ( $T$  étant une constante). Les mots qui ont une relation faible ont une similarité nulle.

Pour la relation forte, deux mots sont fortement similaires s'ils sont sous les conditions suivantes :

- Ils sont appartenant au même synset dans wordnet exemple : *car* et *automobile*.
- Ils sont appartenant aux deux synsets reliés par une relation horizontale dans wordnet par exemple deux mots qui ont des sens opposés comme : *hot* et *cold*.
- Si un mot est un mot composé et l'autre est une partie du mot composé et il existe une relation « is-a » entre le synset du premier mot et du deuxième mot dans wordnet. exemple : *school* et *private\_school*.

Pour la relation Moyen-forte, soit  $T$  un constant utilisé dans la formule pour évaluer la proximité *moyenne forte*. Hirst & St-Onge ont utilisé 8 comme valeur de  $T$  dans leurs expérimentations.

La proximité *moyenne forte* est définie entre deux synsets dans wordnet qui sont reliés par le plus court chemin). La proximité est donnée comme suit :

$$Prox(C_1, C_2) = T - \text{chemin}(C_1, C_2) - k * d$$

Ou :

$D$  est le nombre de changement de directions du chemin.  $T$  et  $K$  sont des constantes. (Dans la pratique  $T = 8$  et  $K = 1$ ).

L'idée est que deux mots sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction. S'il n'y a pas de chemin, ce calcul est égal à zéro. Comparée aux autres mesures de similarité, cette mesure ne donne pas, de ce fait, de bons résultats.

La limite d'utiliser le chemin le plus court est qu'on ne prend pas en considération la position des concepts dans l'ontologie. Intuitivement, deux concepts classés en bas de l'ontologie sont très spécifiques et sont donc à un degré de granularité plus fin que deux concepts classés en haut de l'ontologie. Ainsi les synsets *plant* et *animal* ont la même distance

entre eux qu'*egyptian-cat* et *siamese-cat*, alors qu'intuitivement les deux derniers sont plus proches.

## 5.2. Approches basées sur les nœuds :

Il s'agit de noter le contenu informatif (IC) des concepts de l'ontologie. Dans cette méthode la similarité entre deux concepts est mesurée par rapport à leurs informations c.-à-d. leurs contenus informationnels.

Pour ce faire, il existe deux méthodes, la première utilise un corpus d'apprentissage et mesure la probabilité de trouver un concept ou un de ses descendants dans ce corpus. la deuxième utilisent seulement une ressource sémantique comme wordnet par exemple.

Pour celles qui exploitent un corpus, le calcul de contenu informationnel est basé sur *subsumers* (les ascendants). Supposant qu'un concept *a* est *subsumé* par un concept *b* dans la hiérarchie « is-a » de wordnet. Alors, l'occurrence du *a* dans un texte implique l'occurrence du *b* car *a is-a b* (*a* est un *b*). Par exemple, supposant que *car* est apparait dans un texte, sa implique l'apparition de *motor vehicle* parce que *car* est un *motor vehicle* et on peut dire que sa implique aussi l'apparition de *vehicle* car il est *subsumer* de *motor vehicle*.

Soit *c* un concept, et  $p(c)$  la probabilité de le trouver lui ou un de ses descendants dans le corpus.

Le contenu informatif associé à *c* est alors défini par :

$$IC(c) = -\log p(c)$$

Où  $p(C)$  est la probabilité de retrouver qu'un mot du corpus soit une instance du concept *C* (Un des mots référés par le concept *C* ou par un de ses descendants). Elle est monotone quand on remonte dans la hiérarchie ( $p(A) \leq p(B)$ ) si *A* est plus spécifique que *B* (ex. dans la figure 5.2 ;  $p(\textit{hill}) < p(\textit{geological - formation})$ ).

Cette formule est proposée par Resnik qui inclut le CI dans la mesure de similarité. . Dans les expérimentations de [Resnik, 1995], ces probabilités sont calculées par :

$$p(c) = \frac{\textit{frequence}(C)}{N}$$

Où

*N* est le nombre total de concepts et  $\textit{frequence}(C) = \sum_{w \in \textit{instance}(C)} \textit{count}(w)$ .

Plus un concept est général, plus son contenu informatif est faible, ainsi  $CI(\top) = 0$  parce qu'il est le plus abstrait. A l'inverse, plus le concept est spécifique plus son contenu informatif est important (ex. le contenu informatif de *hill* est plus important que celui de *geological-formation*). L'intuition de la notion de contenu informatif est que la similarité entre deux concepts est la portion d'information qu'ils ont en commun qui, dans le cadre d'une ontologie, peut être déterminée par le concept le plus spécifique qui les subsume (ppac). Cette intuition est indirectement appliquée par les mesures présentées dans la section précédente qui calculent la similarité avec le nombre d'arcs qui séparent deux concepts.

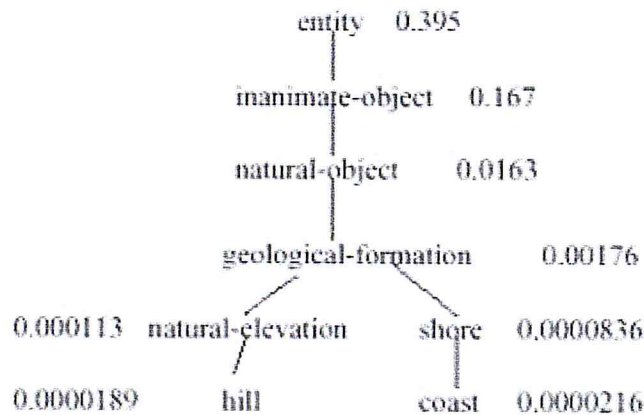


Figure 2.2– Extrait N°01 de WordNet présenté dans [Lin, 1998] avec les probabilités correspondantes aux différents concepts.(source : [Zargayouna, 2005])

En remarque qu'à chaque fois on remonte dans la hiérarchie les probabilités augmentent ce qui signifie que le CI désavantage.

### 5.2.1 La similarité de Resnik :

L'idée du contenu informationnel (CI) remonte à [Resnik, 1995] qui est le premier a pensé de l'utiliser (le CI) pour calculer la similarité entre deux sens. En effet le contenu informationnel mesure la spécificité ou la généralité d'un concept, plus un concept est spécifique plus il est riche en information et plus son contenu informationnel est grand et inversement, plus un concept est générale plus qu'il possède moins d'information et plus son contenu informationnel est petit. Par exemple le concept *car* est plus informatif et doit posséder un contenu informationnel plus grand, dans l'autre côté, le concept *physical entity* est plus générale et exprime moins d'information et doit posséder un contenu informatif plus petit.

La similarité entre  $c1$  (concept  $c1$ ) et  $c2$  (concept  $c2$ ) est la quantité d'information en commun, qui est dans une taxonomie « is-a » le concept le plus spécifique commun  $c$  de  $c1$  et  $c2$ .

Cette intuition est indirectement appliquée dans d'autres mesures telle que celle de [Rada et al. 1989]. Du fait que « si le plus court chemin reliant deux nœud dans le réseau « is-a » est plus long, ca signifie qu'il faut aller plus haut dans la taxonomie pour trouver le ppac des deux nœuds », un exemple donné par [Resnik 1995] est la différence de position de ppac des deux concepts *nickel* et *dime* (soit *coin*) et le ppac des deux concepts *nickel* et *credit\_card* (soit *medium of exchange*) voir la figure en bas.

Le CI est basé sur la probabilité donnée à ce concept :  $CI(c) = -\log(p(c))$   
D'où la similarité de Resnik :

$$Sim_{Res} = IC(lso(S_1, S_2)) = IC(S_3)$$

Ou :

Lso (lowest superordinate) (c1, c2) = CI (c) : le plus petit généralisant.

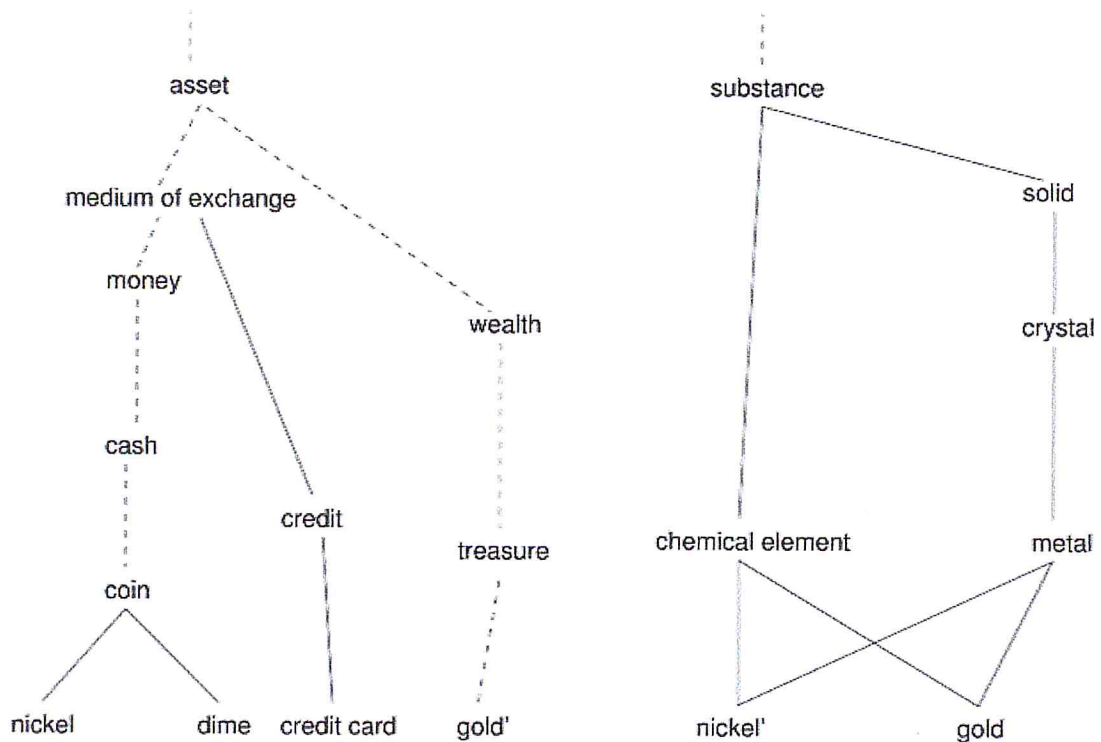


Figure 2.3 Extrait N°02 de wordnet. Les liens foncés représentent le réseau « is-a ». Les liens discontinus indiquent que certains nœuds sont supprimés. Source [Alexander Budanitsky1999]

Dans ces expérimentations, [Resnik 1995] a donné les fréquences aux concepts suite à une analyse du corpus 'Brown' de l'anglais américain qui comporte 1 million de mot. (Le corpus Brown est une collection de textes et d'article de déférente domaine tel que les articles journaux et les articles de science de fiction).

### 5.2.2 La similarité de Jiang et Conrath :

[Jiang et Conrath, 1997] proposent d'utiliser le CI des deux sens car le CI du concept père n'est pas suffisant pour déterminer la similarité. La mesure de Jiang & Conrath pallie les limites de la mesure de Resnik en combinant le contenu informatif du *ppac* à ceux des concepts en prenant en considération le nombre d'arcs.

La mesure est une distance entre *c1* et *c2* inclus la formule de calcul de CI de Resnik, elle est donné par :

$$distance(C_1, C_2) = CI(C_1) + CI(C_2) - (2.CI(ppac(C_1, C_2)))$$

Qui est reformuler après pour qu'elle soit une formule de calcul de similarité :

$$related_{jcn}(c_1, c_2) = \frac{1}{distance_{jcn}(c_1, c_2)}$$

La mesure peut être indéfinie si le dénominateur égale à zéro, et ça pourrait un des deux cas spéciales :

- Premièrement :

$$IC(c_1) = IC(c_2) = IC(ppac(c_1, c_2)) = 0$$

$IC(ppac(c_1, c_2))$  peut avoir la valeur 0 si le  $ppac(c_1, c_2)$  est la racine de la taxonomie, et on sait que le CI de la racine est 0 car le nœud racine est le concept le plus générale.  $IC(c_1)$  et  $IC(c_2)$  peuvent avoir la valeur 0 si les concepts  $c_1$  et  $c_2$  sont jamais marqués leurs apparition dans le corpus, dans ce cas on retourne une similarité de 0.

- Deuxièmement :

$$IC(c_1) + IC(c_2) = 2IC(ppac(c_1, c_2))$$

Qui est peut être le cas spéciale  $IC(c_1) = IC(c_2) = IC(ppac(c_1, c_2))$  et ça quand :  $c_1 = c_2 = ppac(c_1, c_2)$

### 5.2.3 La similarité de Lin :

[Lin, 1998] propose également une mesure de similarité très proche, qui revient essentiellement à une reformulation de la formule de (Jiang and Conrath) :

$$Sim_{Lin} = \frac{2IC(lso(S_1, S_2))}{IC(S_1) + IC(S_2)}$$

Voilà les intuitions de Lin concernant les caractéristiques que devrait respecter la notion de similarité :

La similarité entre deux entités A et B est :

- fonction des caractéristiques communes. Plus les entités ont des caractéristiques en commun, plus elles sont similaires.
- fonction de leurs différences. Plus les entités ont des caractéristiques différentes plus ils sont dissimilaires.
- maximale quand A est identique à B.

#### Définition

Une mesure de similarité est une fonction  $sim : S \rightarrow [0, 1]$  avec S l'ensemble de concepts.

Il est admis qu'une mesure de similarité doit être réflexive et symétrique :



–  $\text{sim}(x, x) = 1$  : réflexivité

–  $\text{sim}(x, y) = \text{sim}(y, x)$  : symétrie

Pour cette mesure il y a un cas spéciale, c'est ou le dénominateur avoir une valeur de 0 ce qui donne une formule indéfinie. Cette situation a pour but que le CI des deux concepts égaux 0. On va simplement retourner une similarité de 0. Car les deux concepts  $c1$  et  $c2$  ne peuvent pas être tous les deux en même temps la racine, sinon le *ppac* ( $c1, c2$ ) n'existe pas.

Alors,  $\text{CI}(c1) = \text{CI}(c2) = 0$  implique que les deux concepts  $c1$  et  $c2$  n'ont pas apparait dans le corpus.

Cette mesure ( $\text{Sim}_{\text{lin}}$ ) peut avoir comme valeur inférieure 0 (pas de similarité) exemple : Soit les deux concepts *rooster* et *voyage*, le *ppac* (*rooster, voyage*) = *entity* (la racine) et puisque le  $\text{CI}(\text{entity}) = 0$  alors  $\text{sim}_{\text{lin}}(\text{rooster}, \text{voyage}) = 0$ . Et elle peut avoir comme valeur supérieure 1 c'est le cas ou  $\text{CI}(\text{ppac}(c1, c2)) = \text{CI}(c1) = \text{CI}(c2)$ .

#### 6. Comparaison avec le jugement humain :

A quel point les mesures de similarités citées précédemment sont fiables ? Est-ce que ils sont capables de donner une bonne mesure ou pas ? Et si on compare deux mesures, est-ce qu'on peut déterminer laquelle la meilleure ?

L'évaluation des approches de mesure de similarité reste une question posée, dans notre travail, on a choisi de les comparer avec le jugement humain. En effet, l'homme a toujours la capacité de spécifier à quel point deux mots sont similaire que les machines ne jamais la possèdent. Il existe d'autre méthode pour évaluer ces approches telle que l'analyse mathématique des résultats obtenus mais la comparaison avec le jugement humain reste toujours la meilleure car l'analyse mathématique repose sur les résultats obtenus après les examiner utilisant des fonctions et formules mathématiques et rien n'assure que l'approche est fiable ou pas sauf les résultats obtenus.

Le jugement humain est une étude faite par [Miller et Charles 1991] sur plusieurs sujets, qui à la fin, donné 30 paire de mots et la similarité entre les deux mots de chaque paire selon les humains.

Les paires de (Miller et Charles) est une étude dérivée de celle de [Rubenstein et Goodenough 1965]. L'étude de (Rubenstein et Goodenough) dans la réalité est un examen pour spécifier la relation entre *la similarité du contexte* et *la similarité du sens (synonyme)*. Rubenstein et Goodenough ont obtenu un *jugement de synonyme (de sens)* par 51 sujets sur 65 paires de mots. Les paires sont classées de *forte similarité* jusqu'au *non similarité (non reliés sémantiquement)* dans un échèle de 0.0 (pas de similarité) à 4.0 (similarité forte). Dans leurs étude, Miller et Charles ont choisi 30 par les 65 originales, 10 pour une similarité forte (entre 3 et 4), 10 pour une similarité intermédiaire (entre 1 et 3), et 10 pour une similarité faible (entre 0 et 1) sur 38 sujets différents.

Les paires de (Rubenstein et Goodenogh) et les résultats obtenus sont présentés dans le tableau2.1

Le tableau2.2 représente les paires et les résultats obtenus par (Miller et Charles)

Tableau2.1 Jugement humain et similarité obtenue par différentes approches : par (Rubenstein et Goodenogh). Source [Alexander Budanitsky, 1999]

##	Pair	Humans	rel <sub>HS</sub>	dist <sub>JC</sub>	sim <sub>LC</sub>	sim <sub>L</sub>	sim <sub>R</sub>
1	cord smile	0.02	0	19.6711	1.38702	0.0900408	1.17616
2	rooster voyage	0.04	0	26.908	0.917538	0	0
3	noon string	0.04	0	22.6451	1.5025	0	0
4	fruit furnace	0.05	0	18.5264	2.28011	0.148152	1.85625
5	autograph shore	0.06	0	22.724	1.38702	0	0
6	automobile wizard	0.11	0	17.8624	1.5025	0.0985543	0.976439
7	mound stove	0.14	0	17.2144	2.28011	0.220406	2.90616
8	grin implement	0.18	0	16.6232	1.28011	0	0
9	asylum fruit	0.19	0	19.5264	2.28011	0.142467	1.85625
10	asylum monk	0.39	0	25.6762	1.62803	0.0706819	0.976439
11	graveyard madhouse	0.42	0	29.7349	1.18057	0	0
12	glass magician	0.44	0	22.829	1.91754	0.0788025	0.976439
13	boy rooster	0.44	0	17.8185	1.5025	0.211185	2.38521
14	cushion jewel	0.45	0	22.9386	2.28011	0.1393	1.85625
15	monk slave	0.57	94	18.9192	2.76553	0.211341	2.53495
16	asylum cemetery	0.79	0	28.1499	1.5025	0	0
17	coast forest	0.85	0	20.2206	2.28011	0.129911	1.50954
18	grin lad	0.88	0	20.8152	1.28011	0	0
19	shore woodland	0.90	93	19.3361	2.5025	0.135051	1.50954
20	monk oracle	0.91	0	22.7657	2.08746	0.182137	2.53495
21	boy sage	0.96	93	19.934	2.5025	0.202764	2.53495
22	automobile cushion	0.97	98	15.0786	2.08746	0.278222	2.90616
23	mound shore	0.97	91	12.492	2.76553	0.498048	6.19744
24	lad wizard	0.99	94	16.5177	2.76553	0.234853	2.53495
25	forest graveyard	1.00	0	24.573	1.76553	0	0
26	food rooster	1.09	0	17.4637	1.38702	0.100578	0.976439
27	cemetery woodland	1.18	0	25.0016	1.76553	0	0
28	shore voyage	1.22	0	23.738	1.38702	0	0
29	bird woodland	1.24	0	18.1692	2.08746	0.138245	1.50954
30	coast hill	1.26	94	10.8777	2.76553	0.532595	6.19744
31	furnace implement	1.37	93	15.8742	2.5025	0.189542	1.85625

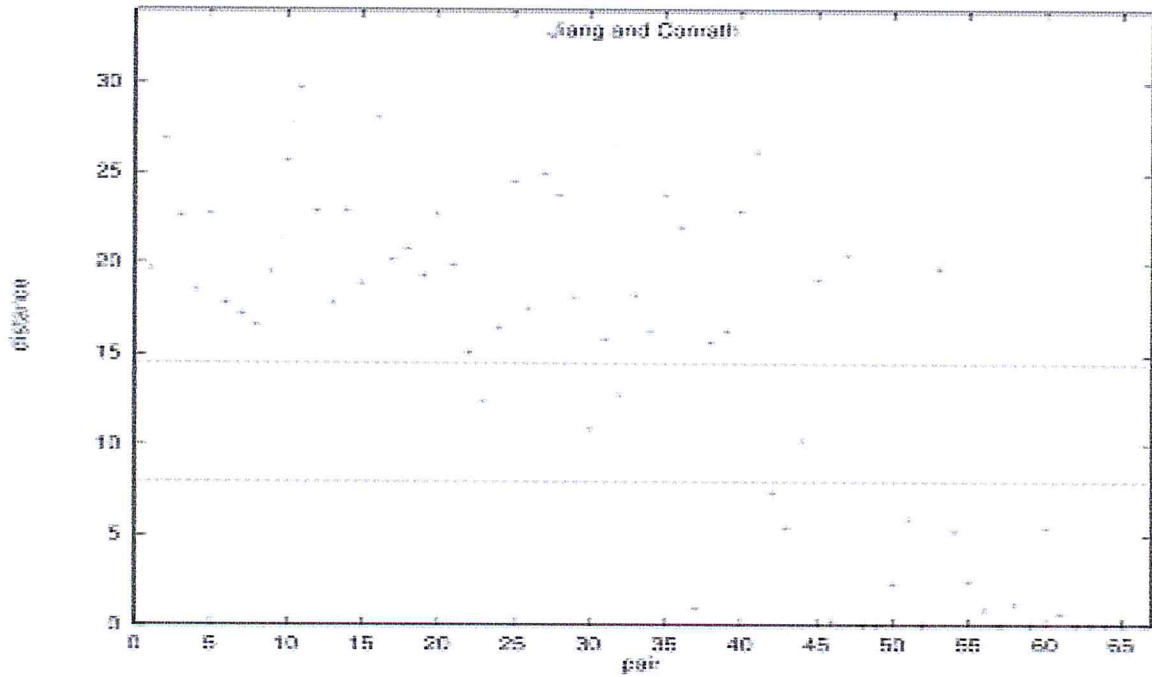
##	Pair	Humans	rel <sub>HS</sub>	dist <sub>JC</sub>	sim <sub>LC</sub>	sim <sub>L</sub>	sim <sub>R</sub>
32	crane rooster	1.41	0	12.806	2.08746	0.581234	8.88719
33	hill woodland	1.48	93	18.2676	2.5025	0.14183	1.50954
34	car journey	1.55	0	16.3425	1.28011	0	0
35	cemetery mound	1.69	0	23.8184	1.91754	0	0
36	glass jewel	1.78	0	22.0185	2.08746	0.144282	1.85625
37	magician oracle	1.82	98	1	3.5025	0.964513	13.5898
38	crane implement	2.37	94	15.6813	2.76553	0.270421	2.90616
39	brother lad	2.41	94	16.3583	2.76553	0.236599	2.53495
40	sage wizard	2.46	93	22.8275	2.5025	0.181733	2.53495
41	oracle sage	2.61	0	26.2251	2.08746	0.162003	2.53495
42	bird crane	2.63	97	7.40301	3.08746	0.705966	8.88719
43	bird cock	2.63	150	5.40301	4.08746	0.766884	8.88719
44	food fruit	2.69	0	10.2695	2.28011	0.227194	1.50954
45	brother monk	2.74	93	19.2087	2.5025	0.208821	2.53495
46	asylum madhouse	3.04	150	0.263035	4.08746	0.991695	15.7052
47	furnace stove	3.11	0	20.5459	2.08746	0.134154	1.85625
48	magician wizard	3.21	200	0	5.08746	1	13.5898
49	hill mound	3.29	200	0	5.08746	1	12.0807
50	cord string	3.41	150	2.27073	4.08746	0.89069	9.25128
51	glass tumbler	3.45	150	5.94251	4.08746	0.792495	11.3477
52	grin smile	3.46	200	0	5.08746	1	10.4198
53	serf slave	3.46	0	19.8021	2.28011	0.34799	5.2844
54	journey voyage	3.58	150	5.21325	4.08746	0.747567	7.71939
55	autograph signature	3.59	150	2.41504	4.08746	0.922084	14.2902
56	coast shore	3.60	150	0.884523	4.08746	0.96175	11.1203
57	forest woodland	3.65	200	0	5.08746	1	11.2349
58	implement tool	3.66	150	1.17766	4.08746	0.913309	6.2034
59	cock rooster	3.68	200	0	5.08746	1	14.2902
60	boy lad	3.82	150	5.39415	4.08746	0.728545	8.29868
61	cushion pillow	3.84	150	0.70044	4.08746	0.974877	13.5898
62	cemetery graveyard	3.88	200	0	5.08746	1	13.7666
63	automobile car	3.92	200	0	5.08746	1	8.62309
64	midday noon	3.94	200	0	5.08746	1	15.9683
65	gem jewel	3.94	200	0	5.08746	1	14.3833

Tableau 2.2 Jugement humain et similarité obtenu par différentes approches : par Miller et Charles. Source [Alexander Budanitsky, 1999]

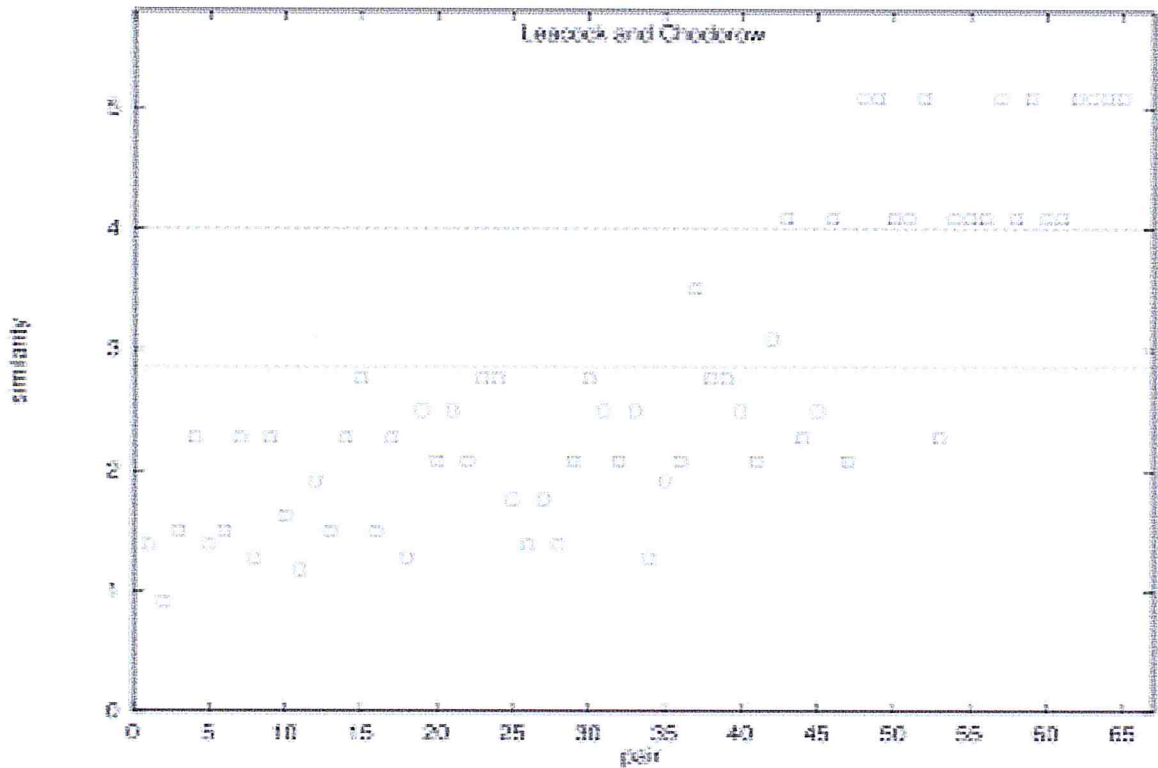
# #	Pair	Humans	rel <sub>HS</sub>	dist <sub>LC</sub>	sim <sub>LC</sub>	sim <sub>L</sub>	sim <sub>R</sub>
1	car automobile	3.92	200	0	5.08746	1	8.62309
2	gem jewel	3.84	200	0	5.08746	1	14.3833
3	journey voyage	3.84	150	5.21325	4.08746	0.747567	7.71939
4	boy lad	3.76	150	5.39415	4.08746	0.728545	8.29868
5	coast shore	3.70	150	0.884523	4.08746	0.96175	11.1203
6	asylum madhouse	3.61	150	0.263035	4.08746	0.991695	15.7052
7	magician wizard	3.50	200	0	5.08746	1	13.5898
8	midday noon	3.42	200	0	5.08746	1	15.9683
9	furnace stove	3.11	0	20.5459	2.08746	0.134154	1.85625
10	food fruit	3.08	0	10.2695	2.28011	0.227194	1.50954
11	bird cock	3.05	150	5.40301	4.08746	0.766884	8.88719
12	bird crane	2.97	97	7.40301	3.08746	0.705966	8.88719
13	tool implement	2.95	150	1.17766	4.08746	0.913309	6.2034
14	brother monk	2.82	93	19.2087	2.5025	0.208821	2.53495
15	lad brother	1.66	94	16.3583	2.76553	0.236599	2.53495
16	crane implement	1.68	94	15.6813	2.76553	0.270421	2.90616
17	journey car	1.16	0	16.3425	1.28011	0	0
18	monk oracle	1.10	0	22.7657	2.08746	0.182137	2.53495
19	cemetery woodland	0.95	0	25.0016	1.76553	0	0
20	food rooster	0.89	0	17.4637	1.38702	0.100578	0.976439
21	coast hill	0.87	94	10.8777	2.76553	0.532595	6.19744
22	forest graveyard	0.84	0	24.573	1.76553	0	0
23	shore woodland	0.63	93	19.3361	2.5025	0.135051	1.50954
24	monk slave	0.55	94	18.9192	2.76553	0.211341	2.53495
25	coast forest	0.42	0	20.2206	2.28011	0.129911	1.50954
26	lad wizard	0.42	94	16.5177	2.76553	0.234853	2.53495
27	chord smile	0.13	0	20.2418	1.62803	0.180828	2.23413
28	glass magician	0.11	0	22.829	1.91754	0.0788025	0.976439
29	rooster voyage	0.08	0	26.908	0.917538	0	0
30	noon string	0.08	0	22.6451	1.5025	0	0

Voici une représentation graphique ci-dessus des résultats obtenus par Rubenstein et Goodenogh. Source [Alexander Budanitsky, 1999]

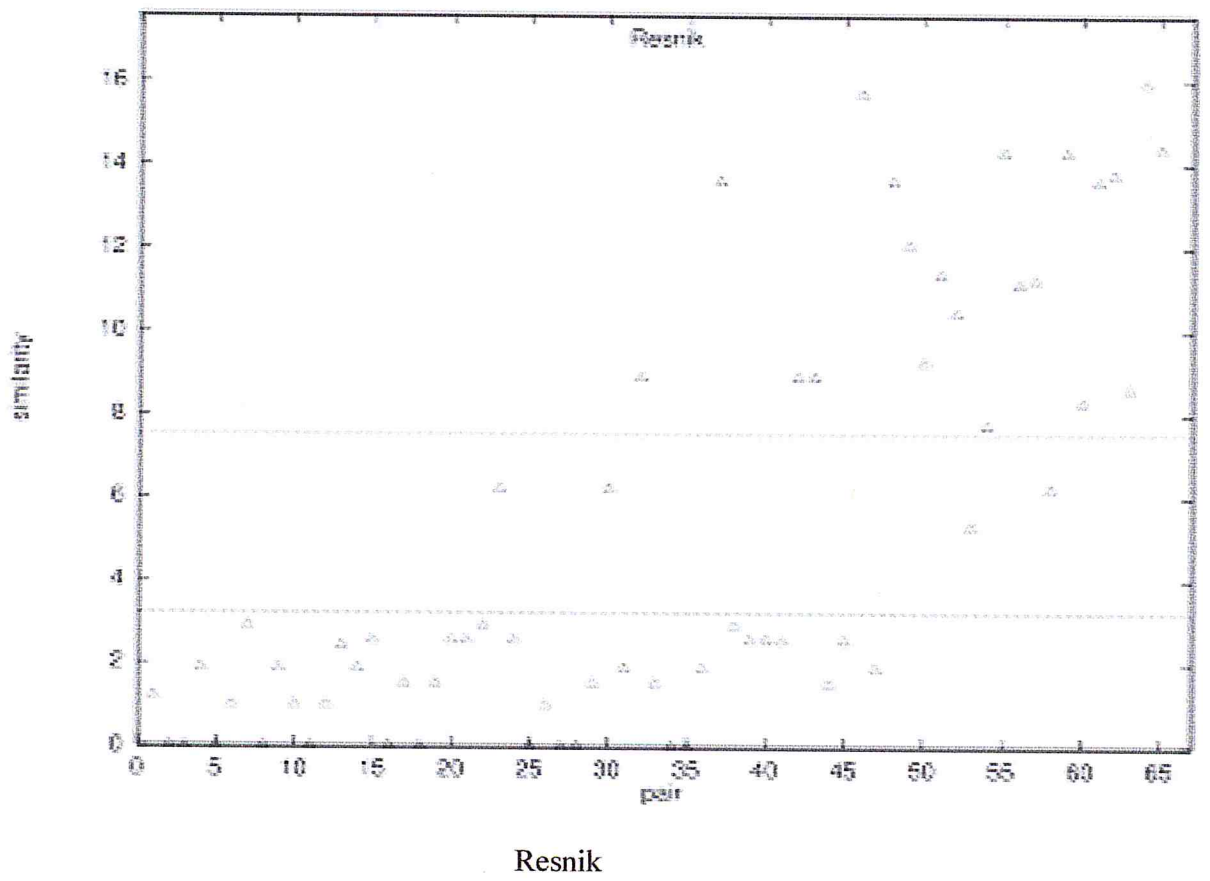
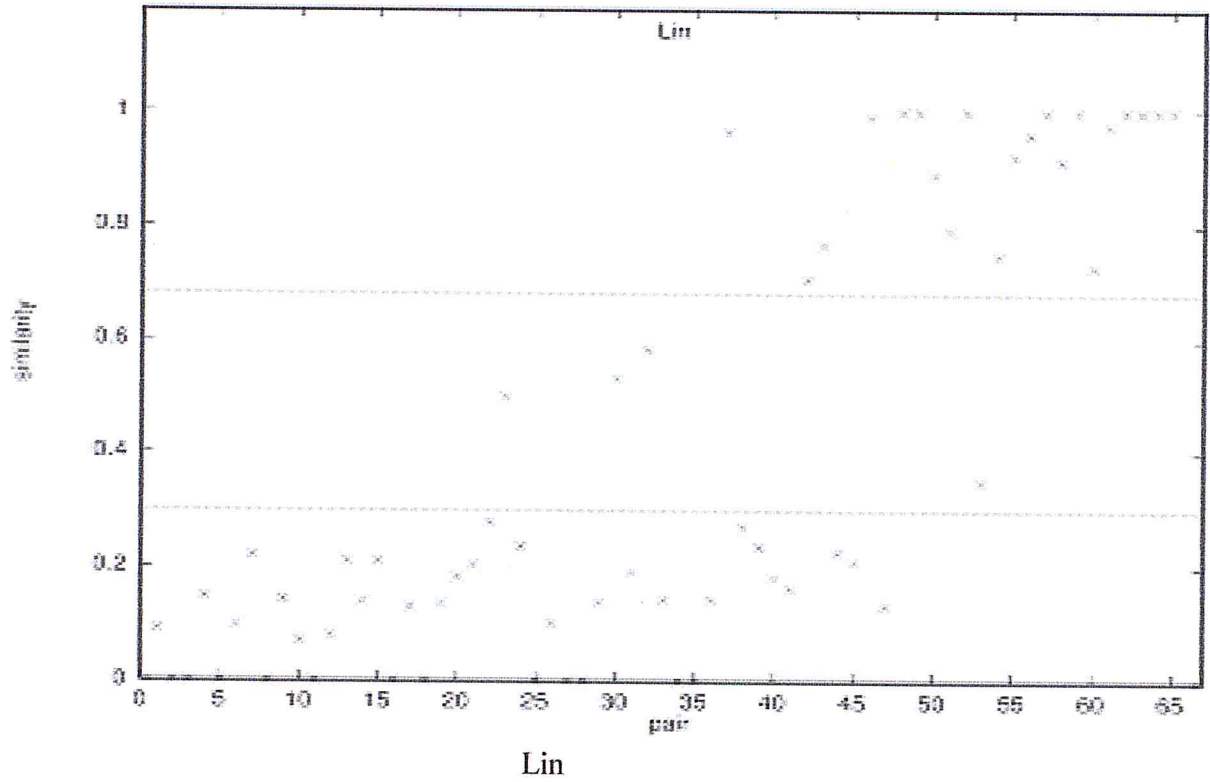
Figure 2.4 représentation graphique des résultats obtenus par Rubenstein et Goodenogh



Jiang ET Conrath



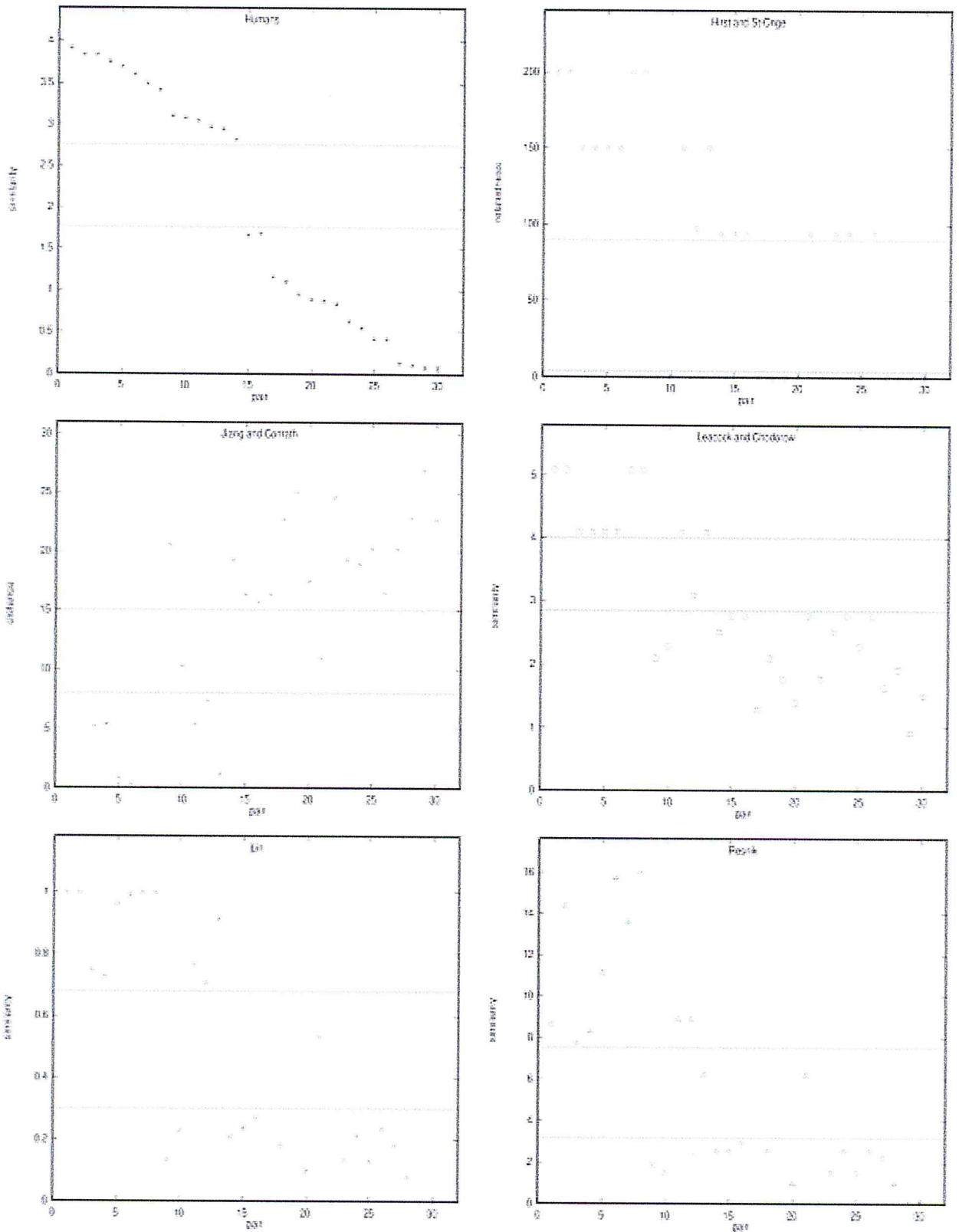
Leacock ET Chodorow



Et voilà une représentation graphique des résultats obtenus par Miller et Charles. Source [Alexander Budanitsky, 1999].

Figure 2.5 représentation graphique des résultats obtenus par Miller et Charles





Les figures ci-dessus montrent une comparaison des méthodes de (Resnik, Jiang et Conrath, et Lin) avec le jugement humain, les résultats pour (Resnik, Lin et Jiang & Conrath) sont pris selon celles de [Alexander Budanitsky, 1999] :

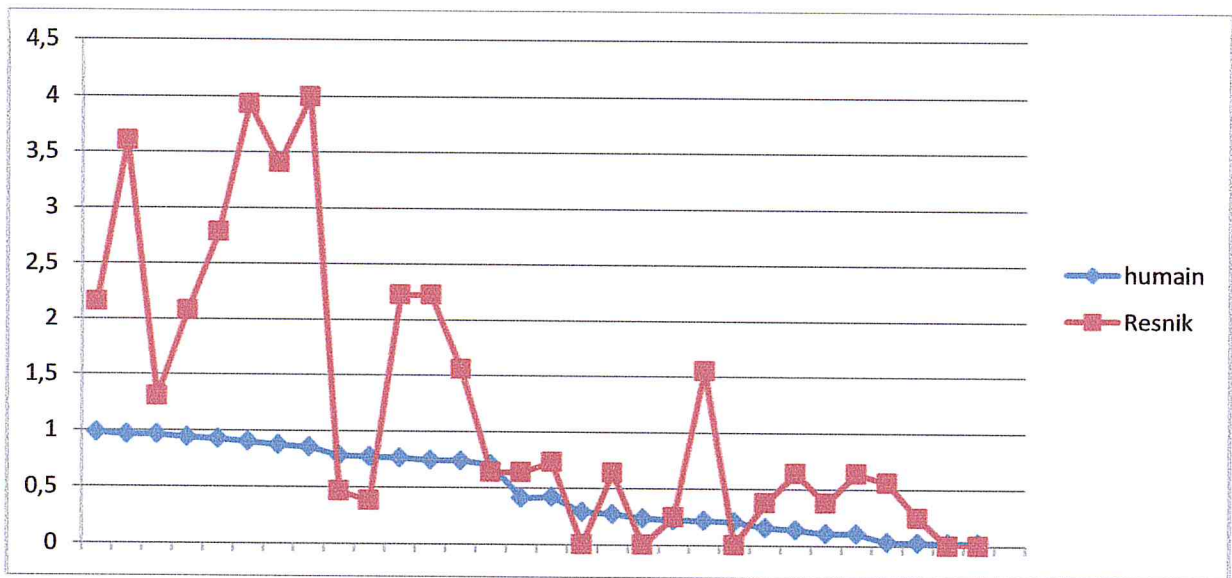


Figure 2.6 comparaison de la mesure de Resnik avec le jugement humain

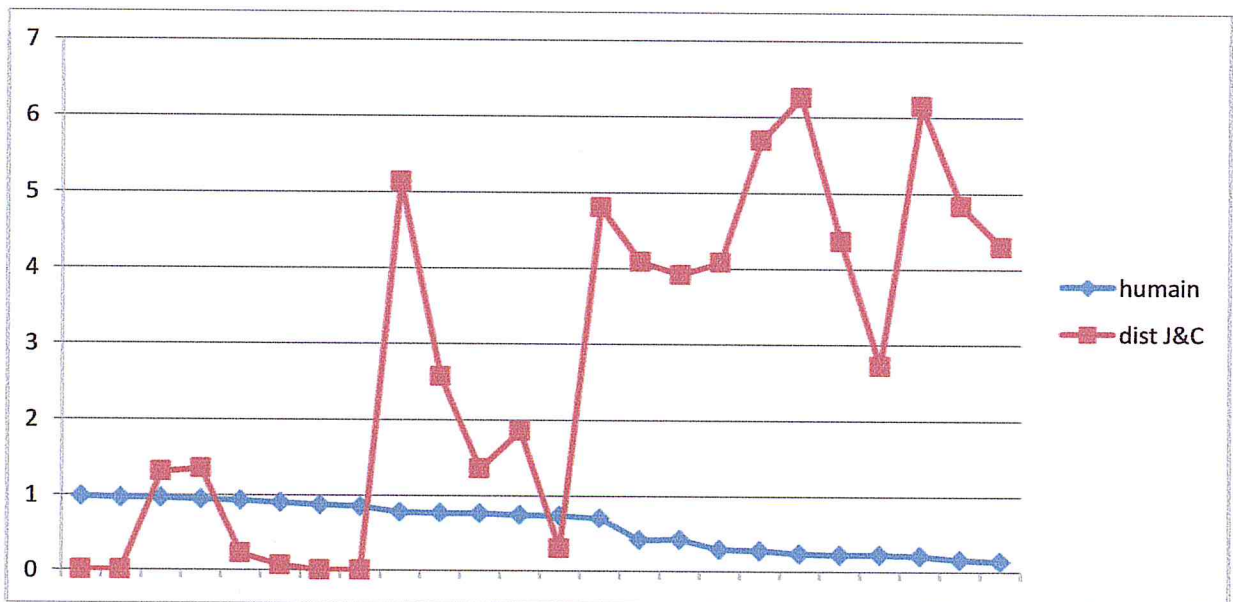


Figure 2.7 comparaison de la mesure de Jiang et Conrath avec le jugement humain

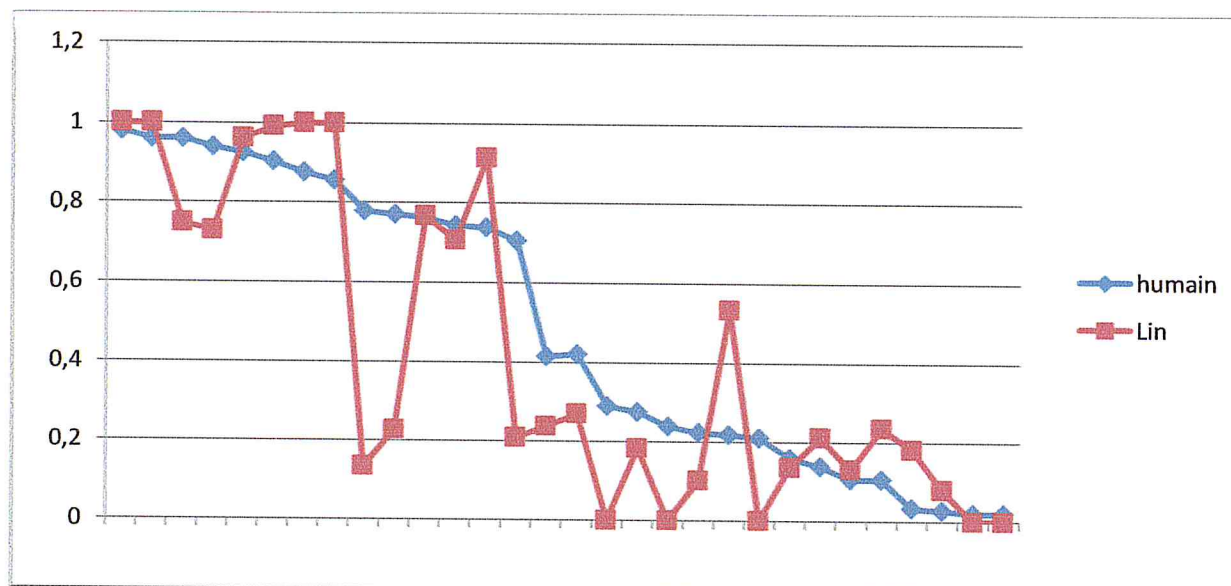


Figure 2.8 comparaison de la mesure de Lin avec le jugement humain

D'après les figures précédentes, il est clair que la mesure de Lin est la plus proche au jugement humain, c'est pourquoi les auteurs la considèrent comme la méthode la plus fiable. De nombreux auteurs comme (HADJ Taieb et al.), (Sebti) couplent leurs formules de calcul de CI (contenu informationnel) par la mesure de Lin a cause de ça fiabilité.

Le tableau 2.3 suivant représente les coefficients de corrélation entre le jugement humain et les différentes mesures. Source [Alexander Budanitsky, 1999].

Measure	Miller-Charles	Rubenstein-Goodenough
Hirst and St-Onge	0.7443990930	0.7861440344
Jiang and Conrath	0.8500267204	0.7812746298
Leacock and Chodorow	0.8157413049	0.8382296528
Lin	0.8291711020	0.8193023545
Resnik	0.7736382148	0.7786845861

Tableau 2.3 coefficients de corrélation entre le jugement humain et les différentes mesures

On remarque que la mesure de Lin, et celle de Resnik ont données un coefficient de corrélation "stable" pour un grand nombre de pair de mots, contrairement avec les autres mesure ou la corrélation n'est pas fixée quand on augmente le nombre de pair de mots.

## 7. Conclusion :

On a présenté dans cette section la similarité sémantique avec les différentes approches de mesure. Pour notre travail, seules les approches basées sur le contenu informationnel nous intéressent : (Resnik, Jiang and Conrath, et Lin).

Ces trois approches utilisent la même formule de calcul du CI (contenu informationnel) introduite par Resnik :  $IC(c) = -\log(p(c))$ . Les auteurs considèrent que le CI peut se calculer autrement, en effet l'utilisation de la probabilité de l'apparition du concept seulement n'est pas suffisante pour mesurer sa richesse en information. Pour cela des études sont faites pour extraire d'autre formule pour le calcul du CI.

Dans la section suivante on va présenter les différentes approches du calcul du contenu informationnel basant sur wordnet.

## ***Chapitre 03***

# ***Les approches de calcul de contenu informationnel***

## 1. Introduction :

Le contenu informationnel (CI) d'un concept est un facteur fondamental dans le traitement linguistique. Il représente la quantité d'information fournie par le concept quand il apparaît dans un contexte. L'idée principale est que les entités (concepts) générale ont une petite valeur du CI que les entités les plus spécifiques. En effet le calcul du CI est basé sur le compte de fréquence d'apparition du concept dans un texte ou dans un corpus. La fréquence associée au concept est incrémentée dans wordnet chaque fois que le concept est observé, ce qui incrémente le compte de fréquence des ancêtres d'un concept (pour l'hierarchie des verbes et noms dans wordnet). Ceci est important car chaque occurrence d'un concept plus spécifique implique aussi l'occurrence des concepts les plus générales (les ancêtres).

De nombreux travaux sont faits pour le calcul du contenu informationnel tel que celle de Resnik [Resnik 1995], Nuno [Nuno 2004].

Dans certaines méthodes, le CI utilise une ressource externe : corpus par exemple. Dans d'autres méthodes le CI utilise seulement la ressource sémantique (comme Wordnet) sans aucune ressource externe.

## 2. les approches de calcul de contenu informationnel :

L'utilisation du contenu informationnel (CI) pour la mesure de similarité entre concepts est introduite pour la première fois par Resnik [Resnik 1995]. Et comme on a dit plus haut, ils existent deux types d'approches de calcul du contenu informationnel, celles basées sur les ressources externes et elles basées sur wordnet.

### 2.1. L'utilisation d'une ressource externe :

[Resnik, 1995], [Jiang *et* Conrath, 1997] et [Lin, 1998] ont proposé d'obtenir le CI suite à une analyse d'un corpus. La méthode utilise la probabilité  $p(c)$  donnée à un concept  $c$  pour calculer son CI. La formule est la suivante :

$$CI(c) = -\log (p(c))$$

Avec :  $p(c)$  est la probabilité que le concept  $c$  ou un de ces instance apparaisse dans le corpus :

$$P(c) = \frac{\text{freq}(c)}{N}$$

Avec :

Freq(c) étant la fréquence d'apparition des instances de  $c$  dans le corpus,

Et  $N$  le nombre total des concepts

Si « $c1$  is a  $c2$  », alors  $p(c1) \leq p(c2)$ . De plus, si la taxonomie possède une seule racine (comme wordnet 2.1) alors sa probabilité est égale à 1.

Une des limites de cette méthode est le corpus utilisé pour calculer le CI d'un concept, car la probabilité donnée à un concept peut changer d'un corpus à un autre.

## 2.2. L'utilisation de la taxonomie "is a" du WordNet :

A cause des limites vues dans l'approche utilisant les corpus, une nouvelle approche fait son apparition, c'est la méthode qui utilise les ontologies pour calculer le CI. Cette méthode de calcul du contenu informationnel exploite la structure hiérarchique et significative de la taxonomie « is a » du WordNet sans faire recours à un corpus externe.

### 2.2.1 CI basé sur hyponymes :

Cette approche est construite par Nuno [Nuno 2004], il considère que le CI d'un concept peut calculer par la fonction d'hyponymes (fonction qui retourne le nombre d'hyponymes) ; plus un concept a d'hyponymes plus il est générale et moins il est riche en information (son CI est petit), et inversement. Sa formule de calcul est la suivante :

$$CI(c) = 1 - \frac{\log(\text{hypo}(c)+1)}{\log(\text{maxwn})}$$

Avec :  $\text{hypo}(c)$  est fonction qui retourne le nombre d'hyponymes de  $c$  (les descendants), et  $\text{maxwn}$  le nombre totale des concepts dans la taxonomie.

Si on considère que la valeur du CI est dans l'intervalle [0..1] alors le CI des feuilles vaut la valeur 1 ( $CI(\text{feuille}) = 1$ ) et il diminue en remontant dans la hiérarchie de wordnet, la racine a donc zéro (0) comme valeur de CI.

### 2.2.2 CI de Zhou :

Zhou [Zhou et al, 2008] a pris la profondeur d'un concept en considération, en effet, deux concepts feuilles peuvent avoir le même CI ( $CI = 1$ ) par l'approche de Nuno même s'ils sont dans des profondeurs différentes. La formule de Zhou est la suivante :

$$CI(c) = k \left( 1 - \frac{\log(\text{hypo}(c) + 1)}{\log(\text{nodemax})} \right) + (1 - k) \left( \frac{\log(\text{deep}(c))}{\log(\text{deepmax})} \right)$$

Avec :  $\text{node}_{\text{max}}$  le nombre totale des concepts dans la taxonomie,  $\text{deep}(c)$  la profondeur de  $c$  dans la taxonomie.  $\text{Deep}_{\text{max}}$  la profondeur maximum dans la taxonomie, et  $k$  un facteur permet de régler le poids des deux entités de la formule.

### 2.2.3 CI de Sanchez :

Après Zhou qui a pensé d'ajouter la profondeur à cette formule, David Sanchez [Sanchez, 2010] a ajouté les ancêtres ainsi que le nombre maximum de feuilles du concept :

$$CI_{\text{Sanchez}} = -\log\left(\frac{|leaves(c)|}{\max\_leaves + 1} + 1\right)$$

Avec :

$Leaves(c)$  c'est le nombre des feuilles de la taxonomie raciné par  $c$  (tous les feuilles descendants de  $c$ )

$Subsumers(c)$  retourne tous les ascendants de  $c$ , et

$Max\_leaves$  le nombre de tous les feuilles de la taxonomie.

Pour *David Sanchez* [Sanchez, 2010], la profondeur d'un concept  $c$ , est le nombre de ses ascendants s'il y a pas d'héritage multiple.

Pour cette méthode, un concept qui hérite de plusieurs concepts (pères directs) est plus spécifique qu'un autre qui hérite d'un seul sachant que les deux sont dans la même profondeur de la même ontologie.

#### 2.2.4 CI de Sebti :

[Sebti et al. 2008] proposent une nouvelle méthode de calcul de CI. Leur méthode inclut la profondeur du concept cible. Un concept situé dans une profondeur élevée est plus informatifs qu'un autre dans une profondeur plus petite car le réseau « is a » de wordnet est organisé suivant une hiérarchie significative.

Voici un exemple expliquant la méthode (source : [Hadj Taieb et al.2012])

On remarque que cette méthode base sur le nombre d'hyponymes directs de chaque concept appartenant au chemin reliant la racine avec le concept cible.

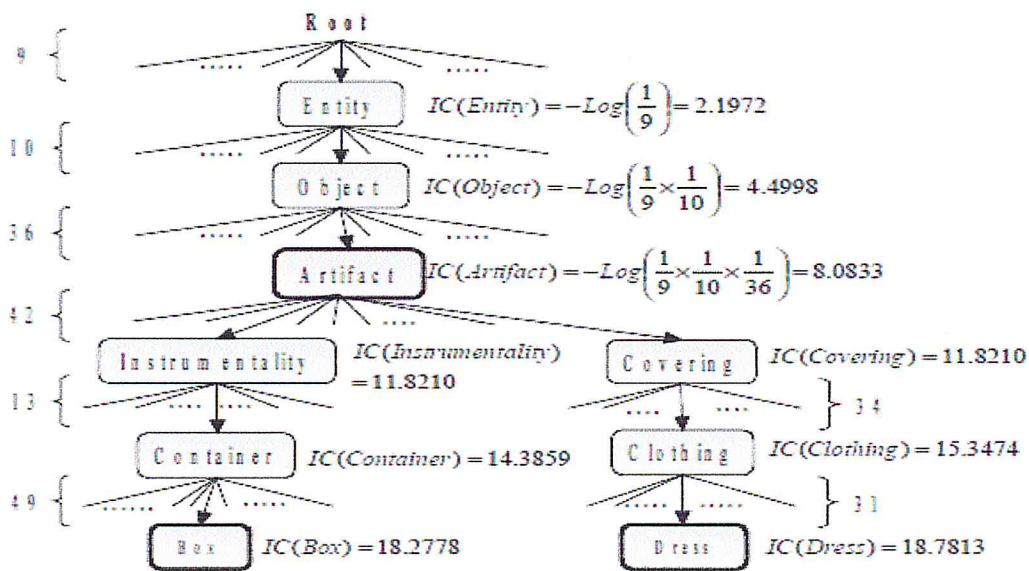


Figure : 3.1. Exemple du calcul du CI pour quelques concepts par la méthode de Sebti

Le CI de Sebti est calculé comme suit :

$$CI(c) = \max_{p \in path(c)} \left( -\log \left( \prod_{c' \in chemin(c)} \frac{1}{hypo(c')} \right) \right)$$



Avec :

- $Chemin(c)$  représente l'ensemble des concepts figurant dans le chemin le plus court reliant  $c$  à la racine de l'ontologie.
- $Path(c)$  représente l'ensemble chemin reliant  $c$  à la racine de l'ontologie.
- $Hypo(c')$  retourne le nombre des hyponymes directs du concept  $c$ .

### 2.2.5 CI de HADJ TAIB :

*Les principes du calcul du CI :*

- Un concept est une accumulation de l'information propagée d'un ancêtre à un autre en ajoutant quelques spécificités à chaque descendant.
- Chaque concept sauf la racine possède un ou plusieurs hyperonymes ne se trouvent pas – en généralement – dans la même profondeur, La profondeur est significative car le passage d'un niveau à un autre plus approfondi engendre le passage des données vers les descendants avec certaines spécificités.

La formule de calcul de CI :

- La première version :

$$CI(c) = \left( \sum_{c' \in \text{Arbre}(c)} Score(c') \right)$$

Ou

$$Score(c) = \left( \sum_{c' \in \text{Hyper}(c) \cap \text{Arbre}(c)} \frac{Depth(c')}{Hypo(c')} \right) \times Hypo(c)$$

**Hypo(c)** : une fonction qui retourne le nombre des concepts subsumés par le concept  $c$ .

**Arbre(c)** : L'ensemble de tous les concepts participants au sous arbre modélisant le contenu informationnel du concept  $c$ .

**Depth(c)**: une fonction qui retourne la profondeur d'un concept dans le thesaurus WordNet.

- La deuxième version

$$CI(c) = \left( \sum_{c' \in \text{Arbre}(c)} Score(c') \right) \times DepthMoy(c)$$

**DepthMoy(c)** : une fonction qui retourne la profondeur moyenne du sous arbre formant le CI du concept  $c$ .

. Soit :

$Niveaux(c) = \{l'ensemble\ des\ profondeurs\ des\ concepts\ \epsilon\ arbre(c)\}$

Et

$$P(Niveaux) = \frac{\text{cardinalité}(\{niveau \in Niveaux(c), niveau=niveau_i\})}{\text{cardinalité}(\text{Arbre}(c))}$$

Alors

$$\text{DepthMoy}(c) = \sum_{niveau_i \in \text{niveaux}(c)} P(niveau_i) * niveau_i$$

### 3. Les limites des approches citées :

Tous les travaux qui existent sur la similarité basée sur le contenu informationnel (CI) assurent que l'ontologie est un bon moyen pour calculer le CI d'un concept [Cross et al, 2012]. Pour cela on va présenter quelques limites sur quelques approches de calcul de CI basée sur Wordnet.

La figure ci-dessus représente un extrait de wordnet 2.1 avec la relation « is a » :

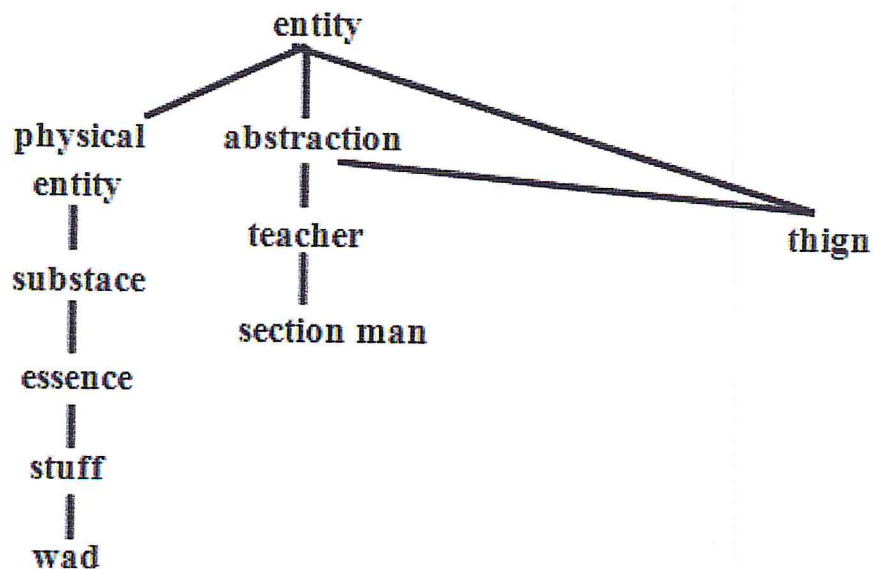


Figure3.2 : extrait N° 03 de wordnet 2.1

Premièrement l'approche basée sur le nombre des descendants, l'auteur de cette méthode [Nuno 2004] considère que le CI d'un concept peut se définir à partir de nombre de ses descendants (hyponymes), ce qui design que le concept qu'il a plus d'hyponymes est plus générale qu'un autre a moins d'hyponymes. Ce qui apparait comme problème pour cette approche est que tous les concepts feuilles ont la même valeur de CI :

$$CI = 1 - \frac{1}{\log(\text{maxont})}$$

Ou : maxont est le nombre totale des concepts dans la taxonomie

Imaginant un concept feuille  $c1$  dont la profondeur est 18, et un autre  $c2$  dont la profondeur 2, appliquant cette approche, le CI des deux concepts est :

$$CI = 1 - \frac{1}{\log(\text{maxont})}$$

Ce qui est illogique.

Revenant à notre figure, on désigne par  $c1$  le concept « *thing* »,  $c2$  le concept « *section man* » et  $c3$  le concept « *wad* ». On appliquant la méthode de Nuno [Nuno 2004] basé sur le nombre d'hyponymes on obtient :

$$CI(c1) = 1 - \frac{-\log(\text{hypo}(c1)+1)}{\log(\text{maxont})} = 1 - \frac{1}{\log(\text{maxont})}$$

$$CI(c2) = 1 - \frac{-\log(\text{hypo}(c2)+1)}{\log(\text{maxont})} = 1 - \frac{1}{\log(\text{maxont})}$$

$$CI(c3) = 1 - \frac{-\log(\text{hypo}(c3)+1)}{\log(\text{maxont})} = 1 - \frac{1}{\log(\text{maxont})}$$

Cette approche A ignoré le “depth “ ou la profondeur, ce facteur est important car on peut trouver deux concepts différents qui ont une même valeur de CI même s'ils sont dans des profondeurs déférentes.

Deuxièmement l'approche de ZHOU [Zhou et al. 2008] qui a ajouté la profondeur à la formule de l'approche précédente, cette approche est plus efficace que la précédente :

$$CI(c) = k \left( 1 - \frac{\log(\text{hypo}(c)+1)}{\log(\text{nodemax})} \right) + (1 - k) \left( \frac{\log(\text{deep}(c))}{\log(\text{deepmax})} \right)$$

Cette méthode utilise un facteur mathématique, le  $k$ , ce qui peut faire des problèmes pour le réglage des poids entre les deux entités de la formule (dans leurs expérimentations, [Zhou et al. 2008] ont pris 0.5 comme valeur de  $k$ )

#### 4. Conclusion :

On a présenté dans cette section quelques approches de calcul du contenu informationnel (CI) ainsi que quelques limites de quelques approches. On peut facilement remarquer que tous les approches utilisent la racine de wordnet (de la taxonomie wordnet) comme point de départ pour calculer le CI des différents concepts (pour le calcul de profondeur d'un concept, pour le calcul de profondeur de la taxonomie wordnet..).

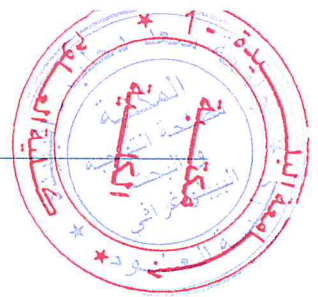
Mais que se passe-t-il si quelqu'un pense à un autre point de départ et ignore la racine de wordnet ?, et si un autre pense à utiliser conjointement la racine de wordnet et une autre entité pour le calcul ?

Dans la section suivante on va présenter notre travail, il s'agit d'une nouvelle approche de calcul de contenu informationnel basée sur wordnet, et qui exploite le réseau « is-a » de WordNet, ainsi que sa comparaison avec le jugement humain.

*Partie II mise en oeuvre*

*Chapitre 04*

*Notre approche*



## 1. Introduction :

L'identification de la similarité dans les ontologies est un concept fondamental qui est adopté par plusieurs techniques telles que la fouille de données (data mining), le Web sémantique et en particulier, le domaine de la recherche de l'information. Dans les chapitres précédents on a parlé de la similarité, on a dit qu'ils existent plusieurs approches pour la mesurée. On a parlé plus essentiellement sur celles basées sur les nœuds (ou le CI). Ces dernières reposent largement sur ce qu'on appelle *le concept*.

Dans une ontologie (Wordnet comme exemple), les concepts sont organisés selon une hiérarchie de spécification/généralisation (dans wordnet la hiérarchie *Is-a*), ou chaque concept peut avoir une –structure- : un concept a un nom, un ensemble de relations, un ensemble des ascendants (hypernymes), et un ensemble des descendants (hyponymes)...

Des travaux précédents, tel que Hadj Taieb [Hadj Taieb et al.2012], considèrent que la profondeur est un coefficient important pour le calcul du CI d'un concept, en effet la profondeur design la distance séparant le concept de la racine de l'ontologie. Il design aussi la spécificité du concept, car un concept plus loin de racine (valeur de profondeur grande) a un arbre d'hypernymes riche on concepts et donc plus d'information.

Zargayouna [Zargayouna, 2005], a parlé dans son article (*Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML*) d'un élément qui peut se placer en bas de l'ontologie et relier (subsumer) par toutes les feuilles, elle a appelé sa *bottom* ou anti-racine en français. Dans ce cas un autre facteur peut faire son apparition, il s'agit de la distance entre le concept et l'anti-racine.

Dans ce qui suit, on va présenter une nouvelle approche de calcul basé sur le contenu informationnel. En plus des facteurs utilisés dans les approches précédentes, notre approche inclut un nouveau facteur qu'on a appelé la *hauteur*

## 2. Notre approche :

Notre méthode pour obtenir le CI d'un concept est basée sur le principe que la structure de la taxonomie wordnet est organisée d'une façon significative c.à.d. un concept fils doit être forcément un synonyme ou dans le même sens que son père, et ça explique le passage de l'information dans wordnet, en effet les concepts plus bas dans l'ontologie sont plus informatifs et leurs valeurs du CI sont plus grandes.

Ce que nous proposons ici c'est d'exploiter d'une façon plus approfondie la structure WordNet et spécialement le réseau « IS A » (Hyperonyme/Hyponyme).

Les idées principales participantes à la création de notre nouvelle méthode de calcul du contenu informationnel sont expliquées dans les paragraphes suivants.

### 2.1 Les principes du calcul du CI :

#### 2.1.1 La relation Hyperonyme/Hyponyme « IS-A » :

La relation C2 (concept C2) *is a* C1 (concept C1), c'est-à-dire C1 est inclus dans C2. Autrement dit, le concept C2 est formé par les données de C1 plus certaines spécificités qui le différencient de son père. Il en résulte donc qu'un concept est une accumulation de l'information propagée d'un ancêtre à un autre en ajoutant quelques spécificités à chaque

descendant. D'où, il est trivial qu'un concept dépend fortement de ses parents directs et ses ancêtres.

- Les hypernymes :

L'ensemble des relations d'hyperonymes directs et indirects d'un concept C forment un sous arbre qui participera à la quantification de son CI.

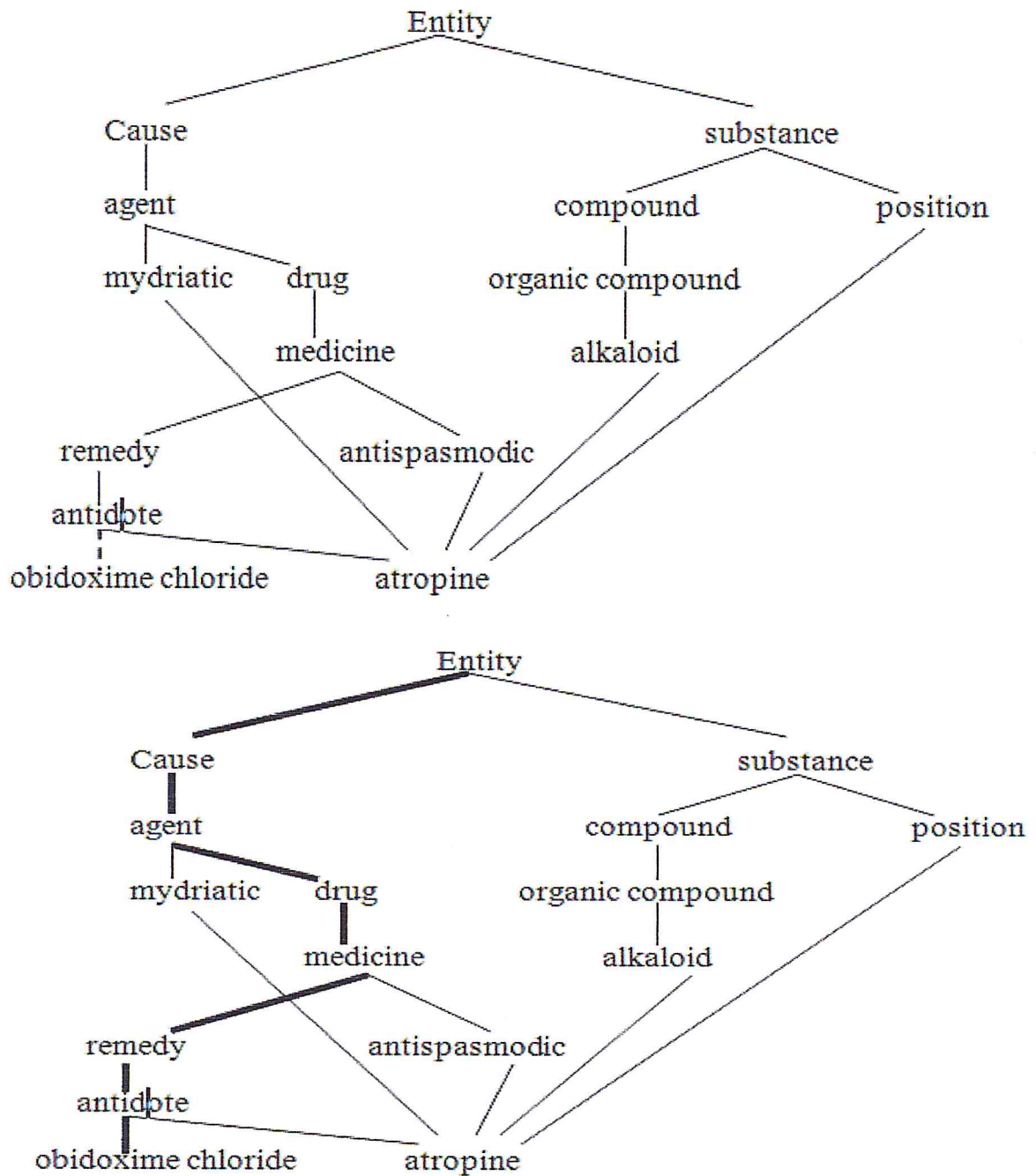


Figure 4.1 Les sous arbre d'hypernymes de (atropine) et de (obidoxime chloride), Extrait de wordnet 2.0 (source : [Hadj Taieb et al.2012])

La figure ci-dessus montre que le sous arbre des hypernymes de « atropine » est plus riche en concepts que celui de « obidoxime chloride », ce qui veut dire que le concept « atropine » avoir l'information de plusieurs concepts (cinq pères directes) donc il est plus riche en information, tandis que le concept « obidoxime chloride » avoir moins d'information (un seul père directe), il est donc moins informatif. Il reste qu'à quantifier la participation de chaque concept appartenant au sous arbre des hypernymes dans la totalité du contenu informatif. Cette participation dépend essentiellement de la profondeur de chacun d'eux, car un concept dans un niveau plus approfondi est plus riche en information et sa valeur du CI doit être plus grande qu'un autre dont la profondeur est plus petite.

- Les hyponymes :

Les hyponymes d'un concept  $c$  peuvent participer à la quantification de son CI, en effet un concept avec plus d'hyponymes exprimes moins d'information et il est loin de contenir l'information complète même s'il a un arbre d'hypernymes riche en concept. Voyons l'exemple suivant :

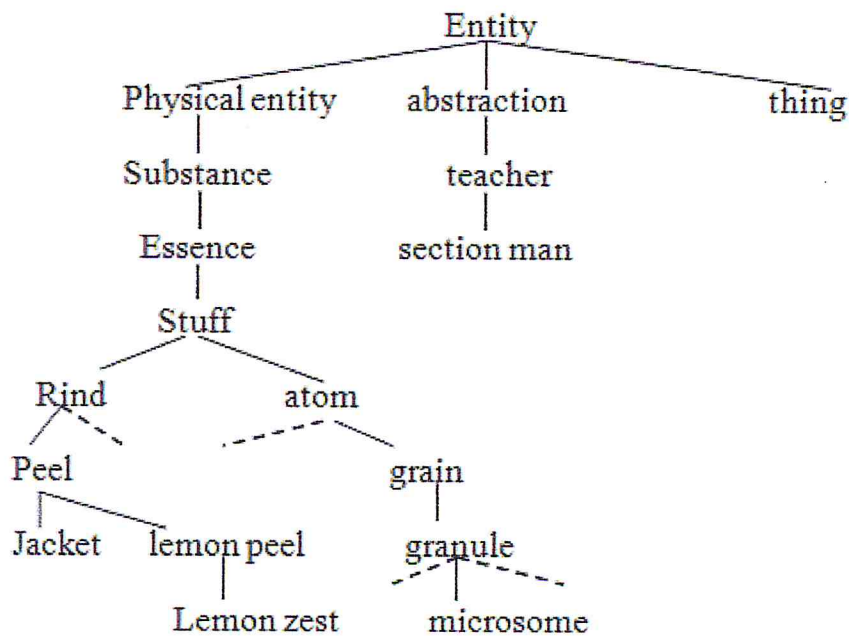


Figure 4.2 Le sous arbre hyponymes de (stuff) Extrait N° 04 de wordnet 2.1

Dans la figure, le concept « thing » avoir l'information complète car c'est un nœud feuille ; il n'a pas d'hyponymes, son CI doit être donc maximum malgré il est enrichie seulement par (*entity*). Dans l'autre part, le concept « stuff » est enrichie par les concepts (*essence, substance, physical entity, entity*), malgré cette richesse en concepts (évidemment en information) dans l'arbre d'hypernymes, il possède beaucoup d'hyponymes (directes et indirectes), et sa explique que le concept « stuff » est plus abstrait et son information n'est pas encore complète et elle doit être encore spécifié par le passage d'information vers les hyponymes. Dans ce cas il doit posséder un CI faible, d'où l'utilité d'hyponymes pour le calcul du CI.



Pour simplifier le travail, on a remplacé Les hyponymes d'un concept  $c$  avec le nombre de feuille qu'il a, en effet un concept qui a un arbre d'hyponymes grand il a forcément un nombre de feuilles grand et inversement,

### 2.1.2 La signification de la profondeur dans la taxonomie « IS A » :

Dans une ontologie (wordnet), chaque concept sauf la racine possède un ou plusieurs hypernymes. Ces hypernymes ne se trouvent pas souvent dans la même profondeur. Le passage d'un niveau à un autre plus approfondis engendre le passage des données vers les descendants avec certaine spécificité à chaque niveau. La transition d'un concept  $c1$  vers un concept  $c2$  en utilisant la relation d'hyponymie ne signifie pas le passage d'une profondeur  $p$  à une profondeur  $p+1$ . Mais, il se peut que malgré cette relation directe entre ces deux concepts, ils se trouvent à des profondeurs non successives dans l'arbre taxinomique « IS A ». Dans la figure 4.1 le concept « atropine » a cinq hypernymes directes. Prenant le cas des concepts « antispasmodic » (profondeur 06) et « position » (profondeur 03), le concept « antispasmodic » est enrichi par les concepts (*medicine, drug, agent, cause, entity*), mais seulement (*substance, entity*) pour le concept « position », il est donc évident que la part de « antispasmodic » est plus grande que celle de « position » dans la participation dans le CI de « atropine », d'où la signification de la profondeur dans le calcul de CI.

### 2.1.3 La hauteur :

Nous définissons un concept 'bottom' ou anti racine ( $\perp$ ); le concept le plus bas de l'ontologie (un concept virtuel qui symbolise la fin de l'ontologie) subsumé par toutes les feuilles, qui sert à évaluer la spécificité de tous les concepts.

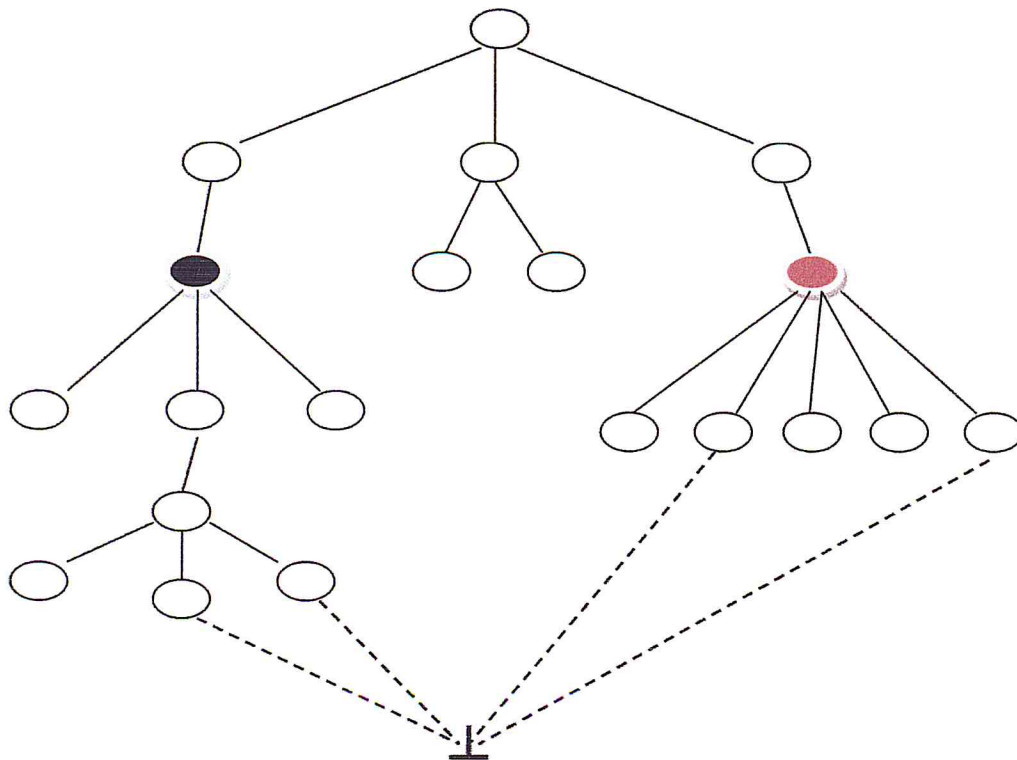


Figure 4.3 représentation de l'anti racine dans un extrait de wordnet

La figure 4.3 montre un extrait wordnet avec une représentation du concept *bottom*. Le principe est que les concepts les plus proches de l'anti racine (*bottom*) sont les plus spécifiques et inversement. Nous définissons une fonction *hauteur(c)* qui sert à déterminer la distance (maximale) en nombre d'arcs entre *c* et *bottom*. Prenant le cas des concepts « rouge » (profondeur 02, 05 feuilles) et « noire » (profondeur 02, 05 feuilles). Dans le cas du concept « rouge », l'information est presque complète, il reste qu'un seul passage vers les nœuds feuilles avec une certaine spécification, la *hauteur* de nœud « rouge » est donc '02'. Cependant, dans le cas de nœud « noir », l'arbre des hyponymes est assez grand ce qui veut dire que le concept « noir » est plus abstrait et il doit être encore spécifié pour que l'information soit majeure. La hauteur maximale de nœud « noir » est '04', il est encore plus loin de l'anti racine malgré que le nombre de ces feuilles est le même que celui de nœud « rouge ». d'où la signification de la hauteur dans le calcul de CI.

### 3. Nouvelle méthode de calcul de CI :

Dans ce paragraphe nous expliquons notre méthode de calcul de CI qui a porté comme nouveauté la hauteur d'un concept dans une ontologie (wordnet), en prenant en considération les facteurs présentés ci-dessus (l'arbre des hypernymes, la profondeur, les hyponymes « feuilles »). Pour ce faire soit :

**Leaves(c)** : une fonction qui retourne les feuilles du *c* ; dans l'exemple précédent *leaves(stuff)* = {lemon zest, micrososome, .....}.

**Hyper(c)** : une fonction qui retourne tous les hypernymes (directes et indirectes) du *c* ; *hyper(stuff)* = {essence, substance, physical entity, entity}.

**Profondeur(c)** : une fonction qui retourne la profondeur du *c* (distance maximale en nombre d'arcs depuis la racine ( $\top$ ) jusqu'à *c*) ; dans l'exemple *profondeur(stuff)* = 04.

**Hauteur(c)** : une fonction qui retourne la distance (maximale) en nombre d'arcs depuis l'anti racine ( $\perp$ ) jusqu'à *c* soit la hauteur de *c* ; dans l'exemple *hauteur(stuff)* = 05.

**MaxLeaves** : le nombre maximum des feuilles dans l'ontologie wordnet ; dans wordnet 2.1 *MaxLeaves* = 62406.

**HautMax** : la hauteur maximale de l'ontologie wordnet ; dans wordnet 2.1 *hautMax* = 19.

- Pour chaque élément appartenant à *hyper(c)* nous calculons un *scorHyper* :

**ScorHyper(c)** : la fonction qui associe à chaque concept de l'arbre d'hypernymes de *c* son profondeur, soit :

$Hyper_i(c) = \{l'élément\ i\ de\ l'ensemble\ des\ hypernymes\ de\ c\}$

$ScorHyper(c) = \sum_{i=0}^n hyper_i(c) * profondeur(i).$

$Distance(c) = profondeur(c) - (\frac{hauteur(c)}{hautMax} + 1)$

L'ajout d'un « 1 » dans la formule c'est pour obtenir une valeur de 'distance' égale à zéro « 0 » pour le concept racine; (le concept racine doit avoir une valeur de CI = 0).

Finalement, nous calculons le CI totale du concept c par la formule suivante :

$CI(c) = -\log\left(\frac{\frac{leaves+1}{maxLeaves+1}}{scorHyper+1}\right) + distance(c)$

La division par *hautMax* est pour la normalisation; en effet pour avoir une valeur de CI = 0 pour le concept racine.

La division par *maxLeaves* est pour la normalisation; en effet pour avoir une valeur dans l'intervalle [0,1] pour la première entité de la formule, pour que les concepts les plus abstraits aient une valeur de CI inférieure que celles des concepts plus spécifiques.

**4. Comparaison avec d'autres approches de calcul de CI :**

	Caractéristique des résultats				
	le modèle du CI dépendante d'un corpus	le modèle du CI indépendante d'un corpus			
	L'approche de resnik	L'approche de seco	L'approche de Zhou	L'approche de Sanchez	Notre approche
Augmentation Depth(c)	NON	NON	OUI Augmentation d'IC	OUI Augmentation d'IC	OUI Augmentation d'IC
Augmentation Hypo(c)	NON	OUI Diminution d'IC	OUI Diminution d'IC	NON	OUI Diminution d'IC
Augmentation Leaves(c)	NON	NON	NON	OUI Diminution d'IC	OUI Diminution d'IC
Augmentation hauteur(c)	NON	NON	NON	NON	OUI Diminution d'IC
compter sur un corpus	OUI	NON	NON	NON	NON

Problème clairsemé de données	OUI	NON	NON	NON	NON
Suggestions pour l'Amélioration	-----	Prendre la topologie de chaque concept en considération, comme, l'arrangement des concepts		Tenir compte le nombre des hyponymes, profondeur des feuilles	-----

Tableau 4.1 comparaison entre les différents modèles de CI

Le tableau ci-dessus donne les valeurs des contenus informationnels pour chaque concept de la liste des jugements humains de [Miller & Charles, 1991] de 60 concepts avec les différentes approches de calcul de CI ainsi que notre approche en utilisant wordnet 2.1.

Les valeurs données avec notre approche ne sont pas normalisées entre 0 et 1 ( $[0,1]$ ).

Les résultats de [Sanchez, 2010], [Sebti et al.2008], [Hadj Taieb et al.2012] sont pris selon celles de [Ferid et khaddem, 2014].

Nouvelle approche de calcul de CI

les concepts	le resnik	le seco	le sanches	le sebol	le hadjriebl	le hadjriebl
car#1	6.0255293	0.6701791	4.3044434	26.593497	34.069904	207.2556
automobile#1	6.0255293	0.6701791	4.3044434	26.593497	34.069904	207.2556
asylum#2	11.416422	0.9381574	4.8023195	22.069949	14.782128	73.910645
madhouse#1	11.821897	1.0	4.805924	22.069949	15.907128	87.489204
midday#1	10.31781	1.0	4.805924	14.876495	25.269993	133.97946
noon#1	10.31781	1.0	4.805924	14.876495	25.269993	133.97946
rooster#1	11.416422	0.9381574	4.8115277	23.660517	32.15455	209.09456
voyage#1	10.030129	0.9019819	4.7767653	23.592619	22.772133	136.6329
gem#3	0.0	1.0	4.7929396	17.449349	12.786334	43.838963
jewel#2	0.0	1.0	4.7929396	17.449349	12.786334	43.838963
journey#1	9.735112	0.6509692	4.1775346	22.494005	21.992699	120.90462
voyage#2	11.821897	0.9019819	4.7767653	25.266594	26.280903	157.65542
boy#1	7.4911937	0.73729706	4.4177437	19.843245	12.878389	49.903755
lad#2	10.909597	1.0	4.8023195	22.241133	14.067067	60.957294
coast#1	9.470511	0.9564059	4.6218634	10.749784	2.2071905	6.6215415
shore#1	9.951473	0.9144722	4.50139	9.651173	2.0443997	5.310979
magician#2	11.821897	0.9144722	4.651827	19.831642	12.799492	49.59416
wizard#2	11.821897	0.9144722	4.651827	19.831642	12.799492	49.59416
journey#1	9.735112	0.6509692	4.1775346	22.494005	21.992699	120.90462
car#1	6.0255293	0.6701791	4.3044434	26.593497	34.069904	207.2556
furnace#1	12.515034	0.7242171	4.331829	17.057102	9.897156	39.583623
stove#1	10.569124	0.9144722	4.64414	20.500122	34.592835	207.354
food#3	11.821897	0.9381574	4.7857203	10.312147	7.264106	25.42437
fruit#2	11.821897	1.0	4.8023195	16.409976	11.849773	59.249967
bird#1	7.655222	0.39591002	2.9817135	18.717903	20.909105	92.26393
cock#5	12.515034	0.9381574	4.8023195	21.975904	22.772912	113.86406
bird#1	7.655222	0.39591002	2.9817135	18.717903	20.909105	92.26393
crane#5	0.0	0.9381574	4.8099304	26.475714	30.183323	190.73194

implement#1	7.013776	0.41250476	2.907055	14.150992	9.437499	33.031246
lad#2	10.905597	1.0	4.8023195	22.241109	14.067067	60.957294
brother#2	10.723275	0.9019918	4.806924	24.690795	23.36258	112.14015
monk#1	10.905597	0.87631476	4.7297658	23.009869	13.991019	60.98442
oracle#1	11.416422	0.87631476	4.732977	24.11991	18.649933	89.519634
cemetery#1	10.569124	0.9281574	4.797955	21.975212	10.998512	49.448303
woodland#1	10.723275	0.7945633	4.6132636	9.152182	9.976696	33.255653
food#1	5.824389	0.35403305	2.3362055	6.8669333	0.8296604	1.6593209
rooster#1	11.416422	0.9281574	4.8116277	23.660517	32.15455	209.00496
coast#1	9.470511	0.8564059	4.6212634	10.749784	2.2071205	6.6215415
hall#2	7.545221	0.7782966	4.49465	15.622799	7.4436406	26.092792
forest#2	10.723275	0.7945633	4.6132636	9.152182	9.976696	33.255653
graveyard#1	10.569124	0.9281574	4.797955	21.975212	10.998512	49.448303
shore#1	8.881473	0.8144722	4.50129	9.651173	2.0443257	5.110975
woodland#1	10.723275	0.7945633	4.6132636	9.152182	9.976696	33.255653
monk#1	10.905597	0.87631476	4.7297658	23.009869	13.991019	60.98442
slave#2	11.821897	1.0	4.797955	21.675455	19.571359	52.599149
coast#1	9.470511	0.8564059	4.6212634	10.749784	2.2071205	6.6215415
forest#2	10.723275	0.7945633	4.6132636	9.152182	9.976696	33.255653
lad#1	8.471993	0.9281574	4.797955	19.843245	12.978389	49.903755
vizard#1	10.030128	0.9281574	4.797955	21.183018	12.937442	50.132397
chord#2	11.416422	0.8564059	4.637584	19.249523	14.922856	74.64428
smile#1	8.877448	0.9019918	4.7469023	13.187252	7.5411134	30.164454
glass#1	9.049299	0.73729706	4.2416525	8.946375	1.4552382	3.6390654
magician#1	10.905597	0.9019918	4.764531	22.610134	19.22327	88.300835
noon#1	10.81781	1.0	4.806924	14.876495	25.269993	139.97946
string#1	10.569124	0.9019918	4.7469023	17.409079	10.112452	40.449906

Tableau 4.2 les valeurs de CI calculés par différentes approches

Le tableau suivant montre les CI des concepts cités plus hauts avec notre formule de calcul de CI :

Le concept	Le CI
car	23.83836
automobile	23.83836
gem	15.7257395
jewel	21.488337
journey	22.279089
voyage	26.069939
boy	19.446882
lad	21.526323
coast	17.136726
shore	15.50242
asylum	20.950958
madhouse	25.280096
magician	22.376791
wizard	21.526323
midday	25.24009
noon	25.24009
food	9.417399
fruit	17.988047
furnace	19.392815
stove	25.05632
bird	16.936367
cock	22.87463
bird	16.936367
crane	11.348269
tool	16.806593
implement	14.845409
brother	23.318684
monk	21.395962
crane	11.348269
implement	14.845409
lad	21.526323
brother	23.318684
journey	22.279089
car	23.83836
monk	21.395962
oracle	23.29271
cemetery	22.715565
woodland	19.265192
food	9.417399
rooster	27.507153
coast	17.136726
hill	16.494873
forest	19.3407
graveyard	22.715565
shore	15.50242
woodland	19.265192
monk	21.395962
slave	18.533894

coast	17.136726
forest	19.3407
lad	21.526323
wizard	21.526323
chord	21.93179
smile	21.498152
glass	15.214738
magician	22.376791
noon	25.24009
string	21.27501
rooster	27.507153
voyage	26.069939

Tableau 4.3 les valeurs de CI calculés par notre approche

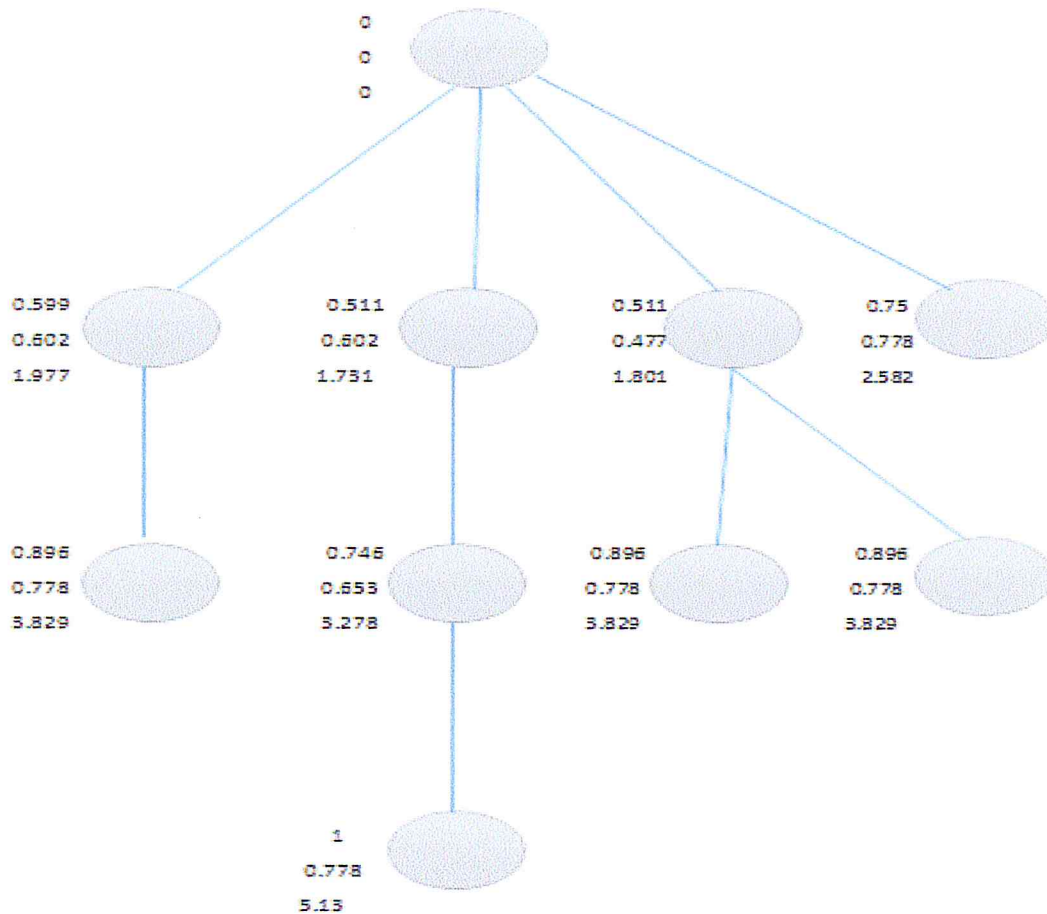


Figure 4.4 exemple de calcul de CI



**Line 01 : résultat par la méthode de Zhou (basé sur les hyponymes)****Line 02 : résultat par la méthode de Sanchez (basé sur les feuilles)****Line 03 : résultat par notre méthode (basé sur la hauteur)**

L'avantage de notre méthode est qu'elle est inclut la hauteur, si deux nœuds ont le même nombre de feuilles et la même profondeur, le plus proche de l'anti racine est celui qui a le CI grand, et comme ça on a balayé le problème de la méthode de Sanchez : si deux nœuds ont le même nombre de feuilles ils ont toujours le même CI.

Le tableau suivant nous donne les valeurs des similarités sémantiques pour chaque couple dans la liste des jugements humains de [Miller & Charles, 1991] de 30 couple de concepts avec les différentes approches couplées avec les trois mesures de calcul des similarités sémantiques basées sur le contenu informationnel [Resnik, 1995],[Jiang et Conrath, 1997],La mesure de [Lin, 1998]. (Source : [Ferid et khaddem, 2014]).

les couples des concepts	ic resnik			ic seco			resnik
	resnik	jiang	lin	resnik	jiang	lin	
car#1--automobile#1	** 6.0263293	0.16595227	1.0	** 0.6708781	1.4905938	1.0	** 4.3044404
asylum#2--madhouse#1	** 11.416422	0.09458886	0.9525519	** 0.9381574	1.0	0.5680921	** 4.8023196
midday#1--noon#1	** 10.31781	0.0969199	1.0	** 1.0	1.0	1.0	** 4.805524
rooster#1--voyage#1	** 0.0	0.046627548	0.0	** 0.0	0.5404072	0.0	** 0.0
gem#2--jewel#2	** 10.405593	-0.095826886	0.0	** 0.7845633	0.3295749	0.7845633	** 4.592835
journey#1--voyage#2	** 5.735112	0.09458886	0.6503135	** 0.6503692	1.1096699	0.8383641	** 4.1775046
boy#1--lad#2	** 7.4511537	0.09169604	0.81409966	** 0.78729708	1.0	0.8487864	** 4.4177407
coast#1--shore#1	** 8.851473	0.10558092	0.5662133	** 0.8144722	1.1676706	0.57490317	** 4.50128
magician#2--wizard#2	** 11.521897	0.09458886	1.0	** 0.8144722	1.227789	1.0	** 4.651827
journey#1--car#1	** 0.0	0.0850272	0.0	** 0.0	0.75651705	0.0	** 0.0
furnace#1--stove#1	** 2.4759674	0.049524995	0.21451658	** 0.1747369	0.78316336	0.22712421	** 1.5881777
food#1--fruit#2	** 1.402681	0.044961717	0.11864697	** 0.1655667	0.56414604	0.17084959	** 1.4677257
bird#1--cock#5	** 7.655222	0.07990389	0.75936044	** 0.38591002	1.0659199	0.5935382	** 2.5917165
bird#1--crane#5	** 7.655222	0.0	2.0	** 0.38591002	1.0659199	0.5935382	** 2.5917165
tool#1--implement#1	** 7.019776	0.121383265	0.91971016	** 0.41250476	2.0915115	0.9268235	** 2.907655
brother#5--monk#1	** 10.905597	0.0	2.0	** 0.97631476	1.0	0.5340803	** 4.7297688
crane#5--implement#1	** 1.4361854	0.17928901	0.40958386	** 0.08231713	0.788429	0.121383151	** 0.9228965
lad#2--brother#2	** 2.4185324	0.052055303	0.22069326	** 0.21123092	0.59145296	0.22211604	** 1.5007834
monk#1--oracle#1	** 2.4185324	0.050242458	0.21669479	** 0.21123092	0.64976115	0.24304898	** 1.5007834

Nouvelle approche de calcul de CI

cemetery#1--woodland#1	**	1.4361394	0.05046208	0.1349016	**	0.05231713	0.6059124	0.09501488	**	0.9223995
food#1--rooster#1	**	0.9090739	0.06316438	0.103605914	**	0.056359116	0.8091538	0.037229006	**	0.93623766
coast#1--hill#2	**	1.4361394	0.06413674	0.16380724	**	0.05231713	0.6441699	0.10071206	**	0.9223995
forest#2--graveyard#1	**	1.4361394	0.05046208	0.1349016	**	0.05231713	0.6059124	0.09501488	**	0.9223995
shore#1--woodland#1	**	1.4361394	0.055131175	0.14673999	**	0.05231713	0.6549997	0.10231959	**	0.9223995
monk#1--slave#2	**	2.4189324	0.04926937	0.21282389	**	0.21123032	0.60057014	0.22515446	**	1.9007304
coast#1--forest#2	**	1.4361394	0.05331173	0.14224072	**	0.05231713	0.63749	0.099719755	**	0.9223995
lad#1--wizard#1	**	2.4189324	0.06217532	0.26143315	**	0.21123032	0.60057014	0.22515446	**	1.9007304
chord#2--smile#1	**	6.435633	0.072130375	0.6346437	**	0.60626655	0.66597997	0.6965692	**	3.8627356
glass#1--magician#1	**	0.9090739	0.052904957	0.09111237	**	0.056359116	0.63174397	0.063760364	**	0.93623766
noon#1--string#1	**	0.0	0.04787632	0.0	**	0.0	0.5257674	0.0	**	0.0

les couples des concepts

	ic ranches#2			ic shou						
	resnik	jiang	lin	resnik	jiang	lin				
car#1--automobile#1	**	3.807241	0.23311629	1.0	**	0.7574051	1.3202974	1.0	**	0.84373686
asylum#2--madhouse#1	**	4.830733	0.20446624	1.0	**	0.3600344	1.1023042	0.97233324	**	0.7755408
midday#1--noon#1	**	4.830733	0.20446624	1.0	**	0.9071905	1.1023043	1.0	**	0.81433106
rooster#1--voyage#1	**	0.0	0.10354301	0.0	**	0.0	0.5625532	0.0	**	0.0
gem#2--jewel#2	**	4.015722	0.13217462	0.34497336	**	0.77039592	1.1223953	0.92769314	**	0.74533335
journey#1--voyage#2	**	2.3624902	0.21449424	0.33302533	**	0.73267314	1.1455319	0.9126314	**	0.81622525
boy#1--lad#2	**	3.7344572	0.23621145	0.37273314	**	0.72176134	1.1453254	0.9093009	**	0.7333737
coast#1--shore#1	**	3.362356	0.2443337	0.3730073	**	0.6303374	1.0632333	0.96323446	**	0.54353374
magician#2--wizard#2	**	4.3663072	0.2403032	1.0	**	0.76034545	1.3153247	1.0	**	0.70526673
journey#1--car#1	**	0.0	0.14556339	0.0	**	0.0	0.6711043	0.0	**	0.0
furnace#1--stove#1	**	0.90311346	0.14742335	0.23343239	**	0.36066967	0.84476344	0.46706042	**	0.5466322
food#1--fruit#2	**	0.9030453	0.11502417	0.13913325	**	0.31313227	0.7236975	0.37644246	**	0.47031743
bird#1--cock#5	**	2.1363246	0.20446624	0.31397426	**	0.5710633	1.1622766	0.73305446	**	0.7462227
bird#1--crane#5	**	2.1363246	0.2014773	0.6057332	**	0.5710633	1.1222781	0.7311356	**	0.7462227
tool#1--implement#1	**	2.1123973	0.40742722	0.325461	**	0.5366306	1.6336902	0.9503495	**	0.6633637
brother#5--monk#1	**	4.4136615	0.20446624	0.54373212	**	0.31327167	1.1222772	0.95316035	**	0.7463465
crane#5--implement#1	**	0.47333245	0.13150303	0.13447717	**	0.22773572	0.3333197	0.1335343	**	0.3731142
lad#2--brother#2	**	1.0373373	0.11473392	0.22339396	**	0.40337676	0.76614602	0.47739543	**	0.6033324
monk#1--oracle#1	**	1.0373373	0.12373227	0.24743996	**	0.40337676	0.8126334	0.49373724	**	0.6033324
cemetery#1--woodland#1	**	0.47333245	0.11333271	0.10673143	**	0.22773572	0.74534732	0.23033333	**	0.3731142
food#1--rooster#1	**	0.24423073	0.13377333	0.07240463	**	0.143334	0.3333336	0.22332771	**	0.23333336

coast#1--hill#2	**	0.47580245	0.10302197	0.11970497	**	0.22771572	0.8187497	0.31044697	**	0.8701142
forest#2--graveyard#1	**	0.47580245	0.11855271	0.10079149	**	0.22771572	0.74504708	0.29009582	**	0.8701142
shore#1--woodland#1	**	0.47580245	0.10371408	0.119603496	**	0.22771572	0.8470439	0.32780074	**	0.8701142
monk#1--slave#2	**	1.0970372	0.10314612	0.23804626	**	0.40937676	0.79712504	0.49252601	**	0.6035224
coast#1--forest#2	**	0.47580245	0.10024547	0.11670529	**	0.22771572	0.8118756	0.31159593	**	0.8701142
lad#1--wizard#1	**	1.0970372	0.116233775	0.22622694	**	0.40937676	0.8100057	0.49351707	**	0.6035224
chord#2--smile#1	**	0.007931	0.17051194	0.49112153	**	0.60225144	1.0000191	0.77894424	**	0.6607936
glass#1--magician#1	**	0.24426073	0.1279078	0.060592614	**	0.148394	0.75747764	0.19901374	**	0.52460336
noon#1--string#1	**	0.0	0.10645789	0.0	**	0.0	0.58405235	0.0	**	0.0

les couples des concepts

	ic sebi			ic hadj saiebi						
	resnik	jiang	lin	resnik	jiang	lin	resnik			
car#1--automobile#1	**	24.315497	0.007614439	1.0	**	24.049904	0.003051417	1.0	**	207.2636
asylum#2--madhouse#1	**	22.059349	0.045523076	1.0	**	14.782128	0.0623649	0.56016225	**	70.910445
midday#1--noon#1	**	14.876495	0.06722018	1.0	**	26.269998	0.00957419	1.0	**	108.97946
rooster#1--voyage#1	**	0.0	0.019237608	0.0	**	0.0	0.012026038	0.0	**	0.0
gem#1--jewel#2	**	20.591928	0.0699921	1.1802678	**	12.042009	0.07490303	0.94010228	**	54.190325
journey#1--voyage#2	**	22.494005	0.00957795	0.9419482	**	21.932652	0.020502462	0.9103422	**	120.90462
boy#1--lad#2	**	19.840245	0.04496173	0.9402017	**	12.878183	0.07108302	0.9388887	**	49.910733
coast#1--shore#1	**	9.651173	0.09102512	0.946149	**	2.0440197	0.4530667	0.96171045	**	5.110975
magician#2--wizard#2	**	18.821642	0.051022112	1.0	**	12.759462	0.0701642	1.0	**	49.59416
journey#1--car#1	**	0.0	0.0092075105	0.0	**	0.0	0.017340097	0.0	**	0.0
furnace#1--stove#1	**	7.7052627	0.028493637	0.4100212	**	0.2048500	0.024221920	0.14606037	**	4.0114765
food#1--fruit#2	**	0.1730539	0.04247352	0.20796916	**	2.4728124	0.060092875	0.2597473	**	4.945669
bird#1--cock#5	**	18.717808	0.045504634	0.9199781	**	20.500105	0.040912012	0.9478527	**	92.26039
bird#1--crane#5	**	18.717808	0.007770465	0.3293403	**	20.500105	0.001591355	0.8049644	**	92.26039
tool#1--implement#1	**	14.150312	0.04837393	0.8927174	**	9.467469	0.007712026	0.959495264	**	88.032246

brother#5--monk#1	** 20.003469	0.041479278	0.974694	** 14.981019	0.06875114	0.90547204	** 40.53442
crane#5--implement#1	** 2.7080503	0.02827225	0.10231283	** 0.5016906	0.025817235	0.026874391	** 0.7975209
lad#2--brother#2	** 11.466462	0.02324627	0.49526926	** 0.05916	0.034246632	0.4305769	** 24.174479
monk#1--oracle#1	** 11.466462	0.02324627	0.49526926	** 0.05916	0.034246632	0.4305769	** 24.174479
semetary#1--woodland#1	** 2.7080503	0.03187304	0.17039788	** 0.5016906	0.049369276	0.05072028	** 0.7975209
food#1--rooster#1	** 1.0596123	0.02904542	0.00194583	** 0.0	0.00021754	0.0	** 0.0
coast#1--hill#2	** 2.7080503	0.04229722	0.2052636	** 0.5016906	0.10985533	0.110130065	** 0.7975209
forest#2--graveyard#1	** 2.7080503	0.03187304	0.17039788	** 0.5016906	0.049369276	0.05072028	** 0.7975209
shore#1--woodland#1	** 2.7080503	0.04229922	0.2052636	** 0.5016906	0.08700671	0.033457394	** 0.7975209
monk#1--slave#2	** 11.466462	0.02009755	0.5021634	** 0.05916	0.05129716	0.53452024	** 24.174479
coast#1--forest#2	** 2.7080503	0.06165109	0.27217295	** 0.5016906	0.08520785	0.03727609	** 0.7975209
lad#1--wizard#1	** 11.466462	0.0322970	0.5583916	** 0.05916	0.058212694	0.6242305	** 24.174479
chord#2--smile#1	** 9.798127	0.046210612	0.6203545	** 10.406654	0.0328774	0.5280318	** 16.413887
glass#1--magician#1	** 1.0596123	0.0028221	0.04922325	** 0.0	0.049369276	0.0	** 0.0
noon#1--string#1	** 0.0	0.000973535	0.0	** 0.0	0.023265401	0.0	** 0.0

Tableau 4.4 les valeurs des similarités sémantiques pour chaque couple dans la liste des jugements humains de [Miller & Charles, 1991] calculés par notre les différentes approches

Le tableau suivant montre les similarités entre concepts cités plus hauts avec notre formule de calcul de CI :

Les paires de mots	La similarité avec Lin (notre CI)
car automobile	0.9267352
gem jewel	0.5708619
journey voyage	0.840773
boy lad	0.9285212
coast shore	0.6926497
asylum madhouse	0.9353827
magician wizard	0.840345
midday noon	0.7522686
food fruit	0.6017082
furnace stove	0.42917043
bird cock	0.7232468
bird crane	1.0179794
tool implement	1.3676295
brother monk	0.84986
crane implement	0.3462368
lad brother	0.47372267
journey car	0.0
monk oracle	0.4753799

cemetery	woodland	0.21603268
food	rooster	0.13989346
coast	hill	0.26966354
forest	graveyard	0.2156448
shore	woodland	0.26085243
monk	slave	0.5320354
coast	forest	0.24862544
lad	wizard	0.49344462
chord	smile	0.78975093
glass	magician	0.13741137
noon	string	0.0
rooster	voyage	0.0

Tableau 4.5 les valeurs des similarités sémantiques pour chaque couple dans la liste des jugements humains de [Miller & Charles, 1991] calculés par notre formule de CI couplé avec Lin

### 5. Expérimentation :

Afin d'évaluer notre nouvelle approche, nous avons utilisé les formules de [Resnik, 1995], [Jiang & Conrath, 1997], et [Lin, 1998] (celles basées sur le CI) afin de mesurer la similarité sémantique entre mots, nous avons utilisé notre méthode de calcul de CI pour quantifier la similarité entre deux concepts  $c1$ ,  $c2$ . Comme nous avons vu précédemment, la mesure de Lin est la plus fiable car elle a donné une bonne corrélation avec le jugement humain.

- Les résultats cités dans les paragraphes suivants sont obtenu par l'utilisation de la mesure de Lin

#### 5.1. Balayage des limites :

Lors de ce paragraphe nous montrons que la méthode proposée pour le calcul du CI fournit des solutions pour les limites trouvées dans les autres méthodes (Nuno, Sebti..). La figure suivante représente un extrait de wordnet 2.1 :

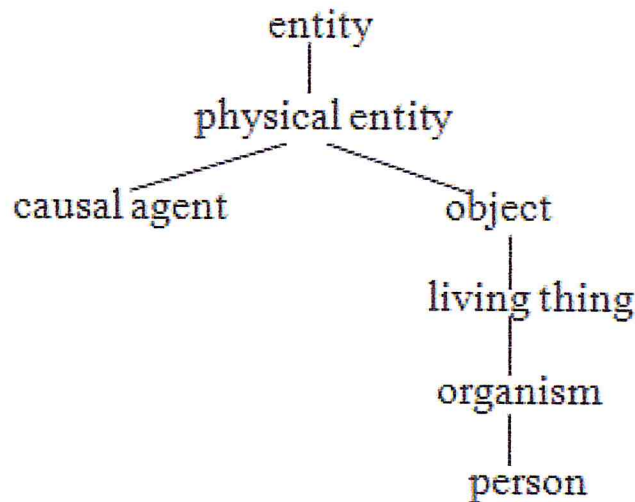


Figure 4.5 extrait N° 05 de wordnet 2.1

Limite1 : le tableau ci-dessus nous donne le CI des concepts dans la figure 4.4 :

Concepts	CI
Entity	0.0
Physical entity	1.57
Object	4.52
Causal agent	5.64
Living thing	8.80
Organism	10.12
Person	12.14

Tableau 4.6 les contenus informationnels calculés avec notre formule

D'après le tableau 4.1, nous remarquons que  $CI(\text{object}) \neq CI(\text{causal agent})$ . Mais, quant à la méthode de Sebti ils sont égaux.

Limite2 : le tableau ci-dessus nous donne les CI de deux concepts dans la figure 4.1

CI (obidoxime chloride)	CI (atropine)
24.89	26.14

Tableau 4.7 les CI des concepts *obidoxime chloride* et *atropine*

La différence remarquable au niveau des sous arbres est bien traduite au niveau de la quantification du CI. De plus, chaque concept appartenant au sous arbre participe au contenu informationnel du concept cible d'où nous avons remédié au problème des concepts neutres dans la méthode de Sebti et al.

Limite3 : voici l'extrait de wordnet 2.1 discuté dans le chapitre 3

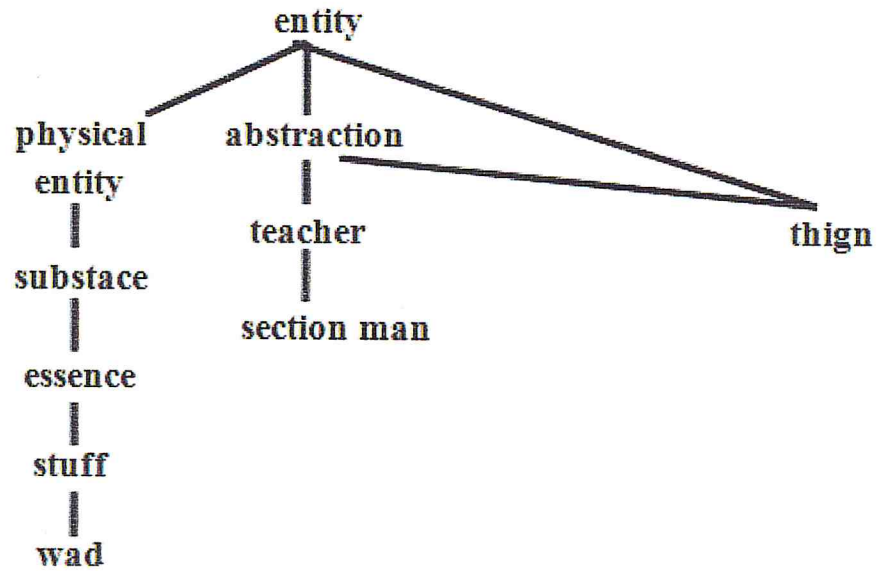


Figure3.2 : extrait de wordnet 2.1

Le tableau ci-dessus nous donne les CI de quelques concepts de la figure précédente :

Concepts	CI
Thing	19.60
Teacher	22.66
Substance	11.31
Section man	26.48
Wad	20.84

Tableau 4.8 les CI des concepts discutés dans la figure 3.2

La déférence entre le CI (section man) et CI (thing) est remarquable d'où nous avons balayé le problème de la méthode de Nuno. Ainsi, CI (teacher) est supérieur à CI (substance) malgré que (substance) est plus approfondie que (teacher). Donc, notre méthode garantit l'influence de la hauteur sur le calcul du CI.

## 5.2 Comparaison avec des estimations humaines de similitude :

Lors de ce paragraphe on va présenter la comparaison de notre approche avec le jugement humain.

Il y a des benchmark utilisés pour la valorisation d'une mesure de similarité sémantique. En effet, Rubenstein et Goodenough (1965) obtiennent des jugements de similitudes de 51 experts concernant 65 paires de mots. Les experts ont été demandés d'affecter un score de 0 à 4 en relation avec leurs similarités au niveau sens. Miller et Charles (1991) ont extrait 30 paires des 65 originaux et ils ont obtenu les avis de 38 experts (score entre 0 et 4).

Dans le but de comparer notre mesure avec celles proposées par les autres auteurs, nous avons utilisé WordNet 2.1, ainsi nous avons couplé notre méthode avec les trois formules de similarité citées plus haut, nous avons obtenu les résultats (coefficient de corrélation) suivants :

Formules de similarité	M&C
Notre avec Resnik	0.74
Notre avec Jiang and Conrath	0.73
Notre avec Lin	0.74

Tableau 4.9 comparaison des résultats obtenus par notre formule avec les différentes formules de similarité

Le tableau ci-dessus montre les coefficients de corrélations entre le jugement humain et les mesures de similarités proposées, notre formule pour le calcul de CI est couplée avec la mesure de Lin, les résultats pour (Resnik, Lin et Jiang & Conrath) sont pris selon celles de [Alexander Budanitsky, 1999]:

Mesure de similarité	M&C	R & G
Lin	0.83	0.82
Jiang & Conrath	0.85	0.78
Resnik	0.77	0.78
Notre	<b>0.74</b>	<b>0.79</b>

Tableau 4.10 Les coefficients de corrélation entre les jugements humains de similarité (Miller & Charles et Rubenstein & Goodenough) et les mesures de similarité proposées.

D'après le tableau ci-dessus, nous pouvons dire que notre méthode de calcul de contenu informationnel nous a permis de générer une nouvelle mesure de similarité sémantique utilisant la ressource sémantique WordNet qui a montré ses performances vis-à-vis les autres mesures et comme vous remarquez, la nouvelle mesure s'améliore pour un ensemble de test plus riche (M&C → R&G).

La figure 4.5 nous donne une représentation graphique de la comparaison entre le jugement humain et notre méthode et celle de Lin avec les paires de mots M&C (30 paires):



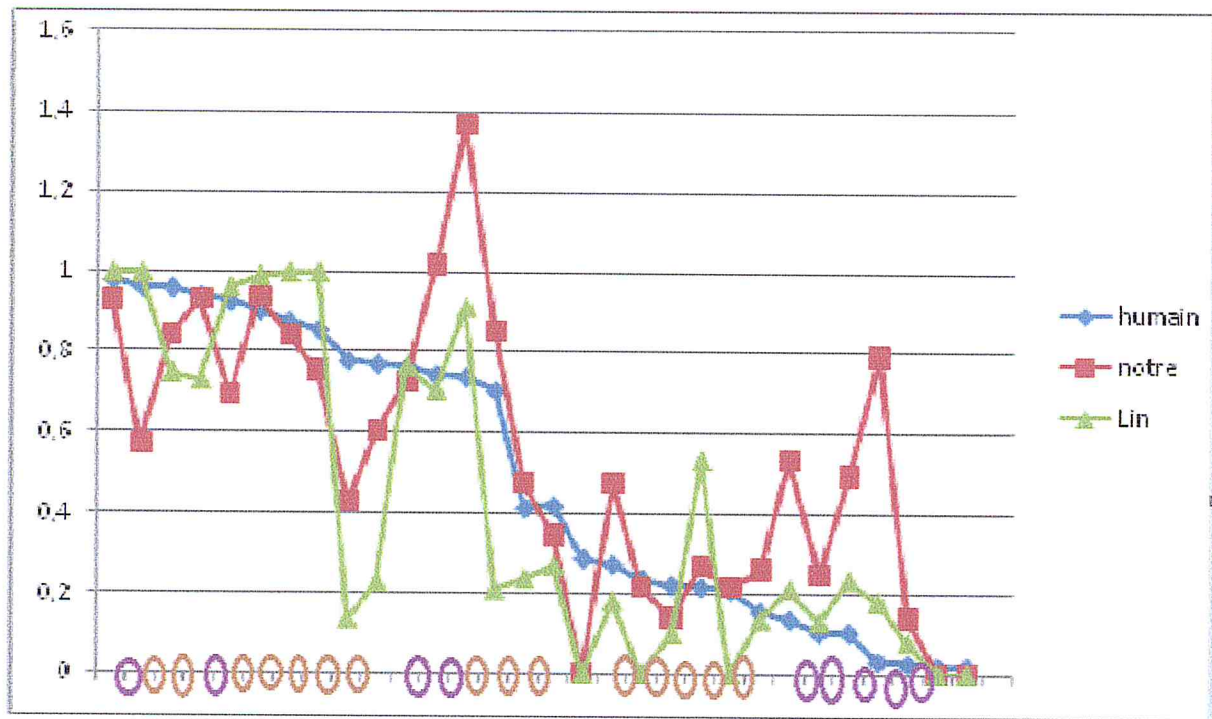


Figure 4.6 Comparaison de notre méthode avec celle de Lin et le jugement humain

Dans la figure 4.6 on voit deux couleurs de cercles, marrons et mauve, les cercles marron représentent les paires de mots où notre méthode est mieux que celle de Lin. Les cercles mauves représentent les paires de mots où celle de Lin est la meilleure. Le reste des paires les deux méthodes sont identiques.

## 6. Implémentation :

### 6.1. Langage utilisé :

Le choix du langage de programmation représente une étape très importante dans la réalisation de n'importe quelle application.

Pour la réalisation de notre travail nous avons utilisé Le langage de programmation JAVA :

Java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton de Sun Microsystems. Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitation (Windows, Linux, Macintosh, Solaris).


Nous avons choisi le langage JAVA pour les raisons suivantes :


- Un langage orienté objet qui met à la disposition du développeur plusieurs paquetages prêt à l'utilisation (java.util, java.io....etc).
- Offre une encapsulation, la généricité et la réutilisation de notre code métier plus facilement car il offre beaucoup de souplesse.
- La disponibilité de la documentation et de l'assistance (forums).


- Java est un langage à haute sécurité.
- Java est un langage simple.
- Java est portable.

## 6.2 Description des fonctions de notre application :

Tout d'abord, on a importé les bibliothèques :

 edu.mit.jwi\_2.3.3\_jdk.jar

 jaws-bin.jar

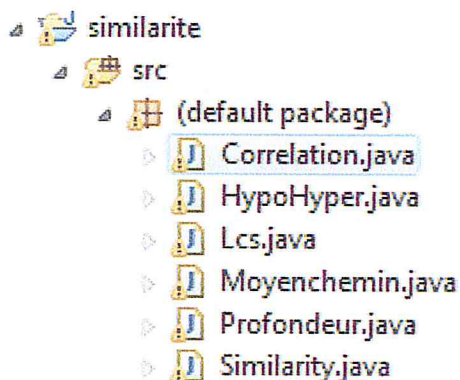
 jaws-bin-1.2.jar

Ces bibliothèques contiennent des packages qui manipulent les fonctions de wordnet.

Pour connecter le langage Java avec WordNet, on a utilisé la fonction suivante :

```
System.setProperty("wordnet.database.dir",
    "C:\\Program Files (x86)\\WordNet\\2.1\\dict");
```

Cette fonction contient l'emplacement du fichier de base (base de données), ainsi que le lien exacte de l'emplacement du fichier dictionnaire dans le dossier «WordNet».



Dans notre projet «*similarite*», on a les classes suivantes :

- a. La classe HypoHyper qui contient les fonctions :
  - la fonction `gethypernymes` : retourne le nombre de concepts dans l'arbre d'hypernymes.
  - La fonction `gethyperform`: retourne les concepts pères d'un concept *c* sous forme de *String*.
  - La fonction `gethyponymes`: retourne le nombre de concepts dans l'arbre d'hyponymes.
  - La fonction `leaves` : retourne le nombre de feuilles racinées par *c*.
- b. La classe Moyenchemin qui contient les fonctions :

- La fonction `getprofondeur` : retourne la profondeur (en nombre d'arcs) du  $c$  dans l'ontologie WordNet.
  - La fonction `gethauteur` : retourne la hauteur (en nombre d'arcs) du  $c$  dans l'ontologie WordNet.
- c. La classe `Lcs` qui contient la fonction :
- La fonction `getLcs` : retourne le concept le plus petit générale de deux concepts  $c1, c2$  sous format de *String*.
- d. La classe `Similarity` qui contient les fonctions :
- La fonction `InformationContent` : retourne la valeur du contenu informationnel du  $c$  calculée par notre formule.
  - La fonction `LinSimilarity` : retourne la similarité entre deux concepts  $c1, c2$  par la formule de Lin couplée par notre de formule de calcul de CI.
  - La fonction `ResnikSimilarity` : retourne la similarité entre deux concepts  $c1, c2$  par la formule de Resnik couplée par notre de formule de calcul de CI.
  - La fonction `JianAConSimilarity` : retourne la similarité entre deux concepts  $c1, c2$  par la formule de Jiang et Conrath couplée par notre de formule de calcul de CI.
- e. La classe `Correlation` qui contient les fonctions :
- La fonction `calculCorrelation` : calcul le coefficient de corrélation entre notre mesure et le jugement humain en utilisant la formule :

$$rp = \frac{(x-\bar{x})*(y-\bar{y})}{\sqrt{(x-\bar{x})^2*(y-\bar{y})^2}}$$

Avec :  $rp$  est le coefficient de corrélation

## 7. Conclusion

Dans ce travail nous avons présenté une nouvelle méthode de calcul du contenu informationnel en utilisant la ressource sémantique WordNet. Cette approche exploite la taxonomie « is a » (hyperonyme/hyponyme) sans faire recours à des corpus externes. Cette méthode a répondu à certaines limites présentes dans les autres méthodes. Ce CI intégré dans une mesure de similarité sémantique (Lin) a montré ses performances avec les jugements humains M&C et R&G.

# *Conclusion générale*

### Conclusion générale :

Pour intégrer la notion de voisinage sémantique, et plus précisément la similarité sémantique; nous avons utilisé l'ontologie de concepts WordNet auxquels sont reliés les termes par des liens sémantique. Dans notre travail nous n'avons pris en considération que les liens de spécialisation/généralisation (le réseau « is-a ») entre les concepts. En nous basant sur les mesures de similarité entre concepts présentées par [Resnik, 1995], [Jiang et Conrath, 1997], et essentiellement [Lin, 1998], nous avons proposé une nouvelle mesure de contenu informationnel (CI) telle que les descendants d'un concept sont considérés, les ascendants, la profondeur, et le nouveau facteur la hauteur. Nous avons alors défini un nouveau calcul du poids des termes (concepts) qui tient compte de la position exacte du concept dans une ontologie.

Une des limites de notre approche, tient au fait que nous avons travaillé seulement sur le réseau de spécialisation/généralisation « is-a » de WordNet. Rappelons que nous nous plaçons dans le cadre de la mesure de la similarité sémantique, pour laquelle on peut supposer qu'il existe certains éléments en communs (propriétés, termes, instances) entre concepts. Pour utiliser le modèle présenté dans cet article, il suffit de disposer d'une structure hiérarchique (plus précisément ontologie) entre concepts correspondant aux liens de spécialisation/généralisation. Cependant, nous sommes conscients que le calcul de la quantité d'information fournit par un concept par restriction sur le lien « is-a » n'est pas toujours bien adapté, les autres types de liens peuvent être aussi importants dans le calcul de la similarité. Nous envisageons de travailler sur la prise en compte d'autres types de liens comme par exemple les liens lexicaux (antonymie par exemple). De plus, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent être plus denses que d'autres. Ces problèmes peuvent être résolus, en partie, en associant des poids aux liens. L'affectation de ces poids peut être basée sur : les types de liens présents, la profondeur du lien dans la taxonomie. Etc.

L'évaluation de notre mesure de similarité est nécessaire pour tester son efficacité. En effet la corrélation que donne une telle mesure est un facteur important pour qu'elle soit fiable ou pas. Parmi les approches qui existent pour tester l'efficacité des mesures de similarité on trouve la comparaison de ces mesures par rapport à un jugement humain, mais il est difficile de mettre en place de telles expérimentations qui porteraient sur un ensemble assez significatif de concepts. Dans notre travail on a testé l'efficacité de notre mesure par coupler notre approche de calcul de CI avec la mesure de Lin, et ensuite la comparaison au jugement humain, avec l'utilisation de deux ensembles de concepts : le premier contient 30 paires de mots sur 38 sujets différents (M&C), le deuxième contient 65 paires de mots sur 51 sujets (R&G).

Comme future travail, nous visons de prendre en compte les différents liens dans une ontologie (WordNet) pour calculer la quantité d'information donnée par un concept. Ainsi, nous voulons que notre mesure soit prise en considération dans une application réelle de traitement automatique de langages et de la recherche d'information.

## Références bibliographiques

- [Alexander Budanitsky, 1999] Alexander Budanitsky, (1999). Lexical semantic relatedness and its application in natural language processing. Computer system research group, University of Toronto, August 1999.
- [Collins, 1975] Collins, A., & Coftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82, 407-428.
- [Cross et al, 2012] Valarie Cross, Anurekha Chennai-Thiagarajan (2012). Measuring information content for an ontological concept. *Computer Science and Software Engineering* Miami University Oxford, OH USA.
- [Fellbaum, 1998] Fellbaum Christiane, A Semantic Network of English Verbs, *WordNet an Electronic Lexical Database*, MIT Press, 1998, pp. 69-103.
- [Ferid et khaddem, 2014] étude de performance des approches de calcul de contenu informationnel (2014). Département d'informatique. Université de Blida, Algérie.
- [Florentina Vasilescu, 2003] Florentina Vasilescu, (2003). Désambiguïsation de corpus monolingues par des approches de type Lesk. Université de Montréal, aout 2003.
- [FURNAS, 1987] G.W.FURNAS, LANDAUERT. GOMEZL. DUMAISS., « The vocabulary problem in human-system communication », *Communications of the Association for Computing Machinery*, vol. 30, 1987.
- [G.A. Miller 1998] Miller George A., Nouns in WordNet, *WordNet an Electronic Lexical Database*, MIT Press, 1998, pp. 23-45.
- [Hadj Taieb et al.2012] Mohamed Ali HADJ TAIEB, Mohamed BEN AOUICHA, et Abdelmajid BEN HAMADOU, (2012). Une nouvelle approche de calcul du contenu informationnel pour mesurer la similarité sémantique utilisant WordNet. Laboratoire MIRACL (MultiMedia, Information Systems and Advanced Computing Laboratory) Université de Sfax, Tunisie.

- [Hirst & St-Onge 1998] Hirst G. & St Onge D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press .
- [Jiang et Conrath, 1997] Jiang, J. J., & Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of the International Conference on Research in Computational Linguistics, ROCLING X (pp. 19-33). Taipei, Taiwan.
- [K.j Miller, 1998] Miller Katherine J., Modifiers in WordNet, WordNet an Electronic Lexical Database, MIT Press, 1998, pp. 47-67.
- [Kozima, 1994] Kozima, H. "Computing Lexical Cohesion as a Tool for Text Analysis". doctoral thesis Computer Science and Information Math , Graduate School of Electro-Comm., Univ. Of Electro-Comm., 1994.
- [Leacock et Chodorow, 1998] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for wordsense identification WordNet: An electronic lexical database (pp. 265-283): MIT Press.
- [Lin, 1998] Lin, D. (1998). An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998 (pp. 296-304). Madison, Wisconsin, USA.
- [Lin, 2003] Lin C. Y., and Hovy E. "Automatic evaluation of summaries using ngram co-occurrence statistics". In Proceedings of Human Language Technology Conference (HLT-NAACL) , Edmonton, Canada, 2003.
- [Miller et Charles 1991] Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1).
- [Nuno 2004] Seco, N., Veale, T., & Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In Proceedings of the 16th European Conference on Artificial Intelligence, ECAI 2004,

including Prestigious Applicants of Intelligent Systems, PAIS 2004 (pp. 1089-1090). Valencia, Spain.

**[Quillian, 1968]**

M. Quillian, *Semantic Memory*, 227–270, MIT Press, 1968.

**[Rada et al. 1989]**

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and Application of a metric on semantic nets. *IEEE Transaction on systems, Man, and Cybernetics*, 19 (1), 17-30.

**[Resnik, 1995]**

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995*. Montreal, Quebec, Canada.

**[Resnik, 1999]**

Resnik, P. "Semantic Similarity in a Taxonomy: An Information-Based Measure and Its Application to Problems of Ambiguity in Natural Language". *J. Artificial Intelligence Research*, vol. 11, pp. 95-130, 1999.

**[Sanchez, 2010]**

David Sanchez, (2010). *Ontology-based information content computation*. Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Spain

**[Sebti et al.2008]**

Ali Sebti et Ahmad Abodollahzadeh, (2008). A new word sense similarity measure in wordnet. Amirkabir university of technology intelligence systems laboratory, Tehran, Iran.

**[Siddharth Patwardhan 2003]**

Siddharth Patwardhan, (2003). *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. UNIVERSITY OF MINNESOTA, August 2003.

**[Smeulders et al., 2000]**

Smeulders A. W. M., Worring M., Santini S., Gupta A., and Jain R., "Content-Based Image Retrieval at the End of the Early Years". *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

**[Tengi, 1998]**

Tengi Randee I., *Design and Implementation of the WordNet Lexical DataBase and Searching Software*, WordNet an Electronic Lexical Database, MIT Press, 1998, pp105-127.



[Zargayouna, 2004]

Zargayouna (a) H. Contexte et sémantique pour une indexation de documents semi-structurés. ACM Conférence en Recherche Information et Applications, CORIA'2004.

[Zargayouna, 2004b]

Zargayouna (b) H., Salotti S. (2004) Mesure de similarité sémantique pour l'indexation de documents semistructurés dans 12ème Atelier de Raisonement à Partir de Cas, Mars 2004.

[Zargayouna, 2005]

Zargayouna, (2005). Indexation sémantique de documents XML. Université Paris 11, 2005.

[Zhou et al. 2008]

Zili Zhou, Lingling Meng, and Junzhong Gu, (2008). A review of information content metric for semantic similarity. Computer Science and Technology Department, Shanghai, 200062, China.

[Wu & Palmer, 1994]

Wu Z. & Palmer M. (1994) Verb Semantics and Lexical Selection, Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics.

## Références Webographie

[w01] :

<http://fr.wikipedia.org/wiki/wordnet>

