Samuel REESE
travail dirigé par Gemma Boleda
au sein du Grup de Processament del Llenguatge Natural,
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (TALP)
Universitat Politècnica de Catalunya
avril – août 2009

WikiNet : Construction d'une ressource lexico-sémantique multilingue à partir de *Wikipedia*

L'obtention de relations sémantiques est une nécessité incontournable en sémantique lexicale de nos jours, utile pour des tâches intermédiaires telles que la désambiguisation, et par là pour des disciplines telles que la traduction automatique par exemple. Afin d'obtenir des ressources suffisamment grandes et denses à un coût raisonnable, il est nécessaire d'obtenir les relations de manière automatique. *Wikipedia* étant une source abondante d'information sémantique multilingue, ce projet a pour but d'en extraire des relations, dans plusieurs langues, puis d'analyser et de comparer les résultats obtenus. Pour ce faire on construit un modèle vectoriel du contenu de l'encyclopédie.

Table des matières

Remerciements

Tout d'abord, je remercie Dieu pour la beauté du langage et l'accès que nous avons au *logos*, à l'expression et au raisonnement, et par là à la communication et à la communion entre êtres humains et avec lui.

Ensuite, je souhaite remercier en premier lieu Gemma Boleda, qui a dirigé ce travail de recherche, m'a donné nombre de conseils utiles et a su me rappeler au bon moment que la durée du stage était finie ; puis Lluís Padró pour son aide et sa patience lorsque j'ai apporté ma contribution à la librairie *FreeLing*. Un grand merci à Montse Cuadros pour ses indications et son aide ; également à German Rigau et Horacio Rodríguez qui m'ont aiguillonné vers un travail utilisant *Wikipedia*.

Enfin, merci à tous les doctorants et stagiaires pour leur aide, pour l'ambiance sympathique, pour la coupe remportée ensemble dans le tournoi de volleyball... Et merci à Enric, Sara, Henoc et Jocabed d'avoir contribué à rendre ce séjour à Barcelone très agréable.

Introduction

Ce travail de recherche m'a permis de prendre connaissance du domaine très spécifique de l'Intelligence Artificielle qu'est le Traitement du Langage Naturel. Au sein de ce domaine il existe un grand nombre de disciplines. Le premier chapitre de ce rapport est donc une présentation permettant d'introduire les notions qui seront évoquées par la suite et de placer dans son contexte le travail effectué.

Le projet mis en œuvre au cours de ce stage avait pour but la réalisation de ressources regroupant des relations sémantiques entre des concepts (ou synsets), extraites automatiquement des *Wikipedia* en anglais, en espagnol et en catalan, et l'établissement de liens entre les ressources lexico-sémantiques obtenues dans les trois langues. L'ensemble a été baptisé WikiNet. La création de telles ressources peut être considérée comme une tâche intermédiaire en traitement du langage naturel, utile pour d'autres tâches telles que la traduction automatique ou l'extraction d'information.

Le deuxième chapitre du rapport décrira donc plus précisément la nature du projet, la méthode utilisée, et les différentes étapes nécessaires à la réalisation du projet ; les résultats obtenus seront présentés dans le troisième chapitre, avec une analyse qualitative et quantitative.

Enfin, un dernier chapitre sera consacré aux conclusions, aux utilisations possibles de la ressource créée et aux prolongements possibles du travail réalisé au cours de ce stage.

Chapitre 1

Positionnement de ce projet vis-à-vis du domaine du Traitement du Langage Naturel

1.1 Traitement du Langage Naturel et Intelligence Artificielle

1.1.1 Un peu d'histoire

Le Traitement du Langage Naturel regroupe toutes les disciplines qui se proposent de réaliser un traitement utile de textes ou de discours exprimés dans une langue humaine. De nos jours, les progrès réalisés dans ces disciplines et l'augmentation de la puissance des ordinateurs ont contribué à rendre ce domaine plus familier, grâce à l'existence de logiciels de traduction automatique, de reconnaissance et de synthèse vocale, ou encore d'extraction d'information (comme par exemple le moteur de recherche Wolfram|Alpha).

Cependant ce domaine existe depuis de nombreuses années, et s'est développé en même temps que les autres domaines de l'Intelligence Artificielle. Il constituerait même en quelque sorte un aboutissement de l'Intelligence Artificielle, dans la mesure où l'aptitude humaine à manier le langage est considérée comme un indicateur fiable de l'existence réelle d'intelligence, ce qui est le parti pris du test de Turing (*Computing machinery and Intelligence*, A. M. Turing, 1950). Des contributions furent apportées, par exemple par Shannon, dès les années 1950 (processus markoviens appliqués au langage ; grammaire formelle), l'une des premières problématiques étudiées étant la traduction automatique.

Il existe un clivage depuis la fin des années 50 entre les tenants d'une approche symbolique et ceux qui préfèrent un traitement stochastique du langage. La première approche s'appuyait sur la théorie du langage, tandis que la seconde s'appuyait sur la théorie de l'information développée par Shannon, et est née de la conférence de Dartmouth College en 1956 (qui est souvent également considéré comme l'évènement fondateur de l'Intelligence Articielle).

Depuis les années 90, l'utilisation de modèles probabilistes est devenue la norme pour la décomposition analytique, la classification en catégories grammaticales, la résolution d'anaphores... Cette approche empiriste a continué à prendre de l'importance au cours de la dernière décennie, la disponibilité de grands corpus textuels et la puissance accrue des ordinateurs permettant l'utilisation de techniques d'apprentissage supervisé et non supervisé.

Pour davantage d'information sur le Traitement du Langage Naturel en général, se référer à [1] ; pour aborder de manière plus spécifique l'approche probabiliste et statistique, l'on peut consulter [2].

1.1.2 Problématiques centrales en Traitement du Langage Naturel

Quelques-unes des disciplines du Traitement du Langage Naturel ont déjà été mentionnées ; voici une liste plus exhaustive des applications envisagées dans ce domaine :

- la traduction automatique ;
- la correction orthographique ;
- la recherche d'information ;
- le résumé automatique de texte ;
- la génération automatique de textes ;
- la synthèse de la parole ;
- la reconnaissance vocale ;
- la classification de documents ;
- les agents conversationnels.

Toutes ces applications sont des tâches de haut niveau, faisant intervenir un nombre important de tâches de traitement du langage de plus bas niveau. Pour pouvoir traiter le langage, il en faut une représentation. On est alors amené à distinguer plusieurs aspects du langage.

La **morphologie** de la langue : elle concerne les parties de mots qui ont un sens, par exemple les marques qui caractérisent le genre et le nombre, ou encore le suffixe *-ment* pour les adverbes.

La **syntaxe** correspond à l'ensemble des règles grammaticales de la langue.

La **sémantique** recouvre le sens des mots de la langue (**sémantique lexicale**), le sens de mots mis en relation avec d'autres (**sémantique compositionnelle**). Par exemple, comprendre le sens de *fin* dans *la fin du XIIIème siècle* et *la fin du jeu* relève tout à la fois de la sémantique lexicale et de la sémantique compositionnelle.

L'étude du **discours** permet d'appréhender des phénomènes qui concernent un énoncé dans sa globalité, et non des phrases prises séparément. Déterminer l'identité de *vieillard* dans

> *"En 1815, M. Charles-François-Bienvenu Myriel était évêque de Digne.* ***C****'était* ***un vieillard*** *d'environ soixante-quinze ans"*
>
> VICTOR HUGO, *Les Misérables*,
> Tome I, Livre premier, Chapitre I

relève de ce niveau d'analyse.

Pour pouvoir accomplir les tâches de haut niveau citées ci-dessus, la problématique essentielle en traitement du langage est souvent la résolution d'ambigüités à l'un des niveaux de la langue. Ainsi :

- la terminaison *-s* à la fin d'un mot peut être la marque du pluriel (*chat****s***) ou non (*relai****s***), ce qui est une instance d'ambigüité au niveau morphologique ;
- la phrase *Jean expédie un vase* ***de*** *Chine* est ambigüe du point de vue syntaxique (le vase est sans doute chinois, mais est-il expédié depuis la Chine ?) ;
- ambigüité sémantique : dans *La* ***pêche*** *est bonne*, *pêche* peut être un fruit ou une activité (ambigüité lexicale).

1.2 Sémantique lexicale et *WikiNet*

1.2.1 Sémantique lexicale

Le projet qui est présenté ici se situe plus précisément dans le domaine de la sémantique lexicale, où l'on s'intéresse au sens des mots pris individuellement. Pour désigner cette notion on peut utiliser le terme de concept (point de vue psycho-linguistique), ou encore de synset. Le mot synset fait référence à *WordNet* (cf. [4]), ressource d'importance capitale en traitement

du langage, développée à l'université de Princeton. C'est une base de données d'information linguistique, où le synset, ensemble de mots ou d'expressions considérés comme synonymes, représente l'unité fondamentale (unité sémantique). On trouve la définition suivante de synset dans la documentation de *WordNet* :

Définition 1 (Synset) *Ensemble de synonymes ; ensemble de mots que l'on peut interchanger dans un contexte donné sans altérer la valeur de vérité de la proposition dont ils font partie.*

Ils comportent un code formé à partir de leur position dans la base de données et d'une lettre indiquant leur catégorie grammaticale (adjectif, nom, adverbe ou verbe), l'ensemble de synonymes qui définit le synset, une définition ("gloss"), et éventuellement un exemple d'utilisation. Quelques exemples :

> 02383458-n : car auto automobile machine motorcar | 4-wheeled motor vehicle ; usually propelled by an internal combustion engine ; "he needs a car to get to work"
>
> 00136205-a : mouth-watering savory savoury tasty | pleasing to the sense of taste
>
> 00191458-r : bewilderedly | in a bewildered manner
>
> 01118553-v : compile | use a computer program to translate source code written in a particular programming language into computer-readable machine code that can be executed

Ces synsets peuvent être reliés entre eux par différentes relations sémantiques : antonymie (sens "contraires" : {*brûlant chaud*}/{*glacial glacé polaire froid algide*} [1]), méronymie (relation de la partie au tout) et holonymie, hyponymie (relation du plus particulier au plus général : {*lézard*}/{*reptile*}) et hyperonymie, ... *WordNet* contient l'essentiel des noms, adjectifs, verbes et adverbes de la langue anglaise, et des " *WordNet*" ont été développés dans plusieurs autres langues, dont l'espagnol et le catalan, ce qui a permis d'utiliser cette ressource pour les trois langues dans le cadre de ce projet. Il est évident que la dimension de ces ressources construites entièrement "à la main" par des spécialistes en linguistique varie en fonction de la langue ; le *WordNet* catalan est ainsi assez petit (12942 synsets), de même que le *WordNet* espagnol (15556 synsets), à la différence du *WordNet* original en anglais (65014 synsets dans la version 1.6 utilisée).

1. Où l'on représente un synset par une liste de synonymes entre accolades.

La problématique du coût explique l'intérêt de la construction automatique de ressources contenant des relations entre synsets telles que celles déjà contenues dans *WordNet*. En effet, même en anglais, le graphe des relations contenues dans *WordNet* n'est pas suffisamment dense pour garantir de bons résultats lors de l'application d'algorithmes de recherche ou d'optimisation. De plus il peut être nécessaire de disposer de telles ressources pour d'autres langues, par exemple des langues comptant bien moins de locuteurs que l'anglais, ou alors pour des domaines spécifiques (textes médicaux, techniques...). Définir des méthodes pour extraire ces relations de manière automatique à partir de corpus textuels devient donc une nécessité (voir [5]).

1.2.2 Modèles vectoriels (*Vector Space Models*)

Comment repérer de manière automatique des relations entre mots parmi les millions de mots qui constituent un corpus ? Une approche mathématique pour répondre à cette problématique consiste à construire un modèle représentant le corpus comme un espace vectoriel. On pourra alors définir la notion de proximité entre les mots du corpus, et dès lors associer à un mot les mots du corpus qui sont les plus proches de ce mot. Cette approche a été mise en œuvre notamment dans la technique dite *Latent Semantic Analysis*, utilisée depuis les années 1990 en Traitement du Langage Naturel. [2]

La méthode utilisée est la suivante : après élimination des mots vides ("stop words" en anglais : ce sont les mots non significatifs tels que *le*, *ça*, *et*...), on choisit un nombre prédéterminé D de mots parmi les mots les plus fréquents du corpus qui correspondent chacun à une dimension du modèle vectoriel, que l'on explorera pour découvrir les relations entre les mots. La coordonnée d'un mot quelconque m suivant une dimension associée à un mot M est alors le nombre d'occurrences de M dans une fenêtre de rayon f (10, 15, 20...) autour de chacune des occurrences de m dans le corpus. En associant à n mots leur vecteur de coordonnées suivant les dimensions du modèle vectoriel, on obtient une "matrice de cooccurrence" de dimension $n \times D$. Telle est donc la première étape de la construction du modèle vectoriel ; pour la mettre en œuvre de manière automatique un index du corpus doit être réalisé, repérant les positions des mots pleins ("content words" ; il s'agit de tous les mots qui ne sont pas des mots vides) et permettant ensuite la construction de la matrice de cooccurrence.

2. Voir par exemple *A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge*, de Thomas K. Landauer et Susan T. Dumais (`http ://lsi.argreenhouse.com/lsi/papers/PSYCHREV96.html`).

Cette matrice étant tout à la fois très grande et plutôt creuse, on utilise le procédé de décomposition en valeurs propres (ou singulières ; *Singular Vector Decomposition*) pour réduire à d le nombre de dimensions de l'espace vectoriel. Intuitivement cela correspond à remplacer les dimensions correspondant à deux mots très "proches" comme par exemple *voiture* et *conduire* (ou si l'on veut, deux mots qui ont tendance à apparaître dans les mêmes contextes) par une seule dimension ; on projette ainsi les mots sur ce nouvel axe qui est une combinaison des axes correspondant à *voiture* et *conduire*. D'un point de vue mathématique, on diagonalise la matrice de cooccurrence et on ne tient compte que des d valeurs propres les plus élevées.

Chaque mot est ainsi représenté par un vecteur dans le modèle vectoriel. Pour deux mots donnés, plus le cosinus entre leurs vecteurs est élevé, plus on considérera que ces mots sont semblables. On pourra alors considérer qu'il existe une relation entre les mots qui sont les plus semblables ; on remarquera cependant que cette méthode générale mesure avant tout la tendance pour un couple de mots à apparaître dans des contextes semblables, plutôt que des relations sémantiques. Cet aspect sera évoqué plus loin, lors de l'analyse des résultats (Sec. 3.1). Par ailleurs, on s'intéresse *a priori* à des mots et non à des synsets ; plusieurs sens d'un mot peuvent ainsi être confondus, ce qui fausse quelque peu les résultats. Pour construire *WikiNet*, on a donc choisi de remplacer les mots du corpus par les synsets correspondants afin de disposer de vecteurs correspondant véritablement à des synsets et non à des mots. C'est pour cette raison qu'une annotation linguistique du corpus est réalisée antérieurement à la modélisation vectorielle (voir Fig. 2.1).

1.2.3 L'attrait de l'encyclopédie *Wikipedia*

L'encyclopédie *Wikipedia* est une ressource connue du grand public, et jouit d'une extrême popularité. Lancée en 2001, elle est devenue en quelques années la plus grande et la plus consultée des encyclopédies, couvrant tous les sujets et constamment mise à jour.

Nous ne mentionnerons ici que quelques aspects de cette encyclopédie qui nous intéressent particulièrement ; pour de plus amples détails, voir [7] qui est une étude approfondie de *Wikipedia* et des travaux scientifiques relatifs à cette encyclopédie.

- *Wikipedia* est un projet **collaboratif**. C'est en quelque sorte un ouvrage collectif de l'humanité entière, ouvert à toutes les contributions. Ce mode de fonctionnement a suscité des critiques au départ, mais petit à petit des règles de conduite ont été définies et *Wikipedia* peut être considéré aujourd'hui comme une source d'information assez fiable, même lorsqu'elle est comparée à des encyclopédies conventionnelles par

exemple. C'est grâce à ce fonctionnement collaboratif que *Wikipedia* a pu croître pour atteindre des proportions gigantesques, et continue à croître à un rythme tout à fait significatif. Il constitue par là une très importante **source d'information sémantique**.

- *Wikipedia* est une encyclopédie **multilingue**, existant dans plus de 250 langues. Le projet a pour ambition, d'après Jimmy Wales, co-fondateur de *Wikipedia*, de distribuer gratuitement une encyclopédie entre les mains de tous, dans la langue de chacun.
- C'est en effet également une ressource entièrement **gratuite** ; qui plus est, le projet étant "open source", il est aisé d'obtenir le contenu intégral de *Wikipedia*, qui est régulièrement mis à disposition sous forme de sauvegardes téléchargeables de la base de données.

On a donc souhaité dans le cadre de ce projet extraire des relations de manière automatique de *Wikipedia*, dans l'espoir de parvenir à des collections de relations potentiellement assez différentes de celles qui existent déjà, et surtout pouvant être obtenues, en suivant le même protocole, dans plusieurs langues.

Chapitre 2

Construction de WikiNet

2.1 But du projet

De nombreuses tâches de traitement du langage ont besoin de l'information sémantique que constitue la donnée de relations entre synsets (par exemple, l'extraction d'information, ou la construction automatique de thésaurus ; voir [8]). L'existence de telles relations permet en effet de constituer un graphe ayant pour sommets les synsets, d'y définir une métrique et d'appliquer des algorithmes généraux sur les graphes aux tâches de traitement du langage considérées.

WordNet contient déjà des relations entre synsets, mais si cette information fournie directement par des linguistes peut être considérée fiable, elle est en revanche très coûteuse. Elle est par ailleurs lacunaire : les graphes qui en résultent se révèlent insuffisamment denses lorsqu'on les utilise pour les tâches de traitement du langage, notamment pour la désambiguisation (voir [9]).

L'extraction automatique de telles relations à partir de corpus textuels est une solution à ce problème, fournissant des graphes beaucoup plus denses, et bien sûr moins coûteux à obtenir.

Un travail d'extraction de ce type (*KnowNet*, voir [5]) a déjà été réalisé à partir d'internet. Il a effectivement permis d'obtenir un ensemble important de relations qui n'étaient pas présents dans *WordNet*. Une analyse des résultats de ce travail a d'ailleurs été effectuée, avant d'entamer la construction de *WikiNet*. Il y sera fait référence dans la partie suivante.

L'encyclopédie en ligne *Wikipedia* représente une alternative intéressante à internet. C'est en effet de loin la plus grande encyclopédie disponible ; et

outre la taille du contenu, le texte des articles est généralement riche en contenu sémantique, bien écrit et à jour des thèmes actuels et des évolutions du langage. Elle présente enfin un avantage tout à fait conséquent : celui de constituer une ressource multilingue, existant dans un grand nombre de langues, et présentant des caractéristiques semblables dans chaque langue, si l'on fait abstraction de la taille, du fait du mode de construction commun (construction collaborative de *Wikipedia*). Ce sont ces caractéristiques qui font de *Wikipedia* une ressource qui intéresse voire fascine les chercheurs en traitement du langage depuis quelques années.

L'on a donc voulu extraire de même des relations de *Wikipedia*, après avoir enrichi le corpus d'une couche d'analyse morphologique, syntaxique et sémantique. La réalisation de cette extraction en anglais, espagnol et catalan a permis dans un deuxième temps de mettre en relation les ressources obtenues dans les trois langues, et d'en faire une analyse comparative, conférant au projet un aspect multilingue. La FIGURE 2.1 illustre ce processus général.

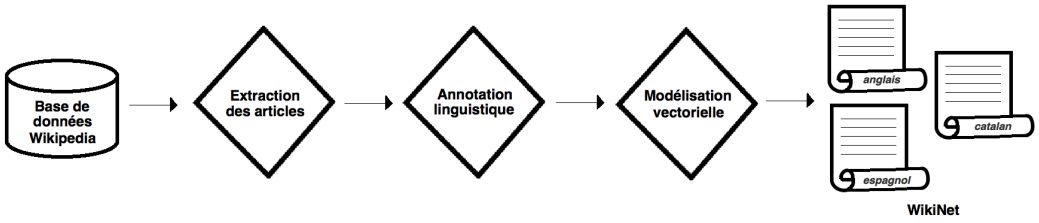

FIGURE 2.1 – Processus général de construction de WikiNet

2.2 Outils et ressources utilisés

Les programmes qui sont présentés dans les paragraphes qui suivent sont tous gratuits. À l'exception de *JWPL*, ils sont également tous "open source".

2.2.1 *Java-based Wikipedia Library*

Afin de pouvoir accéder aux pages de *Wikipedia* dans un programme de traitement, une API (*Application Programming Interface*) est nécessaire. Le programme *JWPL* fournit une telle interface. Après téléchargement d'une sauvegarde de la base de données, une base de données optimisée est construite une fois pour toutes, permettant un accès aux pages en temps pratiquement

constant. Les données sont alors associées à des objets Java, ce qui permet d'y accéder depuis un programme et d'effectuer un traitement à grande échelle de Wikipedia. La FIGURE 2.2 illustre ce fonctionnement.

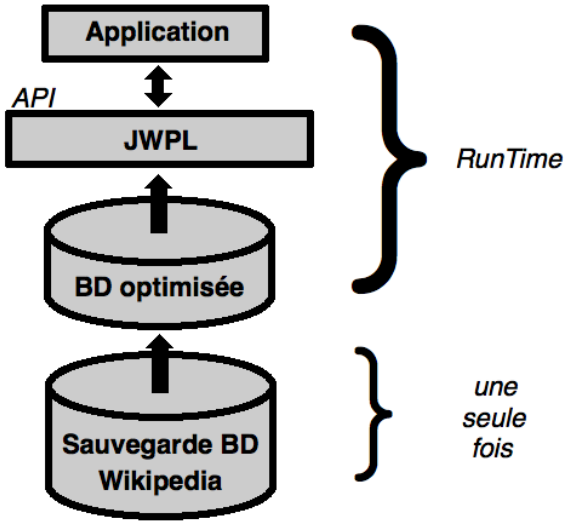

FIGURE 2.2 – Architecture de *JWPL*

2.2.2 *FreeLing*

FreeLing est une librairie qui comporte beaucoup d'outils d'analyse linguistique qui ont permis, dans le cadre de ce projet, d'effectuer le traitement linguistique préalable à l'extraction automatique des relations du corpus.

Voici les principales tâches de traitement linguistique qui ont été effectuées :

- **analyse lexicale** : identification des mots du corpus ; de manière plus générale, **segmentation** du texte, d'abord en mots, puis en phrases ;
- **lemmatisation** : à chaque mot est associé un lemme (forme canonique ou forme du dictionnaire) ;
- **identification des classes grammaticales** : la nature de chaque mot est identifiée (nom, adjectif, verbe, adverbe, interjection, pronom...) ;
- **identification des synsets** : pour les mots présents dans *WordNet*, un synset a été attribué, ce qui suppose la désambiguisation des mots ayant plusieurs sens.

Comme *FreeLing* ne comportait pas encore un module dédié à la désambiguisation, une partie du projet a consisté à intégrer dans *FreeLing* un programme dédié à cette tâche.

2.2.3 *UKB* : Graph-Based Word Sense Disambiguation and Similarity

Le programme qui a été intégré à *FreeLing*, *UKB*, effectue la désambiguisation par application d'une version modifiée de l'algorithme *PageRank* (appelée *Personalized PageRank*) au graphe dont les nœuds sont les synsets de *WordNet* et les arcs sont des relations entre ces synsets. [14] décrit cet algorithme en détail ; on se contentera ici d'une brève description de son fonctionnement.

L'algorithme *PageRank*, appliqué à un graphe G contenant N nœuds et de matrice de transition entre les nœuds M, renvoie le vecteur $\mathbf{Pr}$ solution de

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v}$$

où $\mathbf{v}$ est un vecteur de taille N dont tous les éléments valent $\frac{1}{N}$ et c est un *facteur d'amortissement* (*damping factor*) compris entre 0 et 1, et généralement fixé à 0,85. Ce vecteur représente la probabilité pour chaque nœud d'être le nœud d'arrivée suite à une marche aléatoire dans le graphe suffisamment longue.

Le vecteur $\mathbf{v}$ représente la probabilité de sauter aléatoirement (c'est-à-dire sans suivre les chemins du graphe) sur un nœud quelconque. Si choisit un $\mathbf{v}$ non uniforme, on peut ainsi rendre le saut aléatoire sur certains nœuds plus probable : c'est ce qui a été appelé *Personalized PageRank*.

Le programme *UKB* prend en entrée un graphe de relations entre synsets, y ajoute les mots d'un "contexte" (par exemple les mots d'une phrase) et crée des arcs allant de ces mots aux synsets qui leur sont associés. Le vecteur $\mathbf{v}$ est alors choisi en concentrant la probabilité initiale sur les nœuds nouvellement introduits qui correspondent aux mots du contexte. L'importance relative des synsets du graphe est alors calculée en fonction des mots du contexte.

Le graphe qui a fourni les meilleurs résultats (d'après les créateurs d'*UKB*), et qui a été choisi pour *FreeLing*, contient les relations présentes dans *WordNet*et dans *eXtended WordNet* (une collection de relations construite en partie manuellement et en partie de manière automatique).

2.2.4 *Semantic Vectors*

Le programme *Semantic Vectors* a pour vocation de réaliser des modèles vectoriels de corpus textuels comme ceux décrits dans la Section 1.2.2. Il diffère de son prédécesseur, *Infomap* (initialement utilisé pour ce projet, mais

insuffisamment “extensible” ou *scalable* pour traiter des corpus de la taille de ceux extraits de *Wikipedia* en un temps raisonnable), en ce que la technique *Latent Semantic Analysis* (qui utilise une décomposition en valeurs propres, de complexité temporelle élevée), a été remplacé par *Random Projection* (ou *Random Indexing*). Ce dernier est certes un algorithme approché, mais il permet d'obtenir de bons résultats à un coût bien moins élevé.

2.2.5 Ressources utilisées

Wikipedia Une base de données contenant une sauvegarde des *Wikipedia* en anglais, arabe, catalan et espagnol se trouvait déjà installée dans le centre de recherche du TALP. Elle a été mise à jour le 12 juin 2009.

WordNet La version de *WordNet* utilisée dans toutes les phases du projet est la version 1.6.

2.3 Étapes de la construction

Les Figure 2.3 et 2.4 explicitent chacune des étapes du processus de construction de la ressource. Cette construction s'est faite en deux phases, la première comprenant l'extraction des articles et leur annotation linguistique, la seconde permettant de passer d'un corpus désambiguisé intermédiaire au résultat final, par l'intermédiaire de la modélisation vectorielle.

La première phase a été l'occasion de créer une ressource supplémentaire contenant le texte des articles de *Wikipedia* (tout au moins ceux qui ont pu être traités au cours de la réalisation du projet) accompagné des annotations linguistiques apportées par *FreeLing* (“Wikipedia annoté”). Ce corpus est en lui-même une ressource très utile, pouvant être utilisée pour de nombreuses tâches en traitement du langage naturel (notamment toutes celles qui font intervenir de l'apprentissage, dans le cadre du traitement du langage naturel statistique), mais aussi comme objet d'étude par des linguistes ; le temps nécessaire pour effectuer le traitement linguistique étant relativement long, il est avantageux de disposer de corpus contenant déjà cette information.

En revanche, le corpus désambiguisé est plutôt à usage interne. Il contient les adjectifs, noms, adverbes et verbes du texte des articles de *Wikipedia*, soit tels qu'ils apparaissent dans le texte, soit remplacés par un code de synset lorsqu'ils ont pu être désambiguisés (on rappelle que les *WordNet* catalan et espagnol sont assez petits ; c'est pourquoi les mots qui n'ont pas été désambiguisés ont été laissés dans le corpus “désambiguisé” afin que le modèle vectoriel en tienne compte).

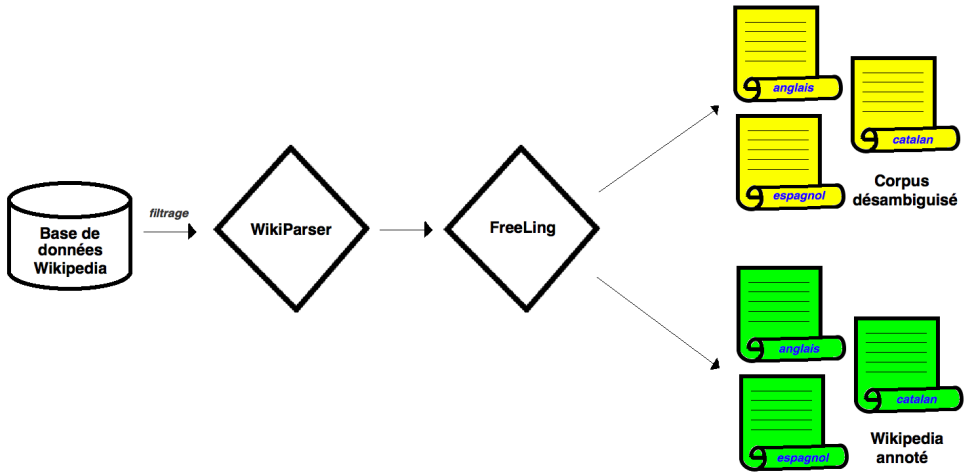

FIGURE 2.3 – De *Wikipedia* à un corpus accompagné d'annotations linguistiques

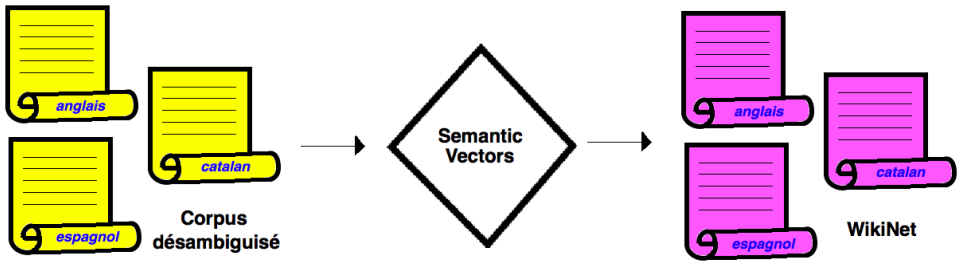

FIGURE 2.4 – Passage du corpus intermédiaire aux fichiers de relations sémantiques qui constituent WikiNet

La TABLE 2.1 donne la taille des corpus ainsi créés au cours du processus de traitement.

Les programmes utilisés ayant déjà été présentés, les paragraphes qui suivent expliquent simplement plus en détail les choix qui ont été effectués et les valeurs des paramètres.

	Wikipedia annoté		Corpus désambiguisé
	Nombre de lignes	Taille (MB)	Taille (MB)
Catalan	49 248 901	950	256
Anglais	183 967 705	3 240	1 032
Espagnol	121 796 841	2 423	639

TABLE 2.1 – Taille des corpus créés au cours du projet. Dans le corpus "*Wikipedia* annoté", le nombre de lignes donne une idée du nombre de mots dans l'ensemble des articles traités, la moyenne étant légèrement inférieure à un mot par ligne (sur chaque ligne les annotations linguistiques apparaissent après le mot auquel elles s'appliquent).

2.3.1 Choix des articles

Comme on l'a vu plus haut, *Wikipedia* est une encyclopédie collaborative, donc ouverte à toutes les contributions. Par ailleurs les pages de *Wikipedia* peuvent être des articles, mais aussi des pages de redirection ou de désambigüisation. La question du choix des articles s'est donc posée : dans l'idéal, on devrait éliminer les articles dont le texte est de mauvaise qualité, ainsi que les pages de redirection ou de désambiguation comportant de longues listes et peu de texte, afin de réduire le bruit (erreurs d'analyse notamment) pour les phases suivantes du projet.

Plusieurs critères ont été envisagés pour filtrer les articles.

- Éliminer les pages de très petite taille aurait été une bonne solution dans la mesure où cela aurait permis d'éliminer nombre de pages inintéressantes :
  - articles courts contenant peu de texte, et qui ont donc plus de chances d'être le résultat de moins de contributions et de ne pas avoir été relues plusieurs fois ;
  - pages de redirection ou de désambiguation ne contenant pas ou peu de texte.

  Malheureusement, le programme d'accès à *Wikipedia* ne comportait pas de méthode pour accéder à la taille des pages ; l'idéal étant de pouvoir filtrer les pages par simple consultation de la base de données, sans devoir lire le contenu textuel de la page, afin de ne pas pénaliser la performance du programme.
- Filtrer les pages selon le nombre de catégories qui leur ont été attribuées.
- Filtrer les pages selon le nombre de liens (entrants ou sortants) avec d'autres articles de *Wikipedia*.

La première option étant éliminée, on a déterminé (empiriquement) que le meilleur critère de filtrage était d'éliminer les pages sans catégorie, qui contenaient un pourcentage élevé de pages de redirection ; les pages sans lien entrant ou sortant étant souvent des articles tout à fait corrects du point de vue du contenu linguistique (en tout cas lorsqu'une catégorie leur a été affectée), ceux-ci n'ont pas été filtrés.

L'annexe A contient le texte des premiers articles de la base de données sans catégorie ou sans lien en espagnol.

2.3.2 Extraction du texte des articles

Seul le texte des articles en langage naturel présente un intérêt linguistique ; or les pages de *Wikipedia* sont écrites en *Wikitext*, c'est-à-dire qu'elles respectent le langage de balisage de *MediaWiki*[1] (*MediaWiki markup format*). Il s'agissait donc d'extraire les articles. L'annexe B.1 présente quelques paragraphes du texte de l'article *Agujero negro* en espagnol, extraits par le programme *JWPL*. On constate qu'il y a beaucoup de bruit dans le texte obtenu, ce qui aurait pour conséquence d'entraîner des erreurs d'analyse dans les parties suivantes du traitement.

Trouver un outil qui permette d'extraire le texte des articles n'a pas été une tâche facile. *MediaWiki* (qui est le parser utilisé par *Wikipedia*) présente les parsers alternatifs qui existent[2]. La plupart permettent de convertir un article aux formats HTML, PDF ou XML. Ils ne sont pas dédiés à l'extraction du texte des articles, et les adapter aurait nécessité un temps de développement important.

De plus, il n'existe pas de grammaire officielle permettant d'établir si un article est bien formé[3], et la grande majorité de ces parsers travaillent en plusieurs passes (bien qu'il existe un projet ayant pour ambition d'écrire un parser en une passe[4]). Afin de gagner du temps, il serait préférable de travailler en une passe.

Un parser a donc été développé en JavaCC, ayant pour objectif de reconnaître les structures principales définies par le langage de balisage de *MediaWiki* (et de permettre l'imbrication de ces structures). L'annexe B.2 contient les mêmes parties de l'article *Agujero negro* que précédemment, extraites par le *WikiParser*.

1. Moteur de wiki initialement conçu dans le but de réaliser *Wikipedia*. Voir `http ://www.mediawiki.org/wiki/MediaWiki`
2. `http ://www.mediawiki.org/wiki/Alternative_parsers`
3. Voir `http ://www.mediawiki.org/wiki/Markup_spec`.
4. `http ://www.mediawiki.org/wiki/One-pass_parser`

Ce parser n'est pas tolérant aux erreurs (oubli de crochets, d'accolades), et surtout lève de temps en temps une exception qui provoque l'arrêt du programme (dépassement de pile). Mais il extrait de manière satisfaisante le texte des articles et est relativement rapide.

L'annexe C fournit la grammaire définie dans le *WikiParser*. La TABLE 2.2 contient des statistiques sur la phase de traitement du contenu de *Wikipedia*, sur le filtrage des articles et sur les erreurs du parser qui ont été mentionnées.

	TOTAL	**FILT**	**PERR**	**PCRASH**
Catalan	160 276	5 832	3 924	7
Anglais	449 983	6 017	11 116	37
Espagnol	224 977	23 267	7 902	2

TABLE 2.2 – **TOTAL** : nombre d'articles parcourus dans la base de données ; **FILT** : nombre d'articles rejetés (filtrés car sans catégorie) ; **PERR** : nombre d'articles que le *WikiParser* n'a pas réussi à extraire ; **PCRASH** : nombre d'articles qui ont causé un dépassement de pile et l'arrêt du programme

2.3.3 Construction de modèles vectoriels des corpus

Les corpus placés en entrée du programme *Semantic Vectors* sont des corpus désambiguisés : la désambiguisation ayant été effectuée à l'étape précédente par *FreeLing*, l'on a remplacé les mots du corpus par leur synset. Les *WordNet* catalan et espagnol étant relativement petits, on a également laissé dans les corpus désambiguisés les noms, adjectifs, verbes et adverbes auxquels *FreeLing* n'avait pas attribué de synset.

Cependant, du fait du bruit inhérent au corpus (par exemple, présence de noms de fichiers qui n'ont pas été filtrés par le parser, de sorte que le "mot" *jpg* apparaît assez souvent dans le corpus), on a dû constituer une liste de mots à éliminer contenant tous les mots qui n'apparaissaient pas dans le dictionnaire de *FreeLing*. Ceci a éliminé des mots tels que *Barcelona* qui auraient autrement pu être gardés.

On a alors construit un modèle vectoriel de dimension 5000 correspondant aux 5000 mots les plus fréquents du corpus. Le rayon de la fenêtre utilisée pour compter les cooccurrences était de 7 mots (dans le cas d'un corpus typique ce rayon est choisi plus grand, mais ici un nombre important de mots du corpus a déjà été éliminé lors de la création du corpus désambiguisé).

2.3.4 Obtention des relations et construction de la ressource multilingue

Les relations entre synsets ont été obtenues en obtenant pour chaque synset la liste des mots du corpus les plus proches de ce synset (comme il a été expliqué dans la SECTION 1.2.2). On a ensuite filtré ces listes pour ne conserver que les synsets. Des fichiers contenant pour chaque synset les relations avec les n synsets les plus proches, où $n \in \{5, 10, 15, 20\}$, ont été construits ; ceux-ci constituent les collections de relations *WikiNet5*, *WikiNet10*, *WikiNet15*, *WikiNet20*.

On a enfin construit, pour chacune des versions ainsi obtenues, des tables contenant les relations qui apparaissent dans plus d'une langue, accompagnées des poids dans chaque langue (0 lorsque la relation n'apparaît pas dans la langue considérée).

Chapitre 3

Analyse des résultats

3.1 Analyse qualitative ; comparaison multilingue

Comme il a été dit à la fin de la Section 1.2.2, la méthode utilisée pour construire les relations entre synsets consiste à sélectionner les synsets dont l'ensemble des contextes dans lesquels ils apparaissent se ressemblent le plus.

Il s'ensuit que cette méthode permettra de détecter certains types relations sémantiques, tandis que d'autres ne seront pas détectées. Par exemple, il se peut que la relation de co-hyponymie *été*/*hiver* soit affectée d'un poids bien plus faible que *été*/*soleil*, dans la mesure où *soleil* apparaîtra plus souvent qu'*hiver* dans le contexte d'*été*.

Ainsi, par l'intermédiaire de la métrique utilisée, les résultats obtenus sont en partie le produit de la méthode qui a permis d'obtenir les relations (algorithme — *Random Indexing*, choix des mots vides, de la taille de la fenêtre de cooccurrence) et du corpus — *Wikipedia*.

Nous allons regarder à présent quelques exemples pour illustrer ce qui précède, et apporter par la même occasion quelques commentaires sur les résultats obtenus. Une analyse quantitative plus rigoureuse sera faite dans la section suivante. Dans tous ces exemples les synsets ont été remplacés par les mots qui les définissent.

Les relations de la Table 3.1 sont quelques-unes des relations qui ont été obtenues dans plus d'une langue ; presque toutes correspondent à des relations sémantiques typiques : antonymie ({*dry*}/{*wet*}), méronymie ({*painting picture*}/{*exhibition expo exposition*}). Mais les relations {*construct fabricate manufacture*}/{*vehicle*}, et plus encore {*film flick movie picture*}/{*begin commence get start*}, illustrent la tendance à détecter des relations séman-

construct fabricate manufacture vehicle
heir heritor inheritor boy son
painting picture exhibition expo exposition
woman man
dry wet
throne king
north south
part region country land nation state
period twelvemonth year yr
bishop bishopric diocese
campaigner candidate nominee election
newspaper paper article
film flick movie picture begin commence get start
book compose indite pen write

TABLE 3.1 – Quelques exemples sélectionnés à la main parmi les relations obtenues.

tiques plus générales : s'il est difficile de nommer la relation sémantique qui existe entre *film* et *begin*, il n'est en revanche pas très surprenant que ces mots apparaissent souvent dans le même contexte.

Les relations obtenues ne sont pas toutes de cette qualité ; certaines relations doivent sans doute être simplement considérées comme du bruit, et surtout certains mots devraient être considérés comme des mots vides : *be*, *have*... Ces mots appartiennent cependant à un synset de *WordNet*, et ont donc été inclus dans le corpus désambiguisé. Il eût été souhaitable d'effectuer un filtrage supplémentaire avant de produire le corpus désambiguisé, afin de les éliminer. Cela apparaît notamment en anglais, comme on pourra le voir dans l'exemple qui suit, le *WordNet* anglais étant plus complet.

Les Tables 3.2, 3.3, 3.4 et 3.5 contiennent, pour quelques mots apparaissant dans la Table 3.1, les cinq premières relations obtenues, dans les trois langues, accompagnées du *similarity score* attribué par *Semantic Vectors*. On peut remarquer que les relations qui paraissent un peu étranges ont généralement un *similarity score* un peu plus faible.

La Table 3.6 indique le nombre total de synsets qui interviennent dans chaque version de *WikiNet*, pour chaque langue. Les nombres de relations obtenues dans plusieurs langues sont présentés dans la Table 3.7. On observe qu'il y en a très peu. Il paraît difficile à ce stade de tirer des conclusions sur ce point. Les résultats sont entachés de bruit, et de nombreuses relations très générales (par exemple celles qui sont de la forme $\{synset\_\ quelconque\}/\{être\}$) apparaissent dans les collections de relations obtenues, or ces relations ont peu de chances d'exister dans plusieurs langues (en tout cas, dès lors que $\{synset\_\ quelconque\}$ est suffisamment peu fréquent). On peut espérer qu'en éliminant ces relations indésirables, la proportion de relations présentes dans plusieurs langues connaîtrait une augmentation significative. Par ailleurs, notons qu'une notion d'ordre est prise en compte : si une relation A-B apparaît dans deux langues dans *WikiNet5*, cela signifie que B est parmi les 5 synsets les plus proches de A dans les deux langues. Si dans une des deux langues il B se trouvait être le sixième synset le plus proche de A, la relation A-B ne serait plus détectée comme commune à ces deux langues.

L'annexe D contient les mots (définissant les synsets) qui correspondent à 50 relations de *WikiNet5* qui apparaissent dans plusieurs langues, choisies au hasard.

Catalan		Anglais		Espagnol	
Relations	*Similarity*	*Relations*	*Similarity*	*Relations*	*Similarity*
interès participació	0.8144401	art	0.71933174	libertad	0.82635534
set	0.8108523	exhibition expo exposition	0.7038002	episodio	0.7818816
exhibir exposar presentar	0.7528544	felid feline	0.69274247	análogo correspondiente	0.77929986
exposició	0.7418806	cashier teller	0.68432975	ola	0.7772615
artista	0.7371184	shopfront storefront	0.6758392	beneficiar	0.7723037

TABLE 3.2 – Cinq premières relations obtenues pour le synset 03079051-n (*pintura quadre tela* | *painting picture* | *cuadro pintura*).

Catalan		Anglais		Espagnol	
Relations	*Similarity*	*Relations*	*Similarity*	*Relations*	*Similarity*
anomenar dir	0.85251516	almost most near nigh virtually	0.82574046	ser	0.90741163
desaparició mort	0.8397927	be	0.80901605	habitar morar ocupar poblar residir vivir	0.90246314
morir	0.8369043	have hold	0.8084334	existir haber	0.89289194
afirmar dir manifestar	0.8336164	consider reckon regard see view	0.8029818	estar haber	0.88812906
polític	0.83252174	component part portion	0.79753333	tener	0.8877335

TABLE 3.3 – Cinq premières relations obtenues pour le synset 06299747-n (*estat nació país terra* | *country land nation state* | *estado nación país tierra*).

Catalan		Anglais		Espagnol	
Relations	*Similarity*	*Relations*	*Similarity*	*Relations*	*Similarity*
vot	0.8387519	campaigner candidate nominee	0.8557742	electoral	0.85082597
partit	0.7743766	place seat	0.8300655	presidencial	0.84870416
sociòleg	0.70213115	obtain	0.76482713	civil	0.8473082
elecció selecció	0.6853241	assemblage assembly gathering	0.75701356	diputado	0.84268
anomenar dir parlar	0.67562616	find get incur obtain receive	0.7427778	articulación	0.7902979

TABLE 3.4 – Cinq premières relations obtenues pour le synset 00118873-n (*elecció* | *election* | *elección*).

Catalan		Anglais		Espagnol	
Relations	*Similarity*	*Relations*	*Similarity*	*Relations*	*Similarity*
poeta	0.82754594	telecasting television video	0.7222763	encuentro	0.79124767
hegemonia	0.79045576	audience	0.70966876	asociación club clube sociedad	0.7520038
assaig experiment experimentació prova test	0.76879865	article	0.7046683	acabar cesar concluir parar terminar	0.7506484
autor escriptor	0.7542504	paper	0.67745465	artículo (04771731-n)	0.73913187
president	0.7483527	earth world	0.6720828	artículo (04739220-n)	0.73511046

TABLE 3.5 – Cinq premières relations obtenues pour le synset 04738466-n (*diari* | *newspaper paper* | *diario periódico*).

	Catalan	Anglais	Espagnol
WikiNet5	4690	11034	5805
WikiNet10	5006	11860	6107
WikiNet15	5241	12396	6355
WikiNet20	5432	12872	6570

TABLE 3.6 – Nombre total de synsets (nombre de nœuds du graphe de relations) intervenant dans chaque langue, pour chaque version de *WikiNet*.

	Catalan	Anglais	Espagnol	Plusieurs langues
WikiNet5	16395	32880	21545	345
WikiNet10	32790	65506	43090	988
WikiNet15	49185	97690	64635	1792
WikiNet20	65580	129257	86180	2745

TABLE 3.7 – Nombres de relations obtenues par langue et nombre de relations qui apparaissent dans plus d'une langue.

3.2 Analyse quantitative

Dans les paragraphes qui suivent, dans le but de mettre en lumière certaines caractéristiques des résultats obtenus, nous allons comparer *WikiNet* avec les ressources sémantiques énumérées ci-dessous.

WordNet contient la définition des synsets et un graphe de relations identifiées à la main par des linguistes ; rappelons que la version utilisée est la version 1.6.

eXtended WordNet a pour vocation d'enrichir *WordNet* en effectuant une analyse syntaxique des définitions qui accompagnent les synsets et une classification sémantique des synsets, puis en extrayant des relations entre synsets ; une partie de ces relations est ensuite révisée manuellement. [1]

KnowNet est une collection de relations sémantiques obtenues en créant des vecteurs de mots à partir d'internet, ceux-ci étant ensuite désambiguisés par l'algorithme *SSI-Dijkstra* (voir [5]).

AzarNet est une collection de relations entièrement aléatoires, construite dans un premier temps pour permettre sa comparaison avec KnowNet dans une phase d'analyse. Elle contient exactement le même nombre de

1. `http ://xwn.hlt.utdallas.edu/index.html`

relations que *KnowNet* pour chaque paire de catégories grammaticales (relations entre un nom et un nom, entre un nom et un adjectif,...).

Top Concept Ontology classifie les synsets dans un arbre ontologique, permettant d'associer à chaque synset ainsi classé une liste d'*attributs* qui sont les étiquettes du nœud où il a été placé et de l'ensemble des ascendants de ce nœud. Au premier niveau, l'arbre contient 3 nœuds : "1stOrderEntity", "2ndOrderEntity", "3rdOrderEntity" ; puis "1stOrderEntity" se divise en "Origin", "Function",... "Origin" se divise en "Natural" et "Artifact", puis "Natural" se divise en "Living", "Human", "Animal", "Plant", "Creature" ; etc. [2]

WordNet Domains a été créé en associant aux synsets des *domaines*, pris dans une liste hiérarchisée d'environ 200 domaines. Premier niveau : "doctrines", "free_time", "applied_science", "pure_science", "social_science", "factotum". Sous "doctrines" se trouvent : "archaeology", "astrology", "history", "linguistics", "literature", "philosophy", "psychology", "art", "religion" ; etc. [3]

3.2.1 Catégories grammaticales

Certaines tâches en traitement du langage naturel s'intéressant exclusivement à une catégorie grammaticale donnée, il peut être intéressant de voir la distribution des synsets participant aux relations parmi les 4 catégories que couvre *WordNet*. La TABLE 3.8 montre cette distribution pour les relations de *WordNet*, *eXtended WordNet*, *KnowNet* et *WikiNet*.

On remarque que *KnowNet* contient proportionnellement plus de noms que les autres collections de relations. La distribution selon les catégories grammaticales des synsets de *WikiNet* est plus proche de celle des relations de *WordNet*. On peut remarquer cependant qu'il y a davantage de verbes, ce que l'on peut attribuer au nombre important de relations faisant intervenir les verbes auxiliaires *être* et *avoir* qui auraient dû être filtrés en amont, comme on l'a vu plus haut.

En considérant cette distribution au niveau de chaque langue, on remarque deux anomalies supplémentaires de *WikiNet* : l'absence des adverbes en catalan et en espagnol qui est due à leur absence dans les *WordNet* de ces deux langues ; et l'absence des adjectifs en anglais (notons que *WikiNet* contient en effet un peu moins d'adjectifs que les relations de *WordNet*), due à une erreur commise lors de l'intégration du programme de désambiguisation *UKB* dans *FreeLing* (le code correspondant aux adjectifs en anglais n'est

2. Voir `http ://www.illc.uva.nl/EuroWordNet/corebcs/topont.html` et [16].
3. Voir `http ://wndomains.fbk.eu/index.html` et [17].

	a	**n**	**r**	**v**	**a/a**	**a/n**	**a/r**	**a/v**	**n/a**	**n/n**	**n/r**	**n/v**	**r/a**	**r/n**	**r/r**	**r/v**	**v/a**	**v/n**	**v/r**	**v/v**	**nbrel**
WN	19,0	62,0	1,8	17,2	16,5	1,8	0,0	0,0	1,1	51,4	9,4	0,1	0,0	8,8	8,1	0,0	1,9	0,0	0,0	0,7	180303
XWN	16,5	61,0	6,6	15,9	2,5	14,7	1,0	0,7	4,6	38,7	3,4	1,2	4,5	15,0	3,2	0,3	2,5	5,5	1,4	0,9	550922
KN5	7,7	87,2	0,3	4,7	1,0	11,1	0,2	0,0	1,7	76,7	1,6	0,0	0,4	6,2	0,5	0,0	0,1	0,5	0,1	0,0	231164
KN10	9,6	82,4	0,7	7,4	1,3	13,1	0,4	0,0	1,9	68,6	1,8	0,0	1,0	9,8	0,8	0,0	0,2	0,9	0,2	0,0	689610
KN15	10,2	79,8	1,0	9,0	1,3	13,7	0,6	0,0	2,0	64,4	1,8	0,0	1,3	11,9	1,1	0,0	0,2	1,4	0,3	0,0	1378286
KN20	10,8	77,9	1,2	10,1	1,4	14,3	0,7	0,0	2,1	61,3	1,8	0,0	1,5	13,3	1,3	0,0	0,3	1,7	0,3	0,0	2358927
WK5	9,0	64,6	1,6	24,8	1,6	5,4	2,3	0,0	5,2	43,9	14,5	0,9	2,0	14,5	7,5	0,6	0,0	0,7	0,5	0,2	70820
WK10	9,1	64,2	1,7	25	1,6	5,3	2,4	0,0	5,2	43,4	14,9	1,0	2,0	14,4	7,6	0,7	0,0	0,7	0,5	0,2	141386
WK15	9,2	64,0	1,7	25,2	1,6	5,2	2,5	0,0	5,2	43,1	15,2	1,1	2,1	14,4	7,5	0,7	0,0	0,7	0,5	0,2	211510
WK20	9,2	63,8	1,7	25,3	1,6	5,2	2,6	0,0	5,3	42,9	15,3	1,1	2,1	14,3	7,5	0,7	0,0	0,7	0,5	0,2	281017

TABLE 3.8 – **a** : adjectif ; **n** : nom ; **r** : adverbe ; **v** : verbe ; **nbrel** : nombre de relations. 4 premières colonnes : pourcentage de synsets de chaque catégorie grammaticale parmi les $2 \times$ **nbrel** synsets de l'ensemble des relations de *WordNet*, *eXtended WordNet*, *KnowNet* et *WikiNet*. Colonnes suivantes : distribution des relations en fonction des catégories grammaticales des synsets (relations adjectif/adjectif, adjectif/nom,...).

pas le même que dans les autres langues).

3.2.2 Taux de recouvrement avec *WordNet*

Construire des collections de relations automatiquement a pour intérêt d'enrichir les graphes de relations entre synsets provenant de ressources sémantiques construites manuellement telles que *WordNet*. On cherche donc à avoir tout à la fois des relations significatives et qui diffèrent des relations présentes dans *WordNet*. La TABLE 3.9 contient le taux de recouvrement avec *WordNet* de *eXtended WordNet* et des différentes versions de *WikiNet*, *KnowNet* et *AzarNet*.

	Taux de recouvrement (%)
eXtended WordNet	8,4
KnowNet5	4,8
KnowNet10	2,3
KnowNet15	1,4
KnowNet20	0,9
WikiNet5	0,3
WikiNet10	0,3
WikiNet15	0,2
WikiNet20	0,2
AzarNet5	0,0
AzarNet10	0,0
AzarNet15	0,0
AzarNet20	0,0

TABLE 3.9 – Taux de recouvrement avec *WordNet* de *eXtended WordNet*, *WikiNet*, *KnowNet* et *AzarNet*.

Les relations obtenues sont donc clairement distinctes de celles présentes dans *WordNet* ; il aurait cependant été plus rassurant d'avoir une nette décroissance de ce taux de recouvrement en allant de *WikiNet5* à *WikiNet20*, comme c'est le cas pour *KnowNet*, car avec chaque version on ajoute pour chaque synset 5 nouvelles relations avec des synsets qui sont censés être moins proches du synset d'origine que ceux des relations déjà présentes. Le taux de recouvrement est en fait assez faible, et suggère à nouveau que le niveau du bruit est élevé.

3.2.3 Proximité des concepts mis en relation

La TABLE 3.10 permet d'apprécier la pertinence des différentes collections de relations à la lumière des classifications faites par *Top Concept Ontology* et *WordNet Domains*. La distance entre deux synsets dans *Top Concept Ontology* est calculée en comptant le nombre d'attributs associés à un seul des deux synsets ; dans le cas où les synsets n'ont été associés qu'à un seul nœud de l'arbre ontologique, cela correspond à la distance entre les deux nœuds dans l'arbre. Pour *WordNet Domains*, on a considéré qu'une relation était vérifiée dans la mesure où les listes d'étiquettes des deux synsets possèdent une étiquette en commun.

Les résultats obtenus dans le cas de *Top Concept Ontology* sont sujets à caution dans la mesure où *Top Concept Ontology* ne contient actuellement que des noms et des verbes : les relations faisant intervenir des adjectifs et des adverbes ne sont donc pas pris en compte.

	Top Concept Ontology		Présence d'une étiquette commune dans *WordNet Domains* (%)
	Distance moyenne	Écart-type	
WordNet	2,9	4,2	81,0
eXtended WordNet	6,3	4,8	39,0
KnowNet5	4,5	5,3	48,8
KnowNet10	6,3	5,4	35,5
KnowNet15	7,1	5,2	29,7
KnowNet20	7,6	5,0	26,3
WikiNet5	8,8	4,1	25,5
WikiNet10	8,8	4,1	25,3
WikiNet15	8,8	4,1	25,3
WikiNet20	8,8	4,1	25,2
AzarNet5	9,2	4,3	9,8
AzarNet10	9,2	4,3	10,2
AzarNet15	9,2	4,2	10,7
AzarNet20	9,2	4,2	10,9

TABLE 3.10 – Évaluation de la qualité de *WordNet*, *eXtended WordNet*, *KnowNet*, *WikiNet* et *AzarNet* à l'aide de la classification des synsets qui a été faite dans *Top Concept Ontology* et dans *WordNet Domains*.

On peut observer que la distance moyenne des relations de *WikiNet* dans *Top Concept Ontology* n'est que très légèrement inférieure à celle des rela-

tions d'*AzarNet*. Cela a du sens dans la mesure où appartenir à des catégories ontologiques proches n'est caractéristique que de certaines relations sémantiques (hyponymie, hyperonymie, co-hyponymie...). De nombreuses relations sémantiques peuvent exister sans qu'il y ait de proximité ontologique entre les deux synsets. Par exemple, on ne s'étonnerait pas de trouver une relation sémantique entre *taxi* et *chauffeur* recevant un *similarity score* plus élevé que la relation entre *taxi* et *voiture*, même si ces deux derniers sont clairement plus proches d'un point de vue ontologique. Ainsi les relations de *WikiNet* ne sont pas essentiellement des relations ontologiques.

En revanche, les étiquettes de *WordNet Domains* montrent que *WikiNet* est tout de même loin d'être une collection de relations aléatoires. Il n'est pas surprenant qu'en extrayant des relations de *Wikipedia* à partir de la cooccurrence des synsets dans des contextes semblables on obtienne des relations entre des synsets qui peuvent être associés à un même domaine. La décroissance de version en version du pourcentage de relations confirmées par *WordNet Domains* existe mais est bien plus faible que pour *KnowNet* ; on remarque cependant que les valeurs de ce pourcentage viennent en quelque sorte prolonger la suite de valeurs obtenues pour les différentes versions de *KnowNet*, et on peut penser que du fait du niveau de bruit, *WikiNet* est globalement une collection de relations de la qualité d'un hypothétique *KnowNet25* ou *KnowNet30*.

Chapitre 4

Perspectives et conclusions

La construction de *WikiNet* est un projet ambitieux, d'une part du fait de la taille de *Wikipedia*, d'autre part parce que l'accès à l'information sémantique qui s'y trouve requiert un effort (temps de développement, temps de calcul). Vu la richesse de l'information contenue dans cette encyclopédie, les bénéfices potentiels pour le traitement du langage naturel sont suffisants pour justifier un tel effort.

L'un des problèmes auxquels ce projet tente d'apporter une solution est la faible densité des ressources sémantiques construites à la main ; or on peut constater que le graphe de WordNet a une densité de 0,0085 %, tandis que le graphe de la partie en anglais de *WikiNet20*, par exemple, a une densité de 0,16 %, ce qui est une amélioration sensible.

Toutefois, deux problèmes importants rendent la ressource peu utilisable en l'état actuel. D'une part, le bruit est considérable ; cela parce qu'il résulte de l'accumulation de plusieurs facteurs : bruit dû au tagger de *FreeLing* (qui attribue les classes grammaticales ; ce bruit est minime, le tagger ayant un taux de réussite de 97 %), bruit dû à la désambiguation (ici le taux de réussite est plutôt de l'ordre de 60 %, 70 % au mieux) ; mais également erreurs commises en cours de projet qui n'ont pas pu être corrigées par la suite (le temps de calcul étant long devant la durée totale du projet), comme par exemple l'oubli de filtrer certaines catégories grammaticales (verbes auxiliaires, certains adverbes) ; enfin le choix de conserver dans le corpus désambiguisé les noms, adjectifs, adverbes et verbes auxquels un synset n'avait pas été attribué (dans le but de parer aux lacunes des *WordNet* espagnol et catalan) a compliqué la tache et sans doute influé sur les résultats.

D'autre part, la petite taille de *WikiNet* pose problème. Si l'on choisit d'augmenter le nombre de synsets pour lesquels des vecteurs sont calculés (en prenant par conséquent des synsets qui apparaissent moins souvent dans le corpus, la barre ayant été fixée à 100 occurrences), on perdra en qualité.

Or le corpus du catalan (la version catalane de *Wikipedia* a pu être traitée dans son intégralité), à moins de 50 millions de mots, est malheureusement assez petit.

Pour ce qui est de la comparaison entre les langues, force est de constater que peu de relations ont été trouvées en commun, dans plus d'une langue. Si l'on parvenait à éliminer une partie du bruit, on aurait ainsi plus de relations significatives dans chaque langue, et on pourrait espérer avoir davantage de relations en commun. Cependant la diversité des langues doit aussi être prise en compte. Deux notions distinctes dans une langue peuvent être confondues dans une autre. Il conviendrait de mettre en correspondance des groupes de synsets d'une langue à l'autre. L'*Inter-Lingual Index*[1] utilisé par la base de données multilingue *EuroWordNet* pour mettre en relation les *WordNet* de nombreuses langues, pourrait être utilisée à cette fin.

1. Voir `http ://www.illc.uva.nl/EuroWordNet/index.html#6`.

Annexe A

Filtrage des articles

Quelques pages de *Wikipedia* en espagnol pour expliquer le choix de filtrer les articles sans catégorie. Les articles apparaissent ici tels quels (ce qui permet notamment de voir les catégories) ; rappelons cependant que nombre d'éléments seront éliminés par le WikiParser : tableaux (`{{...}}`), catégories (`[[Categoría :...]]`)...

- 3 premières pages sans catégorie :

```
108
Achmatherum
#REDIRECT[[Stipa]]


149
Achlaena
#REDIRECT[[Arthropogon]]


150
Achneria
#REDIRECT[[Eriachne]]
```

- 3 premières pages sans lien entrant :

```
47
Francesc Aguilar Villalonga
{{referencias}}
'''Francesc Aguilar Villalonga''' ([[Alaquàs]], [[Provincia de
Valencia|Valencia]], [[España]], [[1942]]) es un pintor,
dibujante y comentarista de arte español.

== Biografía ==

Ha estudiado Dibujo en la Escuela de Artes y Oficios de
Aldaia (Valencia), Dibujo y Modelado en la Academia Fuster
de Valencia y Dibujo y Pintura en la Escuela Superior de
Bellas Artes de San Carlos de Valencia.

Ha realizado exposiciones individuales en [[Mallorca]],
[[Valdepeñas (Ciudad Real)]], [[Valladolid]] y [[Logroño]], ha
participado en certámenes, ferias de arte y exposiciones
colectivas en [[España]], [[Alemania]], [[Francia]], [[Italia]] y
[[Estados Unidos]].

Dentro de un estilo [[neo-impresionismo|neo-impresionista]],
dedica su atención a casi todos los temas, tanto paisajes,
marinas, figuras, naturalezas muertas y retratos. En dibujo
trabaja principalmente la tinta con plumilla y también a la
caña.

Su obra puede verse, entre otras colecciones, en el [[Museo
de La Rioja]] en [[Logroño]], [[España]] y en el [[Florida
Museum of Hispanic and Latin American Art]] de [[Miami]],
[[Estados Unidos]].

Es Académico por la [[Academia Internacionale Greci-
Marino]] de [[Italia]] y Miembro Honorífico del [[Florida
Museum of Hispanic and Latin American Art]] de [[Miami]],
[[Estados Unidos]].

Su biografía ha sido publicada en varios Diccionarios y
Catalogos en España e Italia.
```

```
{{DEFAULTSORT:Aguilar Villalonga, Francesc}}

[[Categoría:Pintores de España del siglo XX]]
[[Categoría:Pintores de España del siglo XXI]]
[[Categoría:Pintores de la Comunidad Valenciana]]
[[Categoría:Nacidos en 1942]]


108
Achmatherum
#REDIRECT[[Stipa]]


150
Achneria
#REDIRECT[[Eriachne]]
```

- 3 premières pages sans lien sortant :

5436
Atención sociosanitaria
La '''atención sociosanitaria''' son los servicios sanitarios específicos para la tercer edad y los enfermos crónicos. En especial, han de buscar el aumento del [[autovalimiento]] (autonomía) del enfermo, paliar sus limitaciones o sufrimientos (en especial, en el momento terminal) y facilitar, además, su reinserción social.

En la planificación de dicha atención se ha de tener en cuenta la tipología de las personas que requieren atención sociosanitaria, el modelo de atención, el catálogo de prestaciones, los recursos y los aspectos organizativos y líneas generales y específicas de actuación.

Estos servicios han de estar convenientemente coordinados con los servicios sanitarios, para garantizar la continuidad de la atención sanitaria. Asimismo, se ha de fomentar la atención de las personas mayores por médicos geriatras en los centros sanitarios.

La prestación sociosanitaria ha estado tradicionalmente discriminada en el sistema de [[Seguridad Social]], ya que se trasladaba dicha carga a las familias. Con el envejecimiento de las sociedades occidentales, se hace necesario el prestar una especial atención a este tipo de atención y, en especial, a que no decrezca la ratio plazas/población mayor (lo que no sucede si se mantienen las mismas plazas que en la actualidad, ya que la población mayor aumenta cada vez más cada año).

[[Categoría:Seguridad social]]

[[ca:Atenció sociosanitària]]

6380

CCIR
'''CCIR''' son las siglas de '''Comité Consultivo Internacional de Radiocomunicaciones''' - ''International Radio Consultative Committee'' - ''Comité Consultatif International des Radiocommunications'', antiguo nombre del comité de normalización de las radiocomunicaciones en la [[UIT]] ahora conocido como '''UIT-R'''.

[[Categoría:Unión Internacional de Telecomunicaciones]]

[[ca:Comité Consultiu Internacional de Radiocomunicació]]
[[da:International Telecommunication Union, ITU Radiocommunication Sector]]
[[de:Comité Consultatif International des Radiocommunication]]
[[en:ITU-R]]
[[it:ITU-R]]
[[ja:ITU-R]]
[[nl:ITU-R]]
[[pl:CCIR]]
[[sv:CCIR]]

6422
Conversión de unidades
La '''conversión de unidades''' es la transformación de una unidad en otra.

Un método para realizar este proceso es con el uso de los [[factores de conversión]] y las muy útiles [[tablas de conversión]].
Bastaría multiplicar una fracción (factor de conversión) y el resultado es otra medida equivalente en la que han cambiado las unidades.

Cuando el cambio de unidades implica la transformación de varias unidades se pueden utilizar varios factores de

conversión uno tras otro, de forma que el resultado final será la medida equivalente en las unidades que buscamos, por ejemplo si queremos pasar 8 metros a yardas, lo único que tenemos que hacer es multiplicar 8(0.914)=7.312 yardas.

== Herramientas de software ==
Los ordenadores de oficina u hogar suelen disponer de aplicaciones de hojas de cálculo o pueden acceder a convertidores gratuitos por medio de Internet.

== Enlaces externos ==

*[http://webs.sinectis.com.ar/alejand/mm/pagina_mm.htm Programa freeware para conversión de unidades]
*[http://illusions.hu/index.php?task=100&lang=0&statpage=86 Programa freeware para conversión de unidades]
*[http://www.metricimperial.com Tablas de conversión]
*[http://www.convertworld.com/es/ Tablas de conversión]
*[http://www.valvias.com/prontuario-conversor-de-unidades.php Conversor de Unidades Online]

[[Categoría:Metrología]]

[[bs:Pretvorba mjernih jedinica]]
[[en:Conversion of units]]
[[fo:Eindarumrokning]]
[[fr:Conversion des unités]]
[[hr:Pretvorba mjernih jedinica]]
[[hu:Mértékegységek átszámítása]]
[[ja:???????]]
[[pt:Tabela de conversão de unidades]]
[[sh:Konverzija mjernih jedinica]]
[[ur:????? ???????]]
[[zh-min-nan:Tan-?i ê o??-s?g-pió]]

Annexe B

Extraction du texte des articles

B.1 Début de l'article *Agujero negro* extrait par *JWPL*

Agujero negro [[Archivo:M87 jet.jpg|thumb|350px|El núcleo de la galaxia elíptica gigante M87, donde -hay evidencia de un agujero negro supermasivo. También se observa un potente chorro (jet) de materia eyectada por los poderosos campos magnéticos generados por éste. Imagen tomada por el Telescopio espacial Hubble.]] Un agujero negro u hoyo negro es una región del espacio-tiempo provocada por una gran concentración de masa en su interior, con enorme aumento de la densidad, lo que provoca un campo gravitatorio tal que ninguna partícula material, ni siquiera los fotones de luz, puede escapar de dicha región. La curvatura del espacio-tiempo o «gravedad de un agujero negro» provoca una singularidad envuelta por una superficie cerrada, llamada horizonte de sucesos.

[...]

Clasificación teórica Según su origen, teóricamente pueden existir al menos tres clases de agujeros negros: Agujeros negros primordiales Aquellos que fueron creados temprano en la historia del Universo. Sus masas pueden ser variadas y ninguno ha sido observado. Según la masa Agujeros negros supermasivos: con masas de varios millones de masas solares. Son el corazón de muchas galaxias. Se forman en el mismo proceso que da origen a las

componentes esféricas de las galaxias. Agujeros negros de masa estelar. Se forman cuando una estrella de masa 2,5 mayor que la masa del Sol se convierte en supernova e implosiona. Su núcleo se concentra en un volumen muy pequeño que cada vez se va reduciendo más. Según el momento angular Un agujero negro sin carga y sin momento angular es un agujero negro de Schwarzschild, mientras que un agujero negro rotatorio (con momento angular mayor que 0), se denomina agujero negro de Kerr. Zonas observables thumb|250px|Visión de un artista de un agujero negro con disco de acreción. [[Archivo:Accretion_disk.jpg|thumb|250px|Impresión de un artista de un agujero negro con una estrella del compañero de cerca que se mueve en órbita alrededor que excede su límite de Roche. la materia en que cae forma un disco de acrecimiento, con algo de la materia que es expulsada en chorros polares colimados altamente energéticos.]] En las cercanías de un agujero negro se suele formar un disco de acrecimiento. Lo compone la materia con momento angular, carga eléctrica y masa, la que es afectada por la enorme atracción gravitatoria del mismo, ocasionando que inexorablemente atraviese el horizonte de sucesos y, por lo tanto, lo incremente. TEMPLATE[VT, Acreción] En cuanto a la luz que atraviesa la zona del disco, también es afectada, tal como está previsto por la Teoría de la Relatividad. El efecto es visible desde la Tierra por la desviación momentánea que produce en posiciones estelares conocidas, cuando los haces de luz procedentes de las mismas transitan dicha zona.

B.2 Début de l'article *Agujero negro* extrait par *WikiParser*

```
Un agujero negro u hoyo negro es una región del espacio-
tiempo provocada por una gran concentración de masa en
su interior, con enorme aumento de la densidad, lo que
provoca un campo gravitatorio tal que ninguna partícula
material, ni siquiera los fotones de luz, puede escapar de
dicha región.

La curvatura del espacio-tiempo o "gravedad de un agujero
negro" provoca una singularidad envuelta por una superficie
cerrada, llamada horizonte de sucesos.
```

[...]

```
 Clasificación teórica .

Según su origen, teóricamente pueden existir al menos tres
clases de agujeros negros:

 Agujeros negros primordiales .

Aquellos que fueron creados temprano en la historia del
Universo. Sus masas pueden ser variadas y ninguno ha
sido observado.

 Según la masa .

;Agujeros negros supermasivos: con masas de varios
millones de masas solares. Son el corazón de muchas
galaxias. Se forman en el mismo proceso que da origen a
las componentes esféricas de las galaxias.

;Agujeros negros de masa estelar. Se forman cuando una
estrella de masa 2,5 mayor que la masa del Sol se convierte
en supernova e implosiona. Su núcleo se concentra en un
volumen muy pequeño que cada vez se va reduciendo más.

 Según el momento angular .
```

Un agujero negro sin carga y sin momento angular es un agujero negro de Schwarzschild, mientras que un agujero negro rotatorio (con momento angular mayor que 0), se denomina agujero negro de Kerr.

Zonas observables .

En las cercanías de un agujero negro se suele formar un disco de acrecimiento. Lo compone la materia con momento angular, carga eléctrica y masa, la que es afectada por la enorme atracción gravitatoria del mismo, ocasionando que inexorablemente atraviese el horizonte de sucesos y, por lo tanto, lo incremente.

En cuanto a la luz que atraviesa la zona del disco, también es afectada, tal como está previsto por la Teoría de la Relatividad. El efecto es visible desde la Tierra por la desviación momentánea que produce en posiciones estelares conocidas, cuando los haces de luz procedentes de las mismas transitan dicha zona.

Annexe C

Grammaire du *WikiParser*

La définition grammaticale proposée d'un article est la suivante :

- Terminaux :

$$
\begin{aligned}
NORMALCHAR &\rightarrow (\sim [\text{“\{”}, \text{“}<\text{”}, \text{“[”}, \text{“]”}, \text{“}>\text{”}, \text{“\}”}, \text{“|”}, \text{“\#”}, \\
&\qquad \text{“}*\text{”}, \text{“}:\text{”}, \text{“;”}, \text{“,”}, \text{“'”}, \text{“}\backslash n\text{”}, \text{“}=\text{”}]) + \\
TAGL &\rightarrow \text{“}<\text{”} \\
TAGR &\rightarrow \text{“}>\text{”} \\
BL &\rightarrow \text{“[”} \\
BR &\rightarrow \text{“]”} \\
BRACEL &\rightarrow \text{“\{”} \\
BRACER &\rightarrow \text{“\}”} \\
APOST &\rightarrow \text{“'”} \\
COLON &\rightarrow \text{“}:\text{”} \\
BAR &\rightarrow \text{“|”} \\
NEWLINE &\rightarrow (\text{“}\backslash n\text{”}) + \\
EQUAL &\rightarrow \text{“}=\text{”} \\
PUNCTUATION &\rightarrow (\text{“\#”}|\text{“}*\text{”}|\text{“;”}|\text{“,”}) +
\end{aligned}
$$

- Non-Terminaux :

$$
\begin{aligned}
Article &\rightarrow WellFormed \quad EOF \\
WellFormed &\rightarrow Treatment* \quad Text \quad WellFormed1? \\
WellFormed1 &\rightarrow Treatment+ \quad Text? \quad WellFormed1? \\
Text &\rightarrow NormalText1 \quad Text1?
\end{aligned}
$$

$$
\begin{aligned}
 & \rightarrow Text1 \\
Text1 & \rightarrow NEWLINE \quad Line? \quad Text1? \\
Line & \rightarrow ListFormatting \quad NormalText? \\
 & \rightarrow NormalText \\
ListFormatting & \rightarrow (PUNCTUATION|BAR|COLON)+ \\
Apostrophe & \rightarrow APOST \quad APOST+ \\
 & \rightarrow APOST \\
Equals & \rightarrow EQUAL \quad EQUAL+ \\
 & \rightarrow EQUAL \\
NormalText & \rightarrow (NORMALCHAR|Apostrophe|Equals) \\
 & \quad (NORMALCHAR|PUNCTUATION|BAR| \\
 & \quad COLON|Apostrophe|Equals)* \\
NormalText1 & \rightarrow (NORMALCHAR|PUNCTUATION|BAR| \\
 & \quad COLON|Apostrophe|Equals)+ \\
NormalText2 & \rightarrow (NORMALCHAR|PUNCTUATION|BAR| \\
 & \quad COLON|NEWLINE|Apostrophe|Equals)+ \\
NormalText3 & \rightarrow (NORMALCHAR|PUNCTUATION| \\
 & \quad Apostrophe|Equals)+ \\
Treatment & \rightarrow (Template|Tag|Bracket) \\
Eliminate & \rightarrow NormalText2 \quad Eliminate1? \\
 & \rightarrow Eliminate1 \\
Eliminate1 & \rightarrow Treatment+ \quad NormalText2? \quad Eliminate1? \\
Template & \rightarrow BRACEL \quad Eliminate \quad BRACER \\
Tag & \rightarrow TAGL \quad Eliminate \quad TAGR \\
Bracket & \rightarrow BL \quad Bracket1 \quad BR \\
Bracket1 & \rightarrow BL \quad Link \quad BR \\
 & \rightarrow Eliminate \\
Link & \rightarrow NormalText3 \quad Link1? \\
 & \rightarrow Link1 \\
Link1 & \rightarrow COLON \quad Eliminate? \\
 & \rightarrow BAR \quad Eliminate?
\end{aligned}
$$

Annexe D

Relations obtenues dans plusieurs langues

Voici 50 relations, choisies au hasard, parmi les 345 relations apparaissant dans plusieurs langues dans *WikiNet5*. Les synsets sont présentés ici tels qu'ils apparaissent dans le dictionnaire anglais de *FreeLing* (il est possible que deux synsets différents y soient associés à la même liste de synonymes, ce qui fait qu'on ne peut plus les distinguer sous cette forme) ; la mention `Synset not found !` correspond à des synsets qui ne sont pas présents dans le dictionnaire de *FreeLing* (les relations correspondantes étaient donc forcément des relations présentes uniquement parmi les relations en catalan et en espagnol de *WikiNet5*).

```
 see
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 advert cite mention name refer
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 can
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 associate connect link relate
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 participant player
 play
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
 bring convey take
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 avoid
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 song
 disc disk saucer
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 choose select take
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 extreme utmost uttermost
 be exist
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 ensue result
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 aim design intent intention purpose
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 intend mean think
 clip time
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 consequence effect issue outcome result upshot
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 leave
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Synset not found!
Synset not found!
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 film flick movie picture
 best
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 construe interpret
 acting performing playacting playing
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 detect discover find notice observe
 be
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 development growing growth maturation ontogenesis ontogeny
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 go locomote move travel
 twelvemonth year yr
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 comprise contain incorporate
 be exist
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 song
 subject theme topic
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 football
 choice pick selection
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 apply employ use utilize
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 end ending
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 get make
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 be
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 act
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 triumph victory
 acquire get
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 woman
 man
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 blood-red carmine cerise cherry-red cherry crimson red
      reddish ruby-red ruby ruddy scarlet
 colour colouring
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
 persist remain stay
 twelvemonth year yr
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 clip time
 cognize know
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 time
 cognize know
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 throne
 king
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 comprise contain incorporate
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 elevate lift raise
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 give
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 component constituent element factor ingredient
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 continue keep proceed
 twelvemonth year yr
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 end
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 accomplish achieve attain reach
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 address cover deal handle plow treat work
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 furnish provide render supply
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 experience
 fashion manner mode style way
```

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 newspaper paper
 article
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 berth office place position post situation slot spot
 be
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 differentiate distinguish mark
 have hold
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 chance opportunity
 be
```

Bibliographie

[1] Daniel Jurafsky, James H. Martin. *Speech and Language Processing : An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2008.

[2] Chris Manning, Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[3] D. A. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

[4] Christiane Fellbaum. *WordNet : an electronic lexical database*. MIT Press, 1998.
`http://wordnet.princeton.edu/`

[5] Montse Cuadros, German Rigau. *KnowNet : Building a Large Net of Knowledge from the Web*. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).

[6] Dominic Widdows. *Geometry and Meaning*. CSLI Publications, 2004.

[7] Olena Medelyan, Catherine Legg, David Milne, Ian H. Witten. *Mining meaning from Wikipedia*. September 2008.
`http://arxiv.org/abs/0809.4530`

[8] Dekang Lin. *Automatic Retrieval and Clustering of Similar Words*. 1998.
`http://acl.ldc.upenn.edu/P/P98/P98-2127.pdf`

[9] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, Robert Schapire. *Adding Dense, Weighted Connections to WordNet*. October 2005.
`http://wordnet.cs.princeton.edu/papers/wordnetplusintro.pdf`

[10] Torsten Zesch, Iryna Gurevych, Max Mühlhäuser. *Analyzing and Accessing Wikipedia as a Lexical Semantic Resource*. Biannual Conference of the Society for Computational Linguistics and Language Technology (pp. 213-221), 2007.

[11] Torsten Zesch, Christof Müller, Iryna Gurevych. *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary*. Proceedings of the Conference on Language Resources and Evaluation (LREC), 2008.

[12] Jordi Atserias, Bernardino Casa, Elisabet Comelles, Meritxell González, Lluís Padró, Muntsa Padró. *FreeLing 1.3 : Syntactic and semantic services in an open-source NLP library*. Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006), ELRA. Genoa, Italy. May 2006.
`http://www.lsi.upc.edu/~nlp/freeling`

[13] Xavier Carreras, Isaac Chao, Lluís Padró, Muntsa Padró. *FreeLing : An Open-Source Suite of Language Analyzers*. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), 2004.

[14] Eneko Agirre, Aitor Soroa. *Personalizing PageRank for Word Sense Disambiguation*. Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece.

[15] Dominic Widdows, Kathleen Ferraro. *Semantic Vectors : A Scalable Open Source Package and Online Technology Management Application*. 2008.
`http://www.lrec-conf.org/proceedings/lrec2008/pdf/300_paper.pdf`

[16] Javier Álvez, Jordi Atserias, Jordi Carrera, Salvador Climent, Antoni Oliver, German Rigau. *Consistent annotation of EuroWordNet with the Top Concept Ontology*. Proceedings of the 4th Global WordNet Conference, Szeged, Hungary (2008).
`http://cv.uoc.es/~grc0_001091_web/files/Alvez-et-al-GWA2008.pdf`

[17] Bernardo Magnini, Gabriela Cavaglià. *Integrating Subject Field Codes into WordNet*. Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000).
`http://wndomains.fbk.eu/publications/lrec-2000.pdf`

[18] H. Zaragoza, J. Atserias, M. Ciaramita, G. Attardi. *Semantically Annotated Snapshot of the English Wikipedia*. 2007.
`http://www.yr-bcn.es/semanticWikipedia`