WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture

François-Régis Chaumartin

WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture

François-Régis Chaumartin

Société Proxem
7 impasse Dumur
92110 Clichy – France
frc@proxem.com

Lattice/Talana – Université Paris 7
30 rue du château des rentiers
75013 Paris - France
fchaumartin@linguist.jussieu.fr

Résumé

Vous connaissez tous WordNet, mais en connaissez-vous tout ? Nous vous proposons ici, d'une part de redécouvrir WordNet (notamment en présentant les spécificités des versions les plus récentes) et d'autre part de découvrir d'autres ressources (lexicales, syntaxiques et sémantiques) qui s'y rattachent. Nous présentons également des techniques d'enrichissement automatique de WordNet, et des applications de TALN l'utilisant.

Mots-clés

WordNet, eXtended WordNet, VerbNet, FrameNet, SentiWordNet, WordNet Domains, WordNet-Affect, SemCor, Wikipédia, SUMO, Cyc, Web sémantique, ontologie

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, l'inférence pour la compréhension automatique de textes, la désambiguïsation lexicale ou la résolution d'anaphores.

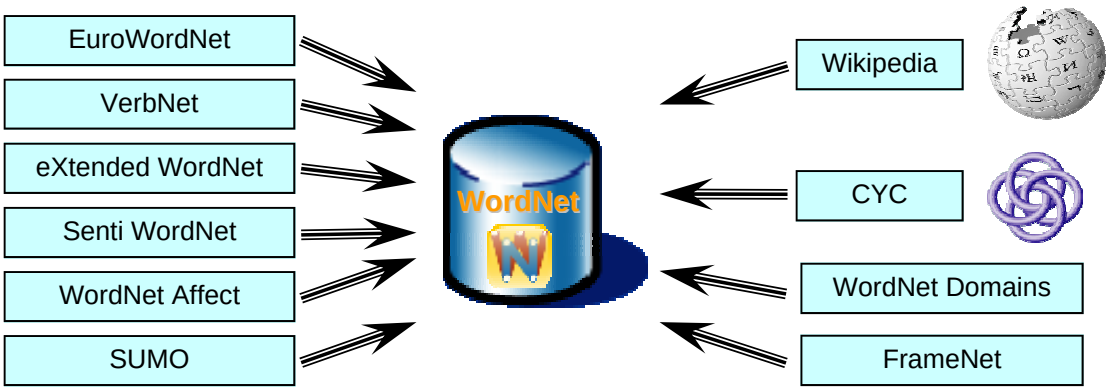


Figure 1 : Ressources disposant d'une traçabilité vers WordNet (liste non exhaustive)

2 WordNet

WordNet (Miller, 1995) 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.

Son but est de répertorier, classifier 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.

S'il n'est pas exempt de critiques (granularité très fine, absence de relations paradigmatiques…), WordNet n'en reste pas moins l'une des ressources de TAL les plus populaires.

2.1 Notion de synset

Le *synset* (ensemble de synonymes) est la composante atomique sur laquelle repose WordNet. Un synset correspond à un groupe de mots interchangeables, dénotant un sens ou un usage particulier. Un synset est défini d'une façon différentielle par les relations qu'il entretient avec les sens voisins.

Les **noms** et **verbes** sont organisés en hiérarchies. Des relations d'hyperonymie (« est-un ») et d'hyponymie relient les « ancêtres » des noms et des verbes avec leurs « spécialisations ». Au niveau racine, ces hiérarchies sont organisées en types de base. Le réseau des noms est bien plus profond que celui des autres parties du discours. A titre indicatif, les deux premiers niveaux de la hiérarchie des noms se constituent des concepts abstraits suivants :

- **ABSTRACTION**: ATTRIBUTE, MEASURE/QUANTITY/AMOUNT, RELATION, SET, SPACE, TIME…
- **HUMAN ACTION**: ACTIVITY, COMMUNICATION, DISTRIBUTION, INACTIVITY, JUDGMENT, LEANING, LEGITIMATION, MOTIVATION, PROCLAMATION, PRODUCTION, SPEECH ACT…

- **ENTITY**: ANTICIPATION, CAUSAL AGENT, ENCLOSURE, EXPANSE, LOCATION, PHYSICAL OBJECT, SKY, SUBSTANCE, THING…
- **EVENT**: GROUP ACTION, NATURAL EVENT, MIGHT-HAVE-BEEN, MIGRATION, MIRACLE, NONEVENT, SOCIAL EVENT…
- **GROUP, GROUPING**: ASSOCIATION, BIOLOGICAL GROUP, PEOPLE, COLLECTION, AGGREGATION, COMMUNITY, ETHNIC GROUP, KINGDOM, MULTITUDE, POPULATION, RACE, RARE-EARTH ELEMENT…
- **PHENOMENON**: EFFECT/RESULT, LEVITATION, FORTUNE/CHANCE, REBIRTH, NATURAL PHENOMENON, PROCESS, PULSATION…
- **POSSESSION**: ASSETS, CIRCUMSTANCES, PROPERTY/MATERIAL POSSESSION, TRANSFERRED PROPERTY, TREASURE…
- **PSYCHOLOGICAL FEATURE**: COGNITION/KNOWLEDGE, FEELING, MOTIVATION/NEED…
- **STATE**: ACTION/ACTIVITY, EXISTENCE, STATE OF MIND, CONDITION, CONFLICT, DAMNATION, DEATH, DEGREE, DEPENDENCY, DISORDER, EMPLOYMENT, END, FREEDOM, ANTAGONISM, IMMATURITY, IMMINENCE, IMPERFECTION, INTEGRITY, MATURITY, OMNIPOTENCE, PERFECTION, PHYSIOLOGICAL STATE, RELATIONSHIP, STATE OF AFFAIRS, STATUS, TEMPORARY STATE, NATURAL STATE…

L'organisation des **adjectifs** est différente. Un sens « tête » joue un rôle d'attracteur ; des adjectifs « satellites » lui sont reliés par des relations de synonymie. On a donc une partition de l'ensemble des adjectifs en petits groupes. Les **adverbes** sont le plus souvent définis par les adjectifs dont ils dérivent. Ils héritent donc de la structure des adjectifs.

La version 3.0, la plus récente (janvier 2007) compte 117 597 synsets et 207 016 lemmes.

2.2 Relations

2.2.1 Relations sémantique (entre synsets)

Le tableau suivant présente un comptage des relations sémantiques de WordNet 2.1 par catégorie.

Relation	Entre	Nombre	Exemple
Hypernym/Hyponym	Verbe / verbe	13 124	EXHALE / BREATHE
	Nom / nom	75 134	CAT / FELINE
Instance Hyponym	Nom / nom	8 515	EIFFEL TOWER / TOWER
Part	Nom / nom	8 874	FRANCE / EUROPE
Member	Nom / nom	12 262	FRANCE / EUROPEAN UNION
Substance	Nom / nom	793	SERUM / BLOOD
Attribute	Adjectif / nom	643	INACCURATE / ACCURACY
Verb Group	Verbe / verbe	1 748	GELATINIZE#1 / GELATINIZE#2
Verb Entailment	Verbe / verbe	409	DREAM / SLEEP
Verb Cause	Verbe / verbe	219	ANESTHETIZE / SLEEP
Adjective Similar	Adjectif / adjectif	22 622	DYING / MORIBUND

Topic Domain	Nom / adjectif	1 108	COMPUTER SCIENCE / ADDRESSABLE
	Nom / nom	4 146	COMPUTER SCIENCE / COMPUTER
	Nom / adverbe	37	
	Nom / verbe	1 236	COMPUTER SCIENCE / CASCADE
Region Domain	Nom / adjectif	75	
	Nom / nom	1 246	FRENCH / FRANCE
Usage Domain	Nom / adjectif	227	
	Nom / nom	563	NEUTRALIZATION / EUPHEMISM
	Nom / adverbe	73	
	Nom / verbe	14	
See Also	Adjectif / adjectif	2 683	BLACK / DARK

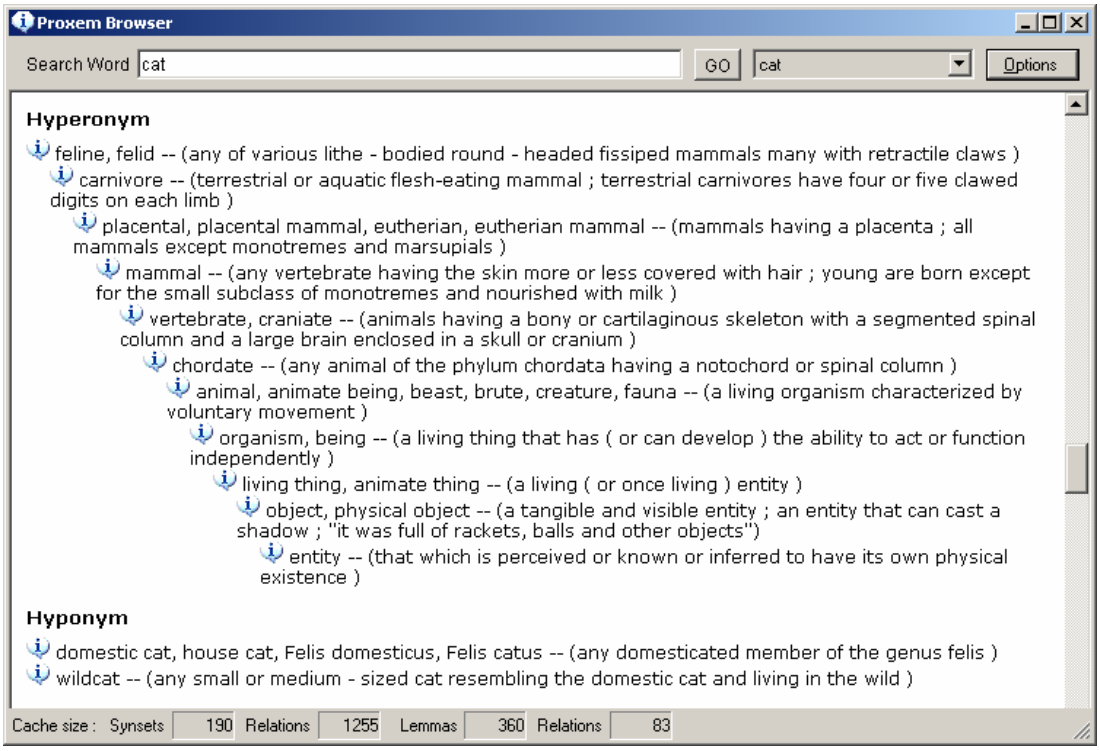
2.2.2 Relations lexicales (entre lemmes)

Le tableau suivant présente un comptage des relations lexicales de WordNet 2.1 par catégorie.

Relation	Entre…	…et	Nombre	Exemple
Usage Domain	nom	nom	379	
See Also	verbe	verbe	582	SLEEP LATE / SLEEP
Adjective Participle	adjectif	verbe	124	APPLIED / APPLY
Antonym	adjectif	adjectif	4 080	GOOD / BAD
	adverbe	adverbe	718	POORLY / WELL
	nom	nom	2 142	WINNER / LOOSER
	verbe	verbe	1 089	DIE / BE BORN
Pertainym	adjectif	nom	4 814	ACADEMIC / ACADEMIA
	adverbe	adjectif	3 213	BOASTFULLY / BOASTFUL
	adjectif	adjectif	38	
Derivation	nom	verbe	21 579	KILLING / KILL
	adjectif	nom	11 401	DARK / DARKNESS
	nom	nom	2 931	AUTOMOBILE / AUTOMOBILIST
	verbe	adjectif	1 508	KILL / KILLABLE
Adjective Cluster	adjectif	adjectif	1 290	STRIDENT / NOISY

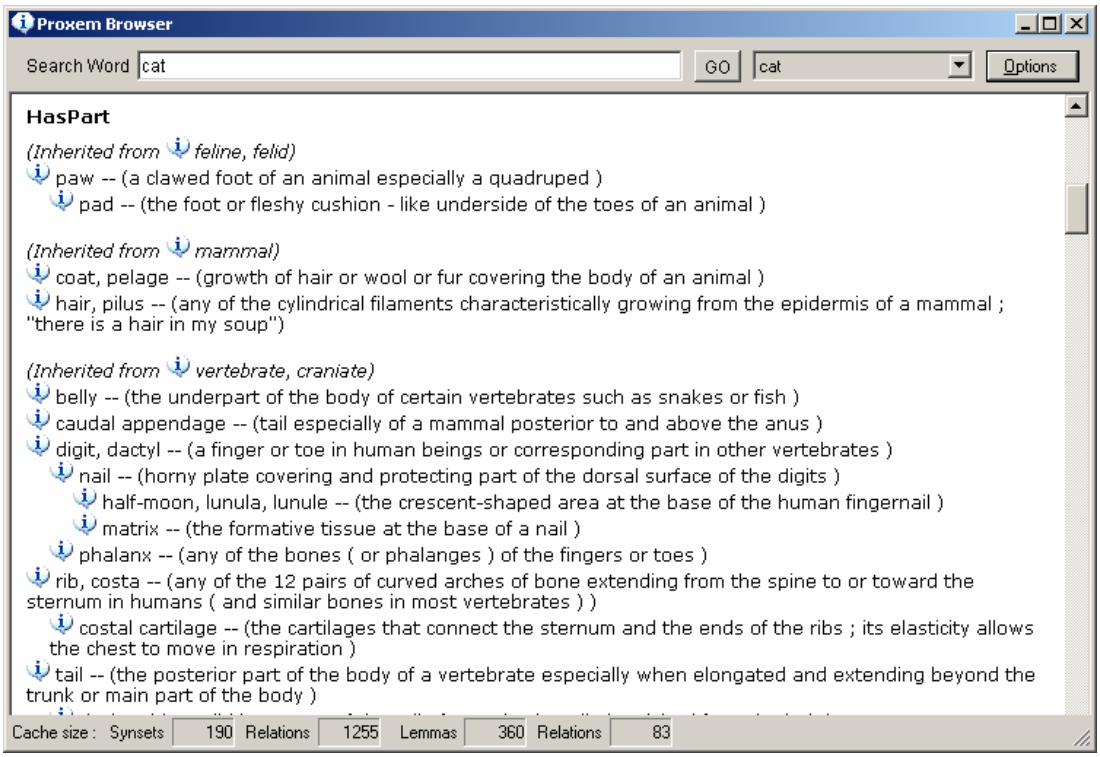
2.2.3 Exemples de relations d'hyperonymie et d'hyponymie

Par exemple, partant du sens le plus général du mot CAT#1 *(*le « chat » félin), on obtient une liste ordonnée d'ancêtres et de descendants, permettant de déterminer qu'un chat est un carnivore, un mammifère, un animal, etc.



2.2.4 Exemples de relations d'holonymie et de méronymie

Grâce à ces relations, on peut déterminer qu'un chat a des pattes, un pelage, une queue…



2.2.5 Notion d'instance hyponyme

La version 2.1 a introduit la notion d' « instance hyponyme », qui désigne une instance (et non une sous-classe) d'un synset (une Entité Nommée). Par exemple, GEORGE WASHINGTON est une instance hyponyme de PRESIDENT OF THE UNITED STATES. De même, le nom TOWER#1 a pour hyponymes SILO, MINARET, PYLON… et TOUR EIFFEL comme instance hyponyme.

2.3 Limites de WordNet

2.3.1 Informations manquantes

WordNet ne précise pas l'étymologie, la prononciation, les formes de verbes irréguliers et ne contient que des informations limitées sur l'usage des mots.

2.3.2 Profusion de sens pour un mot donné

La contrepartie de son importante couverture est que WordNet est très précis dans le sens des définitions. On a une granularité très (trop ?) fine des sens. Par exemple, le verbe *to give* (« donner ») n'a pas moins de 44 sens. Une telle profusion ne facilite pas une tâche de désambiguïsation lexicale.

2.3.3 Absence de relations pragmatiques

WordNet ne matérialise pas d'une façon formelle tout le sens contenu dans les définitions des termes. Par exemple, l'information qu'un chat ne rugit pas figure dans la définition, mais ne se retrouve formalisée dans aucune relation. De même, des relations pragmatiques telles que savon / bain (SOAP#1 / BATH#2) sont absentes de WordNet.

2.4 Mappage entre différentes versions

Il existe une correspondance des identifiants de synsets entre versions de WordNet. Ce mappage est indispensable pour assurer une traçabilité avec la version la plus récente. En effet, plusieurs ressources complémentaires à WordNet, et dignes d'intérêt, ont été définies pour la version 1.7 ou 2.0. Curieusement, le site Web de Princeton n'offre de mappage « officiel » que pour les noms et les verbes. Heureusement, d'autres sites proposent également des correspondances (construites automatiquement) pour les adjectifs et adverbes.

2.5 Corpus étiquetés par rapport à WordNet

A notre connaissance, peu de corpus sont étiquetés manuellement par rapport aux sens de WordNet. Nous pouvons citer le corpus SemCor (un sous-ensemble du corpus Brown), composé de 352 documents, comptant 2000 mots chacun approximativement. Plus précisément, le corpus

SemCor compte au total 676 546 mots (hors ponctuations). 234 135 noms, verbes, adjectifs et adverbes ont fait l'objet d'une désambiguïsation lexicale manuelle par rapport à WordNet 1.6, puis d'un mappage automatique vers les versions suivantes de WordNet (jusqu'à la 2.1). Ce corpus permet par exemple un début d'apprentissage automatique pour des tâches de désambiguïsation lexicale.

2.6 Fréquence des lemmes

WordNet donne une fréquence d'apparition pour chaque lemme définissant un synset. Ce nombre indique combien de fois un mot apparaît dans un sens spécifique. Pour un nom ou un verbe, la somme cumulée des fréquences d'un synset et de ses hyponymes au sein d'un sous-arbre de la hiérarchie permet de calculer son Contenu Informationnel.

2.7 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.

(Pedersen, Patwardhan, Michelizzi, 2004) présentent plusieurs de ces algorithmes de similarité entre mots, et une implémentation basée sur WordNet en Perl appelée WordNet::Similarity.

2.8 WordNets pour d'autres langues que l'anglais

2.8.1 EuroWordNet

EuroWordNet est une base de données pour plusieurs langues européennes. La phase initiale du projet s'est achevée en 1999, avec la conception de la base de données, ainsi que la définition de types de relations, d'un haut d'ontologie (63 éléments partagé par toutes les langues) et d'un Index-Inter-Langues (basé sur la version 1.5 du WordNet de Princeton).

EuroWordNet a produit des wordnets pour le néerlandais, l'italien, l'espagnol, l'allemand, le français, le tchèque et l'estonien. (À notre connaissance, les ressources pour le français ont été fournies par la société MemoData sur la base de son Dictionnaire Intégral.)

Les langues sont reliées ensemble par l'intermédiaire de l'Index-Inter-Langues. Il est ainsi possible de passer des mots dans une langue aux mêmes mots dans n'importe quelle autre langue. EuroWordNet permet donc une recherche d'information monolingue ou multilingue.

Langue	Synsets	Sens de mots	Relations internes à une langue	Relations d'équivalence entre langues différentes
WordNet 1.5	94 515	187 602	211 375	0

Ajouts à l'anglais	16 361	40 588	42 140	0
néerlandais	44 015	70 201	111 639	53 448
espagnol	23 370	50 526	55 163	21 236
italien	40 428	48 499	117 068	71 789
allemand	15 132	20 453	34 818	16 347
français	22 745	32 809	49 494	22 730
tchèque	12 824	19 949	26 259	12 824
estonien	7 678	13 839	16 318	9 004

Plusieurs autres groupes de recherche ont développé des wordnets dans d'autres langues en se basant sur les spécifications d'EuroWordNet (suédois, norvégien, danois, grec, portugais, basque, catalan, roumain, lithuanien, russe, bulgare et slovène).

On peut regretter que, contrairement à la version de Princeton, EuroWordNet ne soit pas distribué librement. Cela explique certainement sa diffusion beaucoup moins importante.

2.8.2 BalkaNet

BalkaNet prolonge la base de données d'EuroWordNet avec d'autres langues européennes : tchèque, roumain, grec, turc, bulgare, et serbe.

	bulgare	tchèque	grec	roumain	turc	serbe
Synsets	21 441	28 456	18 461	19 839	14 626	8 059
Noms	14 174	21 009	14 426	13 345	11 059	5 919
Verbes	4 169	5 155	3 402	4 808	2 725	1 803
Adjectifs	3 088	2 128	617	852	802	324
Adverbes	9	164	16	834	40	13
Lemmes	44 956	43 918	24 366	33 690	20 310	13 295

3 Autres ressources

3.1 VerbNet

VerbNet est un lexique des classes de verbes anglais. C'est un projet mené sous l'impulsion de Martha Palmer (d'abord à l'Université de Pennsylvanie, puis à Boulder au Colorado). VerbNet regroupe par classe les verbes partageant les mêmes comportements syntaxiques et sémantiques. C'est un prolongement des travaux de (Levin, 1993). (Chaumartin, 2006) décrit comment mettre en œuvre WordNet et VerbNet pour implémenter une interface syntaxe-sémantique.

Une classe de verbes regroupe plusieurs verbes, et identifie des rôles thématiques avec d'éventuelles contraintes de sélection. Elle décrit plusieurs constructions typiques (des « *frames* ») des verbes membres. La sémantique de l'action ou de l'événement est également précisée. Des sous-classes permettent de décrire d'éventuelles spécialisations d'une classe. On peut en trouver une description dans (Kipper-Schuler, 2003). La ressource la plus proche pour les verbes français nous semble être le lexique-grammaire du LADL (Gross, 1994).

La version la plus récente (VerbNet 2.1) distingue 237 classes de verbes qui regroupent 4991 sens de verbes. Un verbe membre d'une classe est souvent accompagné d'une précision sur le *synset* correspondant, qui permet d'identifier dans WordNet le sens précis du verbe. VerbNet dispose aussi d'un mappage vers FrameNet. Une API en Java est également disponible.

3.1.1 Structure d'une description de classe de verbes

Chaque fichier de VerbNet décrivant une classe de verbes est représenté en XML, et découpé en sections balisées selon une structure arborescente :

- `<MEMBERS>` décrit les verbes membres qui appartiennent à la classe, en précisant l'identifiant vers le(s) synset(s) correspondant(s) de WordNet,
- `<THEMROLES>` indique les rôles thématiques de la classe :
  - `<SELRESTRS>` précise leurs éventuelles contraintes de sélections,
- `<FRAMES>` indique chacune des constructions typiques, en donnant à chaque fois :
  - `<SYNTAX>` sa syntaxe,
  - `<SEMANTICS>` sa sémantique,
  - `<EXAMPLES>` un ou plusieurs exemples,
- `<SUBCLASSES>` regroupe éventuellement en sous-classes :
  - `<VNSUBCLASS>` les cas particulier d'une classe de verbes.

3.1.2 Un exemple : la classe de verbe "murder"

Par exemple, le fichier *murder.xml* décrit trois constructions typiques :

- *Agent* élimine *Patient* (« Brutus tua Jules César »),
- *Agent* élimine *Patient* avec *Instrument* (« Brutus tua César avec un poignard »),
- *Instrument* élimine *Patient* (« le pesticide tua les insectes »).

Chaque description de classe de verbes déclare des contraintes de sélection sur les rôles thématiques. Par exemple, pour "**murder**", l'*Agent* et le *Patient* doivent avoir un trait *Animé* (en pratique, *Humain* ou *Organisation*) et l'*Instrument* doit être *Concret*.

3.1.2.1 Description de la syntaxe

La deuxième *frame* de la classe de verbe "**murder**" décrit :

- ```
  <SYNTAX>
        <NP value="Agent"/>
        <VERB/>
        <NP value="Patient"/>
        <PREP value="with"/>
        <NP value="Instrument"/>
  </SYNTAX>
  ```

- ```
  <EXAMPLES>
        <EXAMPLE>"Brutus killed Caesar with a knife"</EXAMPLE>
  </EXAMPLES>
  ```

3.1.2.2 Description de la sémantique

Par exemple, pour "**murder"** :

- au démarrage de l'événement, *Patient* est vivant : *alive(start(E), Patient),*
- à la fin de l'événement, *Patient* n'est plus vivant : ***!*** *alive(result(E), Patient).*

3.1.3 Prise en compte de l'héritage entre classes

La balise <SUBCLASSES> déclare les éventuelles sous-classes qui spécialisent une classe de verbe donnée. Une sous-classe permet :

- De raffiner les contraintes de sélection portant sur les rôles thématiques,
- De déclarer de nouveaux rôles thématiques,
- D'associer des verbes à la sous-classe,
- De créer de nouvelles *frames*.

3.2 FrameNet

FrameNet (Baker, Fillmore & Lowe, 1998), projet mené à Berkeley à l'initiative de Charles Fillmore, est fondé sur la sémantique des cadres ("*frame semantics*"). FrameNet a pour objectif de documenter la combinatoire syntaxique et sémantique pour chacun des sens d'une entrée lexicale à travers une annotation manuelle d'exemples choisis dans des corpus sur des critères de représentativité lexicographique. Les annotations sont ensuite synthétisées dans des tables, qui résument pour chaque mot les cadres avec leurs actants sémantiques et arguments syntaxiques.

FrameNet II compte actuellement 825 cadres sémantiques, 10 000 unités lexicales (dont 6 100 complètements annotées) ainsi que 130 000 phrases d'exemples annotés. La totalité des outils et données est (en principe) distribuée librement.

Un mappage entre les verbes de FrameNet II et ceux de WordNet peut être trouvé sur http://www.cs.unt.edu/~rada/downloads.html#verbmap.

3.2.1 Exemple de description du cadre "Crime_scenario"

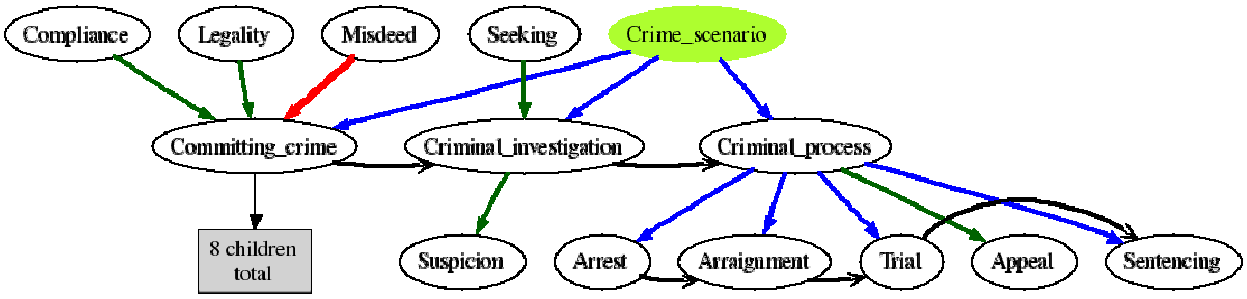
3.2.1.1 Description

A (putative) Crime is committed and comes to the attention of the Authorities. In response, there is a Criminal_investigation and (often) Arrest and criminal court proceedings. The Investigation, Arrest, and other parts of the Criminal_Process are pursued in order to find a Suspect (who then may enter the Criminal_process to become the Defendant) and determine if this Suspect matches the Perpetrator of the Crime, and also to determine if the Charges match the Crime. If the Suspect is deemed to have committed the Crime, then they are generally given some punishment commensurate with the Charges.

3.2.1.2 Frame Elements

Authorities []	The group which is responsible for the maintenance of law and order, and as such have been given the power to investigate Crimes, find Suspects and determine if a Suspect should be submitted to the Criminal_process.
Charge []	A description of a type of act that is not permissible according to the law of society.
Crime []	An act, generally intentional, that matches the description that belongs to an official Charge.
Perpetrator []	The individual that commits a Crime.
Semantic type	Sentient
Suspect []	The individual which is under suspicion of having committed the Crime.

3.2.2 Exemple de relations entres cadres



3.3 eXtended WordNet

3.3.1 Présentation

eXtended WordNet (XWN) est un projet mené en 2003 à l'Université de Dallas, qui enrichit WordNet 2.0. XWN produit une analyse syntaxique de la définition de chaque synset, la désambiguïsation lexicale de chaque mot de la définition, puis un passage en forme logique.

(Moldovan & Novischi, 2002) décrivent comment XWN permet d'améliorer sensiblement les résultats d'un système de Questions-Réponses.

3.3.2 Exemple

Par exemple, le nom COUSIN#1, dont la définition est "*the child of your aunt or uncle*" (« l'enfant de votre tante ou de votre oncle »), a pour analyse syntaxique :

```
(TOP (S (NP (NN cousin) )
        (VP (VBZ is)
                (NP (NP (DT the) (NN child) )
                        (PP (IN of)
                                (NP (PRP$ your) (NN aunt) (CC or) (NN uncle) ) ) ) )
(. .) ) )
```

Ainsi que la forme logique suivante :

```
cousin:NN(x1) -> child:NN(x1) of:IN(x1, x4) aunt:NN(x2) or:CC(x4, x2, x3) uncle:NN(x3)
```

3.3.3 Caractéristiques

Les informations présentes dans XWN sont de qualité *gold* (validé humainement), *silver* (accord entre deux analyseurs syntaxiques) ou *normal*. Si on considère l'analyse des définitions, on a :

Synsets (WN 2.0)	Nombre de définitions	Mots de classe ouverte	Mots mono-sémiques	Qualité *gold*	Qualité *silver*	Qualité *normal*
Noms	79 689	505 946	138 274	10 142	45 015	296 045
Verbes	13 508	48 200	6 903	2 212	5 193	30 813
Adjectifs	18 563	74 108	14 142	263	6 599	50 359
Adverbes	3 664	8 998	1 605	1 829	385	4 920

Du fait de la complexité de la tâche de désambiguïsation lexicale, et de l'absence de validation humain systématique, il est sage de penser que seule les mots étiquetés avec la qualité *gold* sont correctement désambiguïsés (ils ne représentent que 3,2% des mots polysémiques), et que les autres mots contiennent une proportion importante de contresens.

3.4 WordNet Domains

WordNet Domains (Magnini et Cavaglià, 2000) est une extension multilingue de WordNet 2.0, développée à l'Instituto Trentino di Cultura (ITC-irst). La notion de domaine a été employée aussi bien en linguistique qu'en lexicographie pour marquer des usages des mots. Les domaines sémantiques offrent une manière naturelle d'établir des relations sémantiques entre les sens des mots, qui peuvent être utilisée avec profit en informatique linguistique. Dans WordNet Domains, chaque synset est annoté avec au moins une étiquette de domaine (par exemple *Sport*, *Politique*, *Médecine*, *Economie*…), choisie dans un ensemble d'environ deux cents étiquettes organisées hiérarchiquement.

Un domaine peut inclure des synsets de différentes parties du discours et de différentes sous-hiérarchies de WordNet. Par exemple le domaine *Médecine* regroupe des sens de noms tels que DOCTOR#1 (le 1er sens du mot docteur) et HOSPITAL#1, et de verbes comme OPERATE#7.

L'information apportée par ces domaines est complémentaire à celles déjà présentes dans WordNet. Les domaines peuvent créer des regroupements homogènes des sens d'un même mot, avec comme effet secondaire de réduire la polysémie des mots dans WordNet.

3.4.1 Exemple : les domaines associés aux différents sens du nom "bank"

Le mot *bank*, par exemple, a dix sens dans WordNet 2.0. Trois d'entre eux (BANK#1, BANK#3 et BANK#6) sont regroupés au sein du domaine *Economie*, tandis que deux (BANK#2 et BANK#7) sont regroupés avec les étiquettes de domaine *Géographie* et *Géologie*.

Sens	Synset (Définition)	Domaines
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	*Economy*
#2	bank (sloping land ...)	*Geography, Geology*
#3	bank (a supply or stock held in reserve...)	*Economy*
#4	bank, bank building (a building...)	*Architecture, Economy*
#5	bank (an arrangement of similar objects...)	*Factotum*
#6	savings bank, coin bank, money box, bank (a container...)	*Economy*
#7	bank (a long ridge or pile...)	*Geography, Geology*
#8	bank (the funds held by a gambling house...)	*Economy, Play*
#9	bank, cant, camber (a slope in the turn of a road...)	*Architecture*
#10	bank (a flight maneuver...)	*Transport*

3.4.2 Intérêt

L'utilisation de WordNet Domains permet par exemple d'améliorer l'efficacité d'algorithmes de désambiguïsation lexicale et d'expansion de requêtes.

3.5 WordNet-Affect

La détection de connotations affectives dans les textes a des intérêts économiques réels : par exemple, une société peut chercher à détecter, en analysant la blogosphère ou les *news*, s'il se dit du bien ou du mal de ses produits.

Basé sur WordNet Domains, WordNet-Affect (Strapparava & Valitutti, 2004) est une ressource linguistique pour la représentation lexicale de connaissances sur les affects.

Un sous-ensemble de synsets de WordNet appropriés est choisi pour représenter des concepts affectifs. On ajoute des informations additionnelles aux synsets affectifs, en leur associant une ou plusieurs étiquettes qui précisent une signification affective. Par exemple, les concepts affectifs représentant un état émotif sont représentés par des synsets marqués par l'étiquette *Émotion*. Le tableau suivant liste ces étiquettes affectives, avec des exemples de synsets associés :

Etiquette affective	Exemples de synsets associés
Emotion	nom ANGER#1, verbe FEAR#1
Mood	nom ANIMOSITY#1, adjectif AMIABLE#1
Trait	nom AGGRESSIVENESS#1, adjectif COMPETITIVE#1
Cognitive State	nom CONFUSION#2, adjectif DAZED#2
Physical State	nom ILLNESS#1, adjectif ALL IN#1

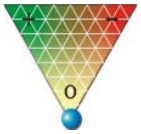
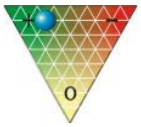
Edonic Signal	nom HURT#3, nom SUFFERING#4
Emotion-Eliciting Situation	nom AWKWARDNESS#3, adjectif OUT OF DANGER#1
Emotional Response	nom COLD SWEAT#1, verbe TREMBLE#2
Behaviour	nom OFFENSE#1, adjectif INHIBITED#1
Attitude	nom INTOLERANCE#1, nom DEFENSIVE#1
Sensation	nom COLDNESS#1, verbe FEEL#3

WordNet-Affect a été développé en deux étapes. La première a consisté à identifier manuellement un premier « noyau » de synsets affectifs. La deuxième étape a permis, en suivant les relations définies dans WordNet, de propager les informations de ce noyau à son voisinage.

3.6 SentiWordNet

SentiWordNet (Esuli & Sebastiani, 2006) est une ressource lexicale permettant le sondage d'opinion. SentiWordNet assigne à chaque synset de WordNet 2.0 trois valeurs : Positivité, Négativité, Objectivité (respectant l'égalité : Positivité + Négativité + Objectivité = 1). Cette ressource a été créée d'une façon semi-automatisées, en mixant des techniques linguistiques et statistiques (utilisation de classifieurs).

Avec cette classification, on a par exemple pour les trois sens de l'adjectif "*estimable*" :

	P = 0 N = 0 O = 1	COMPUTABLE#1 **ESTIMABLE**#3 *may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"*
	P = 0,75 N = 0 O = 0,25	**ESTIMABLE**#1 *deserving of respect or high regard*
	P = 0,625 N = 0,25 O = 0.125	HONORABLE#5 GOOD#4 RESPECTABLE#2 **ESTIMABLE**#2 *deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"*

3.7 SUMO (Suggested Upper Merged Ontology)

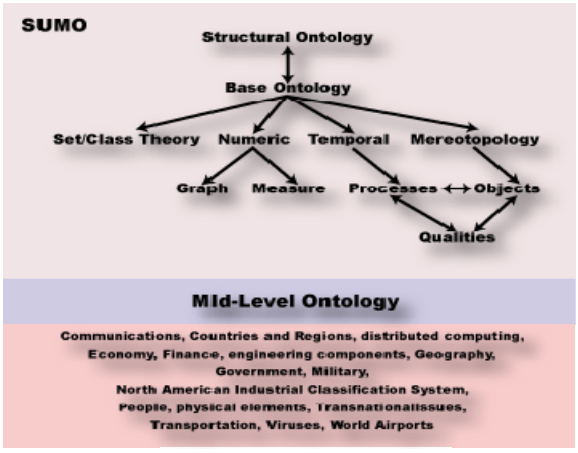
3.7.1 Notion de « haut » d'ontologie

Les ontologies sont des artefacts construits en fonction d'une tâche précise. Force est de constater qu'une ontologie donnée ne semble pas pouvoir être facilement réutilisée pour une tâche autre que celle qui a motivé sa construction originelle.

Il découle de ce constat de nombreuses recherches sur la réutilisabilité du « haut » des ontologies, avec pour argumentaire : puisqu'il est difficile, voire impossible, de réutiliser directement des ontologies, trop proches de vues détaillées qu'on peut avoir sur un domaine, intéressons-nous au « haut » de l'ontologie. Cette *Upper Ontology* répertorie et organise de grandes catégories de la

pensée ou de la société humaine qui devraient pouvoir être réutilisables dans de très nombreuses applications et être alors « génériques ».

L'objectif du groupe *Standard Upper Ontology* est de réfléchir, puis soumettre à la normalisation, la constitution d'un haut d'ontologie qui se voudrait universel pour les grandes catégories d'objets et de pensées. Le résultat est SUMO (*Suggested Upper Merged Ontology*), qui vise à s'imposer en tant que standard, et commence à être utilisée notamment pour le Web sémantique. MILO (*MId-Level Ontologies*) est un ensemble d'ontologies multi domaines, de niveau intermédiaire, créées en se basant sur SUMO.



SUMO (Niles & Pease, 2003) est écrit en langage SUO-KIF, dérivé simplifié de KIF (*Knowledge Interchange Format*), qui est un langage équivalent à la logique du premier ordre. Une traduction vers OWL (le langage du Web sémantique) est également disponible.

L'ensemble compte 20 000 termes et 60 000 axiomes. Il existe un mappage complet de SUMO vers les différentes versions de WordNet (jusqu'à la version 2.1), y compris pour MILO (les ontologies de niveau intermédiaire).

3.7.2 Exemple : le concept « beverage »

Définition : Any food that is ingested by drinking. Note that this class is disjoint with the other subclasses of food, i.e. meat and fruit or vegetable.

Sous-classes : Milk, AlcoholicBeverage, Coffee, Tea

Axiomes (traduits automatiquement en anglais à partir de l'expression en KIF) :

Food is disjointly decomposed into Meat, Beverage

for all beverage *?BEV* holds Liquid is an attribute of *?BEV*

for all drinking *?DRINK* holds if *?BEV* is a patient of *?DRINK*, then *?BEV* is an instance of Beverage

for all Cup *?CUP* holds if contains(*?CUP*, *?STUFF*), then *?STUFF* is an instance of Beverage

for all Tavern *?COMPANY* holds there exist CommercialService *?SERVICE*, beverage *?BEVERAGE* so that *?SERVICE* is an agent of *?COMPANY* and *?BEVERAGE* is a patient of *?SERVICE*

3.8 Cyc

Cyc est un projet d'Intelligence Artificielle lancé en 1984 par Doug Lenat. Cyc vise à regrouper une ontologie et une base de données complètes sur le sens commun, pour permettre à des applications d'I.A. d'effectuer des raisonnements similaires à ceux des humains.

Des fragments de connaissances typiques sont par exemple : « les chats ont quatre pattes » ; « Paris est la capitale de la France ». Elles contiennent des termes (PARIS, FRANCE, CHAT…) et des assertions (« Paris est la capitale de la France ») qui relient ces termes entre eux. Grâce au moteur d'inférence fourni avec la base Cyc, il est possible d'obtenir une réponse à une question comme « Quelle est la capitale de la France ? ».

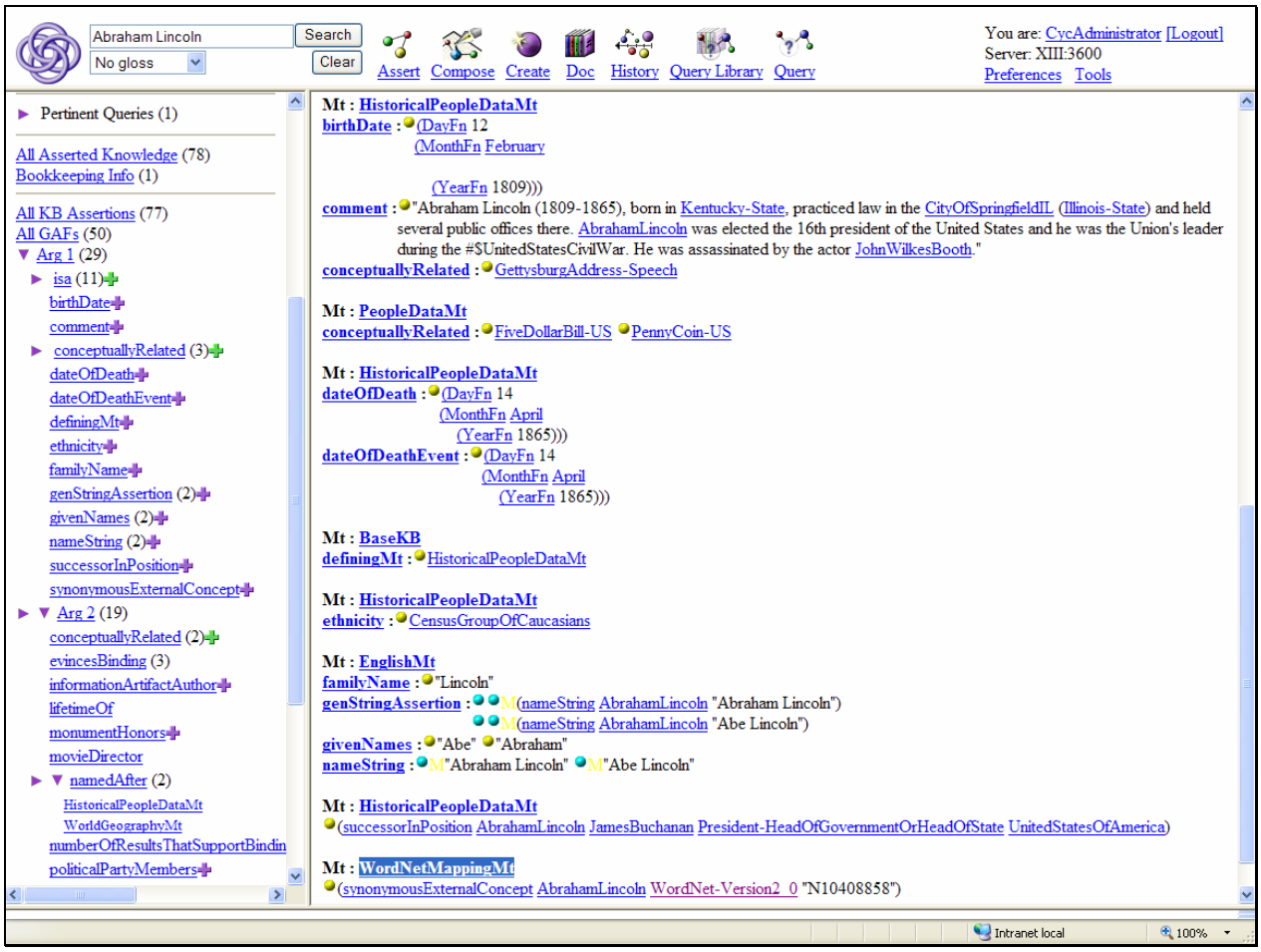


Figure 2 : Interface Web du serveur ResearchCyc (page de description de ABRAHAMLINCOLN)

La base Cyc contient des millions d'assertions (faits et règles) rentrées à la main. Elles sont écrites en langage CycL, qui est un langage logique avec une syntaxe proche de celle de LISP.

La base de connaissance est divisée en plusieurs milliers de micro-théories (Mt), collections de concepts et faits concernant typiquement un domaine particulier de la connaissance. Une micro-théorie est donc un ensemble d'assertions qui partagent le même point de vue : un domaine

particulier, un certain niveau de détail, un certain intervalle de temps, etc. À la différence de la base de connaissance dans son ensemble, chaque micro-théorie doit être exempte de contradictions. Par exemple, Philadelphie était la capitale des Etats-Unis de 1790 à 1800. Dans une micro-théorie couvrant l'intervalle de temps 1790-1800, l'assertion (#$CAPITALCITY #$UNITEDSTATES #$PHILADELPHIA) sera vraie, et dans une micro-théorie couvrant le XX^{ème} siècle, (#$CAPITALCITY #$UNITEDSTATES #$WASHINGTON) sera également vraie.

ResearchCyc 1.0 est la version réservée au monde de la recherche. Elle compte 300 000 concepts et 3 000 000 d'assertions (faits et règles) utilisant 26 000 relations. Des modules en langage naturel permettent de poser des questions et de rentrer de nouveaux faits sans avoir besoin de connaître CycL. La version OpenCyc 1.0 est librement accessible, mais ne contient qu'un sous ensemble de ces règles et assertions.

Les deux versions contiennent à ce jour une correspondance partielle entre les concepts de Cyc et les synsets de WordNet 2.0. Approximativement 11 300 synsets (8800 noms, 2110 verbes, 330 adjectifs et 35 adverbes) sont liés aux concepts de Cyc.

3.9 Wikipédia

Wikipédia est une encyclopédie libre et multilingue écrite de façon collaborative sur Internet avec la technologie wiki. Plusieurs projets visent à établir automatiquement des liens entre la Wikipédia et WordNet.

(Ruiz-Casado, Alfonseca, Castells, 2005) présentent l'implémentation d'un algorithme rapide permettant de réaliser la correspondance entre un article de la *Simple Wikipedia*[1] et le synset correspondant de WordNet. Si aucun synset n'a de lemme en commun avec le titre de l'article, ce dernier est ignoré. Si un seul synset de WordNet a un lemme égal au titre, l'article y est lié sans autre analyse. En cas d'ambiguïté, l'article fait l'objet d'un étiquetage morphosyntaxique (après un filtrage des marqueurs syntaxiques spécifiques à la *Wikipedia*), pour ne conserver que les noms, verbes et adjectifs. Le système analyse les définitions de WordNet, et construit pour chacune d'entre elles un vecteur booléen (contenant « 1 » pour chaque terme en commun avec l'article et « 0 » pour chaque mot en disjonction). L'algorithme calcule alors une mesure de type cosinus entre les vecteurs, et retient le meilleur article, au sens de cette mesure de similarité. Les auteurs revendiquent une précision de 91,11% (83.89% sur les mots polysémiques).

(Chaumartin, 2007) présente une généralisation de ce type d'approche, où WordNet est en plus enrichi avec des nouveaux synsets, avec une identification automatique du bon hyperonyme. La précision de l'appariement entre WordNet 2.1 et un sous-ensemble de la Wikipedia anglaise (autour de 15 800 articles) est de 92% ; en cas de création de nouveau synset, l'hyperonyme est correctement identifié dans 85% des cas.

[1] Une version en anglais simplifié de la Wikipédia.

4 Conclusion

Nous avons présenté en détail WordNet, ainsi que plusieurs autres ressources de nature lexicale, syntaxique et sémantique, qui s'y rattachent. Le fait de mettre en commun plusieurs ressources de large couverture permet d'espérer des progrès dans les applications de TAL. Pour finir, citons quelques projets qui combinent plusieurs de ces ressources.

(Shi, Mihalcea, 2005) revendique la construction d'un analyseur sémantique robuste en langue anglaise, en utilisant WordNet, VerbNet et FrameNet.

Notre projet, ISIDORE[2] (en cours de réalisation), combine WordNet, VerbNet, eXtended WordNet et SUMO. Il vise à extraire des connaissances d'une encyclopédie. Nous espérons disposer fin 2008 d'une indexation sémantique de 15 000 articles de la Wikipedia en anglais.

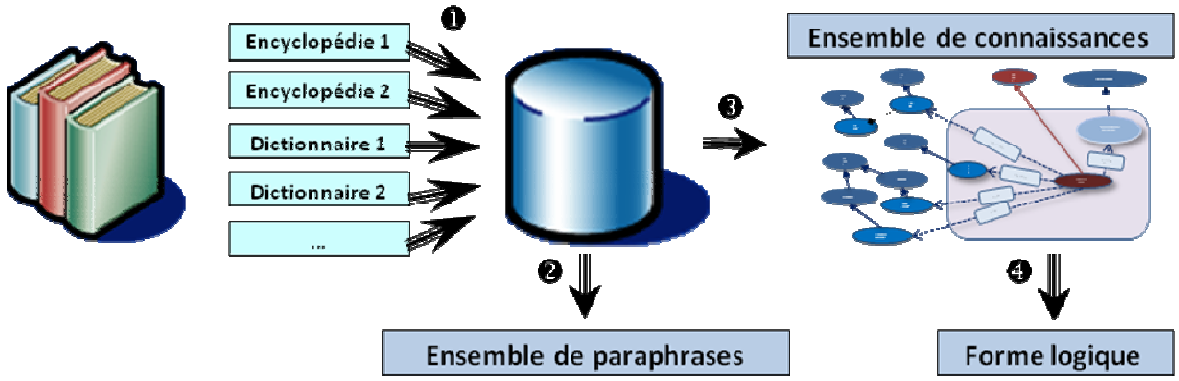


Figure 3 : Architecture d'ensemble du projet ISIDORE

Bibliographie

Andreevskaia A., Bergler S. 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. Actes de *EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italie.

Baker C., Fillmore C., Lowe J. 1998. The Berkeley FrameNet project. Actes de *17th international conference on Computational linguistics.*

Bentivogli L., Forner P., Magnini B., Pianta E. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Genève, Suisse, pp. 101-108.

Chaumartin F. 2006. Construction automatique d'interface syntaxe-sémantique utilisant des ressources de large couverture en langue anglaise. Actes de *TALN 2006*, 729-735.

[2] Saint-Isidore (560-636), patron des informaticiens, fut l'auteur des *Etymologies*, une encyclopédie en 20 livres.

Chaumartin F. 2007. Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement. Actes de *TALN 2007* (à paraître).

Esuli A., Sebastiani F. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Actes de *LREC 2006, fifth international conference on Language Resources and Evaluation*, pp. 417-422.

Gross M. 1994. Constructing Lexicon-grammars. In *Computational Approaches to the Lexicon*, Atkins and Zampolli (eds.), Oxford Univ. Press, pp. 213-263.

Kipper-Schuler K. 2003. *VerbNet: a broad coverage, comprehensive, verb lexicon.* Ph.D. Thesis, University of Pennsylvania.

Levin B. 1993. *English Verb Classes and Alternation: A Preliminary Investigation.* Chicago, IL: University of Chicago Press.

Magnini B., Cavaglià G. 2000. Integrating Subject Field Codes into WordNet. Actes de *LREC-2000, Second International Conference on Language Resources and Evaluation*, Athènes, Grèce, pp. 1413-1418.

Moldovan D., Novischi A. 2002. Lexical Chains for Question Answering, Actes de *COLING 2002*.

Miller G., Fellbaum C., Miller K. 1993. *Five Papers on WordNet*.

Miller G. 1995. Wordnet: A lexical database. Actes de *ACM 38*, pp. 39-41.

Niles I., Pease A. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. Actes de *2003 International Conference on Information and Knowledge Engineering (IKE '03)*, Las Vegas, Nevada.

Pedersen T., Patwardhan S., Michelizzi J. 2004. WordNet::Similarity - Measuring the Relatedness of Concepts. Actes de *Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, San Jose, CA.

Ruiz-Casado M., Alfonseca E., Castells P. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. Actes de *AWIC*, 380-386.

Shi L., Mihalcea R. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. Actes de *CICLing 2005*, Mexico.

Strapparava C., Valitutti A. 2004. WordNet-Affect: an Affective Extension of WordNet. Actes de *4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, pp. 1083-1086.

Valitutti A., Strapparava C., Stock O. 2004. Developing Affective Lexical Resources. In *PsychNology Journal*, 2(1).

Ressources

BalkaNet – http://www.ceid.upatras.gr/Balkanet/

eXtended WordNet – http://xwn.hlt.utdallas.edu

FrameNet – http://framenet.icsi.berkeley.edu/

Global WordNet – http://www.globalwordnet.org

Mappings entre versions de WordNet – http://www.cs.unt.edu/~rada/downloads.html#wordnet et http://www.lsi.upc.es/~nlp/tools/mapping.html

OpenCyc – http://www.opencyc.org/

ResearchCyc – http://research.cyc.com/

SemCor Corpus – http://www.cs.unt.edu/~rada/downloads.html

SentiWordNet – http://sentiwordnet.isti.cnr.it/

Simple Wikipedia – http://simple.wikipedia.org

SUMO – http://www.ontologyportal.org/ - http://ontology.teknowledge.com/

VerbNet – http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html

Wikipedia en anglais – http://en.wikipedia.org

WordNet – http://wordnet.princeton.edu

WordNet Domains & WordNet-Affects - http://wndomains.itc.it/download.html

WordNet::Similarity – http://www.d.umn.edu/~tpederse/similarity.html