Chapitre 3

Ontologie WordNet

Modèle de notre axe de recherche

1. Historique et origine

Machiavel a dit un jour en politique, « la fin justifie les moyens ». L'absence d'un dictionnaire électronique facilement accessible, a fait de WordNet un projet d'étude (développement manuel) au début des années 1980 à l'Université Princeton par une équipe de psycholinguistes et de linguistes, qui se sont basées sur la mémoire lexicale humaine, sous l'impulsion de **G. Miller**. Peu temps après, il a émergé pour devenir ***une ressources lexicale électronique de la langue anglaise***, comprenant plus de 200.000 de mots de classe ouvertes ainsi que plus de 115.000 ensemble de synonymes. [**MIL93**]

Actuellement plusieurs réalisations descendantes de WordNet existent (différentes langues), parmi eux EuroWordNet (EWN, 1996) et ArabicWordNet (AWN), ce dernier est construit selon les méthodes développées pour EuroWordNet. L'approche EuroWordNet maximise la compatibilité à travers les WordNet(s) et se concentre sur un encodage manuel des concepts les plus complexes et les plus importants. [**VOS98**]

Deux éléments ont participé au succès de WordNet :

- La maturité du projet rendue possible grâce à un travail de plus de dix ans.
- Le libre accès aux sources du projet tant pour consultation que pour la modification ainsi que la possibilité de redistribution du produit modifié.

2. Présentation de WordNet

Qu'est-ce WordNet ? Un dictionnaire[1] ? Un thésaurus[2] ? Les dictionnaires contiennent généralement des connaissances sur des lexies[3] alors que les encyclopédies[4]

1 - Dictionnaire : Recueil des mots d'une langue, des termes d'une science, d'un art, rangés par ordre alphabétique, avec leur signification. [*www.mediadico.com*]

2 - Thésaurus : une liste de termes sur un domaine de connaissances, reliés entre eux par des relations synonymiques, hiérarchiques et associatives. Le thésaurus constitue un vocabulaire normalisé. [www.fr.wikipedia.org]

3 - Lexie : C'est une suite de caractères formant une unité sémantique, *un mot*, et pouvant constituer une entrée de dictionnaire. [www.fr.wikipedia.org]

4 - Une encyclopédie peut prendre la forme d'un livre ou plusieurs livres. Elle se présente souvent comme une collection d'articles traitant chacun un thème. [www.fr.wikipedia.org]

contiennent des connaissances éparses, du monde, sur la surface de la terre. Quant aux thésaurus, leur structure est bâti autour des concepts et aident l'utilisateur à acquérir l'unité lexicale la plus appropriée lorsqu'il a un concept à rechercher. WordNet n'est ni un dictionnaire classique ni un thésaurus : il est en fait, un arrangement des traits de chacune de ces deux ressources lexicales. [**FEL98**]

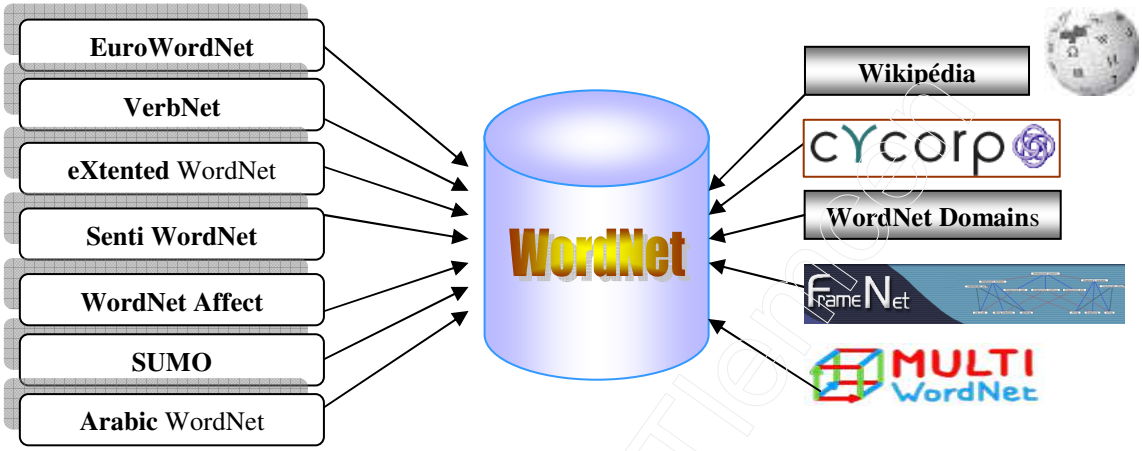

Fig. 28 : Ressources descendances de WordNet

(Liste non exhaustive)

3. Conception & Structure de WordNet

On peut considérer WordNet comme un graphe ou un réseau sémantique, souvent qu'on qualifie d'ontologie légère (Light Ontology), ou chaque nœud représente un concept du monde réel. La conception de WordNet est basée sur les théories de la représentation des connaissances mentales : mémorisation des mots et concepts d'une manière hiérarchique, en utilisant la relation d'inclusion (qui lie, par exemple, des triplets comme «animal », « oiseau», et « Chardonnet »). [**COL98**]

Exemple :

Un concept peut être un objet tel que « *Car* » une entité tel que « *Teacher* » ou un concept abstrait tel que « *art* ». Chaque nœud est constitué d'un ensemble de mots, où chacun représente le concept associé à ce nœud. Un nœud peut être vu comme un ensemble de mots dont chacun représente le même concept.

Exemple :

Le concept « *car* » est représenté par l'ensemble de mots {car, auto, automobile, motocar}.

Dans la terminologie de WordNet cet ensemble est appelé est nommé « Synset ». WordNet offre des descriptions détaillées et précises des mots. Leur structuration sur un axe ontologique a un fondement psychologique. Il résulte de cette approche qu'il arrive parfois que l'on rencontre plus de 20 sens pour un verbe, par exemple le verbe « ***give*** » a 27 sens.

3.1. SynSet

WordNet manipule les unités lexicales non pas par des mots mais par un ensemble de synonymes ou « Synset », groupes de mots ou de phrases qui expriment le même concept. Des différences de sens entre les membres d'un «Synset» se montrent dans différentes restrictions de sélection. Par exemple, « *rise* » (monter) et « *fall* » (tomber / descendre) peuvent choisir comme argument des entités abstraites comme « *temperature* » (température) et « *prices* » (prix).

Un « Synset » est accompagné d'une petite définition dite « ***gloss*** **»** qui décrit un concept du monde réel.

Exemple **:** les nœuds suivant correspondent aux différents sens de "mouse" dans WordNet :

1. Mouse -- (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
2. Shiner, black eye, mouse -- (a swollen bruise caused by a blow to the eye)
3. Mouse -- (person who is quiet or timid)
4. Mouse, computer mouse -- (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad ; on the bottom of the device is a ball that rolls on the surface of the pad ; "a mouse takes much more room than a trackball")
5. Sneak, mouse, creep, pussyfoot -- (to go stealthily or furtively; "…stead of sneaking around spying on the neighbour's house")
6. Mouse -- (manipulate the mouse of a computer)

Notons que WordNet met l'accent sur les liaisons entre les « Synset » (arc du graphe de la Figure 29) pour marquer sa valeur ajoutée faces aux dictionnaires traditionnaires. Chaque lien décrit une relation entre concept du monde réel. Par exemple, les relations telles que : « a spoke **is a part of** a wheel » ou « **a vehicle is a kind of** conveyance »

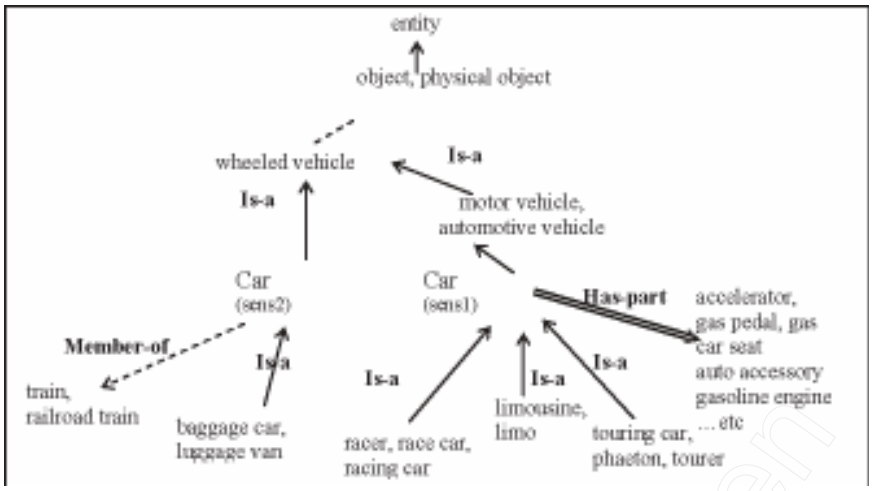

Fig. 29 : Exemple de sous hiérarchie dans WordNet correspondent au concept "car". [BAZ05]

3.2. Organisation

WordNet sépare les données en quatre (04) bases de données, organisées différemment les unes des autres, associées aux catégories de **noms**, **verbes**, **adjectifs** et **adverbes**. 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. [**FEL98**] Le réseau des noms est bien plus profond que les autres parties. Il faut noter que, 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…

WordNet organisent les verbes en taxonomie. Plusieurs propriétés des verbes dépendent de la manière avec laquelle on peut combiner les arguments (sujet, objets direct, objets indirects) [**WID04**].

Notamment qu'elle relation utilise t – on ? « ***is a kind of*** » ?, WordNet fournit une relation équivalente à celle–ci c'est la relation « ***Manner Of*** » nommée : « **Troponymy** »

Les adjectifs et les adverbes sont organisés de paires d'antonymes[1] tels que « good, bad ». Bon nombre d'adjectifs en langue anglaise possède des antonymes, par opposition aux verbes et aux noms. Les adverbes sont souvent définis par les adjectifs dont ils dérivent. C'est ainsi ils héritent de la structure des adjectifs.

3.3. La matrice lexicale

Un sens peut être représenté par plusieurs mots. Et un mot à son tour peut désigner plusieurs sens.

Exemple : le mot « *word* » fait référence en même temps à une *expression* et à un *concept*. Pour éliminer cette ambiguïté : **« *word form* »** sera utilisé pour exprimer (la forme ou l'image) et **« *word meaning* »** sera utilisé dénoter le concept que porte ce mot « *word* ». [**FEL98**]

Dans la matrice lexicale (voir à droite), une entrée dans la cellule de la matrice E_{ij}, suppose le « *word form* » F_i est utilisée pour référencer le concept « *word meaning* » (concept) Mj. [FEL98]

Word Meanings	Word Forms				
	F1	F2	F3	…	Fn
M1	E_{11}	E_{12}			
M2		E_{22}			
M3			E_{33}		
…				…	
Mn					E_{mn}

Tableau 1 : Illustration des concepts de la matrice Lexical [FEL98]

4. Les relations dans WordNet

Deux relations fondamentales interviennent dans WordNet, notamment celle entre les « *word form* » appelés *relations lexicales* (par exemple : la synonymie), et celle qui associent les « *word meaning* » appelés relation sémantiques (par exemple : l'hyperonymie).

Remarquons que la majorité des relations dans WordNet sont des « synset » de la même catégorie, excepté les relations « *pertains to* » et « *attribute* » souvent utilisées entre les adjectifs et les noms. Le tableau 2 illustre un sous ensemble des relations dans WordNet qu'on détaillera par la suite.

1 - Antonyme : nom opposé, (exemple : *black* est antonyme de *White*).

Relation	Description	Exemple
Hypernym	Is a generalization of	Furniture is a hypernym of chai
Hyponym	Is a kind of	Chair is a hyponym of furniture
Troponym	Is a way to	Amble is a troponym of walk
Meronym	Is part/substance/member of	Wheel is a (part) meronym of a bicycle
Holonym	Contains part	Bicycle is a holonym of a wheel
Antonym	Opposite of	Ascend is an opposite of descend
Attribute	Attribute of	Heavy is a attribute of weight
Entailment	entails	Ploughing entails digging
Cause	Cause to	To offend causes to resent
Also see	Related verb	To lodge is related to reside
Similar to	Similar to	Dead is similar to assassinated
Participle of	Is participle of	Stored (adj) is the participle of "to store"
Pertainym of	Pertains to	Radial pertains to radius

Tableau 2 : Quelques relations dans WordNet

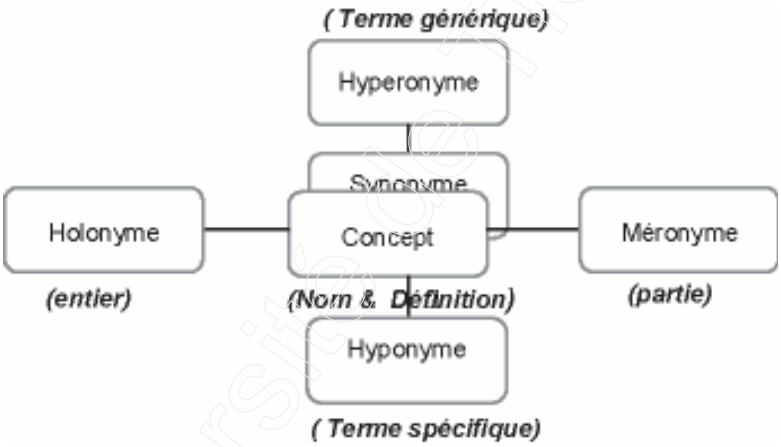

Fig. 30 : Principales relations sémantiques dans WordNet. [BAZ05]

Voilà, une définition des relations les plus importantes dans WordNet :

4.1. Synonymie

La synonymie est une relation liant deux concepts équivalents ou voisins (frêle / fragile). Il s'agit d'une relation symétrique. Christiane Fellbaum, dans [**FEL98**], énonce une définition, généralement attribuée à Leibniz, « *two expressions are synonymous if the substitution of one for the other never changes the truth value of a sentence in which the substitution is made.* » Ce qui ce traduit par : Deux expressions sont synonymes dans un contexte linguistique C, si la substitution de l'une pour l'autre en C, ne modifie pas la valeur de vérité de la phrase dans laquelle la substitution soit faite. Par exemple, la substitution « a

plank » pour « a board » groupé dans un même Synset {board, plank}, sera rarement substituer dans des contextes de menuiserie, bien qu'il y ait d'autres contextes comme un conseil d'administration où cette substitution serait totalement inapproprié. [**FEL98**]

4.2. Antonymie

Une autre relation familière est l'antonymie, qui s'avère être étonnamment difficile de définir. L'antonyme d'un mot « x » n'est pas toujours « Non x ». Par exemple, « riche » et « pauvre » sont des antonymes, mais de dire que quelqu'un n'est pas riche ne signifie pas qu'il doit être pauvre, de nombreuses personnes ne se considèrent ni riche ni pauvre. L'Antonymie, semble être une simple relation symétrique, mais elle est en faite assez complexe, et beaucoup de personnes trouvent des difficultés à reconnaître les antonymes quand ils les voient. **[FEL98]** Antonymie n'est pas une relation sémantique entre « *word meanings* » mais beaucoup plus une relation lexicale entre « *word forms* ». Par exemple, la signification {rise, descend} – (s'élever/descendre) et {fall, ascend} – (monter/tomber) peuvent être conceptuellement opposées, mais ils ne sont pas des antonymes; [rise, fall] sont des antonymes et aussi [asscend, descend]. Ces faits font ressortir la nécessité de distinguer les relations sémantiques entre les « word forms » et les relations sémantiques entre « word meaning ».

4.3. L'Hyperonymie / Hyponymie

Ce sont les relations, les plus importantes, dans notre travail. Elles représentent la relation « ***is a kind of***» ou tout simplement la relation « ***is-a*** ».

Elles sont réservées seulement pour les catégories Nom et Verbes qui se voient organisées sous forme d'une hiérarchie comportant un seul nœud racine. Les nœuds représentant les concepts les plus généraux sont les ancêtres des nœuds représentant les concepts les plus spécifiques. On dit alors que le concept le plus général « Subsume » celui du plus spécifique.

Exemple : dans la figure 29, « Entity » est le concept le plus hiérarchique des noms c'est la racine du nœud, il subsume le concept spécifique « wheeled vehicle » qui à son tour subsume le concept « Car ».

WordNet dispose de 9 hiérarchies pour les noms et 628 pour les verbes. L'hyperonymie représente la relation permettant d'avoir les ancêtres d'un « *word meanings* » à l'opposé de l'hyponymie qui fournit ses descendants. On note, aussi que WordNet permet

l'héritage multiple, ce qui induit qu'un concept peut avoir plusieurs hyperonymes. WordNet représente :

L'hyperonymie par le symbole : @ Exemple : Tree @ → plant

L'hyponymie par le symbole : ~ Exemple : plant ~ → Tree

4.4. Méronymie

C'est une relation liant un concept C1 à un concept C2 qui est en fait une partie de C1 (C1= Fleur / C2= Pétale), un de ses membres (Forêt / Arbre) ou une substance le constituant (vitre / verre). Donc la méronymie est interprétée de trois manières différentes :

Si X is a part of Y (Composant – Objet complet)

Si X is substance of Y (Matière – Objet)

Si X is a member of Y (Membre – collection)

Il existe d'autres relations que la relations que nous allons donner juste une bref définition, telles que :

- **Métonymie** (Holonymie) : relation liant un concept C1 à un concept C2 dont il est une des parties. C'est la relation inverse de la méronymie.
- **Implication** : relation lie un concept C1 à un concept C2 qui en découle (marcher/faire un pas).
- **Causalité** : relation liant un concept C1 à son effet (tuer / mourir).
- **Valeur** : relation liant un concept C1 (adjectif) qui est un état possible pour un concept C2 (pauvre / condition financière).
- **A pour valeur** : relation liant un concept C à ses valeurs (adjectifs) possibles (taille / grand). C'est la relation inverse de Valeur.
- **Voir aussi** : relation entre des concepts ayant une certaine affinité (froid / gelé).
- **Similaire à** : certains concepts adjectifs dont le sens est proche sont regroupés. Un Synset est alors désigné comme étant central au regroupement. La relation « Similaire à » lie un Synset périphérique au Synset central (moite / humide).
- **Dérivé de** : indique une dérivation morphologique entre le concept cible (adjectif) et le concept origine (froidement / froid).

WordNet contient approximativement 117798 mots de nom organisés approximativement en 82115 concepts (Synset) (tableau 1) jusqu'à juillet 2008. Puisque la majorité des noms communs parfois sont des noms propres, aucune tentative curieuse de les

exclure n'est faite. En termes d'exhaustivité le but de WordNet diffère un peu des dictionnaires standards des écoliers. C'est à l'organisation de ces informations que WordNet espère de l'innovation.

Réseau	Formes	Synsets	Paires mot-sens
Noms	117798	82115	146312
Verbes	11529	13767	25047
Adjectifs	21479	18156	30002
Adverbes	4481	3621	5580
TOTAL	155287	117659	206941

Tableau 3. Statistique sur WordNet (juillet 2008)

5. Les verbes dans WordNet (réseau sémantique)

Actuellement, WordNet contient plus de 25000 mots verbes. Les verbes sont divisés en 15 fichiers, selon un critère sémantique, presque chacun représente ce que les linguistes appellent domaine sémantique :

Les verbes de soins de corps et fonctions, changement, cognition, communication, compétition, consommation, contact, création, émotion, mouvement, perception, possession, interaction sociale et les verbes météorologiques.

Pratiquement tous les verbes dans ces fichiers dénotent des évènements et des actions, un autre fichier contient les verbes d'état, tel que *suffice*, *belong*, et *resemble*, qui ne peuvent pas être intégrés dans les autres fichiers. Les verbes dans ce dernier groupe font référence à un état, et ne constituent pas un domaine sémantique et ne partage aucune propriété sémantique.

Plusieurs fichiers prennent leurs noms des verbes les plus hauts, ou "*unique beginners*" qui sont en tête des groupes lexicaux sémantiquement cohérents.

Les verbes sont la catégorie syntaxique la plus importante du langage. Toutes les phrases anglaises doivent contenir au moins un verbe. Les linguistes ont longuement débattues la question de mettre le verbe comme pivot centrale de la phrase.

A cause de la complexité de ces informations, les verbes sont probablement la catégorie syntaxique la plus difficile à étudier.

6. L'hyponymie entre les verbes

La phrase modèle utilisée pour tester l'hyponymie entre les noms, « x est-un y » n'est pas convenable pour les verbes : *to amble is a kind of to walk* ce n'est pas une phrase correcte.

La distinction sémantique entre les verbes est différente des propriétés qui distinguent deux noms dans une relation hyponymique.

Les différentes élaborations qui distinguent un l'hyponyme du verbe de son père immédiat sont résumées dans la relation « de manière que », Fellbaum et Miller la surnommée « Troponymie ». La relation de troponymie entre deux verbes peut être exprimée par la formule *To V1 is to* V2 d'une certaine manière particulière.

7. Polysémie[1]

Bien que les phrases anglaises nécessitent des verbes et non pas nécessairement des noms, le langage a moins de verbes que de noms. Par exemple, le dictionnaire « *Collins English* » liste 43636 différents noms et 14190 différents verbes.

Les verbes sont polysémiques beaucoup plus que les noms : les noms en *Collins* ont une moyenne de 1.74 sens, alors que les verbes ont 2.11 sens.

La haute polysémie des verbes suggère que les sens des verbes sont plus flexibles que les sens des noms. Les verbes changent leurs sens selon les types d'arguments de nom avec lesquels se ils se produisent, alors que les sens de noms tendent à être plus stables avec les différents verbes.

Genter et France ont montré ce qu'ils appellent « la haute mutabilité des verbes », et conclurent que les sens des verbes sont plus facilement changeables parce qu'ils sont moins cohésifs que les sens de noms.

Les verbes les plus fréquemment utilisés (have, be, run, make, set, go, take, …) sont plus polysèmes et leurs sens dépendent légèrement des noms avec lesquels ils entrent en production. Par exemple, les dictionnaires différencient entre les sens de « *have* » dans les phrases comme « I have a Mercedes » et « I have a headache ». La différence est moins due à la polysémie de have que à la nature concrète ou abstraite de ses objets.

Dans le cas des verbes polysèmes comme « beat » (battre), les différences de sens sont déterminées par la sémantique des arguments du verbe plutôt que par les différentes élaborations d'un ou deux composants essentiels communs partagés par la majorité des sens de « beat ». Afin de réduire l'ambiguïté en WordNet, les synsets de verbes pourraient contenir des pointeurs de renvoi aux Synsets Noms qui contiennent des noms choisis par les verbes.

---

1 - Polysémie : est la qualité d'un mot ou d'une expression qui a deux voire plusieurs sens différents.
Exemple : «œil-de-bœuf » désigne une plante en botanique, un animal en zoologie, une pierre en géologie et une fenêtre ronde en architecture.

8. Arabic WordNet (AWN)

8.1. L'écriture arabe [BLA06]

L'arabe est une langue sémitique. Le système d'écriture de la langue arabe a 25 consonnes et trois voyelles longues **: « و، ا، ي »** : (OU, A, iii) on les appelle « **حروف العلّة** » qui sont écrites de droite à gauche et prennent différentes formes en fonction de leur position dans le mot. En plus des voyelles longues, l'arabe a des voyelles courtes qui ne font pas partie de l'alphabet, mais plutôt sont écrites comme des voyelles diacritiques (Fig.30 voyelles en verts) en haut ou en bas d'une consonne pour lui donner le son désiré et par conséquent de générer un mot dans un sens souhaité.

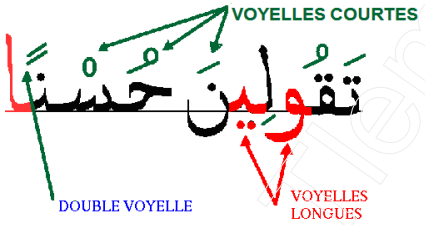

Fig. 31 : Représentation des voyelles arabes[1]

Le terme « arabe classique » renvoie la forme standard de la langue utilisée dans tous les écrits et entendu à la télévision, la radio et dans les discours publics et les serments religieux. Les textes sans voyelles sont considérées comme étant plus appropriée par la communauté de langue arabe puisque c'est la forme habituelle de la vie quotidienne des documents écrits et imprimés (livres, magazines, journaux, lettres, etc.)

Mais quand il s'agit du texte du « Coran », et plus généralement aux collections imprimées des livres scolaires et des dictionnaires arabes sur les supports en papier, les voyelles diacritiques apparaissent dans leurs intégralités. Comme on remarque très souvent dans des livres bien édités, des manuscrits ou bien certains textes imprimés la présence partielle ou au hasard des voyelles diacritiques sur les mots ambigus ou difficiles à lire.

Par exemple, un mot en arabe composé de deux lettres comme « **بر** » c'est à dire, «b» et «r», peut être très ambiguë, sans les voyelles diacritiques (voir Tableau 4), ou par exemple « **علم** ». Notamment pour un écrivain, il peut utiliser les signes diacritiques afin que les lecteurs puissent facilement résoudre toute ambiguïté.

1- **http://www.webarabic.com/portail/apprendre/index.php?rub=ecrire&page=3&section=les%20voyelles%20arabes**

Cependant, bien que la plupart des Arabes peuvent lire des textes avec des voyelles explicitement indiqué, moins ils peuvent écrire des textes en utilisant des voyelles diacritiques correctes. Ainsi, il est très difficile de compter sur des utilisateurs, quel que soit leur origine, pour entrer correctement des mots clés de recherche nécessitant des voyelles diacritiques. [BLA06]

Arabic word	Transliteration	POS	Meaning
بَرَ	**barr** short vowel 'a'	noun	land (as opposed to sea)
بَرَّ	**barr** short vowel '*a*'	adj	reverent, dutiful, kind
بُرُ	**burr** short vowel 'u'	noun	wheat
بِرُ	**birr** short vowel '*i*'	noun	reverence, kindness

Arabe	Translitération	PoS	Sens
عَلَم	'alam	n	flag
عِلْم	'ilm	n	science
عُلِمَ	ulima	v	known
عَلَّمَ	'allama	v	teach
عَلَم	'alam	a	famous

Tableau 4 : Voyelles diacritiques possibles sur « بر » et sur « علـم »

Pourtant, une mauvaise utilisation d'un seul signe diacritique fera un échec lors d'une requête, de recherche d'un document numérique par exemple, « **السكون** » qui indique que la consonne n'est pas suivi par une voyelle, ou la « **الشدّة** » (comme dans « **بـرّ** » 'barr' dans le tableau 4 et « **درّس** » darrasa dans le tableau 5), ce qui indique une double consonne {le premier n'est pas suivie d'une voyelle (ici « **السكون** ») et la seconde est suivie d'une voyelle}.

Arabic word	POS	Pattern	Meaning
دَرَسَ darasa	verb	فَعَلَ fa?ala	study
دَرَّسَ darrasa	verb	فَعَّلَ fa??ala	teach
دَرْس dars	noun	فَعْل fa?l	lesson
دِرَاسَة dira:sah	noun	فِعَالَة fi?a:lah	study
مُدَرِّس mudarris	noun	مُفَعِّل mufa??il	teacher
مَدْرَسَة madrasah	noun	مَفْعَلَة maf?alah	school
تَدْرِيس tadri:s	noun	تَفْعِيل taf?i:l	teaching
تَدَارَسَ tada:rasa	verb	تَفَاعَلَ tafa:?ala	discuss
دِرَاسِي dira:si	adj	فِعَالِي fi?a:li	educational

Tableau 5 : Dérivations de la racine (d r s)

Beaucoup de personnes ont tendance à faire des erreurs sur la position de certains signes diacritiques sur un mot. Cela peut poser un sérieux problème pour les systèmes de recherche d'information et les systèmes informatisés de ressources lexicales qui dépendent de l'entrée, bien formulée, de l'utilisateur. Sinon, on peut assister même à des rejets de requêtes d'utilisateurs.

En particulier, il peut y avoir un rejet total d'une nouvelle ressource lexicale solide tel qu'AWN à moins que cette nouvelle ressource suppose que la plupart des utilisateurs du discours arabe ne sont pas experts en écriture des voyelles diacritiques et par conséquent les ignorent complètement. Ces utilisateurs sont plus à l'aise quand à la lecture de textes sans signes diacritiques, dans les documents écrits de tous les jours y compris les contrats juridiques et commerciaux, journaux, livres ainsi que les dictionnaires sur des supports papiers ou numérisés. En conclusion, on peut dire qu'il est préférable de permettre aux utilisateurs d'entrer des mots en arabe sans voyelles diacritiques mais parallèlement permettre au système de retrouver ces mots avec des voyelles diacritiques pour les besoins de désambiguïsation. [BLA06]

8.2. Description d'AWN

L'Arabic WordNet est une base de données lexicale. Sa conception basé sur Princeton WordNet est construite suivant des méthodes développées pour EuroWordNet est reliée avec l'ontologie SUMO (Suggested Upper Merged Ontology). Arabic WordNet a été développé par DOI / REFLEX (2005-2007) [BLA06] (voir Figure 32).

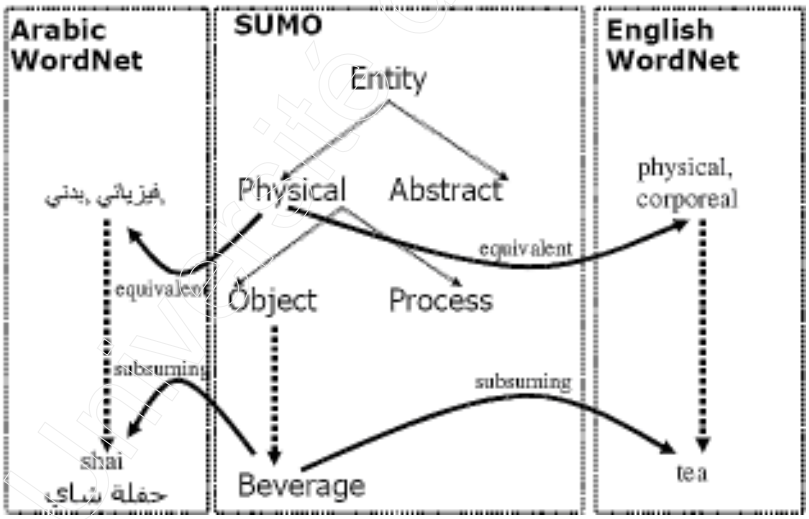

Fig. 32: Mapping de SUMO vers WordNet(s)

(Structure et organisation de l'AWN)

SUMO est en pleine expansion afin d'offrir un fondement solide pour la formalisation sémantique de l'Arabic WordNet (AWN). La base de données AWN[1] est disponible gratuitement.

L'ontologie Arabic WordNet contient 9228 concepts « Synsets » (6252 nominales et 2260 verbales, 606 adjectival, et 106 adverbales), contient 18,957 expressions et 1155 concepts nommés [**BLA06**] le fichier base de l'AWN sous format XML[2] contient les quatre balises :

- Item : Contient les concepts (Synsets), les classes et les instances de l'ontologie.
- Word : Contient les mots arabes vocalisés.
- Form : Contient les Racines des mots arabes « root ».
- Link : Contient les relations entre les concepts.

8.3. Construction d'Arabic WordNet (AWN)

AWN sélectionne les Synsets, en se basant sur des critères [**BLA06**] :

AWN doit être aussi dense que possible par des connexions de chaînes hyperonymie/hyponymie etc. En contreparties, La plupart des Synsets d'AWN doivent correspondre à leurs homologues du WordNet anglais et la topologie entière des deux Wordnets doit être similaire.

Pertinence : Donner la priorité aux concepts les plus fréquents. Ces critères incluront la fréquence des éléments lexicaux (en arabe et en anglais), PoS (Nom, verbes, adjectifs, adverbes…) et la fréquence des racines arabes dans leur corpus de référence respectifs.

Généralités : Les Synsets les plus préférés sont ceux des ontologies de hauts niveaux – WordNet. Pour assurer ces trois critères, deux façons de procéder :

- De l'anglais vers l'arabe : On sélectionne, pour chaque Synset anglais, toutes les variantes correspondantes en arabe.
- De l'arabe à l'anglais : Tous les sens d'un mot arabe doivent être trouvés dans le WordNet anglais, ainsi que chacun de ces sens il faut sélectionner ces Synsets correspondants en anglais.

1 - Disponible gratuitement dans le lien suivant : http://www.globalwordnet.org/AWN/

2 - La base sous format *XML* et une autre *MySQL disponible dans le lien suivant :* http://www.globalwordnet.org/AWN/DataSpec.html**c**

Ces deux étapes doivent être suivies tout au long de la construction d'AWN. Tous les Synsets AWN doivent être validé manuellement (et éventuellement verrouillé, lorsque toutes leurs variantes ont été trouvées), mais il convient d'exploiter, autant que possible, les ressources disponibles pour guider le processus de construction et de validation.

Une fois qu'un nouveau verbe arabe est ajouté à AWN, plusieurs possibilités d'extension sont à considérer : Les extensions des entrées verbales, y compris les dérivés verbaux1, les nominalisations et les noms verbaux2, etc. Nous considérons également les formes les plus productives comme les dérivés des pluriels brisés (**جموع التكسير**)3. Ceux-ci peut-être fait grâce à un ensemble de règles lexicales et morphologiques pour tirer un maximum de profit de ces extensions des itérations courtes seront effectuées. Pour construire AWN, il faut d'abord construire l'ensemble de la base de ses concepts (C.B.) à partir de l'ensemble de la base commune des concepts (CBCs4) d'EWN5 et BalkaNet6. La concentration est faite sur les termes les plus pertinents afin d'obtenir environ 1.000 Synsets nominal et 500 Synsets verbale.

La deuxième étape consiste en une extension verticale de haut en bas de la base des concepts. [**FAR05**] et [**DIA04**]. Certains prétraitements sont nécessaires pour la prochaine étape. Nous citons deux tâches, la préparation et extension.

La préparation : La préparation consiste au traitement des ressources disponibles bilingue et la compilation d'un ensemble de règles lexicales et morphologiques. De l'ensemble des dictionnaires bilingues disponibles, un dictionnaire bilingue homogène (HBIL) a été construit comprenant pour chaque information en entrée Arabe/Anglais, une paire de mot, la racine arabe est ajoutée manuellement, PoS (Part of speech), les fréquences relatives et les sources supportent l'appariement.

---

1 - Exemple : soit le radical ***ktb*** (**كتب**) « écrire », on peut former *les dérivés verbaux :*

- ***kataba*** (**كَتَبَ**) « écrire »,
- *i**kta-ta-ba*** (**اكتتب**) : « copier »

2 - Exemple : **EL**dhikr (**الذِكْر**) : Nom verbal tiré de la racine arabe **dhakara** (**ذَكَرَ**).

3 - Exemple : Arka:n (**أَرْكَانٌ**) : pluriels brisés de *roukn* (**رُكْنٌ**)

4 - CBCs: Set of Common Base Concepts (CBCs) from the 12 languages in EWN and BalkaNet. [TUF04]

5 - EWN : EuroWordNet est un projet visant à construire des ontologies similaires au projet WordNet de l'université de Princeton pour 8 langues européennes dont le français. [http://www.illc.uva.nl/EuroWordNet/]

6 - BalkaNet est un projet européen 2001-2004, visant à développer des Wordnets ***alignés*** pour la région des langues Balkans suivantes : Bulgare, Grec, Romanie, Serbe, Turc et d'étendre le WordNet Tchèque précédemment élaborée dans le projet EuroWordNet.

Dix sept (17) méthodes heuristiques sont utilisées pour le développement d'EWN et sont appliquées à HBIL [FAR05] pour dériver les mots candidats anglais/arabe par un simple mappage des Synsets. Pour chaque mappage, l'information attachée comprend le mot arabe et sa racine, le Synset anglais, POS, les fréquences relatives, l'évaluation du mappage, la profondeur absolue dans WordNet, un certain nombre d'écarts entre le Synset et le sommet de la hiérarchie WordNet et les sources contenants la paire.

Les mots arabes dans les ressources bilingues doit être normalisée et lemmatisée [**DIA04**], [**HAB05**], mais les voyelles et les signes diacritiques doivent être maintenus. Les racines arabes n'ont pas de voyelles.

Extension : Après le prétraitement, l'ensemble des mots marqués arabe/anglais des paires Synset deviennent une entrée à l'étape de validation manuelle. Nous procéderons par blocs d'unités connexes (ensembles de Synsets WordNet connexes, par exemple les chaînes d'hyponymie et l'ensemble des mots arabes connexes (C'est à dire, les mots ayant la même racine) au lieu des unités individuelles (Synsets, sens, mots). [**BLA06**]

Finalement, AWN sera complété par l'ajout d'une terminologie et des entités nommé1, pour combler les lacunes de sa structure qui couvre certain domaine spécifique.

8.4. L'interface Utilisateur

Outre la recherche et la navigation simple et facile sur l'ensemble de la base de données pour les utilisateurs finaux, les lexicographes ont besoin aussi d'une interface d'édition. Une variété de composants hérités sont disponibles, chacun d'eux avec ses avantages relatifs.

Dans [BLA06], William BLACK et Sabri ELKATEB et al, ont choisi d'adapter celle décrite dans Black et Elkateb (2004), car elle peut prendre en charge l'écriture arabe. Toutefois, cette méthode a présenté un tout autre modèle de données, dans lequel les mots arabes étaient directement liés aux écarts représentant les Synsets de WordNet. Elle a également été structurée pour supporter la navigation et la recherche dans un espace de Synset entièrement en anglais et par un simple mappage des mot-Synset pour introduire l'arabe.

Cette nouvelle interface a tenté de mettre les deux langues sur un même pied d'égalité et effectivement être indifférente à la direction d'alignement entre les structures

1 - **Entités nommées** : Sa reconnaissance est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mot) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.
[**http://fr.wikipedia.org/wiki/Entit%C3%A9s_nomm%C3%A9es**]

conceptuelles des deux langues. Par ailleurs, l'éditeur de l'interface communique avec le serveur de base de données en utilisant le protocole SOAP1 (Simple Object Access Protocol). Il s'agit de permettre aux multiples lexicographes de différents sites de maintenir une base de données commune.

9. Conclusion

WordNet est sans doute le précurseur et la référence en matière de base lexicales sémantiques et informatiques devenue peu à peu à caractère ontologique, une description sommaire de ce dernier à était faite et constitue donc un exemple concret de notre présentation des ontologies.

Nous avons tenté à travers ce chapitre, de donner un aperçu global de ce qu'est l'ontologie WordNet. Cette ressource lexicale nous a surpris tant dans sa structure que dans le travail et les efforts investis pour la réaliser. Ce chapitre est un avant plan de ce qu'on prévoit à étudier comme approche à la réalisation d'un WordNet arabe. Le chapitre suivant donne un état de l'art des différentes méthodes de l'apprentissage ontologique.

1 - SOAP : un protocole de communication basé sur XML pour permettre aux applications de s'échanger des informations via HTTP. [**http://igm.univ-mlv.fr/~dr/XPOSE2003/axis_seng/soap.html**]