Ministère de l'Enseignement Supérieur et de la Recherche scientifique

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار - عنابة

Faculté des Sciences de l'Ingénieur

Département d'Informatique

Année : 2016/2017

THESE

Présentée en vue de l'obtention du diplôme de **DOCTORAT en Informatique**

Modélisation sémantique et indexation

Option : système d'information et des connaissances

Par

Bakhouche Abdelaali

Devant le Jury

Nacira Ghoualmi	**Président**	*Prof. Université d'Annaba*
Yamina Tlili-Guiassa	**Directeur de Thèse**	*Prof. Université d'Annaba*
Halima ABIDET- BAHI	**Examinateur**	*Prof. Université d'Annaba*
Mohammed Redjimi	**Examinateur**	*Prof. Université de Skikda*
Yacine Lafifi	**Examinateur**	*Prof. Université de Guelma*
Siham Ouamour	**Examinateur**	*Prof. Université d'Alger*

Année Universitaire : 2016/2017

To translate is one thing
to say how we do it, is another.

Haas W. (1962), Philosophy, vol. 37

Le sens n'a de sens que
si on lui en donne et vice versa

Anonyme, fin du XXe siècle
(Cité par Kleiber, 1999)

A la mémoire de mes parents

A mon épouse.

A mes enfants, que Dieu les protège

A toute ma famille.

Remerciements

Je tiens à exprimer ma profonde reconnaissance à la directrice de cette thèse, Mme. Tlili-Guiassa Yamina, pour m'avoir dirigée, guidée, conseillée, pendant tout le déroulement de ce travail, dont j'espère être à la hauteur de ses attentes.

Des remerciements particuliers à Dr. Didier Schwab pour son accueil au sein de l'équipe GETALP – Laboratoire LIG, Université de Grenoble Alpes. J'ai particulièrement apprécié l'atmosphère scientifique qu'ils m'ont offert et le temps qu'ils m'ont consacré. Je tiens également à remercier à tous son équipe spécialement Andon Tchechmedjiev et Mohammad Nasiruddin.

Je remercie sincèrement Mme. Nacira Ghoualmi, professeur à l'université d'Annaba, qui m'a fait l'honneur de présider ce jury. Je tiens également à remercier Mme. Halima Abidet- Bahi, professeur à l'université d'Annaba, Mr. Mohammed Redjimi, professeur à l'université de Skikda, Mr. Yacine Lafifi professeur à l'université de Guelma et Mme Ouamour Siham professeur à l'université d'Alger pour avoir acceptés d'examiner ma thèse.

ملخص:

التحديد الآلي لمعنى كلمة حسب السياق، هي خطوة رئيسية لعدة تطبيقات لغوية. عدة نماذج اقترحت للغات مختلفة كالإنجليزية والفرنسية ولغات أخرى. لكن الدراسات في هذا المجال بالنسبة للغة العربية بقيت محدودة. العائق الرئيسي هو نقص المصادر اللغوية الإلكترونية كذلك طبيعة اللغة المعقدة. في هذه المخطوطة نقدم خطوات لإنشاء قاعدة بيانات لغوية، نستغل في ذلك مصدر الويب Wiktionnaire وأنطولوجية Wordnet، بعد ذلك نبين خطوات إنشاء نظام تحديد معنى كلمة حسب السياق، نعتمد في ذلك على خوارزمية Lesk كطريقة محلية وخوارزمية مستعمرة النمل كطريقة شاملة. نستعمل الطريقة المحلية لقياس التقارب بين لفظين في حين نستعمل الطريقة الشاملة لنشر هذا القياس على نطاق أوسع. في النهاية نقيم هذا النظام باستعمال مدونة مستخرجة من نصوص عربية.

كلمات مفتاحيه: المعالجة الآلية للغة العربية، التحديد الآلي لمعنى كلمة، خوارزمية Lesk، خوارزمية مستعمرة النمل، الطريقة المحلية / الشاملة، Wiktionnaire، Wordnet

Résumé :

La désambiguïsation lexicale est une tâche fondamentale pour la plupart des applications de traitement automatique des langages naturels. Plusieurs solutions ont été proposées pour certain langues notamment l'anglais ou le français, mais les études dans ce domaine pour la langue arabe restent limitées. Le principal obstacle est le manque de ressources et l'ambigüité de la langue. Dans ce manuscrit, nous présentons les étapes pour construire une base lexicale, en exploitant les ressources Web comme Wiktionnaire et le WordNet qui sont devenues des sources intéressantes pour l'extraction d'information. Nous présentons ensuite la notion d'algorithme local et d'algorithme global pour la désambiguïsation sémantique des textes arabes. Un algorithme local permet de calculer la proximité sémantique entre deux objets lexicaux, cependant l'algorithme global permet de propager ces mesures locales à un niveau supérieur. Nous nous servons de cette notion pour confronter un algorithme à colonies de fourmis à d'autres méthodes issues de l'état de l'art. En les évaluant sur un corpus. Les résultats obtenus nous ont menée à découvrir les facteurs qui ont influencé la performance de ce système qui seront sujets d'éventuelle amélioration dans nos future travaux de recherches.

Mots clés : Traitement automatique de la langue arabe ; Désambiguïsation lexicale ; Algorithme de colonie de fourmis ; Algorithme Lesk ; Algorithme local / global ; WordNet ; Wiktionnaire

Abstract:

The ability to identify the intended meanings of words in context is a central research topic in natural language. Many solutions exist for Word Sense Disambiguation (WSD) in different languages, such as English or French, but research on Arabic WSD remains limited. The main bottleneck is the lack of resources. In this manuscript, we show that it is possible to build a WSD system for the Arabic language thanks to the Arabic WordNet and its connections to the English Princeton WordNet. Given that the Arabic WordNet does not contain definitions for synsets, we construct a dictionary that maps the Princeton WordNet definitions to the Arabic WordNet. We also create an Arabic evaluation corpus and gold standard. We then exploit this dictionary and evaluation corpus to run and evaluate an adapted ant colony algorithm on Arabic text that can use the Lesk similarity measure thanks to definition mapping.

Keywords: Arabic language processing; word sense disambiguation; ant colony algorithm; Lesk algorithm; local/global algorithm; WordNet ; Wiktionary.

Table des matières

Listes des figures

Liste des tableaux

Chapitre 1 :

Introduction générale

1.1 Contexte du travail

Les langues naturelles sont caractérisées par l'omniprésence de l'ambiguïté lexicale qui constitue l'une des sources de richesse et de souplesse des langues. Dans la communication interhumaine, l'ambiguïté ne présente pas un réel problème, mais dans le cadre de traitement automatique de la langue, l'ambigüité lexicale constitue toujours un défi.

L'ambiguïté est la propriété d'un mot, d'une suite de mots ou d'un concept ayant plusieurs significations ou plusieurs analyses grammaticales possibles. On peut rencontrer différents types d'ambigüité en fonction de niveau d'analyse linguistiques (morphologique, syntaxique, sémantique, pragmatique).

« La désambiguïsation sémantique des mots consiste à associer une occurrence donnée d'un mot ambigu avec l'un des sens de ce mot. La résolution de ce problème s'effectue en deux étapes : la discrimination des différents sens du mot, en regroupant les termes similaires, puis l'étiquetage sémantique de chacune de ses occurrences » (Rakho, Pitel, & Mouton, 2008).

Les difficultés liées à la problématique de la désambiguïsation sémantique ont très tôt été identifiées. Toutefois, les solutions qui ont été proposées dans chaque domaine sont également multiples et très diverses, en fonction des besoins et des savoirs afférents à chacune des matières concernées. La définition du problème elle-même ne fait pas l'unanimité. En effet, si un consensus est atteint pour définir la désambiguïsation sémantique comme l'association d'un mot apparaissant dans un contexte avec sa signification ou sa définition qui peut être distinguée des autres définitions qu'on peut attribuer à ce mot, en revanche le même accord n'existe pas pour d'autres tâches.

La désambiguïsation lexicale est une pierre de base pour de nombreuses applications du traitement automatique des langues. Cependant, elle nécessite toutefois des efforts

de développement conséquents, qu'ils s'agissent d'annoter des corpus ou de produire des lexiques et des outils.

Le traitement automatique de la langue arabe est compliqué à cause de ses spécificités dans les différents niveaux (morphologique, syntaxique, sémantique, pragmatique). Grâce à la propagation de l'islam et la diffusion du Coran, l'arabe a connu une expansion dès le septième siècle. Les premiers travaux du traitement automatique de la langue arabe ont débuté vers les années soixante-dix. Ils concernaient notamment les lexiques et la morphologie.

Malgré la disponibilité des outils de traitement des documents arabes et la propagation de la langue arabe sur le Web, les recherches ont abordé des problématiques plus variées comme l'indexation automatique des documents, la traduction automatique, la recherche d'information, …etc.

Notre travail s'inscrit dans le cadre des études visant à la désambiguïsation lexicale de manière automatique et plus précisément pour la langue arabe.

1.2 Motivations

La désambiguïsation sémantique est une tâche qui consiste à déterminer le sens d'un mot dans un contexte. Par exemple, le mot « مكتب » pour le meuble « ما يكتب عليه », pour la salle du bureau « حجرة المكتب » et pour l'équipe de travail « مكتب الإدارة », etc. La majorité des mots d'une langue ont plusieurs sens, qui résultent d'une étymologie complexe, comme c'est le cas pour le mot مكتب ci-dessus. Parfois, les formes sont elles-mêmes homographes, c'est-à-dire que leurs parenté graphique résulte d'un pur accident hors de toute parenté étymologique (exemple le mot souris, petit mammifère rongeur ou dispositif électronique).

La désambiguïsation lexicale est une tâche très importante pour plusieurs applications de traitement automatique des langues naturelles :

- **Traduction automatique (MT)**

La traduction automatique est la discipline où la désambiguïsation lexicale joue un rôle très important afin d'obtenir une bonne traduction des mots ambigus selon leurs contexte. Dans l'exemple qu'Audibert a cité :

« *La traduction en anglais du mot français « mèche » est « lock, wick, fuse » ou bien*

« Drill » suivant qu'il s'agit d'une «mèche de cheveux, de bougie, de pétard » ou de « perceuse » » (Audibert, 2003b).

- **Recherche d'information (RI) et Indexation sémantique**

La désambiguïsation des mots d'une requête peut permettre de filtrer la recherche de manière que les documents retournés soient des documents sémantiquement pertinents. Par exemple, si nous avons besoins des documents qui traitent les sources d'eau, il faut ignorer les documents traitant des sources d'information. Les travaux récents en désambiguïsation lexicale s'inspirent parfois des travaux en recherche d'information.

L'indexation sémantique en Recherche d'information est née du problème de l'ambiguïté des mots de la langue naturelle qui sont utilisés classiquement pour représenter les documents et les requêtes. L'indexation sémantique a pour objet de représenter les documents et les requêtes par les sens des mots permettant ainsi de lever toute ambiguïté, ce qui a pour conséquence d'améliorer les résultats de la recherche (Azzoug, 2014).

L'indexation sémantique requiert des techniques de désambiguïsation des sens des mots pour retrouver les sens corrects des mots dans un document (ou dans une requête donnée). Ces techniques se basent sur l'utilisation de ressources, telles que les corpus d'apprentissage ou les ressources terminologiques(Abdelaali Bakhouche & Yamina, 2012).

- **Génération automatique de textes**

L'objectif de la génération automatique de texte est la production des textes en langage naturel (syntaxiquement et sémantiquement correcte) à partir des représentations formelles d'un contenu. La fonction assurée de cette branche de TALN est la communication entre l'homme et la machine sous forme d'un processus réciproque de la compréhension automatique (il s'agit là d'exprimer au lieu de comprendre) (Roussarie, 2000).

- **la correction orthographique**

Le correcteur orthographique compare les mots du texte aux mots d'un dictionnaire. Si les mots du texte sont dans les dictionnaires, ils sont acceptés, sinon une ou plusieurs propositions de mots proches sont faites par le correcteur orthographique. Le correcteur grammatical vérifie que les mots du texte, bien qu'ils soient dans les

dictionnaires, sont conformes aux règles de grammaire (accords, ordre des mots, etc.) et aux règles de la sémantique (phrase ayant un sens, absence de confusion d'homophones, …etc) (Desmarais, 1994).

- **Résumé automatique de texte**

Un résumé automatique de texte est une forme abrégée et condensée du contenu d'un document. La réalisation de bon résumé nécessite un grand effort d'évaluation, de sélection, d'assemblage et d'organisation des segments textuels selon leur pertinence. Ainsi qu'une bonne compréhension et une bonne gestion des phénomènes de redondance, de la cohérence et la cohésion menée à produire un résumé automatique humainement crédibles (J. M. Torres-Moreno, 2011).

Notre travail s'inscrit dans le contexte de traitement automatique de la langue arabe, En particulier nous nous intéressons à l'étude des approches de désambiguïsation sémantique existantes et la proposition de nouvelle approche.

1.3 Contributions

Le but de notre travail, est d'essayer de mettre en évidence l'utilité de désambiguïsation lexicale de la langue arabe. Nous présentons dans ce qui suit les principales contributions de ce manuscrit :

- **Construction d'une base lexicale :** Les systèmes de désambiguïsation lexicale doivent mettre en relation les occurrences de mots en contexte avec les entrées d'un dictionnaire informatisé ou d'une base de données lexicale. Les informations qui peuvent être exploitées proviennent principalement des mots voisins du mot à désambiguïser (mot-cible). D'autres indices peuvent également être exploités, comme le domaine général du texte dans lequel se situe le mot-cible.

  Les difficultés de la désambiguïsation lexicale sont au moins de deux ordres :

  - La liste des sens du dictionnaire qui sert de référence. En effet, les dictionnaires traditionnels sont plutôt destinés à la consultation par des humains qu'à une utilisation par des machines.
  - Les connaissances qui permettent d'associer les mots du contexte avec le sens adéquat. A l'heure actuelle il n'existe pas des bases de connaissances disponibles qui nous aident à gérer ce problème. La constitution manuelle de telles bases serait d'ailleurs une entreprise gigantesque et sans doute hors de portée pour des décennies.

Pour résoudre ce problème, Nous exploitant d'une part la ressource Web Wiktionnaire qui est devenue une source intéressante pour l'extraction d'information ; et d'une autre part nous utilisons l'ontologie WordNet pour construire une base lexicale arabe.

- **Création d'un système de désambiguïsation lexicale :** Nous montrons la notion d'algorithme local et d'algorithme global pour la désambiguïsation sémantique des textes arabes. Un algorithme local permet de calculer la proximité sémantique entre deux objets lexicaux, tandis que l'algorithme global permet de propager ces mesures locales à un niveau supérieur. Nous nous servons de cette notion pour comparer un algorithme à colonies de fourmis à d'autres méthodes issues de l'état de l'art.
- **Evaluation de système de désambiguïsation par un corpus**. Les résultats obtenus nous mènent à découvrir les facteurs qui ont influencé la performance de ce système qui seront sujets d'éventuelle amélioration dans nos future travaux de recherches.

1.4 Organisation du manuscrit

Le présent manuscrit est découpé en six chapitres, s'intéressant chacun à un aspect particulier de notre travail de recherche.

Nous montrerons dans le second chapitre dans quelle mesure la question du sens se pose dans de nombreuses applications du TALN. C'est la raison pour laquelle nous exposerons les différents niveaux de traitements que comportent de telles applications et les difficultés posées dans chacun de ces niveaux. Nous insisterons particulièrement sur les niveaux sémantiques et pragmatiques de la question. Enfin, nous traiterons quelques relations sémantiques et fonctions lexicales de production.

Nous présenterons ensuite dans le troisième chapitre quelques systèmes reposent sur la notion d'algorithme local et d'algorithme global. L'algorithme local permet de donner une mesure de la proximité sémantique entre deux objets lexicaux (sens, mots, constituants, etc.) tandis que l'algorithme global permet de propager les mesures locales à un niveau supérieur. Cette double typologie nous paraît pourtant centrale et permet de mieux caractériser les propriétés des systèmes de désambiguïsation lexicale. Ainsi un système peut être constitué d'un algorithme local plus ou moins supervisé et d'un algorithme global lui aussi plus ou moins supervisé.

Dans le quatrième chapitre, on donne un aperçu sur les ressources linguistiques existant au niveau international. Nous abordons : Les lexiques, les grammaires syntaxiques, les corpus de textes monolingues bruts et annotés et les corpus multilingues alignés.

Dans le chapitre cinq, nous présentons plusieurs approches de la désambiguïsation sémantique qui ont été abordées en informatique : approches basées sur les connaissances comme celles qui sont basée sur les préférences sélectionnelle, les approches basé sur le chevauchement entre les différentes définitions ainsi que les approches basée sur l'apprentissage automatique supervisé ou non supervisé. Enfin, nous présentons les travaux de recherche qui focalisent sur la langue arabe.

Dans le sixième chapitre, nous présenterons notre approche ACARWSD. Avant introduire les étapes du système, il convient d'en rappeler quelques notions et ressources essentielles. Ensuite nous illustrerons notamment les démarches de la construction d'une base de données lexicale. Enfin nous montrerons par un exemple les étapes de notre système.

Finalement, le chapitre sept décrit quelques expérimentations sur le modèle : les outils utilisés dans l'application et enfin nous utiliserons des métriques pour évaluer notre approche. La conclusion est l'occasion de commenter les différents résultats et présenter nos perspectives.

Chapitre 2 :

L'ambiguïté de la langue

Sommaire

2.1 Introduction

Le traitement automatique du langage naturel signifie l'ensemble des travaux qui visent à modéliser la compréhension des énoncés linguistiques à l'aide de machines pour faciliter la communication. Ce traitement automatique concerne le langage naturel et non pas le langage de programmation. Ce dernier s'intéresse à la manipulation algorithmique. Parmi les difficultés majeures du traitement automatique du langage naturel l'ambiguïté et l'implicité. Dans la présente recherche, nous nous limiterons exclusivement au traitement du langage sous forme écrite.

2.2 Le traitement automatique du langage naturel

D'abord, nous définissons le langage naturel comme un moyen de communication interhumaine tel que le français, l'anglais, l'arabe…etc. par contre les langages formels sont développés par l'être humain en utilisant la logique, le mathématique et l'informatique. Ainsi que, une définition acceptable du TALN serait de le considérer comme :

Le domaine d'étude des techniques automatiques d'analyse (compréhension) et de génération (production) d'énoncés oraux ou écrits (Schwab, 2005).

Deux raisons principales, souvent conjointes dans la réalité, justifient le TALN. Tout d'abord, sur un plan théorique, le TALN permet de vérifier les théories linguistiques ou de manière plus générale, de mieux comprendre comment les humains communiquent entre eux. A cette fin, l'ordinateur est utilisé pour simuler les capacités humaines de compréhension et de production de la langue naturelle (Bouillon, 1998). Il y a d'autres raisons à vouloir s'intéresser au TALN :

- L'exploitation de la richesse de la langue naturelle (support d'information)
- L'acquisition d'une communication Homme-Machine naturelle.

Les résultats ainsi obtenus peuvent ensuite être comparés aux performances humaines. Ensuite, sur un plan pratique, le TALN rend possible la construction des systèmes opérationnels qui débouchent sur des produits commerciaux, largement diffusés. Parmi les plus connues, citons :

- la traduction automatique : application pionnière[1] apparue dès les années 1950 ;
- la correction orthographique ou grammaticale ;
- l'indexation des documents et de la recherche d'information (moteurs de recherche) ;
- la reconnaissance vocale (la dernière version de Windows 10 l'intègre en standard) ;
- la synthèse de la parole (horloge vocale) ;
- la génération automatique de textes ;
- le résumé automatique de textes ;
- le filtrage de texte.

2.2.1 Les connaissances de la langue

Pour traiter automatiquement les langues naturelles, le programme idéal devrait inclure différentes connaissances relatives à la langue qui répondent aux questions suivantes :

Quels sont les différents mots ?

Comment ils se prononcent ?

Comment ils se combinent pour former une phrase ?

Comment le sens des différents mots, contribue au sens de la phrase ?

De plus, ce programme devrait aussi utiliser des connaissances générales sur le monde et les contextes d'utilisation des textes. Ces différentes connaissances sont généralement classées et étiquetées sous les rubriques suivantes : phonétiques et phonologiques, morphologiques, syntaxiques, sémantiques et pragmatiques.

[1] « *A cette époque pionnière, on traduit mot à mot. La méthode va vite révéler ses limites. La langue recèle une infinité d'ambiguïtés, comme nos sympathiques homographes. En français, « le » est soit article, soit pronom, et « savoir », soit un verbe, soit un nom. Ce genre d'écueils va rebuter. En 1960, le logicien Bar-Hillel décrète qu'une traduction automatique est impossible si on ne dispose pas d'une immense banque de données et si le locuteur n'a pas de connaissances extérieures au texte à traduire. En clair, il faudrait que l'ordinateur comprenne. Ce constat d'impuissance donne un coup d'arrêt aux recherches et le rapport Alpac [Automatic Language Processing Advisory Committee], publié en 1964, conduit le gouvernement américain à stopper les financements. Seuls s'entêtent le Geta à Grenoble et le projet canadien, Meteo, qui traduit avec succès depuis 1977 des bulletins météo de l'anglais vers le français* » (Roussel, 2007)

2.2.1.1 ***Phonétiques et phonologiques :***

La Phonétique et la Phonologie sont deux branches de la linguistique qui étudient les sons utilisés dans la communication parlée. Chaque domaine complète l'autre. Le phonéticien traite les propriétés sonores des sons de la langue en utilisant des appareils, tandis que la construction des modèles qui facilitent la compréhension du fonctionnement des sons dans la langue est réalisée par le phonologue. Partons du concret vers l'abstrait et allons de la phonétique vers la phonologie.

2.2.1.2 ***Morphologiques*** :

La morphologie est la discipline de la linguistique qui traite les types et la forme des mots. Le traitement morphologique des mots étudie les relations qui existent entre les différentes formes d'un même mot. Le concept morphème est la plus petite unité portant un sens grammatical. Il est important pour la plupart des études contemporaines de la morphologie, l'étude des morphèmes et de leurs modes de combinaison est l'objet de la morphologie.

2.2.1.3 ***Syntaxiques***

Ces connaissances décrivent la matière dont les mots se combinent en phrases syntaxiquement correctes et encodent ainsi leur régularité structurale (Bouillon, 1998).

2.2.1.4 ***Sémantiques***

D'après les linguistes, la sémantique est ce qui nous exige une conscience et une sensibilisation d'une idée à cause de quelque chose d'inhérent à sa portée, cette définition signifie le convenu indiqué par le mot et ce qui me vient à l'esprit et ne veux pas dire le sens de mot. L'exemple suivant clarifie mieux la définition : Si vous entendez la sonnette sonner, votre esprit se tourne directement vers qu'il y'avait quelqu'un devant la porte qui avait appuyé sur le bouton de la sonnette, Cette transition d'esprit (quelqu'un devant la porte et appuyer sur le bouton) est l'habitude à impliqué par le son de la sonnette du porte.

En bref, la sémantique représente une relation réciproque entre le signifiant (اللفظ "الدال") et le signifié des mots de sens (المعنى "المدلول").

Ces connaissances concernent les sens des mots et la manière dont les sens des mots, se combinent pour former le sens global de la phrase. On verra dans la suite que les

linguistes adhèrent généralement à l'idée de la compositionnalité du sens et admettent, implicitement ou explicitement, que le sens de la phrase est composé du sens de ses différentes parties et de relations syntaxiques qui les lient (Bouillon, 1998).

2.2.1.5 ***Pragmatiques***

Elles comprends toutes les informations générales sur les concepts du monde et celles relatives à la situation de communication, sur le plan psychologiques, sociales et historiques qui déterminent l'émission d'un énoncé à un moment du temps et en lieu donnés (Bouillon, 1998).

2.2.2 Les architectures des systèmes du TALN

On distingue deux types différents d'architecture du système de TALN :

a. L'architecture séquentielle est la plus ancienne, utilise les connaissances de la langue traitée les unes après les autres. Dans pareil système, il existe pour chaque type de connaissance un module informatique différent qui prend en entrée les données que lui sont fournies par le module précédent (cf. Fig.2.1).

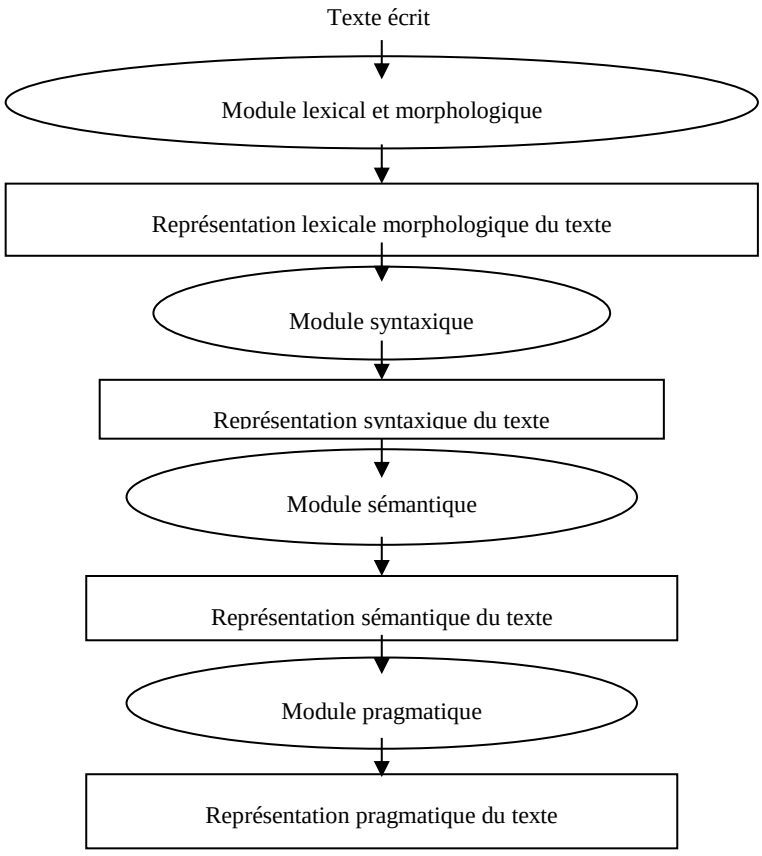


Figure 2-2 Architecture séquentielle

b. D'autres architectures moins hiérarchiques utilisent les différentes connaissances de la langue en même temps, soit par la communication entre les différents modules (système hiérarchiques ou parallèles), soit par l'intégration des différentes connaissances au niveau du mot dans une représentation complexe (Francoise Forest, 1988).

2.3 Les niveaux de traitements

Pour une meilleure compréhension des énoncés de la langue naturelle, on présente dans cette partie les différents paliers du traitement de la langue. L'objectif de cette présentation est de voir l'intérêt de ces niveaux afin de réaliser des applications qui sont constituée par l'ensemble des modules.

Le problème de cette hiérarchie des niveaux est le cumul des difficultés durant la progression et les outils actuels disponibles se font moins performants, et ils sont limités au domaine particulier.

2.3.1 Niveau morphologique

L'identification des items lexicaux des textes se fait dans la phase d'analyse morphologique, ainsi que durant cette phase le mot est décomposé en radical et d'affixe (préfixe et suffixe). Par exemple, *charges* peut être le nom féminin, « charge » au pluriel ou bien le verbe « charger » à l'indicatif ou au subjonctif présent. Cette opération est aussi appelée ***lemmatisation***. Une partie des ambiguïtés peut être levée au niveau syntaxique du processus d'analyse (Schwab, 2005).

2.3.2 Niveau syntaxique

Après la phase d'analyse morphologique, un certain nombre de solutions sont envisageables pour les mots d'une phrase. Une analyse syntaxique permet grâce aux règles, de ne conserver que les solutions qui sont possibles. Par exemple, prenons la phrase « Des charges supplémentaires seront retenues contre l'accusé. ». Le mot, « charges », comme nous l'avons vu dans la partie consacrée à la morphologie, peut être le nom féminin « charge » au pluriel, comme le verbe « charger » à l'indicatif ou au subjonctif présent. La morphologie possible des mots constituant le groupe nominal sujet « des charges supplémentaires » (pour « des » déterminant pluriel et

pour « supplémentaires » adjectif masculin ou féminin pluriel) rendent ici la seule solution possible pour « charges » nom féminin pluriel.

La recherche pour mettre au point des analyseurs fiables, est encore florissante. La tâche est complexe puisqu'il n'existe pas à l'heure actuelle des règles grammaticales pouvant couvrir l'ensemble des phrases correctes dans aucune des langues existantes. Ainsi, deux grandes familles d'analyseurs coexistent :

- ***approche symbolique*** : ces analyseurs se basent sur des règles grammaticales et nécessitent donc une recherche et une implémentation de ces règles.
- ***approche statistique*** : ces analyseurs se basent sur des méthodes d'apprentissage à partir de corpus annotés manuellement ou automatiquement pour produire des règles pondérées.

Quelle que soit la technique utilisée, les analyseurs syntaxiques ne renvoient pas tous un arbre syntaxique complet. Ainsi, deux autres types de résultats sont possibles : renvoyer les relations entre les mots de la phrase ou produire une segmentation en syntagmes.

Une analyse syntaxique ne peut pas toujours lever toutes les ambiguïtés. Ainsi, certaines phrases comme « La petite brise la glace.» ne peuvent être totalement désambiguïsées à ce niveau de traitement comme on peut le constater, deux interprétations syntaxiques sont ici possibles. Dans la première, « petite » et « glace » sont des noms, « brise » est la troisième personne du présent de l'indicatif du verbe « briser » (i.e. une petite fille casse un miroir) tandis que dans la deuxième, « petite » correspond à l'adjectif, « petit », « glace » au verbe « glacer » et « brise » à l'item lexical « brise » (i.e. un léger vent donne froid à quelqu'un ou quelque chose de féminin). Si syntaxiquement, il est absolument impossible de lever l'ambiguïté, des informations de nature sémantique et pragmatique sur cette phrase peuvent permettre d'émettre des préférences (Schwab, 2005).

2.3.3 Niveaux Sémantique et pragmatique

Un découpage est difficile à réaliser aux niveaux sémantique et pragmatique en linguistique. Les niveaux sémantique et pragmatique sont encore beaucoup plus complexes à décrire et à formaliser que les niveaux de traitement précédents. En effet,

ils touchent à l'étude du sens de la phrase dans le contexte général dans lequel elle s'inscrit. Plus que les autres niveaux de traitement, sémantique et pragmatique sont extrêmement liés. Ce lien est dû à la question commune que se posent les deux disciplines « Qu'est-ce que le sens ? ».

2.3.3.1 ***La sémantique***

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

Nombreux sont les ouvrages traitant de sémantique, mais fort rares sont ceux qui se risquent à donner ne serait-ce qu'une esquisse de définition du terme « sens ». En effet, le sens est quelque chose de difficile à décrire car intuitif et souvent considéré dans ces livres comme déjà acquis par le lecteur.

Un bon moyen de faire comprendre ce que signifie un mot est donc d'utiliser une expression équivalente, une paraphrase. Sur cette idée, Alain Polguère (Polguère & Tremblay, 2003) propose comme définition du sens :

Le sens d'une expression linguistique est la propriété qu'elle partage avec toutes ses paraphrases(Polguère, 2003).

Cette définition repose sur la notion d'équivalence entre phrases, les paraphrases. Ces équivalences sont loin d'être rares en langue, c'est même une des caractéristiques essentielles des langues naturelles par rapport aux langages artificiels. Ainsi, pour Polguère, la notion de paraphrase est reconnue comme un concept primitif possédé par un locuteur qui permet de définir la notion de sens.

Le sens d'un énoncé est régi par le principe de compositionnalité sémantique pour lequel « le tout est calculable à partir du sens de ses parties ». Ainsi, un énoncé est directement calculable (dans sa composition lexicale et sa structure syntaxique) à partir de la combinaison du sens de chacun de ses constituants (Polguère, 2003). Par exemple, le sens d'une phrase comme « L'enfant voit la mer.» est calculable à partir :

- des items lexicaux « le », « enfant », « voir », « la », « mer » ;

- des règles syntaxiques et morphologiques du français, utilisées dans la phrase.

Il est souvent spécifié dans la littérature que les locutions transgressent, au moins en partie, le principe de compositionnalité sémantique. De nombreuses théories sur la sémantique ont été élaborées comme la sémantique du prototype, la sémantique distributionnelle ou la sémantique structurale.

2.3.3.2 ***La pragmatique***

La pragmatique considère le sens du point de vue du récepteur et s'intéresse ainsi aux mécanismes de l'interprétation des énoncés. La question du sens d'un énoncé est alors posée de façon différente. De la sorte, un énoncé veut dire :

- Ce que ses récepteurs estiment qu'il veut dire ;
- Ce que ses récepteurs croient que l'émetteur a voulu dire dans/par cet énoncé;
- Ce que ses récepteurs estiment (à tort ou à raison, de façon réelle ou feinte, de bonne ou de mauvaise foi) être, la prétention et l'intention sémantico-pragmatiques du locuteur dans cet énoncé.

Un exemple classique et quasi-quotidien de la pragmatique est la phrase « Peux-tu me donner du sel ? ». Il est clair que le locuteur n'attend pas la réponse « Oui. » à la question littérale qu'il posée mais attend qu'on lui donne le sel (Schwab, 2005).

La pragmatique est donc l'étude du sens des énoncés en contexte, c'est-à-dire l'ensemble des significations que peut lui donner un être humain.

Le niveau pragmatique de la compréhension de textes consiste ainsi à découvrir le sens correct d'un énoncé en fonction des conditions situationnelles et contextuelles dans lesquelles il apparaît. La pragmatique s'occupe en particulier des problèmes de référence, d'anaphore[2], de subjectivité (Schwab, 2005).

[2] Reprise d'un mot ou d'une série de mots au début de phrases ou de propositions successives

2.4 Les difficultés sémantiques du TALN

Le traitement automatique du langage naturel pose plusieurs problèmes, parmi eux on cite : l'ambiguïté du langage, et la quantité d'implicité contenue dans les énoncées de la langue naturelles.

2.4.1 L'ambiguïté

Malgré le développement des outils informatique matériel et logiciel, L'ambiguïté reste l'un des obstacles du traitement automatique du langage naturel. La nature et la spécificité des langues rendent leur traitement automatique difficile. L'ambiguïté de la langue peut être source de calembours[3] ou donner lieu à des quiproquos[4]. Il y a plusieurs interprétations pour un mot dans un énoncée :

Quel prix vaut ce sacrifice ?

A deux interprétations possibles : c'est le prix qui vaut le sacrifice, ou le sacrifice qui vaut le prix. Dans une situation de communication, l'auditeur ne perçoit pas nécessairement l'ambiguïté de la phrase ; c'est en générale le contexte de l'énoncé qui met en évidence le sens du message de l'interlocuteur. Des ambiguïtés peuvent se présenter à tous les niveaux de description de la langue : lexicales (catégorielle et sémantique), syntaxique, sémantique et pragmatique (Bouillon, 1998).

2.4.1.1 *Les ambiguïtés lexicales*

Il y a deux types d'ambiguïté lexicale catégorielle ou sémantique, ce type d'ambigüité est lié au mot lui-même :

Une ambiguïté catégorielle : ce type d'ambigüité apparait quand le même mot a plusieurs formes syntaxiques. Par exemple le mot « porte » peut être un nom, un verbe (troisième personne du singulier du présent du verbe *porter*). La détermination de la catégorie du mot est faite selon le contexte syntaxique. Par exemple dans la phrase *il porte, « porte »* est un verbe puisque seule la séquence (pronom + verbe)

[3] Jeu de mots entre homonymes ou mots de prononciation identique et de sens différents.

[4] Méprise qui fait prendre une chose, une parole ou une personne pour une autre.

est acceptable. La séquence (pronom + nom) est rejetée par la grammaire du français (Bouillon, 1998).

Par contre *l'ambiguïté sémantique* est la propriété de certains mots qui possèdent plusieurs sens. Il existe deux types d'ambiguïté sémantique *les polysèmes* et les *homonymes*.

- *La polysémie* est la propriété d'un mot ou d'une phrase qui a plusieurs sens ou significations différentes. Par exemple le mot *Théâtre* peut signifier l'art, le lieu ou la production littéraire selon le contexte, le mot *Clarté* peut signifier la lumière, la transparence, l'intelligibilité ou la blancheur.
- On parle d'*homonymie*, quand il n'y a pas une relation entre les différents sens d'un mot. Par exemple le mot *avocat* peut signifier un fruit ou un auxiliaire de justice et le mot souris peut signifier un dispositif de pointage ou un animal.

2.4.1.2 ***Les ambiguïtés syntaxiques ou structurales***

« Ce type d'ambiguïté résulte de ce que la même structure de surface peut être dérivée de deux structures profondes différentes. L'ambigüité provient de deux interprétations sémantiques distinctes qui devient être données au même énoncé pour des raisons syntaxiques » (Dubois, 1969).

L'ambiguïté catégorielle conduit à l'ambiguïté structurelle. La phrase *le pilote ferme la porte* a deux analyses syntaxiques possibles (*article- Substantif -verbe-article-Substantif ou article- Substantif -adjectif- pronom-verbe)* car les mots qui la composent sont catégoriellement ambigus. Les ambiguïtés structurelles peuvent aussi provenir d'ambiguïtés de rattachement d'un syntagme prépositionnel ou d'une proposition relative, comme l'illustrent les exemples suivants.

- *Le policier regardait l'espion avec deux jumelles*.
- *Marie frappera l'homme avec un parapluie.*

Dans la première phrase le syntagme prépositionnel *avec deux jumelles* peut être attaché soit au verbe *regardait* soit au syntagme nominal *l'espion*. De même dans la deuxième phrase le syntagme prépositionnel *avec un parapluie* peut être attaché soit au verbe *frappera* soit au syntagme nominal *l'homme*.

Ces exemples présentent des ambiguïtés structurelles réelles perçues par l'homme : chaque analyse syntaxique conduit à une interprétation différente. Mais un système d'analyse automatique purement syntaxique va aussi produire des ambiguïtés structurelles qui ne sont pas reconnues par le locuteur humain.

2.4.1.3 *Les ambiguïtés sémantiques ou logiques*

On parle d'ambigüité sémantique quand une phrase a plusieurs représentations logiques. Elles sont notamment dues à la portée des quantificateurs.

Par exemple :

Tous les enfants mangent une pomme. *(1)*

Cette phrase a deux lectures :

Pour chaque enfant, il y a une pomme. *(2)*

Il y a une pomme pour tous les enfants. *(3)*

L'ambiguïté de (1) se représente facilement par la dualité des deux formules de logique possibles pour traduire (1) :

$\forall x : enfant\ \exists y : pomme\ x\ mange\ y$ (2')

$\exists y : pomme\ \forall x : enfant\ \ x\ mange\ y$ (3')

Dans (2') on dit que la portée du quantificateur $\forall$x est plus forte que celle de $\exists$y, et que dans (3') c'est l'inverse. Le contexte fourni éventuellement les indications qui permettent de lever l'ambiguïté. La nécessité de représenter l'ambiguïté logique est liée au type souhaité de représentation (Bouillon, 1998).

2.4.1.4 *Les ambiguïtés pragmatiques*

« *On parle d'ambiguïté pragmatique pour désigner des cas où l'équivocité ne peut être résolue que par la considération de notre savoir du monde (et en particulier des habitudes et coutumes que l'on rencontre dans ce monde) et où donc la source de l'ambiguïté est dans nos règles d'usage des expressions* » (Lecomte, 2007). L'ambiguïté de la langue naturelle apparaît aussi lorsqu'un énoncé correspond à

plusieurs situations de communications ; un pronom, par exemple peut avoir plusieurs antécédents possibles.

Le professeur a envoyé l'élève chez le proviseur ;

Parce qu'il le trouvait insupportable (« il » est le professeur).

Parce qu'il lançait des boulettes au plafond (« il » est l'élève).

Parce qu'il voulait le voir (« il » est le proviseur).

Dans cet exemple le pronom *il* a trois antécédents syntaxiquement acceptables : le professeur, l'élève et le proviseur.

La description des données linguistiques et la formalisation de ces données sont nécessaires pour résoudre le problème de la désambiguïsation. Les tâches descriptives et de formalisation se heurtent à la nature variée des ambiguïtés et à la diversité des connaissances qui doivent être décrites de manière formelle pour permettre leur utilisation dans des programmes informatiques. D'autre part, les programmes informatiques doivent définir une stratégie de désambiguïsation : ils déterminent quelles informations interviennent et la manière dont elles interagissent (Bouillon, 1998).

2.4.2 L'implicité

La communication entre les êtres humains est réalisée par l'activité langagière et dans ce cas les énoncés de la communication sont compréhensibles parce que l'être humain dispose de plusieurs connaissances du contexte par contre des difficultés majeurs sont posées quand il s'agit de l'interaction homme machine car la machine n'a pas dotée de toutes les connaissances du monde et de son fonctionnement. Pour cette raison la majorité des énoncés restent incompréhensible motivant l'ajout d'une base de connaissances qui donne accès à un savoir sur le contexte statique et dynamique.

La compréhension des énoncés par la machine pose d'autres problèmes qui sont due au manque des connaissances des figures de style (ellipses[5], métaphores[6], …). Pour

[5] Omission d'un ou de plusieurs mots qui, dans une phrase, ne sont pas utiles à la compréhension.

[6] Procédé qui consiste, par analogie, à donner à un mot un sens qu'on attribue généralement à un autre.

résoudre ce type de problèmes les chercheurs proposent des approches concernant la restriction des textes analysés en domaine particulier (textes scientifiques, politiques, économiques,...). L'objectif de ces approches et de lever l'ambiguïté sémantique d'une part et de représenter formellement les connaissances nécessaires à la compréhension des énoncés du domaine considéré d'une autre part.

2.5 Les relations sémantiques et les fonctions lexicales

2.5.1 Les relations sémantiques lexicales

L'organisation du lexique sur le plan pragmatique est basée sur les relations sémantiques. On présente dans cette section quelques relations qui sont réparties en deux classes, les relations d'hiérarchie (hyperonymie/hyponymie, holonomie/ méronymie) et les relations symétriques (synonymie/antonymie).

2.5.1.1 *Les relations d'hiérarchie*

Les relations hiérarchiques organisent le lexique de manière hiérarchique. Elles sont transitives et unidirectionnelles. On cite parmi ces relations : *hyponymie/hyperonymie* et *méronymie/holonymie*. Les relations hiérarchiques vérifient la propriété suivante :

$$R(A,B) \equiv \overline{R(B,A)} \qquad (1.1)$$

Cette figure montre un extrait de l'hiérarchie sémantique des lexies françaises

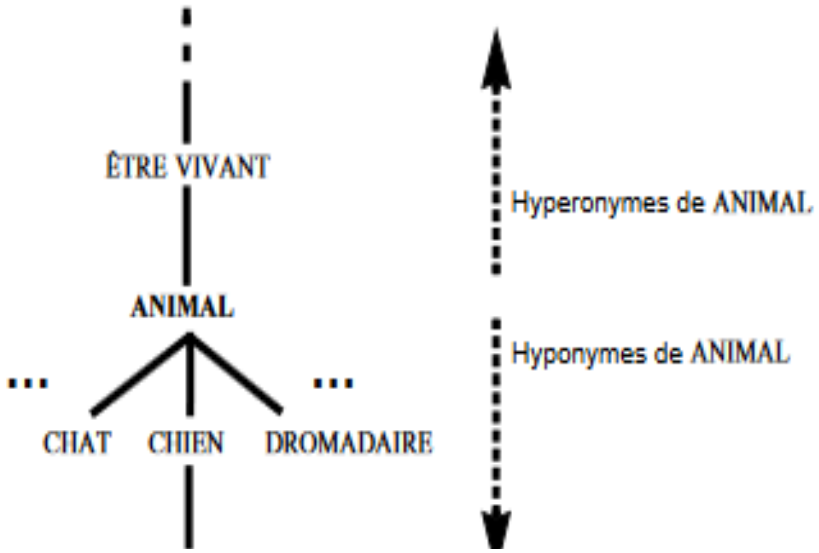


Figure 2-3 Extrait de la hiérarchie sémantique des lexies françaises (centré autour de ANIMAL) (Polguère, 2003)

1. ***La méronymie et holonymie*** : la relation de méronymie est la relation sémantique qui lie la partie au tout. On dit que le mot X méronyme de mot Y si X représente une partie de Y. Par exemple "جدار" ("مسكن"\" جدار") est méronyme de"مسكن "), "صفحة"\"كتاب", «يوم"\"سنة". l'holonymie est la relation inverse.

2. ***L'hyponymie et l'hyperonymie*** :

   On parle d'hyponymie quand il y a une relation entre un item lexicale et un autre plus générale par exemple "حيوان\حصان", "فواكه\تفاح" tandis que l'hyperonymie est la relation inverse.

2.5.1.2 ***Les relations symétriques***

La synonymie et l'antonymie sont des relations symétriques. Ces relations vérifient la propriété de symétrie. Ainsi, si R est une relation symétrique entre deux items *C1* et *C2*, alors :

$$R(C1, C2) \equiv R(C2, C1) \qquad (1.2)$$

Par exemple, si *C1* est antonyme de *C2*, alors *C2* est antonyme de *C1*.

Dans cette section nous allons aborder les deux relations sémantiques (la synonymie et l'antonymie) dans la langue arabe.

A. *La synonymie*

On dit que deux mots sont synonymes, quand ils n'ont pas la même forme mais ils expriment le même sens ou un sens très proche. Il doit savoir que les synonymes ne sont pas complètement des mots identiques, sauf s'ils ont le même opposé et la même utilisation en contexte, ce qui n'est pas généralement le cas. Donc on ne peut pas dire que la synonymie est la similitude complète et absolue, mais c'est la similitude de la plupart des traits sémantiques (Schwab, 2005).

Les synonymes offrent les avantages suivants :

- Offrir à l'utilisateur un lexique très riche et plusieurs termes pour désigner le même sens, donc l'utilisateur a l'occasion de sélectionner ce qui est adéquat avec le contexte.
- Stimuler la jouissance et réduire l'ennui du lecteur, car la diversité offre à l'auteur la possibilité de choisir ses termes loin de toute ambiguïté sémantique et donc il peut fixer le sens voulu.

B. *L'antonymie*

On dit que deux mots sont antonymes, lorsqu'ils expriment des sens qui s'opposent entre eux. Elle est capable d'enrichir les dictionnaires très fortement, l'opposé du mot éclaire le sens de son opposé bien qu'il n'est généralement pas complet. Il y a des raisons derrière l'antonymie des mots, citons (Abdelaali Bakhouche & Yamina, 2011, 2012) :

- ***Ressemblance*** *Exemple* **:** (بشرة) pour la peau de l'être humain et pour la plante
- ***Opposition*** *Exemple* : (عسس) pour l'avènement de la nuit (إقبال الليل) et pour retour de la nuit (إدبار الليل)
- ***Différents sens de même champ sémantique* :** *Exemple* (السرحان) pour le lion et pour le loup
- ***Classe de parole*** : *Exemple* (أجم) pour le verbe (اقترب) *s'approcher* et pour le nom (كبش بلا قرون) *bélier qui n'a pas des chélicères*

- ***La brièveté et la métaphore*** **:** *Exemple* « مكتب » pour le meuble « عليه يكتب ما », pour la salle du bureau « المكتب حجرة » et pour l'équipe de travail « مكتب الإدارة ».

Les aspects de cette notion sont :

- ***L'antonymie complémentaire :*** la différence entre les deux mots est totale, il n'y a pas d'aspects ou des degrés de rapprochement entre les deux :

(أنثى ″ féminin″- ذكر ″masculin″), (ميت ″*mort*″- حي ″*vivant*″), (متزوج ″*marié*″- أعزب ″*célibataire*″), (امرأة ″femme″- رجل ″ homme″).

- ***L'antonymie scalaire :*** entre les deux mots existent des aspects de rapprochement sémantique, exemple : (صعب ″difficile″- سهل ″facile″) il y a entre les deux des nuances de facilitation de difficulté, d'autres exemples

(قريب ″proche″- بعيد ″ loin″), (حار ″ chaud″- بارد ″ froid″), (قوي ″fort″- ضعيف ″ faible″)

- ***L'antonymie des conversifs :*** entre deux mots accompagnes exemple :

(أب ″père″- ابن ″fils″), (زوج ″époux″- زوجة ″épouse″), (مهزوم ″vaincu″- فائز ″gagnant″)

- ***L'antonymie duale :*** il s'agit des antonymes spatiaux comme : (تحت ″ sous″- فوق ″sur″), (شرق ″est″- غرب ″ouest″), (شمال ″nord″- جنوب ″sud″), les antonymes temporels comme : (نهار ″jour″ - ليل ″nuit″) et culturels comme (سؤال ″question″ - جواب ″réponse″).

L'antonymie est un phénomène linguistique intéressant dans l'éclaircissement du sens, tel que l'opposé du mot éclaire son sens, bien que l'antonymie ne soit généralement pas complète que dans des rares cas, mais il l'est dans certains traits.

Exemple : (ميت) ″mort″ son opposé est (حي) ″vivant″, donc l'opposé est désigné dans un seul trait qui est le trait de vie « الحياة », mais pour le mot « ميت » il y a plusieurs traits comme « الحياة » ″ la vie″, « الجنس » ″ le sexe″, « النوع » ″ le type″….

L'antonymie est l'un des types de la synonymie négative (les synonymes jouent un rôle de l'aspect positif pour le champ sémantique, alors que les antonymes jouent le rôle de l'aspect négatif). Et de ce fait l'antonymie montre les aspects sémantiques, d'un côté et enrichi le lexique, d'un autre côté, et surcroît les liaisons sémantiques entre les champs.

2.5.2 Les fonctions lexicales

Le concept des Fonction lexicales a été apparu au début des années 60. Zholkovsky et Mel'čuk [1965, 1966, 1967] ont présenté les premières descriptions de la fonction lexicale dans le cadre de la théorie linguistique Sens-Texte[7] (TST) en russe, Zholkovsky et Mel'čuk [1970] en français et Mel'čuk et Zholkovsky [1970] en anglais (Jousse, 2010).

« Les fonctions lexicales ont été définies dans le cadre de la TST pour décrire les relations sémantiques lexicales au moyen d'un outil formel conçu sur le modèle des fonctions mathématiques » Schwab (2005).

La fonction lexicale ***f*** est une relation entre une unité lexical *X* (l'argument de *f)* et un ensemble des unités lexicales {*Y*1, *Y*2, . . ., *Yn*} appelé la valeur de l'application de ***f*** à l'unité ***X***.

La fonction lexicale ***f*** est telle que :

- l'expression *f* (*X*) représente l'application de *f* à l'unité X : *f* (*X*) = {*Y*1, *Y*2, . . ., *Yn*}.
- chaque élément de la valeur de *f* (*X*) est lié à X de la même façon et remplit (à peu près) le même rôle.

[7] « La première publication présentant les fondements de la THÉORIE SENS-TEXTE [TST] remonte à plus de trente ans : ëolkovskij et Mel'.uk [1965]. La théorie linguistique proposée par I. Mel'.uk et ses collègues de Moscou peut être décrite succinctement à partir des cinq caractéristiques suivantes :

- La TST rend compte de l'association que tout locuteur d'une langue L est capable de faire entre un sens donné de L et l'ensemble des énoncés paraphrastiques de L exprimant ce sens.
- La TST est universelle, c'est-à-dire qu'elle repose sur des principes généraux s'appliquant à toutes les langues.
- La TST est linguistique en ce sens qu'elle permet, à partir des principes généraux sur lesquels elle repose, de construire des modèles linguistiques spécifiques pour chaque langue humaine.
- La TST permet de construire des modèles calculables.
- La TST est formelle. C'est-à-dire qu'elle utilise des langages formels pour :
  - représenter les énoncés linguistiques ;
  - encoder les règles de manipulation des représentations linguistiques — de telles règles servent à modéliser la correspondance Sens <=>Texte. » (Polguère, 1998)

$$\frac{f(X_1)}{X_1} \approx \frac{f(X_2)}{X_2} \qquad (1.3)$$

Aussi, afin de montrer que la fonction lexicale de synonymie modélise le rapport qu'entretiennent entre les termes "شجاع " et "جريء " d'une part et "سفينة " et "باخرة " d'autre part.

Mel'ˇcuk note :

$$\frac{\text{شجاع}}{\text{جريء}} \approx \frac{\text{سفينة}}{\text{باخرة}} \qquad (\mathbf{1.4})$$

De même, la fonction lexicale d'antonymie formalise le rapport qu'entretiennent entre les termes "اليقين" et "الشك" d'une part et "الحياة" et "الموت" d'autre part.

$$\frac{\text{اليقين}}{\text{الشك}} \approx \frac{\text{الحياة}}{\text{الموت}} \approx \frac{\text{الموت}}{\text{'الحياة}} \qquad (1.5)$$

Il existe autant de fonctions lexicales qu'il existe de liens lexicaux et chaque fonction lexicale est identifiée par un nom particulier. Deux classes de fonctions lexicales sont identifiées : les fonctions lexicales paradigmatiques et les fonctions lexicales syntagmatiques.

2.5.2.1 *Les fonctions lexicales paradigmatiques*

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

- Synonymie (*Syn*) :

  Syn ("الكتابة") = "التسجيل ", "التدوين", "التأليف",. . .

- Antonymie (*Anti*) :

  Anti (الجهل) = "العلم ", "المعرفة", ...

- Générique (*Gener*) : Les génériques sont les équivalents des hyperonymes.

  Gener ("أرض ") = "كوكب " ; *Gener* ("تفاح") = "فواكه"

Dérivés syntaxiques : Ces fonctions associent à un item sa contrepartie nominale (Sub-Stantification S_0), verbale (verbalisation V_0), adjectivale (adjectivisation A_0) ou adverbiale (*Adv*0) : S_0 ("قتال ") = "القتل", S_0 ("يغذي") = "غذاء", ...

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

- $\subset$ pour indiquer une inclusion du sens de l'argument dans la valeur de la fonction. On trouve ce cas avec les rapports d'hyperonymie :

  $Syn_{\subset}$ ("غراب") ="طيور" ;

- $\supset$ pour indiquer une inclusion du sens de la valeur dans l'argument de la fonction : $Syn_{\supset}$ ("طيور ") ="غراب " ;

- $\cap$ pour indiquer une intersection de sens. Ainsi $Syn._{\cap}$ ("يلعب") ="يرفه" puisqu'on peut jouer sans s'amuser et s'amuser sans jouer.

2.5.2.2 ***Les fonctions lexicales syntagmatiques***

Collocations dans toutes les langues, certaines combinaisons d'items lexicaux prévalent sur d'autres sans qu'il ne semble y avoir de motif logique.

Par exemple, on parle de « نوما عميقا » plutôt que de «نوما كثيفا» ou « نوما كاملا» pourtant aucune raison (du moins en synchronie) ne semble expliquer cette préférence. On parle, dans ces cas, de phénomène de collocation.

L'énoncé *AB* (ou *BA*) formé des items lexicaux *A* et *B* est une collocation si, pour produire cette expression, le locuteur sélectionne *A* librement d'après son sens alors qu'il sélectionne *B* pour exprimer un autre sens en fonction de *A* (Pol03). On appelle ***A*** base de la collocation et ***B*** collocatif. On peut citer comme exemples de collocations :

"سفينة"$_{(=A)}$ "الصحراء"$_{(=B)}$, "الآلي"$_{(=B)}$ "الإعلام" $_{(=A)}$,"الناطق"$_{(=A)}$ "الرسمي"$_{(=B)}$.

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

- *Bon* qui marque une évaluation positive :

Bon ("choix") = "bon" ;

Ou leurs opposés :

- *AntiBon* qui marque une évaluation négative :

AntiBon ("choix ") = "mauvais" ;

2.6 Conclusion

Malgré le développement des outils informatique matériel et logiciel, Le traitement automatique du langage naturel demeure un domaine d'expertise compliqué. De nombreux progrès restent à accomplir pour bâtir des systèmes capables de soutenir la comparaison avec l'humain, mais l'état des connaissances en permet aujourd'hui de proposer de nombreuses solutions efficaces à des problèmes et des demandes réels.

Une des limitations de tous les systèmes de traitement un peu sophistiqués est que ceux-ci font appel à une somme importante de connaissances d'expert : lexiques, règles de grammaires... ceci explique en partie pourquoi il n'existe pas de système de traitement qui soit à la fois complet (i.e. intégrant tous les niveaux de traitement) et indépendant du domaine (i.e. capable de traiter avec une même efficacité n'importe quel type de texte). Il existe une autre raison, moins visible, qui limite l'avancée des progrès en TALN, et qui est pour un bon nombre de phénomènes. L'état de la connaissance linguistique est insuffisamment formalisé pour pouvoir être utilisée par les concepteurs de systèmes de TALN.

La sémantique occupe une position importante dans le traitement de la langue naturelle. Il n'est pas facile de réaliser des traitements profonds des textes sans informations suffisantes sur la sémantique des termes et les relations sémantiques entre les mots constitutifs des textes, pour cela les chercheurs du traitement automatique de la langue n'ont pas écarté les questions de Synonymie et d'Antonymie et d'autres relations sémantiques, en raison de leurs valeurs ajoutées pour résoudre plusieurs problèmes comme l'analyse automatique des textes, la compréhension des textes, la traduction automatique,...

Le Traitement automatique de la langue est donc un domaine de recherches interdisciplinaires qui vise à traiter automatiquement les langues naturelles. Malgré ses

applications nombreuses dans plusieurs langues (français, anglais,....), les travaux réalisés concernant la langue arabe, se sont révélés insuffisants. Il se heurte à un problème principal : l’ambiguïté de la langue.

Chapitre 3 :

Algorithmes locaux et globaux pour la désambiguïsation lexicale

Sommaire

3.1 Introduction

De nombreux systèmes de désambiguïsation lexicale reposent sur la notion d'algorithme local et d'algorithme global (Gelbukh, Sidorov, & Han, 2003). L'algorithme local permet de donner une mesure de la proximité sémantique entre deux objets lexicaux (sens, mots, constituants, etc.) tandis que l'algorithme global permet de propager les mesures locales à un niveau supérieur (Schwab, Goulian, & Guillaume, 2011). Cette double typologie est intéressante qui peut caractériser les propriétés des systèmes de désambiguïsation lexicale.

Dans ce chapitre nous présentons plusieurs mesures de similarité sémantique locales qui ont été abordées en informatique : approches basées sur les arcs (distances), approches basées sur les nœuds (le contenu informatif) ainsi que les approches basées sur une représentation vectorielle et les approches basées sur les traits. Enfin, nous présentons quelques algorithmes globaux stochastiques pour la désambiguïsation lexicale (Algorithme génétique, Recuit simulé, La méthode recherche Tabou, chaines lexicales).

3.2 Mesures de similarité sémantique locales

L'évaluation de la proximité entre les concepts reliés à des termes d'un texte pose des difficultés dans plusieurs applications : traduction automatique, désambiguïsation sémantique, résumé automatique, recherche d'information, indexation automatique, etc.

On présente dans cette section quelques mesures de similarité conceptuelle. Les détails de ces mesures sont présentés par (Patwardham, 2003) où ces différentes mesures sont comparés par rapport à des évaluations faites par des sujets humains. Les deux premières mesures sont fondées sur la notion de contenu informationnel, Tandis que les autres basées sur les traits et la représentation vectorielle.

3.2.1 Approches basées sur les arcs (distances)

Dans une ontologie les mesures de similarité sont représentées par les distances entre les concepts. L'estimation de la similarité sémantique entre les différents objets dans une ontologie est basée sur les mesures qui déterminent la structure de l'ontologie.

L'évaluation des distances entre les concepts nécessite un graphe de spécialisation des objets. La distance entre les objets dans ce graphe représente le chemin le plus court qui peut déterminer le nœud commun ou bien un ancêtre qui réunit deux objets via des descendants communs. On cite quelques travaux dans ce contexte :

3.2.1.1 ***Mesure de Wu & Palmer***

La similarité est une notion définie entre deux concepts dans la même hiérarchie conceptuelle ou par leur position par rapport à la racine. pour trouver la similarité entre deux objets X et Y de l'ontologie, on calcul les distances (N1 et N2) qui séparent les nœuds X et Y du nœud racine et la distance qui sépare le concept subsumant (CS) de X et de Y du nœud R (Wu & Palmer, 1994).

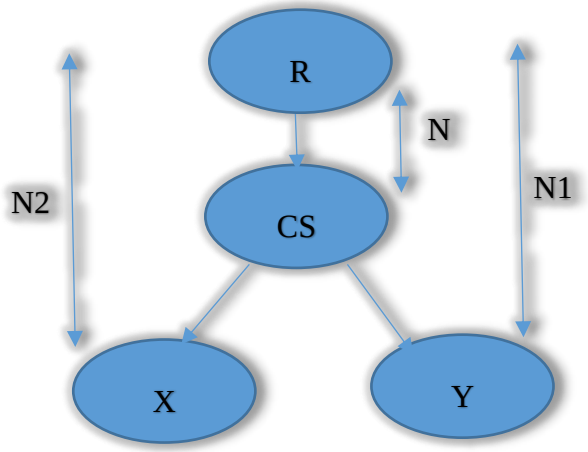


Figure 3-1 Exemple d'un extrait d'ontologie

La mesure de Wu & Palmer est définie par la formule suivante :

$$Simc(X,Y) = \frac{2N}{N1+N2} \qquad (3.1)$$

L'avantage de cette méthode est la facilité d'implémentation par rapport à d'autre méthodes Lin (Lin, 1998), un autre avantage est que cette méthode a une bonne performances que les autres méthodes de similarité, Cependant cette méthode possède une détection limite de la similarité entre deux concepts par rapport à leur distance avec la racine, et donc l'incapacité de détecter les mêmes similarités que la similarité conceptuelle symbolique. Un autre inconvénient est l'obtention d'une similarité plus

élevée entre un concept et son voisinage par rapport à ce même concept et un concept fils, ce qui la rend inconvenable pour la recherche de l'information (Zargayouna & Salotti, 2004).

3.2.1.2 *Mesure de Rada*

La similarité dans un réseau sémantique basée sur les liens taxonomiques a été introduit par Rada et al (Rada, Mili, Bicknell, & Blettner, 1989). Le principe de cette méthode consiste à utiliser des liens hiérarchiques pour calculer la similarité entre les concepts. Dans une taxonomie l'évaluation de la similarité sémantique est basée sur le calcul de la distance entre les concepts en basant sur le plus court chemin. Selon les chercheurs, cette approche est applicable pour tous les liens de type hiérarchique et qui nécessite une adaptation pour d'autres types de liens.

3.2.1.3 *Mesure d'Ehrig*

Ehrig et al. (Ehrig, Haase, Hefke, & Stojanovic, 2005) ont suggéré une méthode de mesure de similarité pour les ontologies. Trois couches nécessaires pour calculer cette mesure : les données, l'ontologie et le contexte. La première couche est consacrée pour mesurer la similarité des entités au niveau des données en considérant les valeurs de données de type simple ou complexe (entiers, caractères), tandis que la deuxième est réservée pour mesurer les relations sémantiques entre les entités dans l'ontologie. Finalement la couche du contexte spécifie comment les entités de l'ontologie sont utilisées dans un certain contexte externe, plus spécifiquement, le contexte de l'application.

3.2.1.4 *La mesure de Hirst-St.Onge*

La méthode de Hirst-St.Onge (Hirst & St-Onge, 1998) est basée sur les relations de WordNet pour calculer la similarité entre deux concepts. Ils ont proposé trois classes haut (partie-de), bas (sous-classe), horizontal (antonyme), pour calculer la similarité entre les mots, ils ont utilisé le poids du chemin le plus court entre les concepts. Il est calculé en fonctions de ces classifications qui indiquent les changements de direction

$$Sim\ (c1, c2) = S - C - K\ x\ d \qquad (3.2)$$

Tels que *S* et *K* sont des constantes, *C*'est la longueur du chemin le plus court en nombre d'arcs et *d* est le nombre de changements de direction.

Dans cette mesure, si deux concepts sont liés par un petit chemin et qui ne change pas la direction alors ils sont proche sémantiquement tandis que le poids est nul si les deux concepts ne sont pas liés.

L'inconvénient de cette mesure est qu'elle ne donne pas de bon résultat par rapport aux autres mesures parce qu'elle traite tous type de relation. (Zargayouna & Salotti, 2004).

3.2.1.5 *La mesure de Zargayouna*

L'idée principale de la méthode (Zargayouna & Salotti, 2004) est inspirée de celle de méthode de Wu & Palmer. La méthode donne une importance à la relation père-fils au détriment des autres relations de voisinage. L'adaptation de la méthode est réalisée par l'équation ci-dessous qui mesure le degré de spécialisation d'un concept et sa distance par rapport à l'anti-racine.

$$\text{sim}(c1, c2) = \frac{2 * N(c)}{(dist(c1, c) + dist(c2, c) + 2 * N(c) + S(c1, c2))} \quad (3.3)$$

$$spec(c1, c2) = 2 * prof_b(c) dist(c1, c) + dist(c2, c) \quad (3.4)$$

Où N(c) correspond au nombre maximum d'arcs qui séparent le plus petit ancêtre commun du concept « virtuel » représentant l'anti-racine.

3.2.2 Approches basées sur les nœuds (le contenu informatif)

Ces méthodes sont inspirées des mesures entropiques de la théorie de l'information. L'entropie est calculé par la formule ci-dessous dont P(.) représente la probabilité pour identifier une instance d'un concept.

$$E(c) = -log\big(P(c)\big) \quad (3.5)$$

La probabilité d'un concept est le rapport des instances de *c* sur le nombre total des instances. Pour éviter le problème de la fiabilité des distances des arcs, les auteurs de de ce type des approches ont associé des probabilités des concepts, ce qui a donné une nouvelle façon de calculer la similarité. Parmi les mesures basées sur le contenu informationnel on peut citer : mesure de Resnik, mesure de Lin et mesure de Seco.

3.2.2.1 ***Mesure de Resnik***

Resnik (Resnik, 1995) a introduit la notion du contenu informationnel pour calculer la similarité entre deux concepts en utilisant la quantité d'information commune entre eux par la formule suivante :

$$Sim(X,Y) = Max\left[E\left(CS(X,Y)\right)\right] = Max\left[-log\left(P\left(CS(X,Y)\right)\right]\right) \quad (3.6)$$

Dont le concept le plus spécifique est représenté par CS(X, Y) qui subsume les deux concepts X et Y. P est la probabilité de trouver une instance du concept *CS*.

3.2.2.2 ***Mesure de Lin***

La mesure de Lin (Lin, 1998) prend en compte le contenu informatif commun entre les deux concepts et leur propre contenu par cette formule :

$$Sim(c_1, c_2) = \frac{2.\log p\,(c)}{\log p\,(c_1).\log p\,(c_2)} \quad (3.7)$$

3.2.2.3 ***Mesure de Seco***

Seco et al. (Seco, Veale, & Hayes, 2004) ont défini le contenu informatif en basant sur l'hypothèse que, plus un concept a de descendants, moins il est informatif. Ils ont calculé le contenu informatif par l'utilisation des relations sémantiques hyponymes

$$IC_{wn}\,(c) = \frac{\log(\frac{hypo(c)+1}{max_{wn}})}{\log(\frac{1}{max_{wn}})} = 1 - \frac{\log\left(hypo(c)\right)+1}{\log(max_{xn})} \quad (3.8)$$

3.2.3 Approches hybrides

Plusieurs auteurs ont proposé de combiner l'approche basée sur les arcs par une approche basée sur les nœuds en formant une approche hybride. Parmi ces travaux on cite : la mesure de Jiang et Conrath et la mesure de Leacock et Chodorow.

3.2.3.1 *Mesure de Jiang et Conrath*

Jiang et Conrath (Jiang & Conrath, 1997) ont proposé une nouvelle mesure pour améliorer celle de Resnik. Cette mesure combine la technique basée sur les distances et la technique basée sur le contenu informationnel. La mesure de la distance de similarité est réalisée par la combinaison de l'entropie de contenu informationnel du concept spécifique et ceux des autres concepts. L'amélioration des résultats fait par le compte des arcs à partir des calculs basés sur les nœuds. Cette mesure est représentée par la formule suivante :

$$sim(c_1, c_2) = \frac{1}{distance(c_1, c_2)} \quad (3.9)$$

Dont la distance entre c_1 et c_2 est calculé par la formule suivante :

$$distance(c_1, c_2) = CI(c_1) + CI(c_2) - \left(2 * CI\left(PPG(c_1, c_2)\right)\right) \quad (3.10)$$

3.2.3.2 *Mesure de Leacock et Chodorow*

Leacock et Chodorow (Leacock & Chodorow, 1998) ont combiné les deux approches basées sur les arcs et celles qui ont basées sur le contenu informationnel, ils ont exploité les liens de l'ontologie Wordnet et précisément (is-a) pour trouver le plus court chemin entre les synsets. Cette mesure est représentée par la formule suivante :

$$sim(X; Y) = -\log\left(\frac{cd(X;Y)}{2*M}\right) \quad (3.11)$$

Sachant que M est la longueur du chemin le plus long qui sépare le concept racine, de l'ontologie, du concept le plus en bas (la profondeur global). On dénote par CD (X, Y) la longueur du chemin le plus cours qui sépare X et Y.

3.2.4 Approches basées sur une représentation vectorielle

Ces approches sont utilisées dans le domaine de la recherche d'information pour représenter des documents dont le nombre des termes qui constituent l'ensemble du corpus représente la dimension de l'espace vectoriel, et les vecteurs dans cet espace correspondent aux documents, dont chaque composante de vecteur correspond à la fréquence d'un terme dans le document. Parmi ces approches : L'indice de Jaccard, Similarité de cosinus et Similarité de Dice.

3.2.4.1 *L'indice de Jaccard*

Le coefficient de Jaccard évalue la similarité entre les concepts par la formule ci-dessous qui est un rapport du nombre de mots communs et la déférence du nombre de mots total et celui de mots communs :

$$sim(D_1, D_2) = \frac{\sum_i^k v_{1i} v_{2i}}{\sum_i^k v_{1i} v_{1i} + \sum_i^k v_{2i} v_{2i} - \sum_i^k v_{1i} v_{2i}} \qquad (3.12)$$

3.2.4.2 *Similarité de cosinus*

Cette approche est basée sur la représentation vectorielle. L'idée principale de cette technique est de calculer le cosinus entre les vecteurs qui représentent les documents. Le résultat de cette opération représente la similarité entre les documents. La formule est définie par le produit scalaire des vecteurs divisée par le produit de la norme des deux vecteurs.

$$Sim(X, Y) = \cos(X, Y) = \frac{x.y}{\|x\|_1 \|y\|_2} \qquad (3.13)$$

3.2.4.3 *Similarité de Dice*

La mesure de Dice, est définie par le double rapport entre le nombre d'attributs communs et le nombre total des attributs de chaque mot, La mesure de Dice est donc définie par la formule suivante :

$$Sim(X, Y) = \frac{2xy}{\|x\|_2^2 . \|y\|_2^2} \qquad (3.14)$$

3.2.5 Approches basées sur les traits

Cette approche est basée sur la théorie d'ensemble, deux concepts sont similaire si le nombre de ses traits communs est plus grand que celui des restants.

3.2.5.1 *Mesure de Tversky*

Le principe de base de la méthode de Tversky (Tversky & Gati, 1982) est de calculer le nombre des propriétés communes et distinctives entre deux concepts, plus le nombre des propriétés communes est important que celui des propriétés distinctives plus les concepts sont similaire.

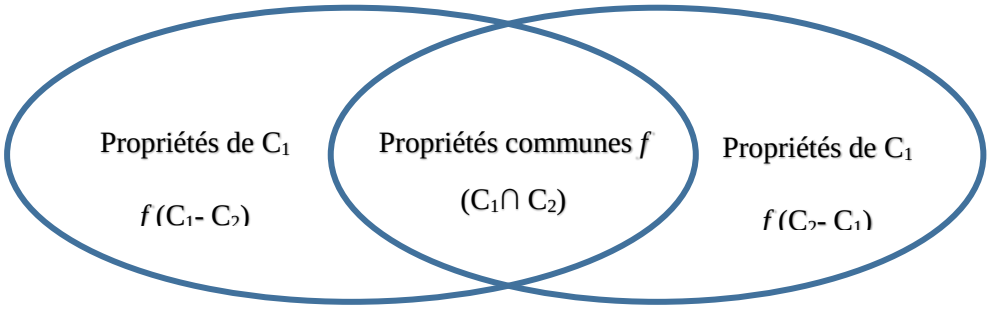


Figure 3-2 Similarité entre concepts selon Tversky (Tversky & Gati, 1982)

$$Sim(c_1, c_2) = \frac{f(C_1 \cap C_2)}{f(C_1 \cap C_2) + \alpha f(C_1 - C_2) + \beta f(C_1 - C_2)} \qquad (3.15)$$

Sachant que C_1 représente l'ensemble des propriétés de c_1 et C_2 représente l'ensemble des propriétés de c_2 ; $\alpha \geq 0$, $\beta \geq 0$ sont des paramètres qui pondèrent les différences et f est une fonction qui quantifie l'information que porte chaque concept. Dans ce cas :

f ($C_1 \cap C_2$) mesure l'information commune portée par les deux concepts.

f (C_1- C_2) et f (C_2- C_1) mesurent l'information spécifique que porte chaque concept par rapport à l'autre.

3.3 Algorithmes globaux stochastiques pour la désambiguïsation lexicale

Nous présentons dans cette partie quelques méthodes globales stochastiques pour la désambiguïsation lexicale : algorithme génétique, recuit simulé, la méthode recherche Tabou et chaines lexicales.

3.3.1 Algorithme génétique pour la désambiguïsation lexicale

3.3.1.1 *Principe de l'algorithme*

Les algorithmes génétiques, initiés dans les années 1970 par John Holland, sont des algorithmes d'optimisation. Ils combinent une stratégie de « *survie des plus forts* » avec un échange d'information aléatoire mais structuré. Pour un problème dans lequel une solution est inconnue donc un ensemble de solutions possibles est créé aléatoirement.

Cet ensemble est représenté par la population dont les variables sont déterminées par les gènes, ces derniers se combinent pour former des chromosomes et par la suite des individus. Un individu est lié à une solution, dont l'évaluation de cet individu est dépend de la correspondance avec la bonne solution au problème. La convergence vers la meilleure solution peut être démontrée par un processus de sélection naturelle inspiré de Darwin.

Trois étapes principales pour réaliser un algorithme génétique :

- Sélection
- Reproduction
- Mutation

L'algorithme génétique fut développé par Holland (Holland, 1975).

Début

1 : Générer une population aléatoire de n chromosomes

2 : Evaluer la fitness des chromosomes avec la fonction f (Gelbukh et al.)

---

3 : **Répéter**

4 : Calcule la fonction fitness f (Gelbukh et al.), pour tout chromosome x

5 : Appliquer l'opération de sélection

6 : Appliquer l'opération de croisement avec une probabilité PC

7 : Appliquer l'opération de mutation avec une probabilité PM

8 : Ajouter les nouveaux chromosomes à la nouvelle population

9 : Calcule la fonction fitness f (Gelbukh et al.), pour tout chromosome x

10 : Appliquer l'opération de remplacement

11 : Jusqu'à la satisfaction des conditions de terminaison

Fin

Figure 3-3 Algorithme génétique de base

L'algorithme génétique a été utilisé pour la désambiguïsation sémantique par Gelbukh et d'autres chercheurs dans (Gelbukh et al., 2003), dans cet algorithme la population est représentée par les chromosomes. Les allèles de chromosomes sont représentés par des indices du vecteur d'une configuration, donc les allèles possibles pour un indice sont les différents sens du mot ambigu.

3.3.1.2 *Avantages*

- L'obtention des bons résultats pour des problèmes très complexe
- Adaptation rapide à de nouveaux environnements.
- Plusieurs approches sont proposées.
- Coévolution, parallélisme et distribution facile.
- Les représentations facilitent la compréhension.

3.3.1.3 *Inconvénients*

- La sélection des bons éléments est difficile lorsque les performances des individus sont similaires.

- Coûteux en temps de calcul, puisqu'ils manipulent plusieurs solutions simultanément (Berthiau & Siarry, 2001).
- Aucune garantis de solution optimale en un temps fini.

3.3.2 Recuit simulé pour la désambiguïsation lexicale

3.3.2.1 *Principe de l'algorithme*

La méthode de recuit simulé est un algorithme d'optimisation (Hammersley, 2013). Elle a été proposée par des chercheurs Kirkpatrick, Gelatt et Vecchi (Kirkpatrick, Gelatt, & Vecchi, 1983). Le principe général du recuit simulé tel qu'il a été proposé par Metropolis en 1953 est de simuler le comportement de la matière dans le processus du recuit qui est largement utilisé dans la métallurgie. Le but est d'atteindre un état d'équilibre thermodynamique, cet état d'équilibre (où l'énergie est minimale) représente (dans la méthode du recuit simulé) la solution optimale d'un problème ; L'énergie du système sera calculé par une fonction coût (ou fonction objectif). La méthode va donc essayer de trouver la solution optimale en optimisant une fonction objective, pour cela, un paramètre fictif de température a été ajouté par Kirkpatrick, Gelatt et Vecchi.

L'application de cette méthode dans le domaine de désambiguïsation lexicale a été faite par Cowie et al. en 1992 (Cowie, Guthrie, & Guthrie, 1992). Elle réalise des changements aléatoires dans la configuration de l'espace de recherche, ensuite elle évalue le changement. Si ce dernier est bénéfique, elle le conserve.

Cette évaluation est représentée par l'équation suivante :

$$P\ (conservation) = e^{\frac{-\Delta s}{T}} \qquad (3.16)$$

$$\Delta s\ =\ score(C') - score(C) \qquad (3.17)$$

Dont

P : probabilité de conservation du changement

T : paramètre de température

C' : la configuration modifiée

C : la configuration avant modification

Le but de l'acceptation des configurations inférieures étant de ne pas converger sur un maximum local, comme le ferait une descente par gradient. La températureTdiminue géométriquement après un certain nombre d'itérations pour tout de même garantir une convergence (A. Tchechmedjiev, 2012).

1. Choisir ; aléatoirement, une solution initiale x du system à optimiser et évaluer de la fonction objectif $f = f(x)$;
2. Choisir une température initiale « élevée » T.
3. Perturber cette solution pour trouver une nouvelle solution $x' = x + \Delta x$;
4. Calculer $\Delta s = f(x') - f(x)$;
5. Accepter ou refuser la solution x′, en appliquant une certaine ″règle d'acceptation″
6. Sauver le meilleurs point rencontré ;
7. Si ″l'équilibre thermodynamique″ du system à la température T est atteint
   Alors abaisser légèrement la température T ;
   Sinon aller à l'étape 3) ;
8. Si ″le système est figé″ (par exemple, la température T est inférieure à une température seuil voisine de 0),
   Alors aller à l'étape 9) ;
   Sinon aller à l'étape 3) ;
9. Solution = meilleurs point trouvé ; arrêt du programme.

Figure 3-4 Algorithme du recuit simule

3.3.2.2 ***Avantages***

- Très simple et très rapide à mettre en place.
- Convergence vers un optimum global, la prédiction de la future à partir du présent ne nécessite pas la connaissance du passé. Plus clairement cette métaheuristique ne nécessite pas de mémoire (passé) afin de trouver les espaces de recherche locaux suivants (futur).
- Le recuit simulé peut trouver la meilleure solution si on le laisse chercher indéfiniment.

3.3.2.3 *Inconvénients*

- Non utilisation de mémoire bride.
- Il faut déterminer les paramètres à la main : température initiale, modification élémentaire en testant divers valeurs.

3.3.3 La méthode de recherche Tabou

L'idée principale de la méthode tabou est d'orienter d'autres méthodes pour trouver les bonnes solutions. La recherche Tabou est une méthode métaheuristique utilisée pour la résolution des problèmes d'optimisation, destinée principalement à guider d'autres méthodes afin de trouver de meilleures solutions à partir d'une solution initiale obtenue par l'une des heuristiques.

3.3.3.1 *Principe de base de la Recherche Tabou*

- La méthode Tabou permette d'explorer totalement les critères d'évaluation ainsi que l'historique de la recherche grâce à l'utilisation de mémoire flexible.
- Un mécanisme de contrôle basé sur l'alternance entre les conditions qui restreignent (restriction Tabou) et qui libèrent (critère d'aspiration) le processus de recherche.
- L'incorporation des stratégies dites d'intensification et de diversification de la recherche :
    - La stratégie d'intensification est une exploration locale poussée Elle soutient la recherche dans la région des meilleures solutions trouvées récemment.
    - La stratégie de diversification est un balayage de tout l'espace des solutions. Elle oriente la recherche dans de nouvelles régions.

Etape 1 : Initialiser une solution d'une manière aléatoire.

Etape 2 : Générer une liste des solutions.

Etape 3 : Adopter la meilleure solution selon le critère d'aspiration et les restrictions Tabou.

Etape 4 : Si une condition d'arrêt est atteinte, stop. Sinon, retour à Étape 2.

Figure 3-5 Algorithme générale de la recherche tabou

3.3.3.2 *Avantages*

- elle est utilisée dans les problèmes d'optimisation combinatoire : elle a été testée avec succès sur les grands problèmes classiques.
- Offre des économies de temps de résolution pour des programmes de grosse taille.
- Algorithmes faciles à implémenter.

3.3.3.3 *Inconvénients*

- Absence de garantie de résultat.
- Demande en ressources importantes si la liste des tabous est trop imposante.
- Difficulté de prévoir la performance (qualité et temps).

3.3.4 Chaines lexicales

Une chaine lexicale signifie la liaison de plusieurs mots entre eux dont le nombre de mailles représente la densité de la chaine tant dis que l'étendu de la chaine dans le texte représente la longueur ; la notion d'activité d'une chaine lexicale dépend de la couverture d'un point dans un texte. Les mots dans un texte sont reliés à des mots antérieurs ou à des concepts, cette relation constitue une chaine cohésive. L'idée principale de cette méthode a été proposée par Halliday et Hasan en 1976 (Halliday & Hasan, 1976) et en suite elle a été améliorée par Morris et Hirst en 1991 (Morris & Hirst, 1991).

```
REPEAT
    READ next word
    IF word is suitable for lexical analysis THEN
        CHECK for chains within a suitable span
        (up to 3 intermediary sentences, and no limitation on
         returns):
            CHECK thesaurus for relationships.
            CHECK other knowledge sources
            (situational, general words, proper names).
        IF chain relationship is found THEN
            INCLUDE word in chain.
            CALCULATE chain so far
            (allow one transitive link).
        END IF
        IF there are words that have not formed a chain for a suitable
         number of sentences (up to 3) THEN
            ELIMINATE words from the span.
         END IF
         CHECK new word for relevance to existing chains that
         are suitable for checking.
         ELIMINATE chains that are not suitable for checking.
    END IF
END REPEAT
```

Figure 3-6 Algorithme de construction des chaines lexicales (Morris & Hirst, 1991)

Parmi les inconvénients de cette méthode est le problème d'exactitude et de précision du fait de sa nature gloutonne (Navigli, 2009). Brazilay et Elhadad (Barzilay & Elhadad, 1999) ont proposé une approche pour améliorer la précision et pour une meilleur performance Silber et McCoy (Silber & McCoy, 2000) proposent un algorithme de construction de chaines lexicales linéaire.

3.4 Conclusion

Dans ce chapitre nous avons présenté quelques méthodes de similarité sémantique locales : les méthodes basées sur les arcs (distances), les méthodes basées sur les nœuds (le contenu informatif), les méthodes basées sur une représentation vectorielle et les méthodes basées sur les traits. Ainsi que nous avons présenté quelques algorithmes globaux stochastiques pour la désambiguïsation lexicale (Algorithme génétique, Recuit simulé, La méthode recherche Tabou, chaines lexicales).

On constate que les mesures de similarité sémantique locales sont très utiles pour les systèmes de traitement automatique du langage naturel, et également pour la construction automatique des ressources lexicales. Au niveau global une possibilité d'amélioration se situe au niveau du temps de convergence et des performances en général. Notre travail est basé principalement sur la combinaison des mesures de similarité aux niveaux local et global. Ces combinaisons peuvent se faire en utilisant des mesures différentes pour essayer de capter différents aspects utiles à la désambiguïsation.

Chapitre 4 :

Ressources linguistiques

Sommaire

4.1 Introduction

Aujourd'hui, les ressources linguistiques sont construites pour être utilisé par des programmes informatiques. Elles sont de plus en plus utilisées pour accompagner le travail de modélisation linguistique par des méthodes statistiques, qui tiennent une position de plus en plus importante dans les applications de (TALN). Dans ce chapitre, on donne un aperçu sur les ressources linguistiques existantes au niveau international. Nous abordons : Les lexiques, les grammaires syntaxiques, les corpus de textes monolingues bruts et annotés et les corpus multilingues alignés.

4.2 Définitions

Le concept de ressource linguistique est souvent employé sans que sa définition soit vraiment posée. On va voir quelques définitions provenant des acteurs experts de la matière.

Définition trouvée sur le site d'ELDA

« Les ressources linguistiques sont tous les types de données relatives à la langue, accessibles dans un format électronique, et utilisées pour le développement des systèmes de traitement de la parole et du texte dans des applications en technologies de l'information » (Cailliau, 2010)

« Une ressource linguistique est un ensemble de données comportant des connaissances linguistiques exploitables par un traitement automatique en particulier. » (Cailliau, 2010).

Définition en provenance du manuel de GATE :

« Language Resource (LR) : refers to data-only resources such as lexicons, corpora, thesauri or ontologies. Some LRs come with software (e.g. Wordnet has both a user query interface and C and Prolog APIs), but where this is only a means of accessing the underlying data we will still define such resources as LRs. » (Kenter & Maynard, 2005).

« Processing Resource (PR): refers to resources whose character is principally programmatic or algorithmic, such as lemmatises, generators, translators, parsers or

speech recognisers. For example, a part-of-speech tagger is best characterised by reference to the process it performs on text. PRs typically include LRs, e.g. a tagger often has a lexicon; a word sense disambiguator uses a dictionary or thesaurus. » (Cunningham & Scott, 2004).

Nous souhaitons aussi faire la distinction entre ressources linguistiques et ressources langagières. Ce dernier terme englobe des ressources bien plus vastes que les ressources linguistiques comme par exemple des dictionnaires électroniques ou papier, des exercices de langue, des correcteurs orthographiques, etc. Le terme de ressources langagières est donc beaucoup plus général et s'applique sur tout document (au sens large) ou logiciel qui traite la langue.

4.3 Les lexiques

Pour traiter les langues naturelles, il est indispensable de disposer de différentes informations linguistiques relatives aux mots : Quels sont-ils ? Comment s'écrivent-ils ? Quelles sont leurs différentes formes (flexionnelle, dérivationnelle ou compositionnelle) ? Quelles informations morphologiques leurs sont liées ?

Le lexique ou le dictionnaire comporte l'ensemble de connaissances associe aux mots. L'ensemble (ou les ensembles) d'informations, liées au mot, en constituent l'entrée lexicale ou (les entrées lexicales). Le lexique contient donc l'ensemble des données et des règles qui vont permettre aux systèmes d'identifier les mots bien formés et leur associer toutes les données pertinentes pour la suite de traitement.

4.3.1 Les informations lexicales

On peut distinguer deux types d'informations lexicales intralexicales et interlexicales.

4.3.1.1 *Les informations intralexicales*

Ces informations définissent les contextes variés dans lesquels les mots apparaissent ; Elles sont de nature morphologique, syntaxique, sémantique et phonétique.

4.3.1.2 *Les informations interlexicales*

Elles construisent des relations entre les mots sur plusieurs axes morphologique, paradigmatique (synonymie, antonymie, mèronymie) et syntagmatique (collocations, idiomes). Ce type d'information a été écarté car il complique la structure du lexique.

4.3.2 Les lexiques monolingues

Les lexiques monolingues sont des ressources intéressantes pour les analyses linguistiques (morphologiques, syntaxique, sémantique) des documents.

4.3.2.1 *Le lexique BDlex*

Le lexique BDLex est utilisé pour le traitement morphologique ainsi que phonétique. Il inclut 450.000 formes fléchies générées à partir d'environ 50000 formes canoniques avec des connaissances sur la prononciation et la morphosyntaxe.

Ce lexique contient d'autres informations concernant les statistiques lexicales qui sont représentées sous forme d'indice de fréquence.

Graphie	Prononciation		Morpho syntaxe			
ORTHO	PHONO	FPH	CS	VS	M	LIEN
prendre	pRa~dR	@	V		inf	=
prennent	pREn	@t"	V	3P	pi	prendre
petites	p@tit	@z"	J	FP		Petit
Un	9~	n"	D	MS	di	=
Avion	avjo~		N	MS		=

PHONO : représentation phonologique, FPH : fonctionnement phonologique de la finale,
CS : catégorie syntaxique, VS : variation syntaxique, M : mode,
LIEN : entrée lexicale (lemme) dont la forme est dérivée.

Figure 4-1 Structure lexicale des entrées de BDlex (Nguyen, 2006)

4.3.2.2 *Les Ressources MHATLex*

Ces ressources se ressemblent aux celles de BDLex au niveau de lexique et de l'aspect morphosyntaxique. Tandis qu'elles se diffèrent sur le plan prononciation: de façon plus clair et plus pratique à la reconnaissance vocale, MHATLex permet une meilleur modélisation de la prononciation avec les variabilités libres et contextuelles (Pérennou & De Calmes, 2000).

Les ressources MHATLex sont représentées en trois paliers (syntaxique : S, phonologique des mots : W, phonétique : P)

Il existe deux formes de représentation d'un mot au palier phonologique

- la représentation d'entrée (représentation W) où les mots sont simplement importés du lexique,
- la représentation de sortie (W' ou phonotypique) où les mots ont une représentation phonotypique qu'impose leur contexte dans la phrase.

Ces ressources comportent des mots fléchis (parmi lesquels les mots canoniques).

MHATLexSt (& BDLex) MHATLexW : comporte environ 50 000 entrées (canonique) & 440 000 entrées (fléchis).

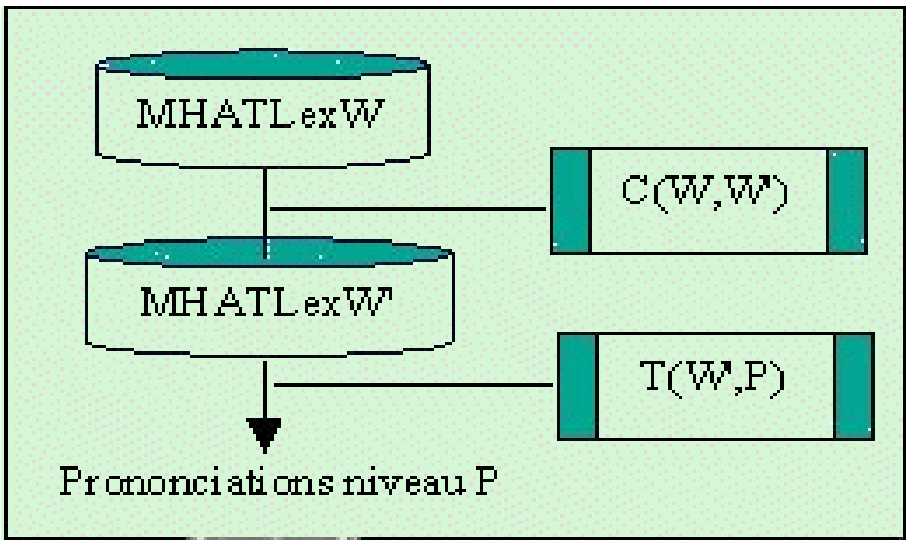


Figure 4-2 Ressources MHATLex

MHATLexW' : comporte environ 81 000 entrées (canonique) & 854 000 entrées (fléchis)

La représentation des mots se fait selon les attributs suivants : leur graphie, leur prononciation leurs attributs morpho-syntactique et l'indicateur de fréquence.

Seule la partie relative à la prononciation change selon le lexique (sauf si l'usager génère son propre lexique en se passant de quelques attributs).

Quatre lexiques peuvent être générés dans MHATLex :

- MHATLexW : c'est actuellement la ressource lexicale centrale permettant de générer tous les autres lexiques.
- MHATLexW' (ou MHATLexPht) : donne les représentations des mots pour chaque contexte pertinent.
- MHATLexSt : avec forme standard et simplifiée de la prononciation.
- BDLex (ou BDLex50) : forme déjà distribuée par ELDA[8]

4.3.2.3 ***Wordnet***

Miller et ses collègues (Miller, 1995) ont implémenté la base de données lexicale WordNet en 1985 au niveau de laboratoire des sciences cognitives de l'université de

[8] http://www.elda.org/catalogue/fr/speech/S0004.html

Princeton, qu'est une base de données lexicale concernant la langue anglaise, Ils ont basé sur la théorie psychologique du langage. La première version a été apparue en 1991.

La mise en relation par différente façon du contenu sémantique et lexical est le but principal de WordNet. La base de données électronique est téléchargeable sur un système local. L'avantage de Wordnet est sa disponibilité pour plusieurs langages.

Beaucoup de ressources linguistiques ont été construit à partir de WordNet. Des programmes issus du monde de l'Intelligence Artificielle ont également établi des passerelles avec WordNet. (Chaumartin, 2007).

Figure 4-3 Ressources disposant d'une traçabilité vers WordNet (Chaumartin, 2007)

Wordnet comporte 200000 paires de mots-sens. Chaque mot représente un concept lexical. L'organisation des mots est réalisée en groupe des synonymes ou synsets. Ces derniers sont organisés sous forme d'un graphe dont les nouds représentent les concepts et liens représentent les relations sémantiques (hyponymie, antonymie, métonymie,...).

Tableau 4-1 Nombre des mots, synsets et sens

Partie de discours	Synsets unique	chaines	Paires de mots-Sense
Noms	117798	82115	146312
Verbes	11529	13767	25047
Adjective	21479	18156	30002
Adverbe	4481	3621	5580
Totale	155287	117659	206941

Tableau 4-2 Des informations de polysémie

Partie de discours	Monosémies	Polysèmes	Polysèmes
	Mots and Sense	Mots	Sense
Nom	101863	15935	44449
Verbes	6277	5252	18770
Adjective	16503	4976	14399
Adverbe	3748	733	1832
Totale	128391	26896	79450

Wordnet joue un rôle intéressant pour plusieurs travaux qui nécessite un étiquetage sémantique ou qui visent l'accès aux textes par le sens.

Sense 2
dictionary, lexicon
=> wordbook
=> reference book, reference, reference work, book of facts
=> book<<<<
=> publication
=> work, piece of work
=> product, production
=> creation
=> artifact, artefact
=> object, physical object
=> entity
=> whole, whole thing, unit
=> object, physical object
=> entity

Figure 4-4 Exemple de hiérarchie hyperonymique dans WordNet

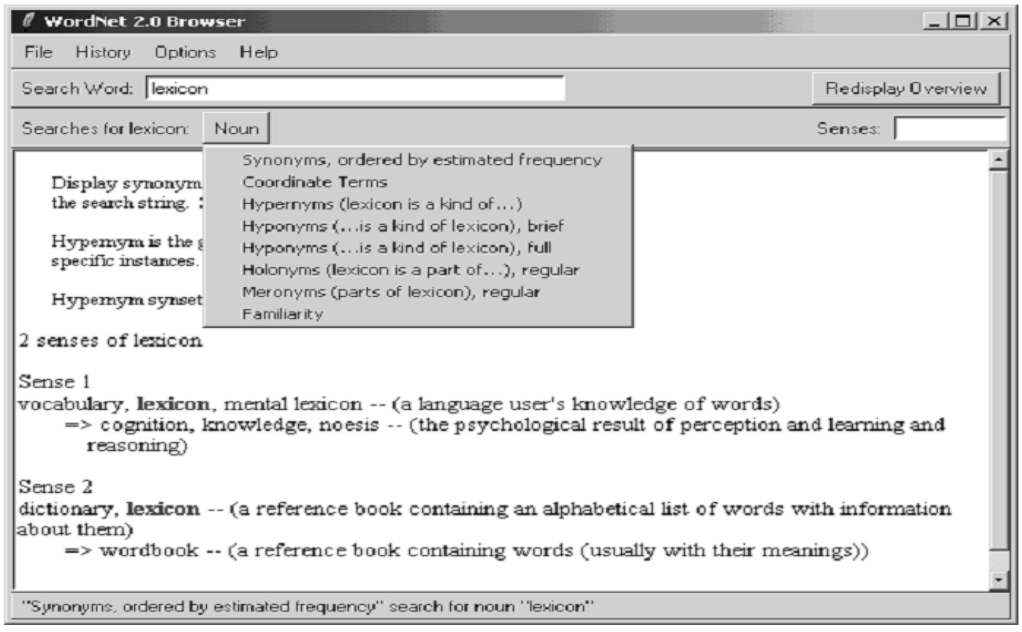


Figure 4-5 Exemple de consultation de Wordnet

4.3.2.4 ***FrameNet***

Le projet FrameNet a débuté il y a plus de dix ans (Baker, Fillmore, & Cronin, 2003; Lönneker-Rodman & Baker, 2009). Le but de ce projet est d'étudier les liens entre les unités lexicales (paires mot-sens) et leur cadre sémantique, en utilisant des corpus notamment le British National Corpus (BNC). Pour chaque unité lexicale, on trouve ses définitions et des exemples annotés (voir Figure 4.6), elle illustre toutes ses possibilités combinatoires, et liée à un cadre sémantique, qui peut être partagé par d'autres unités lexicales (Baker, 2009).

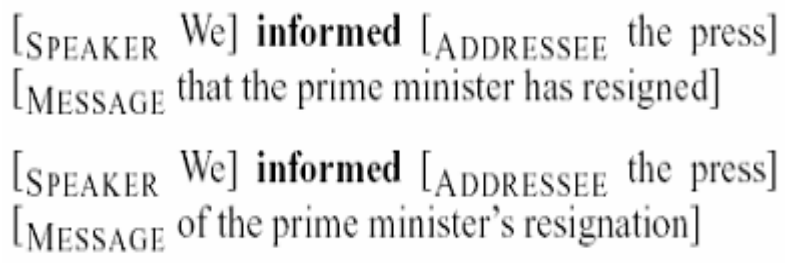


Figure 4-6 FrameNet – Exemples annotés du cadre sémantique du verbe « inform »

Le but de ce projet est la description des possibilités combinatoires de chaque acception d'un mot en même temps sur tous les plans linguistiques (leurs valences) grâce à un système semi-automatique d'étiquetage qui permet l'affichage des résultats (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006). Les exemples de la section précédente sont tirés des annotations de la base de données FrameNet.

FrameNet contient actuellement plus de 13492 unités lexicales, dont plus de 8392 (62%) sont complètement annotées dans 12.4 UL/ cadres sémantiques, et exemplifiés dans plus de 202101 phrases annotées. FrameNet contient également un réseau de relations entre les cadres. La base est disponible sous licence par le biais de son site Internet[9]. D'autres projets dérivés pour l'allemand, l'espagnol et le japonais (Baker, 2009).

[9] https://framenet.icsi.berkeley.edu/fndrupal/IntroPage

Tableau 4-3 Frame

Total Frames	1218
Lexical Frames	1084 (89%)
Non-Lexical Frames	134 (11%)
FEs in Lexical Frames	10470
FEs/Lexical Frame	9.7
Frame Relations	1838
Frame Element Relations	10438

Tableau 4-4 L'état des unités lexicales par une partie du discours

Lexical Units	In FN	Finished	>10 Annotation
Nouns	5486	2701	2210
Verbs	5157	2861	2289
Adjectives	2378	1368	1062
Other POS	471	65	129
Total	13492	6995	5690

Tableau 4-5 Distribution des unités lexicales par des ensembles d'annotation par unité lexicale

Annotation	1–9	10–14	15–19	20–29	≥30	Total
Full text	3580	262	143	113	156	4254
Lexicographic	2789	1221	1012	1571	1799	8392

4.3.3 Les lexiques multilingues

Deux types de lexiques multilingues sont distingués : les lexiques qui s'occupent de la correspondance entre deux langues et les lexiques qui s'intéressent au développement d'un mécanisme générique qui permettent la mise en parallèle d'informations lexicales pour un nombre a priori arbitraire de langues.

4.3.3.1 *Lexiques bilingues*

Les entrées lexicales monolingues sont enregistrées dans deux dictionnaires de mots qui fournissent leur information grammaticale (représentée comme une liste d'attributs) et un lien à un concept du dictionnaire de concepts. Chaque entrée lexicale est une forme fléchie de mot, ce qui n'est pas très efficace pour les langues fortement flexionnelles comme le français (Nguyen, 2006).

Tableau 4-6 Lexiques Multilingues

Lexique multilingue de base (MEMODATA)
Entrées : 30 000 pour chaque langue Langues : français, anglais, italien, allemand, espagnol Format : ASCII ou ANSI avec séparateurs entre les entrées. Support : CD-ROM Les mots sont associés par leur sens. Les catégories lexicales sont : noms (5 * 18 000), verbes (5 * 8 000), adjectifs (5 * 6 000), adverbes (5 * 1 500).
Lexique hollandais-français (LanTmark)
Vocabulaires génériques et spécialisés pour le transfert. Vocabulaire général : 26 000, Administratif : 32000, Traitement de données : 10 000 Format : ASCII : jeu de caractères ISO 8859-1. Support : disquette, cartouche QIC 150 MB Ce lexique général hollandais-français est réparti selon les catégories suivantes : noms (14.000), verbes (6.000), adjectifs (5.000), adverbes (1.000). Le vocabulaire administratif est réparti selon les catégories suivantes : noms (30.000), verbes (2.000). Le vocabulaire traitement de données dispose de 10 000 noms de transfert. Chaque entrée contient des informations sur le domaine, désambiguïsation sur la langue source, des traits, des actions sur la langue cible.

Lexique anglais-français (LanTmark)
Vocabulaire général pour transfert Entrées : 33287 Format : ASCII Support : disquette, cartouche QIC 150 MB Ce lexique anglais-français est réparti selon les catégories lexicales suivantes : noms (env. 14 000), verbes (env. 7 000), adjectifs (env. 5 000), adverbes (env. 1 000). Chaque entrée contient une information sur le domaine, désambiguïsation sur la langue source, des traits, des actions sur la langue cible.
THAMUS Dictionnaires bilingues
Domaines techniques Langues : Anglais => Italien Domaine : Informatique, 15 700 entrées, formes canoniques Les dictionnaires techniques bilingues disposent d'une codification morphologique qui permet de générer toutes les formes fléchies grâce à un logiciel écrit en langage C. Les mots composés contiennent une codification morphologique sur la tête des mots.

4.3.3.2 ***Wikipédia***

Wikipédia est une encyclopédie collective établie sur Internet, universelle, multilingue (291 langues mi-2015) et fonctionnant sur le principe du wiki, c.-à-d. un site Web dont les pages sont modifiables (gratuitement) par tout ou partie des visiteurs du site (Sadat & Terrasa, 2010). Wikipédia est considéré comme une source de références générales d'Internet. Le nombre de visiteurs est en augmentation continu, il y a d'environ 500 million visiteurs chaque moi.

Jimmy Wales ainsi que Larry Sanger ont lancé Wikipédia le 15 janvier 2001. Bien que son contenu ait alors été principalement en anglais, il est rapidement devenu multilingue, grâce en partie au lancement des versions internationales de Wikipédia. Toutes ces versions sont semblables les unes aux autres, mais il existe des différences notamment dans le contenu et les techniques d'édition.

Depuis sa création, le contenu de Wikipédia a connu une progression quasi exponentielle, du moins pour sa version anglaise. Wikipédia comporte aujourd'hui 160 éditions différentes en langues « actives » (plus de 100 articles), et totalise en septembre 2006 plus de 5 millions d'articles, dont 1,4 millions en anglais, 460 000 en Allemand, et 355 000 en Français (Firer-Blaess, 2007).

Figure 4-7 Resultat de la recherche du mot Wikipédia

Wikipédia en arabe a été lancée le 9 juillet 2003 et compte, au 9 septembre 2015, 381 437 articles et 1 062 202 utilisateurs enregistrés, ce qui en fait la 21eme édition linguistique de Wikipédia par le nombre d'articles et la 9eme par le nombre d'utilisateurs enregistrés, parmi les 287 éditions linguistiques actives.

Tableau 4-7 Éditions de Wikipédia

Langue	Nombre d'articles	Moyenne d'articles par jour depuis le 20/11/14	Nombre de pages
anglaise	5 059 439	955,69	38 288 535
allemande	1 898 626	283,01	5 445 907

Langue	Nombre d'articles	Moyenne d'articles par jour depuis le 20/11/14	Nombre de pages
française	1 716 205	359,32	7 918 648
russe	1 283 225	278,83	4 673 345
espagnole	1 226 854	205,64	5 353 089
portugaise	906 195	132,17	4 004 217
chinoise	857 570	143,97	4 547 262
arabe	405 250	153,24	2 519 897
turque	259 842	54,97	1 395 227
basque	230 043	60,2	602 032
bulgare	212 210	Entrée	469 644
slovaque	208 128	Entrée	447 394

4.3.3.3 ***Wiktionnaire***

Wiktionnaire est un dictionnaire multilingue, universel et librement diffusable. Il a été lancé le 15 janvier 2001 par Jimmy Wales et Larry Sanger. Son nom provient de deux mots «wiki» et «dictionnaire».

L'objectif du projet Wiktionnaire est de créer un dictionnaire multilingue au contenu gratuit. Il contient des thésaurus, des rimes, des traductions, des prononciations audio, des étymologies et des citations (Navarro et al., 2009). Depuis mai 2013 le Wiktionnaire est disponible dans plus de 150 langues avec plus de 16 000 000 entrées. L'édition anglaise contient Le plus grand nombre d'entrées, avec 3 380 000 entrées.

La version arabe de Wiktionnaire est lancée le 24 mai 2004, elle contient maintenant plus de 60,000 articles.

4.4 Grammaires électroniques

Une grammaire électronique est un objet qui se situe à la charnière entre la complexité et la diversité linguistiques et les besoins de régularité et d'efficacité pour le traitement informatique (Marie-Hélène Candito, 1999).

Elle contient des informations détaillées et exhaustives sur la sous-catégorisation des foncteurs syntaxiques, c'est-à-dire sur le nombre et le type de leurs arguments (Gardent, Guillaume, Falk, & Perrier, 2005). Les grammaires permettent d'analyser et de modéliser la structure syntaxique de la phrase, et ainsi de préciser les relations existantes entre ses composants, ce qui constitue un apport d'information très important pour l'accès au sens. Les recherches menées en TAL ont déclenché le besoin de grammaires à large couverture, c'est à dire elle permettent de modéliser toute la langue, ou bien une partie importante de celle-ci. Ces grammaires doivent décrire, pour une langue (Nguyen, 2006):

- les dépendances locales : accord, sous-catégorisation des prédicats, expressions semi-figées, restrictions modifieurs-modifié, clitiques, etc.,

- les dépendances « moyennes » : pronominalisation, contrôle des infinitives, association négative, quantifieurs flottants, etc.,

- les dépendances à distance : questions, relatives, constructions disloquées, etc.,

- les alternances syntaxiques : passif, impersonnel, causatives, etc.,

- les phénomènes de coordination et de comparaison.

Plusieurs projets ont été entrepris et menés à bien pour la création de grammaires, principalement pour les langues indo-européennes, qui sont les plus étudiées.

4.4.1 Exemples

Depuis plusieurs années une équipe de chercheur a présenté un bon travail sur l'application de la grammaire électronique des programmes de traitement.

Le développement des lexiques multilingues librement accessibles nécessite des modèles syntaxiques stables basés sur l'unification, qui inclut des outils de test et de mise à jour et permettent la création de grammaires neutres quant à leur application.

On présente quelques grammaires qui sont citées dans le travail de Abeillé (Abeillé, 1998) :

La grammaire Lexicale Fonctionnelle (LFG) représente la grammaire d'unification qui a été présentée par les chercheurs J. Bresnan et R. Kaplan (Kaplan & Bresnan, 1982). Elle consiste à rétablir les fonctions grammaticales, qui sont liées aux syntagmes dans l'arbre syntaxique de surface. Ainsi elle décrit ces fonctions comme des structures de traits qui contiennent toutes les fonctions des éléments subordonnés.

G. Gazdar et d'autres chercheurs (Gazdar, 1985; Gazdar & Pullum, 1982; Sells, 1989) ont proposé une grammaire Syntagmatique Généralisée (Generalized Phrase Structure grammar, ou GPSG), qui consiste à établir et d'enrichir les règles syntagmatiques en décomposant les catégories en traits et en décrivant des principes de partage de valeurs de traits entre composants d'un même syntagme. Ce type de grammaire ressemble à la grammaire hors contexte du langage formel.

Les grammaires catégorielles avec unification ont été proposées par Baschung et Kray-Baschung (Baschung, 1990; Kray-Baschung, 1992) . L'intérêt de ce travail est d'enrichir les grammaires catégorielles en analysant les catégories en traits. Ils ont décrit les items lexicaux sous forme de catégories complexes pour déterminer leurs propriétés combinatoires et que l'analyse syntaxique conduit à un problème de déduction logique.

A. Joshi (Joshi, 1987; Joshi & Schabes, 1997) ont présenté la grammaire d'arbres adjoints (TAG) qui met la grammaire sous la forme d'une structures arborescentes dans le lexique mais élémentaires (arbres lexicalisés) enrichies par des traits. Ce type de grammaire ressemble à la grammaire sensible au contexte du langage formel.

C. Pollard et I. Sag (Pollard & Sag, 1987) ont proposé un autre type de grammaire qui sont les grammaires syntagmatiques guidées par les têtes qui combinent les grammaires syntagmatiques généralisées et celle de Chomsky. Dans cette dernière, l'expression des contraintes des connaissances linguistiques est représentée sous forme des structures hiérarchisées qui décrit partiellement les objets linguistiques.

Pour l'anglais, le projet XTAG a développé et distribué (sous licence GPL) une grammaire électronique « réutilisable » à grande échelle (Prolo, 2002; Tateisi, Torisawa, Miyao, & Tsujii, 1998). XTAG utilise le formalisme LTAG (Lexicalized Tree Adjoining Grammar) (Marie-Helene Candito, 1996). La grammaire se compose des familles d'arbres initiaux représentant les cadres de sous-catégorisation, ainsi que des arbres auxiliaires permettant l'adjonction des modifieurs dans les syntagmes.

4.5 Les corpus

Selon Sinclair (Sinclair, 1996) la communauté linguistique considère qu'un corpus est "*Une collection de données Langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de langage*".

Une autre définition plus souple a donc été proposée par Gibbon et al (1998) :

"*A corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house*" (Gibbons, Gerrard, Blanton, & Russell, 1998).

4.5.1 Corpus de textes bruts et étiquetés

L'étiquetage d'un corpus est le processus qui consiste à associer aux mots d'un texte les informations grammaticales correspondantes comme la partie du discours, le genre, le nombre, le sens, etc. à l'aide d'un outil informatique. Ces ressources sont cruciales pour les études ultérieures comme le découpage du texte en groupes syntagmatiques, son analyse syntaxique, l'élaboration de concordances, etc. ; elles peuvent également être employées par des applications « finales » comme le résumé automatique de textes ou l'extraction de terminologie.

Les corpus de la langue anglaise américaine sont les premiers qui ont été construits. Le corpus de Brown est le plus ancien et le plus connu de ces corpus, il rassemble un million de mots étiquetés manuellement. Par sa mise dans le domaine public, ce corpus a joué un rôle moteur pour les recherches basées sur les corpus. Son équivalent pour l'anglais britannique est le corpus de Lancaster-Oslo-Bergen (LOB).

Un autre corpus de l'anglais britannique qui est le BNC (British National Corpus) fournit une ressource très importante. Il contient 100 millions de mots (dont 90%

relèvent de la langue écrite et 10% de la langue orale). Ce corpus contient des textes de fiction et des textes informatifs venant de livres,revues périodiques, discours, etc.

4.5.2 Corpus arborés : Treebanks

Les Treebank jouent un rôle très important pour la description de la syntaxe d'une langue, ainsi que pour l'entraînement ou la validation des systèmes d'analyse automatique. Ces corpus peuvent avoir des utilisations multiples en TAL, par exemple pour l'évaluation des étiqueteurs et des parseurs, pour l'induction de grammaires ou de préférences syntaxiques, pour la construction automatique d'étiqueteurs ou de parseurs probabilistes, ou pour l'enrichissement des dictionnaires électroniques (extraction de collocations, extraction de cadres de complémentation). De tels corpus permettent également d'obtenir une documentation détaillée sur les annotations attendues pour les principales constructions rencontrées dans les textes mais négligées dans les grammaires (Nguyen, 2006).

4.5.2.1 *Exemple des corpus arborés :*

On présente quelques corpus arborés qui sont cité dans le travail de Nguyen (Nguyen, 2006) :

- Les corpus de la langue anglaise sont les premiers corpus qui ont été construits. Parmi eux on note Les projets Treebanks, qui sont des exemples de création de grands corpus annotés. L'université de Pennsylvanie aux États-Unis a fait construire le Penn Treebank, distribué par le LDC, qui contient environ 4 millions de mots, dont 2 millions de mots sont analysés pour la structure prédicat-argument.
- Le corpus Suzanne est un corpus de taille moins importante par rapport aux corpus précédemment décrits. Il contient 128 000 mots extraits de corpus Brown qui sont annotés sémantiquement.
- L'annotation des mots par la grammaire de dépendance a été réalisée dans le corpus Bank of English, qui contient 200 millions, qui est considéré comme une référence importante depuis 1995, mais ce corpus n'est malheureusement pas disponible publiquement.

* Un corpus journalistique français contient un million de mots annotés syntaxiquement a été réalisé par la coopération de trois équipes : TaLaNa dirigée par A. Abeillé à l'Université de Paris 7, l'équipe LATL (Laboratoire de Genève) et l'équipe RALI (laboratoire de Montréal).

4.5.3 Corpus multilingues alignés

Les corpus multilingues alignés sont des ensembles de paires de documents dans des langues différentes étant des traductions mutuelles. Dans un corpus alignés, nous distinguerons les documents sources auxquels seront associés des documents cibles. Ces corpus peuvent être utilisés pour la création de bases de données multilingues, la consolidation des lexiques, la construction et la validation de mémoires de traduction.

Pour les langues occidentales comme l'anglais et le français, la création des corpus est facile par rapport aux autres langues moins pratiquées, ou pratiquées dans les pays moins développés technologiquement, en raison de manque de ressources linguistiques électroniques, et pour les langues n'employant pas l'alphabet latin, les codages des textes dans les différentes langues considérées peuvent s'avérer incompatibles.

Il existe plusieurs travaux visant à la création de corpus de textes parallèles, dont les principaux sont :

- Corpus BAF est un ensemble de paires de documents anglais et français dont les phrases ont été "alignées". Il contient environ 400 000 mots dans chaque langue et qui sont classés en 4 catégories : institutionnels, scientifiques, techniques et littéraires.

- Le corpus Hansard (français-anglais) est l'un des premiers corpus de textes parallèles, et le plus connu, il a été construit en 1980 et qui contient 50 millions mots extraits des délibérations du Parlement canadien.

- Le corpus de textes trilingues (français, anglais, espagnol) de l'International Télécommunications Union CCITT Handbook (13,5 millions de mots) et de l'International Labour Organisation (5 millions de mots) ;

- Dans le cadre du projet MULTEXT financé par la Commission européenne (LRE 62-050) un corpus a été construit, ce corpus comprend des données brutes, étiquetées et alignées des questions écrites et des réponses du Journal Officiel de la Communauté

Européenne. Ce corpus contient environ 5 millions de mots en allemand, anglais, espagnol, français et italien.

4.6 Conclusion

Dans ce chapitre, nous avons vu trois types de ressources linguistiques (lexique, grammaire, corpus) avec quelques exemples.

Le lexique pour le TAL doit être plus qu'une base de données pourvue d'un mécanisme d'interrogation. Il ne se limitera pas à une description statique des mots mais il sera pensé comme un objet dynamique capable de gérer la productivité linguistique.

Les grammaires s'intéressent de la construction des mots et de la structure des phrases, tandis que les corpus jouent un rôle important pour la construction des dictionnaires et des grammaires.

Chapitre 5 :

Travaux connexes

Sommaire

5.1 Introduction

Le processus de désambiguïsation lexicale (Word Sense Disambiguation – WSD) consiste à sélectionner les sens corrects d'instances contextualisées des mots ambigus, parmi l'ensemble de leurs sens possibles ou sens candidats (Apidianaki, 2008) .

La désambiguïsation lexicale automatique est une étape importante dans la plupart des applications de traitement automatique des langues. Nous citons par exemple la recherche d'information, si l'on recherche dans une base de données conséquente des informations à l'aide d'une requête textuelle, la désambiguïsation permet de donner un sens aux mots de la requête et donc de retourner une information plus ciblée (Claude de Loupy, El-Beze, & Marteau, 1998).

Le domaine de désambiguïsation a connu une expansion de travaux cherchant à résoudre le problème de l'ambiguïté. Dresser un panorama exhaustif d'un état de l'art reste une tâche difficile, on présentera dans ce chapitre un aperçu général sur le domaine de la désambiguïsation sémantique automatique en se référant de l'état de l'art présenté dans (Ide & Véronis, 1998) qui introduit le domaine en détail et qui a été mis à jour par Audibert en 2003 (Audibert, 2003a) et par Kolhatkar (Kolhatkar, 2009) et Navigli en 2009 (Navigli, 2009).

Dans ce chapitre, nous présentons plusieurs approches de la désambiguïsation sémantique qui ont été abordées en informatique : approches basées sur les connaissances comme celles qui sont basées sur les préférences sélectionnelle, les approches basées sur le chevauchement entre les différentes définitions ainsi que les approches basées sur un corpus étiqueté ou non étiqueté. Enfin, nous présentons les travaux de recherche qui focalisent sur la langue arabe.

5.2 Désambiguïsation lexicale

Le processus de désambigüisation lexicale (Word Sense Disambiguation) consiste à sélectionner les sens corrects d'instance contextualisé des mots ambigus, parmi l'ensemble de leur sens possibles (ou sens candidats).

Pour choisir le sens d'un mot le plus approprié en contexte, on a besoin d'un modèle sophistiqué qui peut le sélectionner parmi les sens proposés, ainsi que l'existence d'un

dictionnaire pour décrire les sens possible. Le résultat de cette tâche peut être exploité par plusieurs applications comme la traduction automatique, la recherche d'information, l'indexation sémantique ainsi que il peut servir de métadonnées pour l'étiquetage sémantique de texte, ce qui permet la création de ressources enrichies par des informations sémantiques.

Un consensus est atteint pour définir la désambiguïsation sémantique comme la sélection des sens corrects d'instance contextualisé des mots ambigus, parmi l'ensemble de leur sens possibles. En dépit des difficultés liées à la problématique de la désambiguïsation sémantique qui ont très tôt été identifiées. Des solutions ont été proposées dans chaque domaine ont également été multiples et très diverses, en fonction des besoins et des savoirs afférents à chacune des matières concernées.

Il faut d'abord déterminer l'ensemble des sens possibles pour chaque mot de la langue. [Kelly et Stone, 1975] montre que l'affectation d'un sens particulier à un mot polysémique dans un contexte donné n'est pas chose facile. Néanmoins, la désambiguïsation sémantique est réalisée à partir de sens prédéfinis, grâce à diverses ressources linguistiques. La sélection d'un sens particulier d'un mot est basée sur le contexte d'apparition des occurrences de chaque mot et les bases de connaissances externes qui permettent de corréler les mots en contexte avec leur sens.

C'est sur la nature de la base de connaissance que survient ici le désaccord,

- certaines méthodes privilégiant des ressources d'un ordre plutôt lexical pour fournir ces données,
- d'autres préfèrent des informations sur le contexte provenant de corpus aux unités lexicales préalablement désambiguïsées.

Parmi les différentes méthodes de détermination de la signification des mots en contexte, on trouve celles qui correspondent aux restrictions, à savoir les méthodes qui se fondent sur des critères linguistiques pour effectuer la tâche qui leur est confiée. Par ailleurs, la désambiguïsation sémantique s'inscrit ici dans le contexte d'un processus d'enrichissement du texte qui lui est soumis, et doit de ce fait permettre la sélection de l'information lexico-syntaxique la plus riche et la plus précise, ce qui implique l'utilisation d'une ressource lexicale bien structurée. De plus, le texte qui est soumis à la désambiguïsation sémantique est libre et susceptible d'atteindre un volume

important, d'où une nécessité de robustesse. Ces exigences limitent donc l'horizon des systèmes de désambiguïsation sémantique auxquels nous nous intéressons.

5.3 Approches basées sur les connaissances

L'émergence d'un grand nombre de ressources telles que les dictionnaires électroniques, les thésaurus et les lexiques informatiques a donné une autre orientation pour la résolution du problème de l'ambiguïté. Les approches basées sur les connaissances tentent d'extraire de manière automatique les informations dont on a besoin pour la tâche de désambiguïsation.

Plusieurs formes de ressources sont utilisées pour la désambiguïsation sémantique, citons les dictionnaires électroniques des langues classiques, comme par exemple Robert, ainsi que les thesaurus et les ontologies qui contient d'autres informations utiles comme les relations sémantiques entre les mots (Vasilescu & Langlais, 2003).

La désambiguïsation d'un mot ambigu dans un contexte donnée, est basée sur l'extraction des mots cooccurrents présents dans ce contexte. Parallèlement, pour chaque sens du mot polysémique, on extrait du dictionnaire la liste des mots présents dans chacune des définitions correspondantes. Enfin, la désambiguïsation consiste à choisir parmi les sens possibles, celui dont la définition possède le plus grand nombre de mot commun avec le contexte (Jacquet, 2005).

5.3.1 Approches basées sur les préférences sélectionnelle (restrictions)

Le principe général de ces approches est basé sur les critères syntaxiques pour déterminer le choix du sens, en prenant en considération *les préférences sélectionnelle* ou *les restrictions sélectionnelle* (Manning & Schütze, 1999). Il s'agit des types sémantiques qui caractérisent les arguments d'un verbe. Par exemple le cas du verbe *manger*, qui se construit avec des arguments relevant du domaine de la nourriture.

Les travaux de Sophie Piron (Piron, 2004) sont basés sur WordNet pour l'identification des préférences sélectionnelle. Elle a intégré à son corpus des annotations qui détermine l'information sélectionnelle des arguments du verbe *entendre*. Elle a basée sur la physiologie de la perception auditive qui caractérise la particularité de

l'information insérée, pour distinguer trois modes auditifs environnemental, phonétique et musical. Ces distinctions ont été appliquées à la catégorisation des compléments du verbe *entendre*.

Wagner et d'autre chercheurs (Wagner, Schmid, & Im Walde, 2009) regroupent les verbes ayant les préférences sélectionnelle et de sous-catégorisation similaires. Cette méthode montre 57.06% de précision. D'autres travaux identifient les préférences sémantiques (Mirella Lapata & Brew, 2004) et les marques de sous-catégorisation (Maria Lapata & Brew, 1999) des verbes apparaissant dans plusieurs classes de Levin (Levin, 1993).

5.3.2 Approches basées sur le chevauchement

Les ressources linguistiques ont donc constitué une source alternative d'informations exploitables pour la désambiguïsation lexicale. Lorsqu'ils sont disponibles sur support électronique, ces ressources sont alors directement exploitées par les méthodes automatiques de désambiguïsation. Les méthodes qui exploitent des informations d'une ressource externe sont caractérisées comme des méthodes dirigées par les connaissances (Apidianaki, 2008). La première méthode de ce type a été proposée, et qui a eu une grande influence sur les méthodes qui ont suivi, est celle de Lesk (Lesk, 1986). Le principe de base de cette méthode est de mesurer le chevauchement entre les différentes définitions, dans le dictionnaire, d'un mot ambigu et les définitions de ses voisins immédiats, dans une fenêtre de 10 mots. La désambiguïsation a donc lieu en choisissant, pour le mot ambigu et les mots qui l'entourent, les définitions qui se recoupent le plus. L'idée principale de la méthode de Lesk a été reprise et élaborée dans de nombreux travaux qui ont suivi.

Une autre méthode qui a été proposée par Wilks et autres (Y Wilks, Fass, Guo, MacDonald, & Plate) permet d'estimer la similarité entre entrées de sens et contextes, même s'ils n' y a pas de mots communs. L'objectif de cette méthode est l'expansion des entrées ou du contexte par la collection des données d'cooccurrence à partir des définitions de sens. Cette expansion se fait par l'inclusion de mots liés aux mots présents dans les contextes et les entrées de sens. Un vecteur de mots est alors construit pour l'entrée de chaque sens et un autre pour le contexte (la phrase où le mot apparaît), en ajoutant les vecteurs des mots liés à chacun des mots de l'entrée ou du contexte,

respectivement, et en excluant le mot ambigu. Le sens retenu est celui dont le vecteur est le plus similaire au vecteur du contexte.

Vasilescu et al. (Vasilescu & Langlais, 2003) analysent de façon détaillée les paramètres déterminant la performance des méthodes de désambiguïsation basées sur l'algorithme de Lesk et exploitant les informations de WordNet. Cette analyse s'effectue en comparant les variantes de l'algorithme, variantes relatives à la manière dont le contexte des mots ambigus est considéré, la manière dont les sens sont décrits, leur pondération ainsi que les mots du contexte pris en compte.

Cowie (Cowie et al., 1992) a enrichit les définitions dictionnairiques du 'LDOCE' avec des informations de domaine, en traitant le code de domaine attribué à un sens comme un mot faisant partie de sa définition. La sélection correcte des sens des mots repose sur l'idée que les sens lexicaux qui apparaissent dans la même phrase ont plus de mots et de codes de sujet en commun dans leurs définitions que ceux appartenant à des phrases.

Patwardan et al. (Patwardhan, Banerjee, & Pedersen, 2003) prolongent l'étude de Banerjee et Pedersen (Banerjee & Pedersen, 2002): en considérant le recouvrement des définitions comme une mesure de similarité sémantique, ils procèdent à la désambiguïsation en utilisant d'autres mesures de similarité sémantique sur la base des informations de WordNet.

Naskar et Bandyopadhyay (Banerjee & Pedersen, 2002), en revanche, utilisent l'algorithme de Lesk dans un système de désambiguïsation basé sur 'Extended WordNet' (Harabagiu, Miller, & Moldovan, 1999), où les définitions des synsets sont étiquetées par des informations sémantiques et morphosyntaxiques.

D'autres études ont donc cherché à l'améliorer en utilisant d'autres indices (Walker, 1986). Véronis et Ide (Veronis & Ide, 1990) ont prolongé la méthode de Lesk en créant un réseau de neurones à partir des définitions du Collins English Dictionary. Wilks et al. (Yorik Wilks, 1993) ont beaucoup réfléchi pour utiliser de façon optimale les ressources électroniques afin d'identifier les sens des mots polysémiques.

5.3.3 Approches basées sur l'algorithme de densité conceptuelle

La Densité Conceptuelle (DC) est une mesure de corrélation entre le sens d'un mot et son contexte. Elle a été présentée dans le domaine de traitement automatique des langages naturelles par Agirre et Rigau (Agirre & Rigau, 1996) puis reformulée par Rosso et al (Rosso, Masulli, Buscaldi, Pla, & Molina, 2003). Cette dernière est ensuite adaptée à la désambiguïsation des toponymes[10] par Buscaldi et Rosso (Buscaldi & Rosso, 2008).

La formule de calcul de DC est :

$$CD(c;m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0;2}}}{descendant_c}$$

Tel que *m* est le nombre de nœuds (synsets) pertinentes dans la sous-hiérarchie composée des lieux du contexte. Dans la méthode de Buscaldi et Rosso (Buscaldi & Rosso, 2008), la densité conceptuelle est calculée pour chaque référent candidat du toponyme ambigu. Ensuite, le référent qui maximise cette valeur (c.-à-d. la densité conceptuelle) est celui qui sera attribué au toponyme ambigu.

L'explication détaillée de cette méthode est donnée par (Buscaldi & Rosso, 2008) mais il est suffisant de dire que la densité conceptuelle est une quantification d'une certaine proximité entre les toponymes du contexte. C'est-à-dire que cette heuristique résout les toponymes ambigus par les référents les plus proches les uns aux autres.

5.3.4 Approches basées sur l'algorithme de marche aléatoire

Depuis quelques années, un certain nombre d'auteurs ont tenté à décrire le sens lexical en s'appuyant sur un objet mathématique, les graphes, et en particulier, les graphes issus de dictionnaires (Newman, 2003; Véronis, 2004a). Cependant, ces expériences ont privilégié des dictionnaires de synonymes ou une analyse en termes de synonymie ou de distance sémantique. Loiseau et d'autre chercheurs (Loiseau, Gréa, & Magué, 2011), ont proposé une analyse des graphes de dictionnaire de langue monolingue.

[10] **La toponymie** (du grec tópos, τόπος, lieu et ónoma, ὄνομα, nom) est une branche de l'onomastique qui étudie les toponymes, c'est-à-dire les noms propres désignant un lieu.

Ceux-ci, à la différence des dictionnaires de synonymes, ne codent pas une seule relation sémantique, à savoir la synonymie, mais des relations sémantiques beaucoup plus hétérogènes. Cette complexité explique la difficulté à construire des graphes sémantiquement interprétables à partir de ces dictionnaires. Cependant, ils ont montré que des graphes construits sur les dictionnaires de langue manifestent des propriétés sémantiques remarquables, qui relèvent d'un cadre sémantique onomasiologique. Ils ont insisté sur la prise en compte des propriétés textuelles des dictionnaires et ont proposé une analyse contrastive de plusieurs graphes.

5.4 Approches basées sur corpus

Les premières recherches pour la tâche de désambiguïsation sémantique ont commencé en 1949 dans le domaine de la traduction automatique. Dans son Mémorandum, Weaver (Voorhees, 1993; Weaver, 1949) introduit le besoin de désambiguïsation lexicale dans la traduction par ordinateur, mais les corpus sont utilisés très tôt pour des travaux concernant le sens des mots (Eaton, 1940; Palmer, 1933; Thorndike, 1948; Zipf, 1945). Les corpus étant de grands ensembles d'exemples, leurs utilisations vont de pair avec les méthodes empiriques et statistiques. Mais par la suite, les traitements statistiques sur corpus sont dénigrés au profit des règles linguistiques formelles. Pourtant, pendant cette période, plusieurs corpus de taille importante sont constitués comme le corpus Brown (Ku & Francis, 1967), le Trésor de la Langue Française (Imbs & QUEMADA, 1971), le corpus Lancaster-Oslo-Bergen (Johansson, 1980), etc. Dans le domaine du traitement automatique des langues, en 1966, le rapport ALPAC (Pierce & Carroll, 1966) préconise une utilisation intensive des corpus pour créer des grammaires et des lexiques à grande couverture. L'utilisation des corpus continue toutefois à se faire rare. Dans ce contexte, l'utilisation des corpus et des méthodes empiriques faite par Weiss (Weiss, 1973) et par Kelly et Stone (Kelly & Stone, 1975), dans le domaine de l'extraction automatique de connaissances pour la désambiguïsation lexicale, semble innovante. Le principe de base de ces approches consiste à l'acquisition de connaissances à partir de ces corpus en étudiant les différents exemples et on utilisant des méthodes statistiques.

Kelly et Stone (Kelly & Stone, 1975) travaillent sur un corpus d'un demi-million de mots. Avec l'aide de leurs étudiants, ils construisent manuellement un modèle de désambiguïsation pour 1815 mots du corpus possédant une fréquence de plus de 20

occurrences. Les modèles sont construits en observant un contexte de quatre mots à gauche et à droite pour chacune des occurrences des 1815 mots dans le corpus. Chacun de ces modèles contient un ensemble de règles basées sur la morphologie du mot, sur la présence de mots saillants dans le contexte et sur la classe sémantique (déterminée manuellement) des mots du contexte (Audibert, 2003b).

Deux types de corpus sont utilisés, les corpus étiquetés et les corpus non étiquetés.

5.4.1 Approches basées sur corpus étiquetés

Les corpus étiquetés sont des outils très important pour plusieurs applications de traitement automatique des langages naturels. Le cas le plus utilisé est l'étiquetage morphosyntaxique qui permet la segmentation de texte en mots en basant sur les critères grammaticales plus que le découpage par repérage de caractères délimiteurs et attache à chaque segment sa catégorie morphosyntaxique et ses informations linguistiques (Pincemin, 2004). Un autre type d'étiquetage est celui de l'étiquetage sémantique qui consiste à associer à chaque mot d'un texte une étiquette correspondant au « sens » qu'a ce mot dans le contexte particulier où il apparaît (Véronis, 2004b).

Weiss (Weiss, 1973) est l'un des chercheurs qui ont travaillé sur les corpus étiqueté pour la désambiguïsation sémantique. Son travail est basé sur deux types de règles. Les règles qui cherchent l'existence de mots saillants dans un contexte de taille donnée et les autres règles cherchant la présence d'un mot saillant à une position donnée par rapport au mot ambigu.

Black (Black, 1988) a réalisé un étiquetage manuel sur cinq mots, il a étiqueté 2000 occurrences pour chaque mot dont 75% sont utilisés pour réaliser l'apprentissage et 25% pour tester les méthodes de désambiguïsation. Il a fait une comparaison entre trois méthodes de désambiguïsation. La première méthode est très proche de celle de Weiss. La deuxième contient des catégories thématiques manuellement déterminées. La troisième inclut les catégories de sujet (par exemple *sport, scientific*) du dictionnaire LDOCE.

Yarowsky (Yarowsky, 2000) a utilisé l'apprentissage supervisé sur un corpus pour effectuer la désambiguïsation sémantique d'un nombre limité des mots. Ce travail est basé sur deux étapes. La première est une étape d'apprentissage sur une partie de corpus tandis que la seconde est consacrée à la désambiguïsation sur le reste de corpus.

Le bon sens de chaque occurrence des mots ambigus est déterminé quand l'apprentissage est effectué.

Dans les deux dernières décennies, la communauté de traitement automatique de la langue a connu une progression importante dans les approches basées sur l'apprentissage automatique pour la classification automatique des sens des mots (Agirre & Martinez, 2000; Bruce & Wiebe, 1994a, 1994b; Claude de Loupy et al., 1998; C de Loupy, El-Bèze, & Marteau, 1998; El-Beze, Michelon, & Pernaud, 1999; Escudero, Màrquez, & Rigau, 2000a, 2000b; Mooney, 1996; Ng, 1997; Ng & Lee, 1996; Pedersen, 2001, 2002; Pedersen & Bruce, 1997; Pedersen, Bruce, & Wiebe, 1997; Yarowsky, 2000; Yarowsky, Cucerzan, Florian, Schafer, & Wicentowski, 2001).

Les méthodes d'apprentissage supervisé permettent de produire des règles à partir d'un corpus généralisant ce qui a pu être appris aux données inconnues. La première étape des approches supervisées est l'apprentissage dans laquelle les connaissances nécessaires pour la désambiguïsation sémantique sont extraites à partir d'un corpus étiqueté manuellement selon des bases de connaissances tel que WordNet. La seconde étape est la désambiguïsation sémantique qui consiste à associer à chaque mot des nouveaux textes une unité lexicale adéquate en se basant sur les connaissances obtenues durant la première étape.

Ce type des approches est considéré comme une classification supervisée qui nécessite un corpus étiqueté manuellement pour l'apprentissage ainsi que pour l'évaluation de leur performance.

Les corpus étiquetés sont des ressources précieuses riches en connaissances concernant les langues. Parmi les corpus qui ont été réalisés pour la langue anglaise celui de Semcor (Fellbaum, 1998) qui contient 250000 occurrences étiquetés à partir de WordNet. Les textes de ce corpus sont extraits du corpus Brown et de la nouvelle « The Red Badge of courage ». un autre corpus qui est le corpus de DSO (Ng & Lee, 1996) contenant 192 800 mots des 121 noms et 70 verbes les plus répondus et les plus ambigus de la langue anglaise. L'étiquetage de ce corpus est basé sur les lexies de WordNet. Ce corpus inclut des extraits du corpus Brown. On cite également certain corpus qui sont considérés comme des références de campagnes d'évaluation Senseval-1 (Kilgarriff & Rosenzweig, 2000), Senseval-2 et Senseval-3. La pauvreté des corpus étiqueté pose un grand problème pour l'extraction des connaissances dans

le but de la désambiguïsation lexicale. Pour résoudre ce problème d'une part les études sont menées sur un nombre limité de mots et d'autre part les chercheurs ont travaillé pour réaliser ce type de ressource mais ses travaux restent limités pour quelques langues comme l'anglais, le français,…etc.

Une estimation a été proposée par Mihalcea et Chklovski (Mihalcea & Chklovski, 2004) montre que l'étiquetage de 20000 mots nécessite 80 années-humains (c'est-à-dire l'équivalent de travail de 80 humain pendant un an).

5.4.2 Approches basées sur corpus non étiquetés

Plusieurs études sont menées pour résoudre le problème de manque des corpus lexicalement étiqueté. Ces méthodes sont basées sur des corpus brut non étiqueté pour induire de sens des mots ambigus(Pereira, Tishby, & Lee, 1993).

Une autre proposition est basée sur l'espace vectoriel (Schütze, 1992, 1998) qui a été utilisé dans le domaine de la recherche d'information (Salton, Wong, & Yang, 1975) dans laquelle les mots sont représentés par des vecteurs dans un espace, les vecteurs les plus similaires sont regroupés en paquets dont chaque paquet représente un sens.

Le problème majeur de ces méthodes est le manque des informations concernant la correspondance des sens (Yorick Wilks & Stevenson, 2000). Pour résoudre ce problème, Pedersen et Bruce (Pedersen & Bruce, 1997) ont proposé une méthode dans laquelle ils ont fait la correspondance entre des groupes de sens avec un sens d'un lexique donné, mais les résultats obtenus sont insuffisant.

5.5 Approches hybrides

Les approches hybrides sont basées sur plusieurs sources de connaissances. Premièrement l'étiqueteur morphosyntaxique (tagger) son intérêt est d'éliminer les autres sens qui ne correspondent pas à celui de la classe grammaticale donner par l'étiqueteur. Deuxièmement, des sources de connaissances extraites du LDOCE, sont utilisées dans ces approches. D'abord l'utilisation des définitions du dictionnaire avec les mots qu'elles contiennent. Ensuite l'utilisation des indications de domaine et de registre de langue qui se trouve dans la majorité des définitions. Enfin l'utilisation des

restrictions de sélection qui détermine à certaines catégories sémantiques (science, sport, politique, économie, etc.).

On cite dans cette catégorie des approches celle de Wilks et Stevenson (Yorick Wilks & Stevenson, 1998) qui ont combiné un dictionnaire et l'apprentissage supervisé sur un corpus pour la désambiguïsation sémantique. Ils ont utilisé le dictionnaire LDOCE (Longman Dictionary of Contemporary English).

5.6 Approches basées sur les méthodes de l'apprentissage

L'idée principale de ces approches consiste à désambiguïser un mot ambigu à partir de calculs statistiques en basant sur les instances de ce mot dans le corpus et les différents contextes dans lequel il est employé. Ensuite le but est de créer des corrélations entre les sens de ce mot et les différents emplois du mot polysémique. Ces approches exploitent les différentes relations entre les mots comme les relations de cooccurrences, les relations syntaxiques, les relations thématiques. Plusieurs techniques sont utilisées dans cette catégorie.

Le modèle des réseaux bayésien représente d'une façon graphique les relations d'influence entre les différentes variables d'un modèle. La relation d'influence entre deux variables aléatoires est représentée par un arc, et l'absence de cette dernière marque l'indépendance entre les variables. Le sens de l'arc détermine le sens de l'influence que nous souhaitons modéliser. Un réseau bayésien permet de décrire qualitativement et quantitativement les relations d'influence entre les VA données par un graphe dirigé au moyen d'un ensemble de distributions de probabilités conditionnelles.

Une liste de décision (Rivest, 1987) est une description des objets sous forme d'une série ordonnée de paires "attribut-valeur". Les listes de décisions peuvent être interprétées comme une série ordonnée de règles du type "si ... alors". La classification de cette liste est effectuée en testant chaque règle sur le corpus d'apprentissage. Lors de la désambiguïsation, le sens donné à l'occurrence du mot polysémique sera celui associé à la première règle de la liste s'appliquant à cet exemple. Lorsqu'aucune règle ne s'applique, une valeur par défaut est retournée.

Yarowsky (Yarowsky, 1995, 2000) insiste sur le fait que ses listes de décision sont hiérarchiques, c'est-à-dire qu'une règle appartenant à une première liste de décision

peut pointer sur un sens comme dans toute liste de décision classique, mais elle peut aussi pointer sur une autre liste de décision où la liste nommée promise peut pointer sur la liste promise L_N qui elle-même peut pointer sur la liste promise L_M. L'intérêt d'une telle structure est de pouvoir notamment construire des règles conditionnelles de type « si x et y alors z ».

5.7 Bref aperçu des approches de désambiguïsation de la langue arabe

Plusieurs approches qui s'intéressent au problème de la désambiguïsation lexicale de la langue arabe ont été présentés au cours des dernières années. Citant le travail de Mona Diab, qui a utilisé une approche d'apprentissage supervisé. Elle exploite dans cette approche un corpus parallèle arabe-anglais pour annoter un texte arabe qui présente une précision absolue de 56,9% (Diab, 2004).

Menai et Wodjan ont proposé une approche pour résoudre le problème de désambiguïsation lexicale basé sur les algorithmes génétiques. Les sens de chaque mot ont été récupérés à partir de l'Arabe WordNet (AWN). GA AWSD est utilisé pour trouver la correspondance la plus appropriée des mots aux sens extraits d'AWN dans le contexte. Les auteurs ont montré que GA est plus performant que d'autres méthodes. Ils ont obtenu une précision de 78,9% (Menai, 2014a, 2014b; Menai & Alsaeedan, 2012).

Le système présenté par Elmougy et al. (Elmougy, Taher, & Noaman, 2008) utilise l'algorithme de routage avec un classifieur de Bayes naïf pour résoudre l'ambiguïté des mots non-diacritiques dans la langue arabe. Il a atteint un taux de précision de l'ordre de 76,6%.

Pour lever l'ambiguïté des mots arabes, Zouaghi et al. ont testé l'algorithme de Bayes Naïf, les listes de décision et l'algorithme à base des exemples sur des échantillons étiquetés et une grande quantité de données dans le corpus utilisé. Pour ces méthodes, les résultats obtenus atteignent des taux de précision de 48,23%, 43,86% et 52,02%, respectivement (Zouaghi, Merhbene, & Zrigui, 2011).

Une autre méthode pour WSD a été proposée par Zouaghi et al. Leur approche a évalué les variantes de l'algorithme Lesk en utilisant le AWN pour effectuer WSD en arabe. La version modifiée de l'algorithme Lesk atteint une précision de 67%.

La méthode présentée par Nameh et al. est une méthode d'apprentissage supervisé pour WSD est basé sur le produit scalaire des vecteurs. Le système extrait deux ensembles des caractéristiques : l'ensemble des mots qui sont apparus fréquemment dans le texte et l'ensemble des mots voisins au mot ambigu. Cette approche permet d'obtenir une précision de 77,1% (Nameh, Fakhrahmad, & Jahromi, 2011).

Shah et al. ont proposé une approche pour la prédiction précise de texte arabe composé de deux lemmes et analyses morphologiques. La précision de leur système a été signalée à 90% (Shah et al., 2010).

Malgré les études importantes dans ce domaine, les résultats obtenus restent encore insuffisants.

5.8 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art sur la désambiguïsation lexicale, qui nous intéresse particulièrement dans notre travail de recherche.

Nous avons dressé l'état de l'art des méthodes existantes : les méthodes basées sur les connaissances comme celles qui sont basées sur les préférences sélectionnelle, les approches basées sur le chevauchement entre les différentes définitions ainsi que les approches sur un corpus étiqueté ou non étiqueté et celles qui sont basées sur l'apprentissage automatique supervisé ou non supervisé. Enfin, nous présentons les travaux de recherche qui focalisent sur la langue arabe.

Les travaux connexes présentés dans ce chapitre, qui utilisent des modèles basés sur les connaissances, ont montré que les performances des systèmes de désambiguïsation s'améliorent considérablement lors de la combinaison de deux méthodes. Le chapitre suivant sera dédié à expliquer les étapes de l'approche proposée.

Chapitre 6 :

ACARWSD : Notre approche proposée

Sommaire

6.1 Introduction

Dans ce chapitre, nous présentons notre approche qui est basée sur la notion d'algorithme local et d'algorithme global pour la désambiguïsation lexicale des textes arabe. Un algorithme local permet de calculer la proximité sémantique entre deux objets lexicaux. L'algorithme global permet de propager ces mesures locales à un niveau supérieur (D. S. J. G. A. Tchechmedjiev, 2013). Cette approche a besoins d'une base lexicale et comme la langue arabe ne dispose pas beaucoup de ces ressources. On a exploité d'une part la ressource Web Wiktionnaire qui est devenue une source intéressante pour l'extraction d'information et d'autre part on a utilisé l'ontologie WordNet pour construire notre base lexicale pour cette langue. Nous avons utilisé l'algorithme LESK comme un algorithme local et l'algorithme à colonies de fourmis comme un algorithme global. En les évaluant sur un corpus de référence, nous montrons que l'efficacité des algorithmes à colonies de fourmis rend possible l'amélioration automatique du paramétrage et en retour leur amélioration qualitative.

6.2 Motivations

La plupart des mots ont plusieurs significations. Pourtant, quand une personne entend une phrase contenant un mot ambigu, elle la comprend généralement (sans même percevoir l'ambiguïté) sur la base d'une signification particulière de ce mot.

Tout se passe comme si, à un moment donné du processus humain de compréhension de la langue, la bonne signification du mot était automatiquement sélectionnée en fonction du contexte parmi toutes les significations possibles de ce mot. Nous pouvons désigner par désambiguïsation lexicale cette tâche, consistant à choisir la bonne signification d'un mot polysémique dans un contexte donné.

La désambiguïsation automatique est un enjeu important dans la plupart des applications de traitement automatique des langues TALN. On peut citer comme exemple les applications de recherche d'information (Sanderson, 1994; Stokoe, Oakes, & Tait, 2003; Uzuner, Katz, & Yuret, 1999; Zhong & Ng, 2012), de traduction automatique (Carpuat & Wu, 2007; Chan, Ng, & Chiang, 2007; Vickrey, Biewald,

Teyssier, & Koller, 2005), de reconnaissance de la parole et des caractères (Sproat, Hirschberg, & Yarowsky, 1992; Yarowsky, 1992, 1997), (cf. Ide & Véronis, 1998).

6.3 Partie 1 : Construction de la base lexicale

Les ressources linguistiques électroniques jouent un rôle très important en traitement automatique du langage naturel. Elles sont utilisées dans plusieurs applications linguistiques notamment la traduction automatique, l'indexation des textes, le résumé automatique…etc. L'objectif de cette partie est de créer une base lexicale pour la langue arabe qui ne dispose pas beaucoup de ces ressources. Nous exploitons les ressources Web comme Wiktionnaire et le WordNet qui sont devenues des sources intéressantes pour l'extraction d'information. Dans ce travail, nous cherchons à extraire automatiquement des relations sémantiques notamment les synonymes et les antonymes à partir du Wiktionnaire arabe ensuite nous utilisons WordNet pour construire une base lexicale arabe (Abdelali; Bakhouche & Yamina Tlili-Guiassa, 2013).

6.3.1 Extraction des relations sémantiques à partir du Wiktionnaire arabe

Les dictionnaires sont des sources intéressantes pour l'extraction automatique des différentes connaissances lexico-sémantiques, dans ce sens plusieurs travaux ont utilisé ces ressources pour extraire des relations sémantiques, dont le but était de créer des grands réseaux sémantiques (Abouenour, Bouzoubaa, & Rosso, 2008; Atserias, Climent, Farreres, Rigau, & Rodriguez, 2000). D'autres études utilisent le dictionnaire pour la désambigüisation du sens du mot, ainsi que pour l'analyse des textes (Apidianaki, 2008). La différence entre les ressources en quantité et en nature pose un problème d'extraction d'information, et pour résoudre ce problème des chercheurs ont combiné plusieurs dictionnaires (Ide & Véronis, 1995).

Quelques études récentes sur l'acquisition d'ontologies et sur l'extraction de relations sémantiques à partir des ressources comme WordNet qui est une base de données lexicales développée au laboratoire des sciences cognitives de l'université de Princeton (Nichols, Bond, & Flickinger, 2005), cette ressource a été utilisée par les auteurs (Abouenour et al., 2008) dans le cadre de question/réponse pour construire une

ontologie. Il existe d'autres sources, qui ont été utilisées dans la création ou l'enrichissement des bases de données lexicales telles que des corpus ou des ressources collaboratives comme Wikipédia et Wiktionnaire. Ces deux dernières fournissent des dictionnaires électroniques au contenu gratuit, avec des définitions, des exemples et des informations sur la partie du discours (POS), des traductions, des prononciations et l'étymologie, ainsi que des informations sur les relations sémantiques (synonymes, antonymes...) (Meyer & Gurevych, 2012), qui ont été construits manuellement par des gens non professionnels sur le Web, aujourd'hui le Wiktionnaire contient environ 5 millions d'entrées dans 170 éditions linguistiques (Meyer & Gurevych, 2012).

Les Wiktionnaires ont été exploité pour l'extraction des relations sémantiques et les comparer avec celle extraites des autres ressources (Pérez, Oliveira, & Gomes, 2011). Par ailleurs, le Wiktionnaire a été utilisé pour l'enrichissement des ressources lexicales existantes ainsi que dans la création automatique des nouvelles ressources lexicales (Weale, Brew, & Fosler-Lussier, 2009; Zesch, Müller, & Gurevych, 2008).

Dans cette section, nous décrivons la méthode d'extraction des synonymes et des antonymes à partir de Wiktionnaire arabe pour construire une base lexicale. Notre approche est composée de 3 parties, la première est une phase de préparation (prétraitement et extraire les définitions), la deuxième phase est l'analyse des vocabulaires de ces définitions ainsi l'extraction des relations sémantiques. La dernière phase consiste à créer une base lexicale. La Figure 6.1 montre l'architecture de notre approche.

6.3.1.1 ***Prétraitement et Extraction des définitions***

L'extraction des informations lexico-sémantiques disséminées dans les bases de connaissances collaboratives nécessite des outils d'accès automatiques à son fichier XML (Figure 6.2). Ces outils sont disponibles pour les langues comme l'anglais,

l'allemand (Zesch et al., 2008) et le portugais (Pérez et al., 2011), dans cet objectif nous avons développé un outil pour analyser la structure du fichier XML pour la langue arabe. Qui peut supprimer les lettres latines, les chiffres, les caractères spéciaux,…

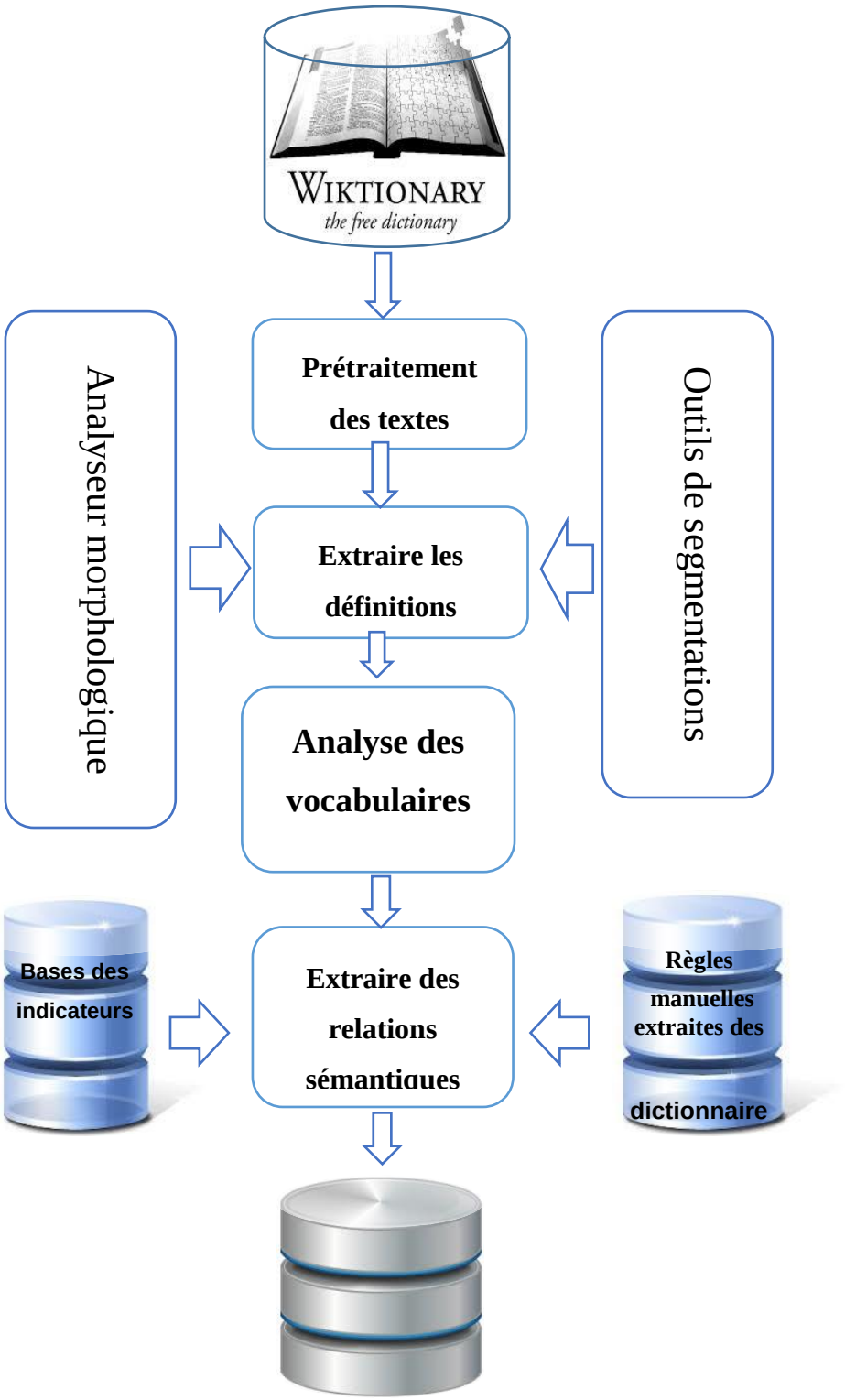


Figure 6-1 Extraction des relations sémantiques à partir du Wiktionnaire Arabe

<page>
<title>حَاسُوْب </title>
.....
<text xml:space="preserve">wikipedia
= عَرَبِيَّة=
==المعاني == # [[حَوَاسِيب]] اسم مذكر يُجمع جمع تكسير على الحَاسُوْب
آلة]] الكترونية [[تقوم بعمليات]] حِسَابِيَّة [[سريعة بواسطتها يمكن للآلة القيام بمهمات مختلفة
على]] البَيَانَات والمَعْلُومَات [[.اِسْتِخْدَامٌ حَدِيْثٌ
=== المرادفات===
'"1": [[حَاسِبَة]]
'"2": [[آلي حَاسِب]]
.....
</page>

Figure 6-2 Les informations contenues dans le fichier XML pour l'entrée حَاسُوْب

Cet outil peut exporter toutes les définitions et les transformer en format convivial où chaque ligne contient l'entrée, sa partie de discours et sa définition. Comme le montre l'exemple suivant (Tableau 6-1) :

Tableau 6-1 Définitions obtenus à partir de l'entrée حَاسُوْب

مذكر يُجمع جمع تكسير على حَوَاسِيب	**اسم**	**الحَاسُوْب**
آلة الكترونية تقوم بعمليات حِسَابِيَّة سريعة بواسطتها يمكن للآلة القيام بمهمات مختلفة على البَيَانَات والمَعْلُومَات اِسْتِخْدَامٌ حَدِيْثٌ	**اسم**	**الحَاسُوْب**
حَاسِبَة	**صفة**	**الحَاسُوْب**
حَاسِب آلي	**اسم**	**الحَاسُوْب**

6.3.1.2 *Analyse des vocabulaires*

Pour extraire les relations sémantiques à partir des définitions. Nous avons utilisé l'outil de segmentation AraSeg (Mouelhi, 2008), pour découper les textes des définitions en unités lexicale : paragraphe, phrase, syntagme, mot graphique, mot-forme, morphème,…etc. ensuite l'analyseur morphosyntaxique AraMorph (Buckwalter, 2002) utilise les différentes ressources de connaissance de la langue

arabe pour extraire les lemmes et les classes grammaticales des mots constituant les définitions, le résultat de ces outils est présenté dans le tableau suivant.

Tableau 6-2 Exemple de résultat de l'analyseur morphosyntaxique

Mot	Lemme	Catégorie grammaticale
آلة	**آلة**	**اسم، مبتدأ**
الكترونية	**إلكترون**	**اسم، خبر**
تقوم	**قام**	**فعل**
عمليات	**عمل**	**اسم، مجرور**

6.3.1.3 ***Extraction des relations sémantiques***

L'extraction des relations sémantiques est basée sur plusieurs informations morphosyntaxiques, marqueurs et des règles inspirées du dictionnaire. La phase actuelle est très importante car elle affecte les résultats du système. Elle a pour but de déterminer la relation sémantique entre les entrées de Wiktionnaire et les mots constituant leurs définitions. Ce système fait l'extraction des relations à base des indicateurs et des règles conçues manuellement. Les indicateurs et ces règles ont été inspirés à partir des dictionnaires (المحيط ,لسان العرب). Les indicateurs sont présentés sous forme d'une base contenant par exemple les pronoms personnels comme (هي، هو،...) et autres indicateurs comme (ضده، مرادفه ،معناه...). Cependant les règles représentent les structures des phrases.

Nous avons obtenus 8321 relations sémantiques induites à partir de 25037 définitions. Les relations sont représentées sous forme de triplets (A, R, B) dont A est un mot dans la définition, B est l'entrée de Wiktionnaire et R est le nom de la relation.

Tableau 6-3 Exemple de représentation des relations sémantiques

L'entrée de Wiktionnaire «Représenté par le caractère B »	Relation Représenté par le caractère « R»	Mot dans la définition « Représenté par le caractère A»
الحاسوب	Synonyme	**آلة حاسبة**
العَيْن	Synonyme	**مُعَايَنَة**
العَيْن	Antonyme	**نَهْر**
……	**……**	**……..**

6.3.1.4 *Création de la base lexicale*

La dernière étape consiste à mettre en relation les mots et les connaissances sémantiques trouvées dans l'étape précédente afin de construire la structure générale des données. La base lexicale est représentée par le modèle entité-association telle que les entités représentent les mots et les associations représentent les relations sémantiques entre eux. On travaille toujours sur la base et sa représentation et pour des raisons d'analyse on représente la base sous la forme suivante (Figure 6.3).

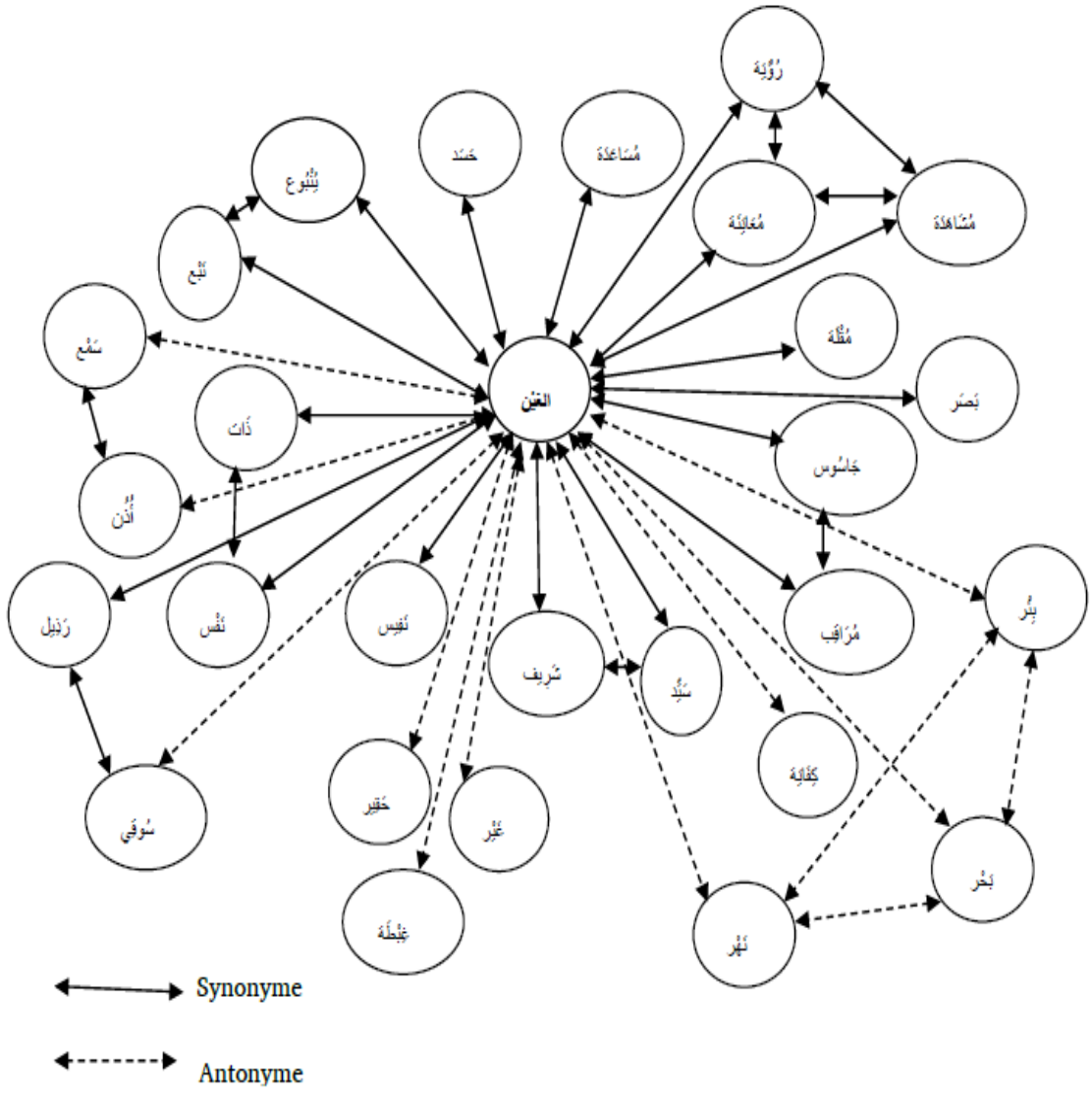


Figure 6-3 Exemple d'une base lexicale

6.3.2 Génération de dictionnaire à partir de Wordnet

Plusieurs ressources linguistiques ont été constituées à partir de WordNet (voir la section 4.3.2.3). Des programmes issus du monde de l'Intelligence Artificielle ont également établi des passerelles avec WordNet.

Nous abordons cette section par la définition de dictionnaire multilingue ainsi la matrice lexicale. Nous présentons ensuite les étapes de création d'une base lexicale à partir de WordNet (intersection des synonymes, correspondance des mots, correspondance sémantique).

6.3.2.1 *Qu'est-ce qu'un dictionnaire multilingue*

Un dictionnaire multilingue est un dictionnaire dans lequel des expressions dans une langue (dite langue source ou départ) sont traduites dans une autre langue (dite langue cible ou d'arrivée). Dans notre approche un dictionnaire est une base lexicale contenant l'ensemble des mots de la langue arabe avec ses sens expliqués dans la langue anglaise (voir la figure 6.5).

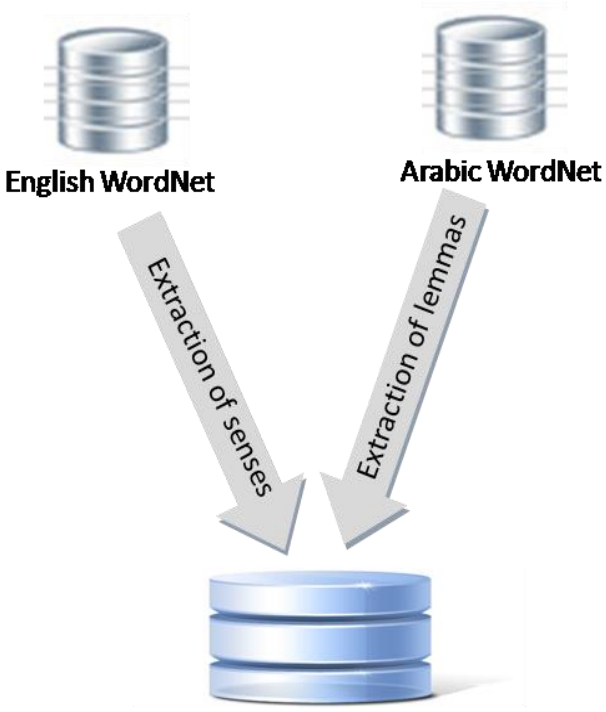


Figure 6-4 Génération de base lexicale arabe

Mots	Translitération	Sens
سحور	suhu : r	Light meal taken before starting a new day of Ramadan
إفطار	ifta : r	Meal at the end of daily fasting during Ramadan
مدفع إفطار	Midfa' ifta : r	Gun announcing the end of daily fasting during Ramadan
عمرة	Umra	Visit to the holy shrines in Mecca and Medina out of the time of pilgrimage

Figure 6-5 Mots arabes avec ses sens en anglais

6.3.2.2 ***La matrice lexicale multilingue***

Le réseau multilingue WordNet a été construit en langue anglaise ensuite il a été réutilisée facilement par des langues latines (français, italien,…) car il y a beaucoup de similarités entre elles. Mais son réutilisation reste difficile pour les autres langues comme (Arabe, chinois, japonais,…).

La création de la matrice lexicale multilingue à partir du réseau WordNet est inspirée de travail de Magnini et al (Magnini, Strapparava, Ciravegna, & Pianta, 1994). Cette matrice est considérée comme une matrice lexicale bidimensionnelle implémentée dans Wordnet. En ajoutant une troisième dimension à la matrice qui représente une autre langue (la langue arabe dans notre cas) La figure 6.6 montre les trois dimensions de la matrice :

- Les mots sont représentés par $\mathbf{W_i}$
- Les sens sont représentés par $\mathbf{M_i}$
- Les langues sont représentées par $\mathbf{L_k}$

En outre les relations principales (lexicales et sémantiques) sont montrées dans la section précédente (voir figure 6.3).

Pour développer la matrice multilingue, il est nécessaire de correspondre le lexique de la langue et le sens M_i, en criant un ensemble de synset pour l'arabe.

Le résultat est une redéfinition des relations lexicales notamment les relations sémantiques. Dans ce cas la dimension des sens est considérée constante par rapport aux langues et des mots de chaque langue.

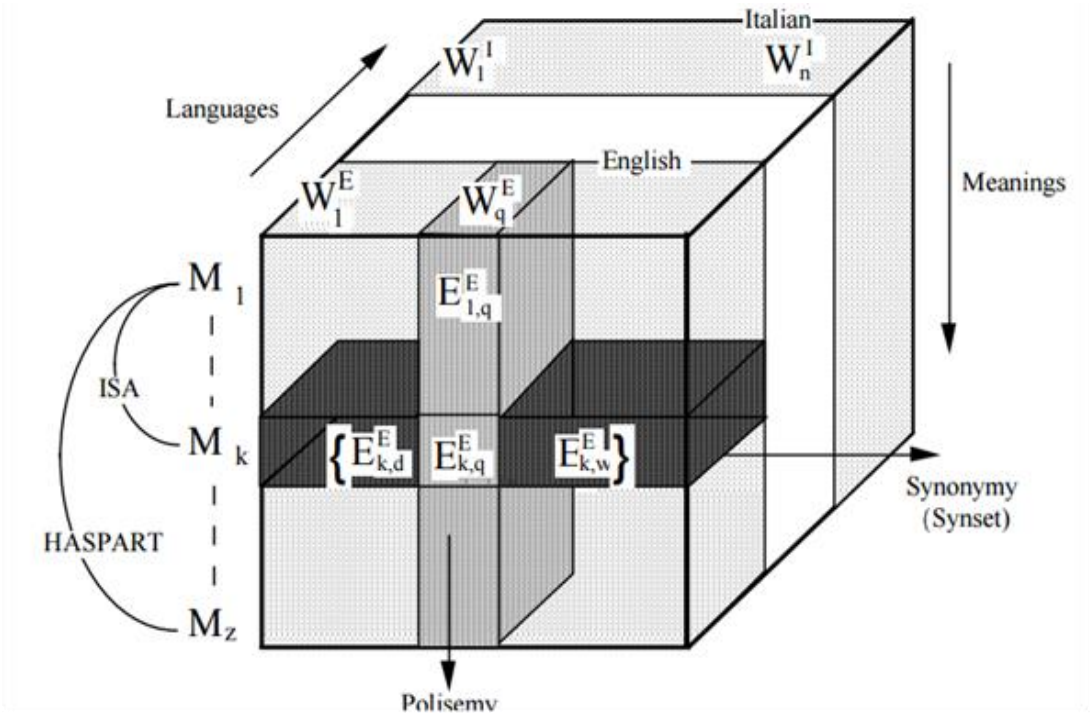


Figure 6-6 La matrice lexicale multilingue (Magnini et al., 1994)

6.3.2.3 *Création automatique de la base de connaissances lexicales*

La tâche principale pour créer la base de connaissance lexicale basée sur Wordnet est de trouver une correspondance correcte entre les mots arabes et les synsets définis en anglais.

La création automatique de base de connaissance lexicale pose deux problèmes principaux :

- L'extraction des informations des mots arabes à partir des sources disponibles ;

- La correspondance entre les mots arabes et ceux de l'anglais.

Pour résoudre le premier problème, on a utilisé WordNet Arabe et un dictionnaire de la langue arabe en format électronique comme (المحيط ,لسان العرب).

Pour le deuxième problème, on a réalisé trois niveaux :

- Le premier niveau ne fonctionne que sur synset de WordNet et dans ce cas le programme cherche à trouver l'équivalence entre les mots arabe et anglais ;
- Tandis que le deuxième niveau est consacré à la comparaison des définitions extraites d'un dictionnaire et les gloses de Wordnet en utilisant des méthodes statistiques ;
- Les techniques de traitement automatiques de la langue sont utilisées dans le troisième niveau pour faire la comparaison.

Les sections suivantes expliquent la procédure et l'algorithme de chaque niveau.

Intersection des synonymes

A ce niveau l'intersection des synonymes des mots des deux langues est considérable. L'idée est d'exploiter les sources d'information suivantes :

a. Synset anglais est ses relations avec taxonomie Wordnet ;
b. Un dictionnaire bilingue pour les deux langues ;
c. Un dictionnaire pour les synonymes de la langue arabe.

L'algorithme est conçu pour obtenir le synset des mots arabes avec des sens comparable à celles de l'anglais comme dans l'exemple suivant :

Synset WordNet = {registration, enrollment}

La traduction en arabe est prise de deux mots de synset :

Registration : **1.** تسجيل، توثيق، تدوين،

Enrollment : **1.** تسجيل،

2. إدراج، إعداد

Avec une simple intersection des ensembles des mots arabes, nous pouvons déduire que Registration [1] et Enrollment [1] ont la même signification ; nous constatons que la Synset arabes = {تسجيل} correspond à celle en anglais {Registration, Enrollment}. Dans les cas plus complexes, il peut être nécessaire d'utiliser le dictionnaire des

synonymes arabe pour définir correctement la correspondance entre les différents sens d'un seul mot.

Correspondance des mots

A ce niveau la correspondance faite par la comparaison entre les gloses Wordnet et les définitions de dictionnaire arabe. L'algorithme de similarité entre les définitions est basé sur une méthode statistique de présence des mots communs dans les deux définitions. Cependant ; il existe une complication supplémentaire : les définitions sont données en deux langues différentes, en utilisant correctement les données de WordNet (synset, mots, concept,…).

Correspondance sémantique

Avec cette méthode, il est nécessaire d'extraire les informations sémantiques pour les définitions de dictionnaire et les gloses de Wordnet. La tâche pour les gloses est simple car les vocabulaires utilisés sont limitées (environs 70000 racines) et la construction syntaxique est simple. Un autre niveau d'analyse consiste à faire une analyse syntaxique superficielle des définitions pour une représentation sémantique simple.

Pour ce type d'analyse, on utilise seulement les données de catégorie syntaxique des mots et le type de sous-catégorie des verbes disponible dans le dictionnaire papier. L'algorithme doit établir un degré de similarité entre les deux formes, qui contient les données dérivées respectivement de l'arabe et de l'anglais.

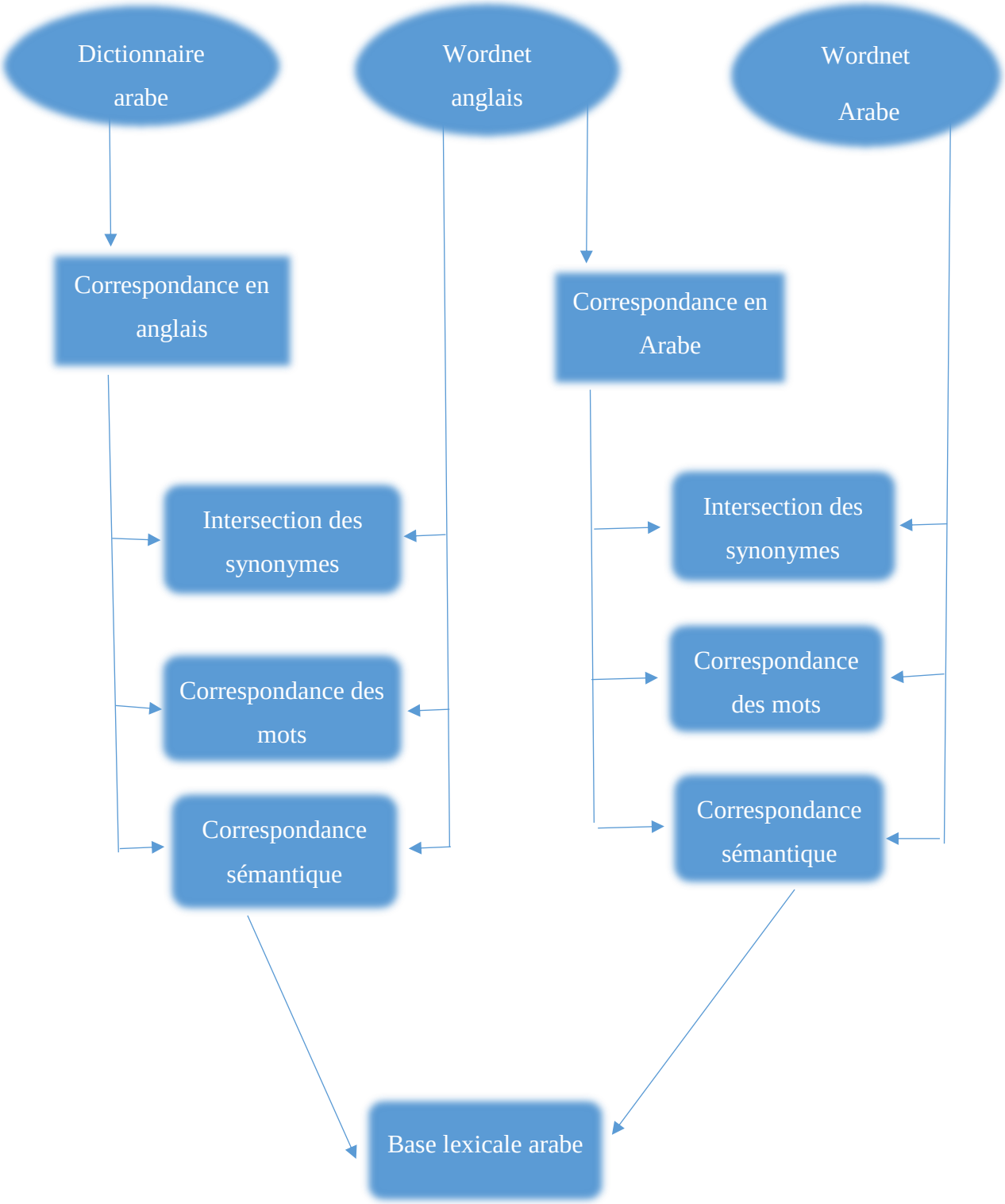


Figure 6-7 Procédure de génération de base lexicale arabe

6.3.3 Algorithme local : algorithme Lesk

L'algorithme de Lesk (Lesk, 1986) est une méthode de désambiguïsation bien connue qui consiste à compter le nombre de mots communs entre les définitions d'un mot (généralement trouvées dans un dictionnaire électronique) et les définitions des mots de son contexte. Le sens retenu correspond à la définition pour laquelle on compte les mots les plus communs avec le contexte (Vasilescu, 2003).

Cette idée simple a donnée des résultats intéressants et s'est avérée meilleure que les autres méthodes.

6.3.3.1 ***Principe :***

La méthode de désambiguïsation proposée par Lesk en 1986, fait partie de la catégorie des méthodes de désambiguïsation basées sur les connaissances (contexte).

A la différence des méthodes supervisées utilisant des corpus annotés (où chaque occurrence d'un mot polysémique dans le corpus est annotée par une étiquette de sens), les méthodes non supervisées font appelles à d'autres types de ressources, d'habitude de nature lexicale et sémantique disponibles par l'intermédiaire d'un dictionnaire électronique. L'avantage de ces approches réside principalement dans le fait qu'elles nécessitent seulement des corpus non annotés.

Par exemple**,** si on considère la cooccurrence des mots anglais « *pine* » et « *cône* » dans le même contexte, un programme de désambiguïsation automatique serait capable de choisir le sens *arbre* du mot *pine* en comptant les intersections entre les différentes définitions de sens des deux mots :

"Pine":

1. Kind of evergreen tree with needle-shaped leaves …;
2. Waste away through sorrow or illness …;

"Cone":

1. Solid body which narrows to a point …;
2. Something of this shape whether solid or hollow …;

3. Fruit of certain evergreen tree …

Dans ce cas, le nombre maximal de mots communs (*evergreen*, *tree*) est donné par l'intersection entre les définitions 1, respectivement 3 de *pine* et *cône*, ce qui détermine le choix du sens correspondant pour le mot *pine* soumis à l'analyse. Ses programmes traitent de manière séquentielle les mots à désambiguïser (à un moment donné, on compare les définitions des sens d'un mot cible avec toutes les définitions de chaque mot du contexte).

Ces approches suggèrent cependant que, une fois la décision prise sur le sens d'un mot, seulement la définition de ce sens soit prise en compte pour les désambiguïsations ultérieures, des autres mots.

6.3.3.2 *Algorithme de LESK simplifié (Vasilescu & Langlais, 2003):*

On calcule le score en comptant les superpositions entre l'entité descriptive du sens candidat *D, (s_i)* et les mots du contexte *w* (et non plus leurs définitions). Soit *C (t)* la fenêtre de contexte formée par le sac de mots *w*, en forme de base, alors la représentation en pseudo code de l'algorithme devient :

```
Pour chaque mot à désambiguïser t
Best_score = 0
Best_candidate = s1
Sup = 0
Déterminer C (t) le contexte de t
```

Pour chaque sens candidat s_j de t

Extraire du dictionnaire la définition D (s_j)

Calculer le nombre de superpositions sup = $\left| D(\text{Baker et al.}) \cap C(t) \right|$

Si Best_score < sup

Best_score = sup

Best_candidate = s_j

Attribuer à t le sens donné par Best_candidate

Figure 6-8 Algorithme du LESK

6.4 Partie 2 : Désambiguïsation lexicale des textes arabes

Nous avons abordé dans la première partie de ce chapitre les étapes de construction de la base lexicale, nous allons présenter dans cette partie les étapes de création de notre approche de désambiguïsation lexicale des textes arabes qui sont inspirées des travaux de Schwab et al (Schwab et al., 2013).

6.4.1 Algorithme global : Algorithmes à colonies de fourmis

Les algorithmes à colonies de fourmis forment une classe des métaheuristiques récemment proposé pour des problèmes d'optimisation difficile. Ces algorithmes s'inspirent des comportements collectifs de dépôt et de suivi de piste observé dans les colonies de fourmis. Une colonie d'agents simples (les fourmis) communiquent indirectement via des modifications dynamiques de leur environnement (les pistes de phéromones) et construisent ainsi une solution à un problème en s'appuyant sur leur expérience collective (Dréo, Pétrowski, Taillard, & Siarry, 2003).

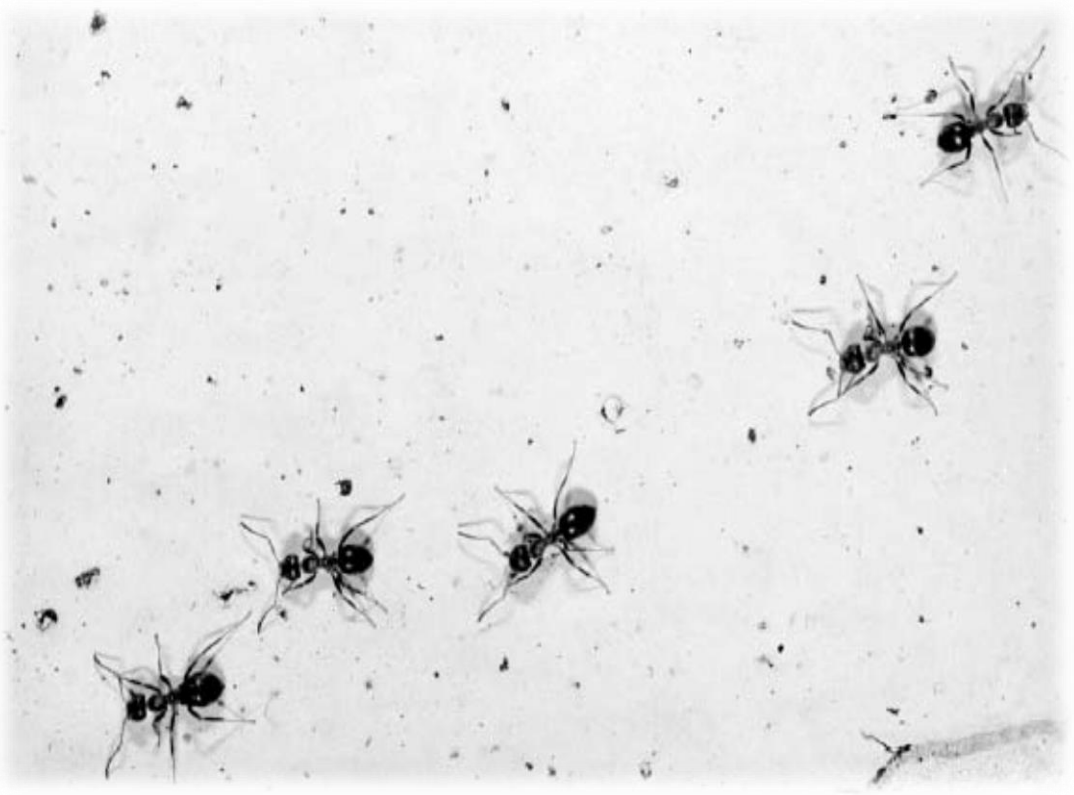


Figure 6-9 Des fourmis suivant une piste de (Belfadel & Diaf, 2014)

Cette méthode a été inventée en 1996 par Marco Dorigo de L'Université Libre de Bruxelles, l'idée initiale provient de l'observation des fourmis. Celles-ci ont la capacité de trouver le chemin le plus court entre leur nid et une source de nourriture en contournant les obstacles qui jonchent leur chemin.

On peut faire l'expérience suivante : on remarque que les fourmis se déplacent "en ligne", suivant le chemin des fourmis précédentes. Si on place un obstacle sur cette ligne (une pierre par exemple), de manière que le contournement de l'obstacle par un côté soit nettement plus court que l'autre. Au bout d'un certain temps on verra que toutes les fourmis contourneront l'obstacle par le côté le plus court.

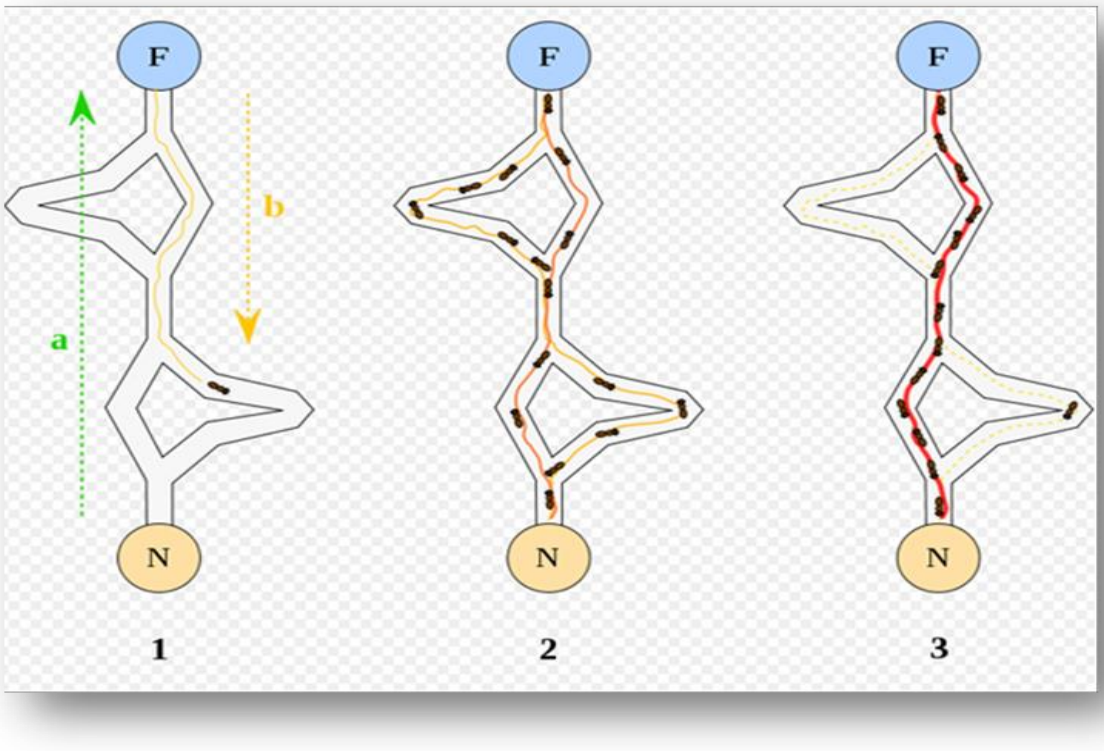


Figure 6.10 Choix du plus court chemin par une colonie de fourmi

Le phénomène s'explique de la manière suivante : les fourmis déposent en marchant des marqueurs chimiques appelés *phéromones*. Les autres fourmis suivent le chemin tracé par ces phéromones. Plus un chemin est marqué, plus elles ont tendance à le suivre. Lorsque l'on pose l'obstacle, le chemin de phéromones est coupé, et les fourmis choisissent au hasard l'un ou l'autre côté pour contourner l'obstacle.

Comme un chemin est plus court, plus de fourmis auront franchi l'obstacle en passant par ce côté. Par exemple, sur 6 fourmis, 3 fourmis pourront être passées par le côté court alors que les 3 fourmis ayant passé par le côté long n'auront pas encore fini de contourner l'obstacle. Une fourmi arrivant en sens inverse verra donc plus de phéromone sur le chemin qui passe par le côté court, et prendra donc ce chemin.

L'algorithme général basé sur une colonie de fourmis peut être décrit comme suit :

1) Initialiser les traces.

2) Répéter en parallèle pour chacune des p fourmis et tant qu'un critère d'arrêt n'est pas satisfait :

a. Construire une nouvelle solution à l'aide des informations contenues dans les traces et une fonction d'évaluation partielle.

b. Évaluer la qualité de la solution.

c. Mettre à jour les traces.

d. Application avec une probabilité donnée d'une recherche locale à ces solutions (la meilleure).

6.4.2 Détail de l'approche

Cette section décrit en détail les étapes de l'approche proposée (Abdelaali Bakhouche, Yamina, Schwab, & Tchechmedjiev, 2015; Schwab et al., 2013) (voir la figure 6.14).

6.4.2.1 *Prétraitement du texte*

Les corpus textuels constituent la matière première des applications de traitement automatiques de la langue. Mais souvent les textes contiennent des mots mal orthographiés ou collés, des incohérences typographiques, des phrases grammaticales, des caractères bizarres ou dans un encodage différents de celui attendu par le programme, etc. Les textes en quelque sorte « bruités ». Les textes doivent subir un « nettoyage » et de normalisation appropriée. Avant de pouvoir être traité par les programmes de traitement automatiques des langages naturels. Cette phase appelée prétraitement du texte est donc essentielle. Si elle est négligée ou réalisée de façon trop simpliste. Les systèmes risquent de fausser leur résultats (J.-M. Torres-Moreno, 2011).

Au cours de cette phase, nous allons mener trois phases la segmentation, la suppression des mots inutiles et la stemmatisation (voir figure 6.11).

A. *Segmentation*

La segmentation est une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type. La segmentation d'un texte informatisé est l'opération de délimitation des segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme, mot.

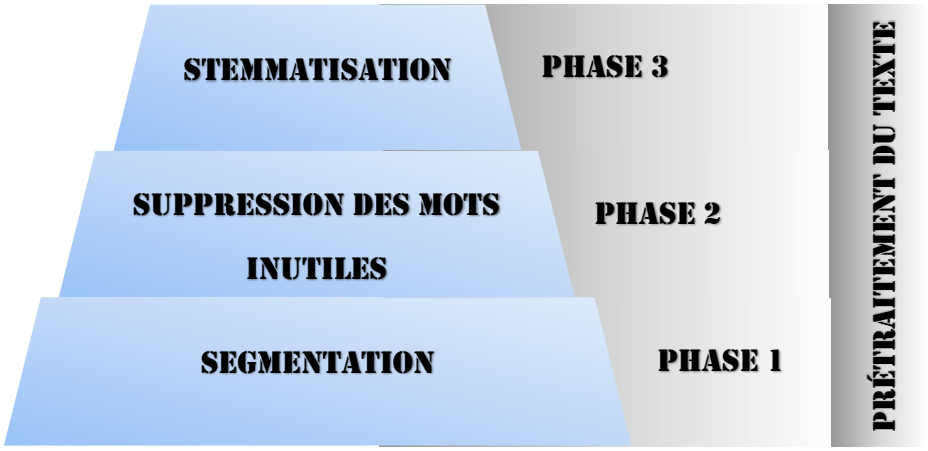


Figure 6.11 Les différentes phases du processus de prétréitémént du texte

Tableau 6-4 : Exemples de segmentation de mots dans la langue arabe

	Avant la segmentation	Après la segmentation
Exemple 1	وللمكتبات /walilmaktabat/ « Et pour les librairies »	و+ل+ال+مكتبة+ات /wa+li+al+maktaba+at/ Et+pour+les+librairies+pluriel
Exemple 2	وسنفعلها /wasanaf'aluhaa/ « et on va la faire »	و+ س+ ن+ فعل+ ها /wa+sa+na+f'alu+ha/ « et+on+nous+faire+elle »

B. *Suppression des mots inutiles*

Les mots les plus fréquents n'apportent pas, généralement une grande quantité d'informations. On les appelle mots ou termes vides de sens. En système d'information, les termes vides comme les articles, les conjonctions, les chiffres, la ponctuation, et les symboles spéciaux peuvent être supprimé lors d'un filtrage. Les mots sont appelés en anglais stop-words. Typiquement environs de 26,037 des mots vides dans la langue arabe. Ils sont réunis dans le site web Sourceforge[11].

C. *Stemmatisation*

La stemmatisation est une technique morphologique largement utilisée pour la préparation des textes dans une recherche documentaire. Elle consiste à rechercher la racine lexicale ou stem[12] pour des mots en langue naturelle, et ceci, par l'élimination des affixes qui leur sont rattachés, en d'autre terme regrouper sous un même identifiant des mots dont la racine est communes.

Tableau 6-5 Exemple de stemmatisation de mots dans la langue arabe

	Avant la stemmatisation	Après la stemmatisation
Exemple	صادق، الصدق الصديق، صدق، يصدق	les mots sont des flexions du mot "صدق"

6.4.2.2 ***Environnement***

L'environnement des fourmis est un graphe. Ce graphe est simple, sans information linguistique externe. Il est organisé en fonction des éléments du texte. Chaque mot est un sommet, ce sommet possède autant de fils que le mot correspondant possédant de sens différents. Un fils est également un nid et tous les fils d'un sommet associé à un mot sont en compétition. Un vecteur correspondant au sens associé (voir figure 6.12).

[11] http://sourceforge.net/projects/arabicstopwords

[12] La forme du mot après l'enlèvement de toutes les affixes

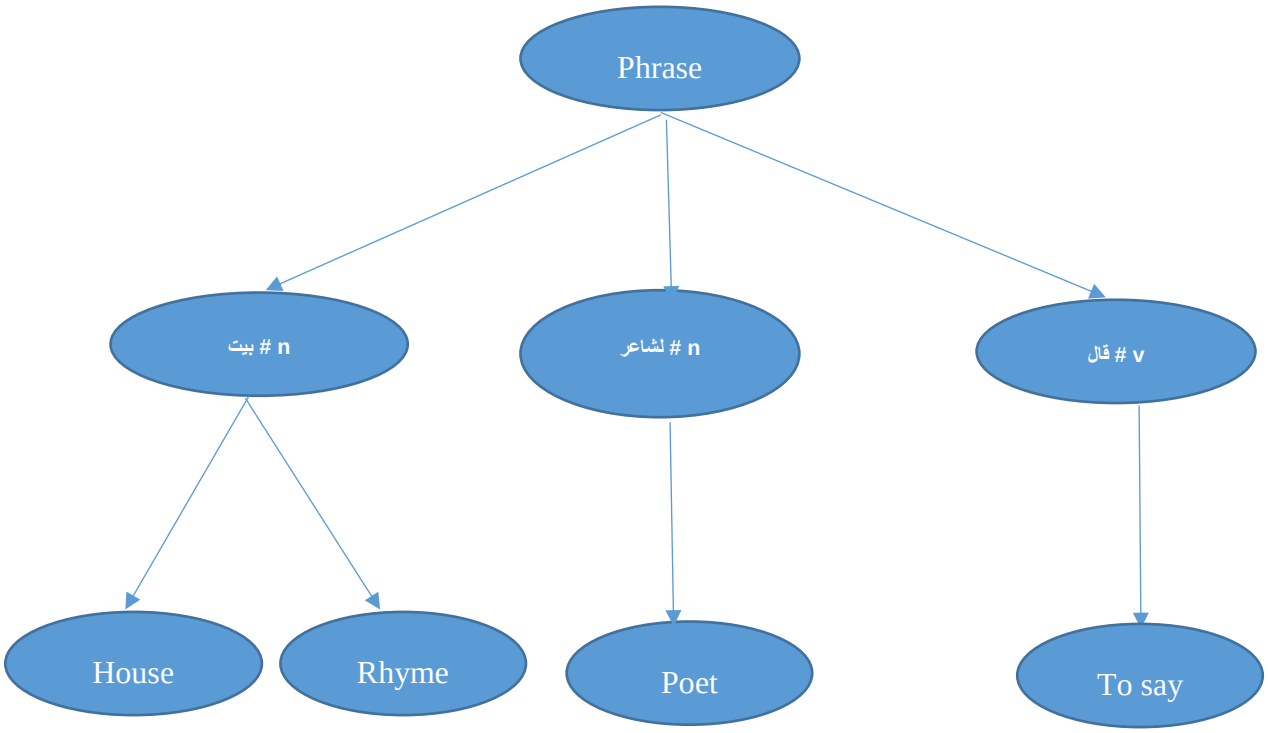


Figure 6-12 Exemple d'environnement utilisé dans l'éxpérience.

6.4.2.3 *Types des nœuds*

Du point de vue d'une fourmi, il existe deux types de nids : les nids amicaux et les nids inamicaux. Les nids inamicaux correspondent aux autres sens du mot et les fourmis qui en sont issues constituent des concurrentes pour l'accès aux ressources.

Les nids amicaux sont tous les nids des autres mots. Les nids amicaux d'une fourmi peuvent l'inciter à leur céder une partie des ressources qu'elle transporte. Les nids inamicaux au contraire constituent des sources de ressources pour les fourmis, c'est à dire qu'une fourmi peut décider de récupérer une partie des ressources d'un nid inamical dès lors que le niveau des ressources de ce nid est positif.

6.4.2.4 *Déplacement des fourmis*

Une fourmi peut se déplacer en empruntant les arêtes et dans certaines circonstances, elle peut construire de nouveaux liens que nous appelons des ponts. En dehors des informations morphosyntaxiques, chaque sommet contient les attributs suivants :

- un niveau de ressource R

- un vecteur V. Chaque arête contient un niveau de phéromone. Le principal objet des phéromones est de représenter la popularité des arêtes sur lesquelles elles sont présentes.

L'environnement par lui-même évolue de différentes manières :

6.4.2.5 *Vecteur de définition*

Le vecteur d'un sommet est légèrement modifié à chaque passage d'une nouvelle fourmi. Seuls les vecteurs des nids ne varient pas (ils ne peuvent être modifiés). Un sommet qui est aussi un nid est initialisé avec le vecteur du sens du mot qui lui est associé, tous les autres sommets se voient attribuer un vecteur nul (voir figure 6.13) ;

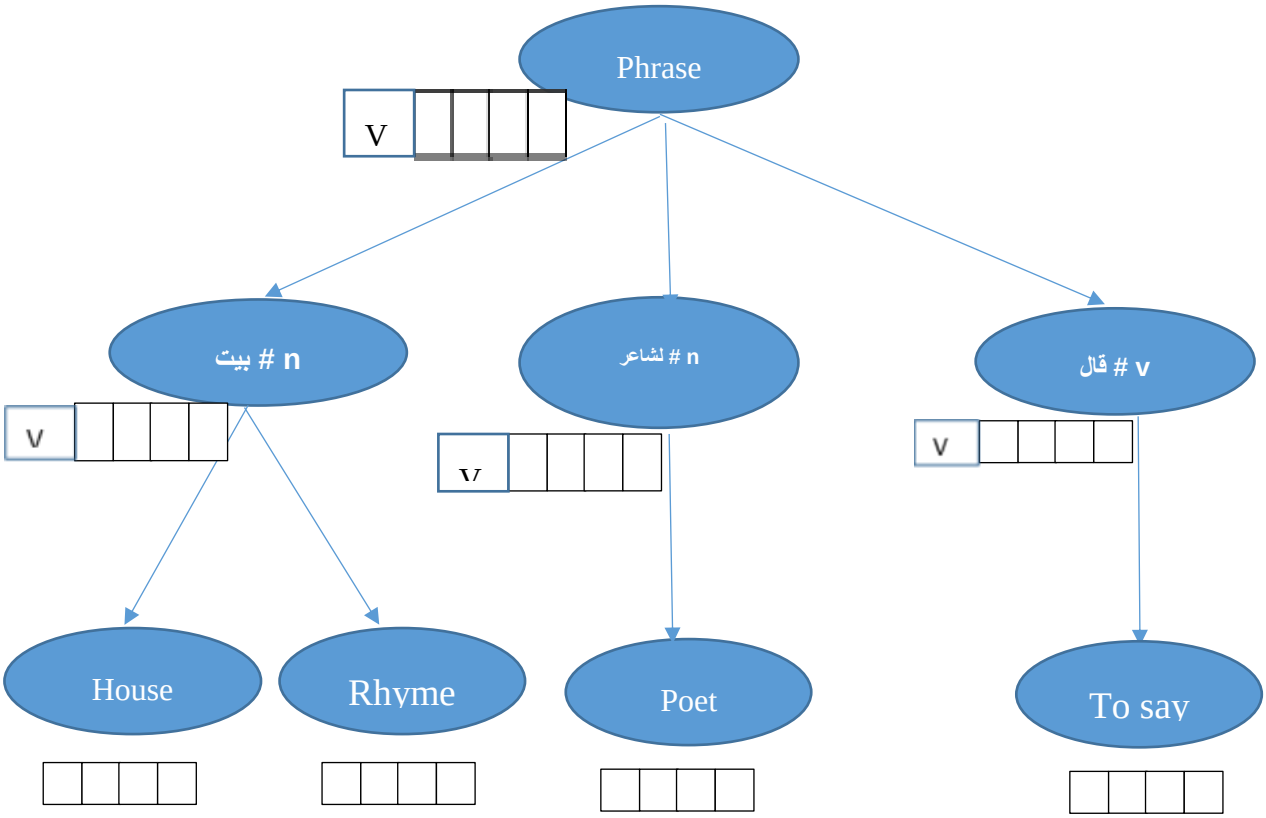


Figure 6.11 Etat de l'environnement au départ

6.4.2.6 *Energie*

Les ressources tendent à être redistribuées entre les nids qui réinvestissent à leur tour pour la production de fourmis. Les nids ont une quantité de ressources initiale.

6.4.2.7 *Phéromone de passage*

Les niveaux de phéromone des arcs sont modifiés par le passage des fourmis. Nous utilisons un facteur d'évaporation δ qui assure qu'avec le temps le niveau de phéromone d'un arc tend à décroître vers zéro si aucune fourmi ne l'emprunte. Seuls les ponts (arcs créées par les fourmis) peuvent disparaître si leur niveau de phéromone atteint zéro.

6.4.3 L'algorithme

Au sein de l'arbre syntaxique, chaque sens de mot correspond à un nid, et chaque sens de mot possède un vecteur de définition. Chaque fourmi dans l'algorithme correspond à un processus indépendant. Le processus global est itératif, et à chaque étape :

1. l'âge de chaque fourmi augmente d'une unité. Si l'âge d'une fourmi est supérieur à une limite, cette fourmi meurt ;
2. chaque fourmi choisit son comportement :
    - soit chercher de l'énergie ;
    - soit chercher à revenir à leur fourmilière mère ;
3. chaque fourmi se déplace en fonction de son comportement, dépose des phéromones et modifie légèrement le vecteur de chaque sommet traversé. Elle peut également déposer ou récupérer des ressources et créer un pont ;
4. l'environnement est mis à jour (phéromones).

L'algorithme peut être considéré comme un processus itératif sans fin. Cependant, l'utilisateur peut choisir une condition d'arrêt qui soit un nombre d'étapes donné ou une condition sur la durée (en nombre d'étapes consécutives) de la stabilité du vecteur de la racine.

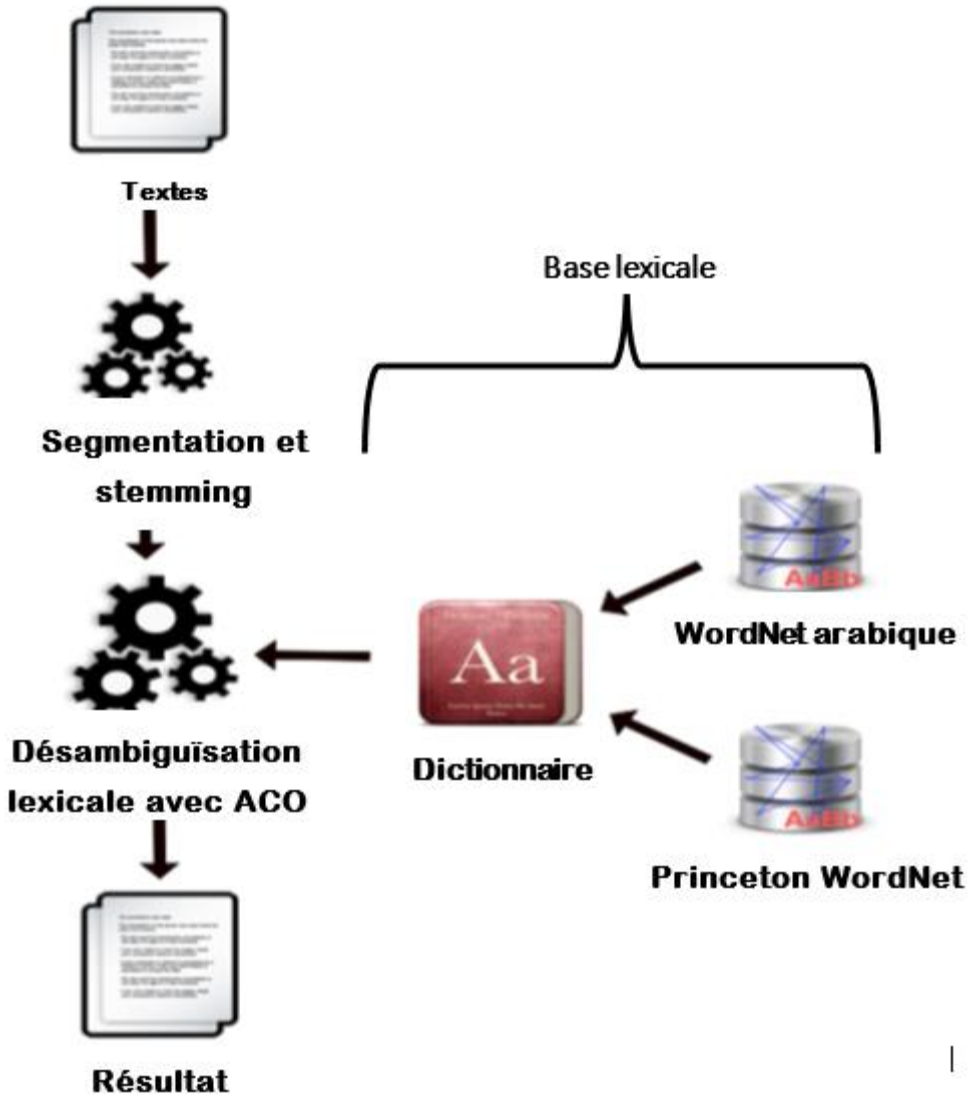


Figure 6-12 Schéma de l'approche proposée pour WASD

6.5 Exemple illustré

La meilleure façon de comprendre le déroulement de l'algorithme est par le biais d'un exemple ; nous illustrons les étapes de l'algorithme par le chemin emprunté par une fourmi sur l'expression suivante :" قال الشاعر بيتا " .

Nous suivons le chemin d'une fourmi qui est né dans la fourmilière بيت / Rime - correspondant à la deuxième signification pour le mot.-بيت

1. Étape 1. La fourmi commence à chercher de la nourriture pour prendre de l'énergie à son nid. elle débute à explorer le graphe à partir du nœud qui correspond au nom بيت # n.
2. Étape 2. Sur le chemin vers ce nœud, la fourmi dépose sa phéromone, et quand elle arrive, elle prend une quantité d'énergie. Elle dépose également une partie de son odeur sur le nœud. Cette odeur est marquée par le dépôt de deux composants pris au hasard dans son nid de mère.
3. Étape 3. La fourmi choisit alors de suivre un autre chemin où elle dépose sa phéromone. elle dépose également deux composantes de vecteur du nid mère au hasard et prend une quantité d'énergie. A ce stade, la fourmi a une quantité totale d'énergie, et une décision pseudo-aléatoire a été pris de retourner au nœud mère.
4. Étape 4. Comme une fourmi ne peut pas aller sur un nœud d'où elle vienne, sauf si cela est sa seule chance, notre fourmi a le choix entre aller à قال # v- ou الشاعر # n.
5. Étape 5. Les fourmis venant de nœud الشاعر / Poet, une partie de leur vecteur sur الشاعر # n'ont plus en commun avec le nœud de vecteur que nos fourmis, بيت # v-.
6. Étape 6. Notre fourmi est donc plus susceptibles d'aller à الشاعر # n. Elle le fait tout en déposant sa phéromone le long du chemin. La fourmi quitte son odeur sur le nœud de destination, mais ne prend pas l'énergie, car c'est un nœud de retour. Maintenant, sa seule option est d'aller à الشاعر / Poet -.
7. Étape 7. Pour la fourmi, le الشاعر / Poet-ami est un nœud de potentiel. Il a donc la possibilité de construire un pont à sa fourmière mère. Par conséquent, la fourmi choisit d'emprunter et de déposer son phéromone. Elle arrive à son nid de mère et elle dépose toute son énergie. La fourmi retourne alors en mode de recherche et reprend ses voyages vers le nœud où il va mourir et elle dépose son énergie transportée.

6.6 Conclusion

La réalisation de la plupart des applications de traitement automatique des langues naturelles (TALN) nécessite un ensemble minimal de ressources lexicales, et la

réussite de ces réalisations dépend fortement des données lexicales en liaison avec le logiciel de traitement.

Par conséquent, la qualité, la consistance et la normalisation des ressources lexicales est une condition préalable et importante pour le développement d`applications robustes et de large couverture. La structure et la représentation des ressources lexicales sont nécessaires pour élaborer des lexiques réutilisables à large couverture.

Ce chapitre a établi une démarche pour la construction des bases de données destinées à crier un système pour lever l'ambigüité des mots arabe par la prise en considération de la sémantique qui est le plus grand défi qui nous confronte dans le traitement automatique des langues humaines. Pour motiver notre travail d'un point de vue pratique, le chapitre suivant décrit l'expérimentation de notre modèle.

Chapitre 7 :

Expérimentations et évaluations

Sommaire

7.1 Introduction

Dans ce chapitre, nous présentons l'implémentation de notre approche présentée dans le chapitre précédent, puis l'évaluation et la discussion des résultats obtenus.

Comme nous l'avons mentionné dans le chapitre précédent de ce manuscrit, la mise en œuvre de notre système nécessite comme entrées : une base lexicale et un corpus de textes pour évaluer notre approche. La création de base lexicale constitue elle-même une tâche difficile et cela dû à la complexité liée à la variabilité sémantique des concepts qui dépendent fortement du contexte, et elle nécessite elle-même comme entrées, d'autre ontologie comme WordNet.

En ce qui concerne le corpus de textes arabes, nous avons utilisé un corpus existant collecté auprès d'un ensemble de sites de la presse arabe.

7.2 Environnement technologique

Nous avons implémenté en Java (version 1.6) les différents modules de notre approche correspondants à nos différentes propositions présentées dans le chapitre précédent. Le choix de Java est motivé par la disponibilité et la gratuité de diverses APIs java qui nous ont servis à la réalisation et à l'évaluation de notre approche proposée :

- **L'API JWNL28 (Java WordNet Library) :** est utilisée pour accéder à l'ontologie WordNet. Cette API nous a permis de lemmatiser les mots du texte en se basant sur les relations morphologiques d'un mot dans WordNet d'une part et d'exploiter les principales relations entre concepts de WordNet, telles que la synonymie et la taxonomie (is-a) des noms et verbes dans WordNet, d'autre part.
- **L'API JWS29 (Java WordNet Similarity) :** offre la possibilité d'évaluer la similarité sémantique entre deux concepts dans WordNet. Cette API implémente les différentes mesures de similarités sémantiques
- **AraMorph** est un analyseur morphologique développé par Tim Buckwalter (Buckwalter, 2002) pour le compte du LDC. Il prend comme entrée un texte sous forme d'un fichier texte et donne comme sortie une liste de mots et toutes les solutions possibles de ses lemmes, la vocalisation, la morphologie, la catégorie, le

glossaire et une analyse statistique qui donne le nombre des lignes, le nombre des mots arabes et le nombre de mots non arabes dans le texte (Zaidi–Ayad, 2013).

Il existe deux versions d'AraMorph, celle en PERL, développé par *Buckwalter et* celle en JAVA, traduite par *Pierrick Brihaye* accessible en ligne. Le projet inclut des classes *Java* permettant l'analyse morphologique de fichiers textuels en arabe et ce, quel que soit leur encodage. A cet effet, il a proposé 3 fichiers de test dans les principaux encodages utilisés pour la langue arabe : UTF-8, ISO-8859-6 et CP1256. Ce projet inclut également des classes compatibles avec l'architecture de Lucene, ce qui permet l'analyse, l'indexation et l'interrogation de documents en arabe.

Exemple d'un résultat pour le mot « كتاب » le fichier de sortie `results.txt` qui est, lui, en `UTF-8` :

```
Processing token :     كتاب
Transliteration:       ktAb
Token not yet processed.
Token has direct solutions.
SOLUTION #3
Lemma:  kAtib
Vocalized as:  كُتَّاب
Morphology:
        Prefix: Pref-0
        Stem: N
        Suffix: Suff-0
Grammatical category:
        Stem: كُتَّاب     NOUN
Glossed as:
        Stem authors/writers

SOLUTION #1
Lemma:  kitAb
Vocalized as:  كِتاب
Morphology:
        Prefix: Pref-0
        Stem: Ndu
        Suffix: Suff-0
Grammatical category:
        Stem: كِتاب     NOUN
Glossed as:
        Stem: book
SOLUTION #2
Lemma:  kut~Ab
Vocalized as:  كُتَّاب
Morphology:
        Prefix: Pref-0
        Stem: N
        Suffixe: Suff-0
Grammatical catégorie:
        Stem: كُتَّاب     NOUN
Glossed as:
        Stem kuttab (village school)/Quran School
```

7.3 Description de l'application

Le diagramme de classe est le point central dans un développement orienté objet. En analyse, l'objectif de ce diagramme est de décrire la structure des entités manipulées par les utilisateurs. En conception, le diagramme de classes représente la structure d'un code orienté et met en évidence d'éventuelles relations entre ces classes.

On montre dans cette partie le digramme de classe d'extraction des relations sémantiques de Wiktionnaire (voir les détails des classes principales du code source dans la partie Annexe).

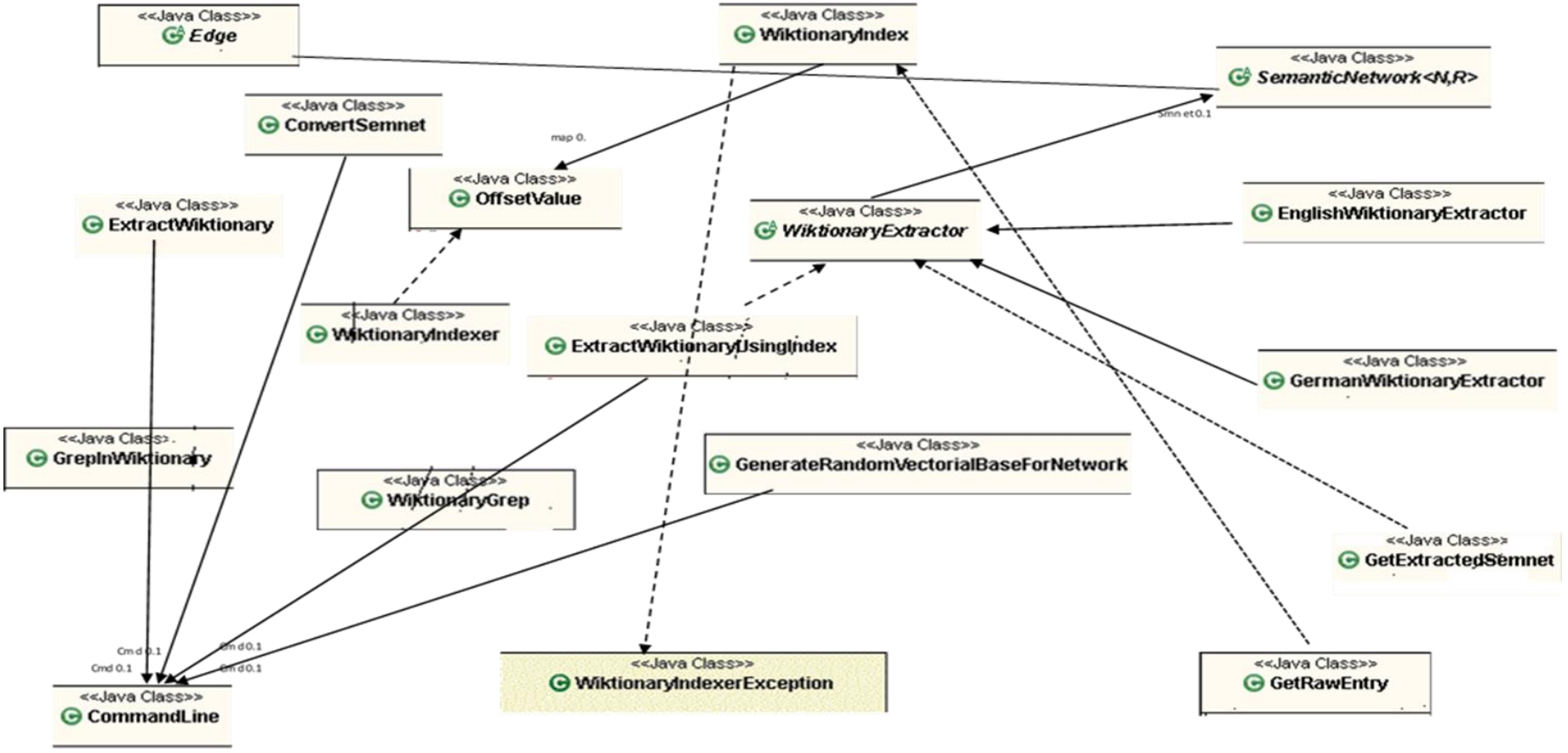


Figure 7-1: Diagramme de classe

7.4 Statistiques de la Base lexicale extraite de Wiktionnaire

Une fois le processus d'extraction est appliqué à des définitions extraites du Wiktionnaire arabe, nous avons obtenu 8321 triples relationnels, répartis dans le Tableau 7.1. Ce tableau présente, le nom de la relation sémantique, le nombre des relations extraites de chaque type. Le pourcentage de chaque relation. D'après le résultat on remarque que la synonymie représente la relation la plus importante.

Tableau 7-1 Resultat d'extraction des relations sémantiques de winktionnaire arabe

Relations	Nombre de relations	Pourcentage
Synonymie	6784	81%
Antonymie	1537	19%

7.5 Statistiques de dictionnaire généré de WordNet

Les statistiques principales du dictionnaire généré dans ce travail sont résumées dans le tableau 7.2. La première colonne de ce tableau donne les parties du discours, tandis que les autres colonnes indiquent le nombre de parties arabes de la parole et le nombre de définitions anglaises des sens. La figure 7.3 montre le dictionnaire anglais-arabe basée sur WordNet. On voit que les mots arabes ne correspondent pas suffisamment à tous les sens possibles anglais.

Tableau 7-2 Statistiques du dictionnaire

	Arabic POS	English sens
noms	9641	14,680
verbes	2777	6084
adjectif	662	762
collocations	12,905	263
total	25,985	21,789

7.5.1 Structure de dictionnaire :

La base lexicale a été organisée sous la forme d'un fichier en format XML. Chaque entrée de la base est représentée sous la forme suivante :

Chaque unité est stockée dans la base sous forme d'élément composé de :

Attribut nommé tag comporte la valeur d'unité lexicale avec le type (v : verbe, n : nom, a : adjectif) séparé par le signe %.

Elément feuille « sens ». À son tour, se compose de :

Elément feuille « ids » qui contient la sémantique des unités. Comme le montre la figure 7.2, le verbe « طالب » a plusieurs sens : « claim », « necessitate » et « demand » :

```
<word tag="طالب%v">
<sense>
<ids>claim%2:32:02:: </ids>
</sense>
<sense>
<ids>necessitate%2:42:00:: </ids>
</sense>
<sense>
<ids>demand%2:32:00:: </ids>
</sense>
</word>
```

Figure 7-2 Les différents sens du verbe « طالب » dans la base lexicale

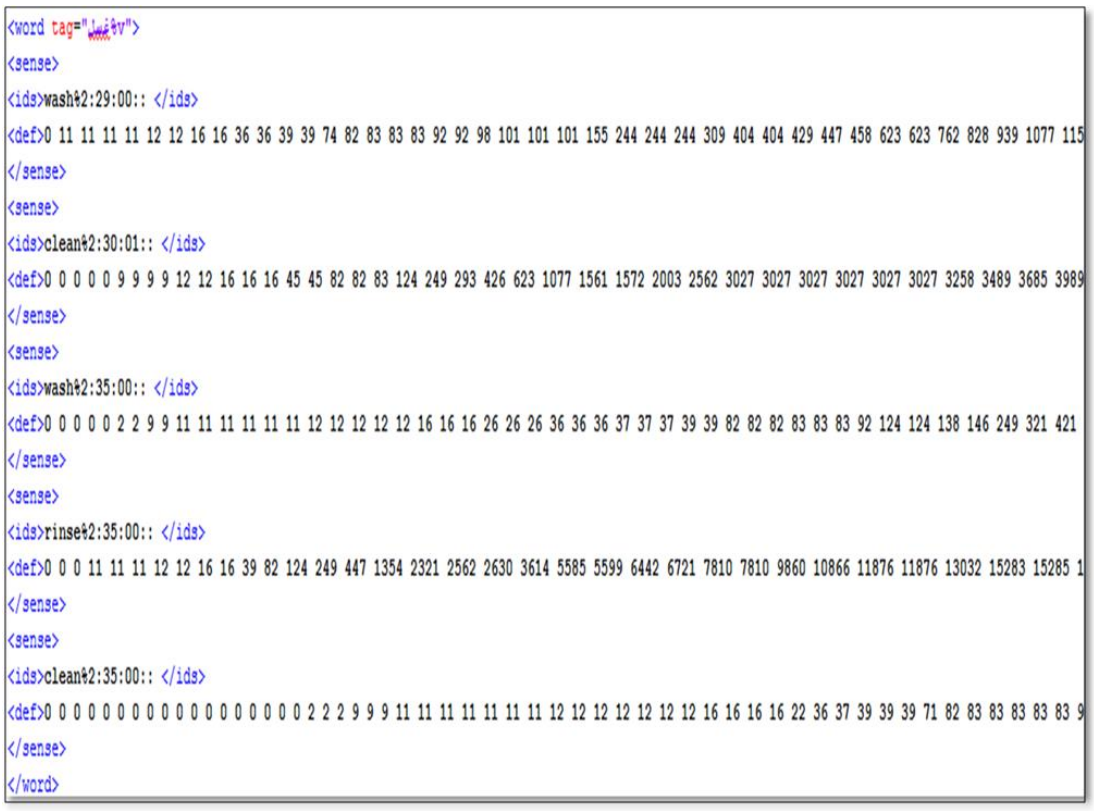


Figure 7-3 Une partie de la base lexicale

7.6 **Corpus d'évaluation**

Afin de développer des corpus annotés pour une langue comme l'arabe, nous avons essayé de suivre les démarches testées et prouvées pour des corpus similaires et pour des langues ayant plus de ressources disponibles et notamment la langue anglaise comme le montre le récent travail de recensement de corpus dans la langue arabe (Zaghouani, 2015). L'adaptation des procédures d'annotation existantes permet un gain en temps de recherche, même s'il existe toujours des adaptations à faire pour se conformer aux particularités de la langue et à la nouvelle annotation à créer. Le tableau 7.3 montre les domaines de corpus et la distribution des articles de chaque domaine, Ainsi que la figure 7.4 illustre un texte de corpus de domaine économie.

Tableau 7-3 Statistiques du corpus

Domaines	nombre d'articles
Culture	2782
Économie	3468
International	2035
locale	3596
Religion	3860
Sports	4550
Total	20291

مع تزايد الحاجة الى ترشيد الطرق، وتعظيم الاستفادة من الموارد المتاحة، يتم اجراء العديد من المراجعات والدراسات التي تكشف عن تكاليف أنماط التنمية والرفاهية التي أشيعت خلال العقود الثلاثة الماضية. فاذا كانت الطفرة النفطية خلال حقبة السبعينيات قد مكنت من دخول مرحلة من الرفاهية الاقتصادية والاجتماعية، فإنها بذات الوقت ولدت عددا من القيم الاستهلاكية والاتكالية بين صفوف عامة الناس، كما لم ترافق هذه المرحلة تطوير معايير وقيم انتاجية الفرد والمجتمع. وحتى بالنسبة لفئات عديدة من رجال الأعمال، ارتبطت تلك المرحلة بالقيام بأنشطة المضاربات وذات المردود الربحي السريع. الا انه بطبيعة الحال كانت هناك شرائح من رجال الاعمال سعت ولا تزال الى تجسيد دورها التنموي السليم من خلال ما تتعهده من انشطة ومشروعات اقتصادية منتجة. اننا نتذكر اليوم هذه الحقائق ونحن نتابع سعي الدول الى تطوير قنوات وانماط تجسد من خلالها المشاركة الشعبية في تحمل تكاليف التنمية والحفاظ على المستويات الجيدة للرفاهية الاجتماعية. إن أبرز العوائق التي يصطدم بها هذا السعي هو الثقافة الاقتصادية الاستهلاكية والاتكالية التي أفرزتها الحقبة الماضية وكرستها بين فئات لا يستهان بها سواء من رجال الاعمال او المواطنين. وهذا ما يفسر اسلوب الحذر في تطوير تلك القنوات والانماط خشية عدم نجاحها او سوء فهمها مما يولد انعكاسات اجتماعية سلبية.

Figure 7-4 Exemple d'un texte de corpus de domaine économie

7.6.1 Prétraitements des textes du corpus

L'objectif de cette phase consiste à sélectionner uniquement les données potentiellement utiles dans la base.

L'ensemble des données est ensuite soumis à des prétraitements, afin de les transformer et de gérer les données manquantes ou invalides.

L'étape suivante dans cette phase consiste à formater ces données pour les rendre compréhensibles aux algorithmes :

- Chaque document dans l'ensemble de données arabe est traité pour supprimer des chiffres, des tirets et des signes de ponctuation.
- Nous avons normalisé les lettres 'ء' '(hamza)', " (aleph mada) 'أ' (aleph avec hamza sur le dessus) 'ؤ' (hamza sur w) " إ (alef avec hamza sur le fond) et 'ئ' (hamza sur toit). L'idée derrière cette normalisation est que tous formes de Hamza sont représentées dans les dictionnaires comme une seule forme.
- Nous avons normalisé la lettre «ى» à «ي», et la lettre " ة à '.' La raison derrière cette normalisation est qu'il n'y a pas une seule convention pour l'orthographe 'ى' ou 'ي' et 'ة' ou " quand ils apparaissent à la fin d'un mot arabe.
- Tous les textes non-arabes et les mots vides ont été enlevés.

Tous les résultats de l'algorithme ont été obtenus en utilisant un PC avec processeur Intel Core Duo CPU 2,66 et 6.GB RAM de mémoire principale sur Windows 7. L'application était développée en langage Java.

Pour étudier l'effet de l'algorithme ACARWSD, nous avons mené six groupes d'expériences, pour chaque groupe et pour chaque catégorie de texte, un tiers des articles a été spécifié au hasard et utilisé pour les essais et les articles restants ont été utilisés pour la formation. Le tableau 7.4 illustre les domaines abordés par ces textes et la distribution des mots tels que décrits dans les textes.

Tableau 7-4 Les six articles dans la base lexicale

Texte	Domain	Mots	annoté
D001	Culture	800	242
D002	Economie	2130	1133
D003	International	600	325
D004	Local	755	432
D005	Religion	295	295
D006	Sport	450	298
Total	–	5030	2555

Ces textes courts ont été construits et chaque terme (nom, adjectif, et verbe) a été manuellement marqué. Une marque (tag) est un terme qui nomme un sens particulier. Par exemple, le terme banc peut être annoté comme مكتب/desk, مكتب /office, مكتب / administration office, etc. Dans la base de données des vecteurs, chaque sens de mot est associé au moins une marque. Utiliser une marque est généralement plus simple qu'un numéro de sens, spécialement pour les personnes qui ont la charge d'un tel marquage.

Figure 7.5 Exemple d'un graphe du texte

7.7 Métriques

Les mesures utilisées pour l'évaluation quantitative des résultats de notre méthode de WSD sont les taux de ***précision*** et de ***rappel*** :

$$Précesion = \frac{\text{nombre de prédictions correctes}}{\text{nombre de prédictions faites par le système}}$$

Ce critère permet de mesurer la pertinence des réponses retournées par la méthode évaluée.

$$Rappel = \frac{\text{nombre de prédictions correctes}}{\text{nombre de nouvelles instances}}$$

Ce critère permet de mesurer l'exhaustivité des bonnes réponses dans l'ensemble des réponses générées.

Le taux de rappel indique la proportion des nouvelles instances d'un mot ambigu qui sont correctement désambiguïsées, tandis que la précision indique, quant à elle, la proportion des prédictions de désambiguïsation faites par le système qui sont correctes. Nos résultats sont ensuite comparés aux résultats issus d'une méthode de base (Baseline). Les résultats de cette méthode correspondent aux deux scores, précision et rappel, comme nous allons l'expliquer dans le paragraphe suivant. Pour faciliter la comparaison entre nos résultats et la Baseline, nous avons également eu recours à la ***f-mesure*** (Van Rijsbergen, 1979), score combinant précision et rappel en une mesure unique. Il s'agit, plus précisément, de la moyenne harmonique du rappel et de la précision, calculée selon la formule suivante :

$$\text{f} - \text{mesure} = \frac{2 * (\text{précision} * \text{rappel})}{\text{précision} + \text{rappel}}$$

7.8 Évaluation pratique

Dans cette section, nous décrivons la tâche d'évaluation en utilisant les estimations des paramètres et la comparaison avec d'autres algorithmes.

7.8.1 Sélection des paramètres

Dans ce système, nous utilisons certains paramètres (EnergPrise, MaxEnerg, EvaPhero, InitEnerg, VieFourmi, TailleVecteur, DepotVector, NbrCycle) que nous avons besoins pour obtenir un meilleur score dans le corpus d'évaluation (voir le tableau 7.5).

Tableau 7-5 Paramètres de ACARWSD

Natation	Description	Plage des valeurs
EnergPrise	Énergie prise par une fourmi	1–30
MaxEnerg	La quantité maximale d'énergie qu'une fourmi peut porter	1–60
EvaPhero	Vitesse d'évaporation de la phéromone entre deux cycles	0.0–1.0
InitEnerg	quantité initiale d'énergie sur chaque nœud	5–60
VieFourmi	La durée de vie de fourmi	1–30
TailleVecteur	Longueur de vecteur d'odeur	20–200
DepotVector	Le pourcentage de la composante du vecteur d'odeur	0–100
NbrCycle	Nombre de cycles de la simulation	1–100

7.8.2 Tests et configurations expérimentales

L'examen des résultats de l'évaluation va nous permettre d'identifier les points forts de notre méthode. La première démarche à envisager est l'examen global de notre

méthodologie, qui valide ou infirme l'utilité d'un traitement basé sur des méthodes de propagation des mesures locales à un niveau supérieur. Pour ce faire, nous confrontons les mesures de similarité avec les résultats de la méthode dite « globale ».

7.8.3 Analyse des résultats

L'application des méthodes globales sur des textes met en œuvre l'ensemble des processus et enrichissements disponibles (analyse morphosyntaxiques, désambiguïsation sémantique, synonymie,). Au niveau des contraintes paramétrables, cette méthode globale privilège la précision en imposant la présence de l'unité lexicale focus, mais sans négliger le rappel (voir le tableau 7.6).

Tableau 7-6 Estimation pour le texte

Natation	**Plage des valeurs**	**Estimation pour le texte**
EnergPrise	1–30	9.0
MaxEnerg	1–60	22.142
EvaPhero	0.0–1.0	0.3577
InitEnerg	5–60	32.0
VieFourmi	1–30	27.0
TailleVecteur	20–200	135.637
DepotVector	0–100	0.9775
NbrCycle	1–100	-

Tableau 7-7 Environnement de simulation pour le texte d001

Notation	Valeurs
Nombre de nœud	415
Nombre de chemin	4096
Nombre de fourmi	1026
N° de cycle	99

Le résultat F-mesure généré par l'algorithme contre six ensembles de données. Dans chaque ensemble de données, nous considérons arbitrairement 70% des documents de l'entrainement et 30% pour les tests.

L'algorithme ne donne pas les mêmes significations sur chaque exécution, donc la répétition de l'exécution est nécessaire au moins deux fois pour obtenir le sens exact. Nous avons noté qu'après plusieurs cycles (environ 70), les résultats obtenus sont similaires, comme le montre la figure 7.5, et nous pouvons obtenir les mêmes résultats entre deux exécutions.

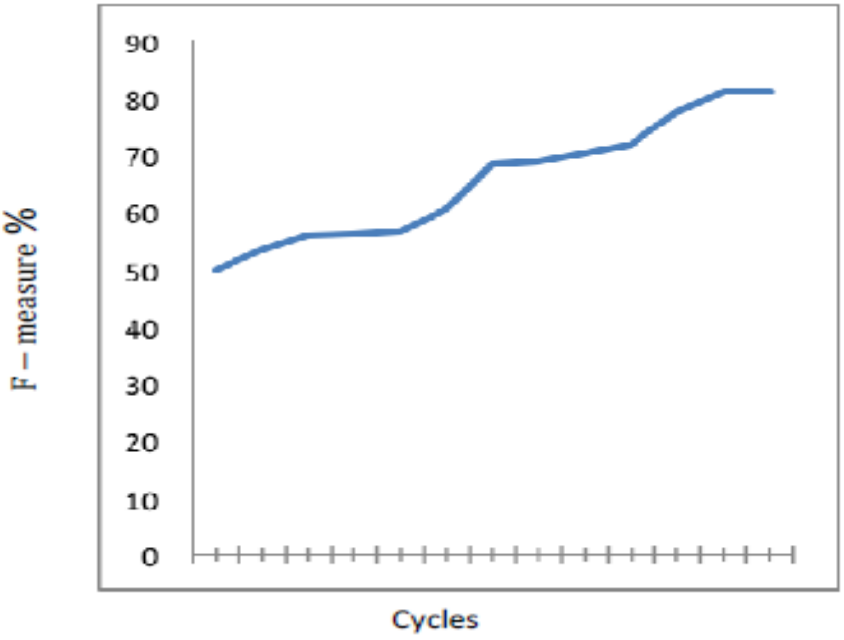


Figure 7-6 Le résultat F-mesure généré par l'algorithme

```
d001 d001.s001.t001 augmentation%1:04:00:: !! lemma=زيادة#n
d001 d001.s001.t002 need%1:26:00:: !! lemma=حاجة#n
d001 d001.s001.t003 adulthood%1:28:00:: !! lemma=رشد#n
d001 d001.s001.t004 idealization%1:04:00:: !! lemma=تعظيم#n
d001 d001.s001.t005 benefit%1:21:00:: !! lemma=استفادة#n
d001 d001.s001.t006 resource%1:21:00:: !! lemma=مورد#n
d001 d001.s001.t007 leave%2:42:01:: !! lemma=أجاز#v
d001 d001.s001.t008 orphan%2:40:00:: !! lemma=يتم#v
d001 d001.s001.t009 proceeding%1:04:00:: !! lemma=إجراء#n
d001 d001.s001.t010 many%3:00:00:: !! lemma=عديد#a
d001 d001.s001.t011 follow-up%1:04:00:: !! lemma=متابعة#n
d001 d001.s001.t012 report%1:10:03:: !! lemma=تقرير#n
d001 d001.s001.t013 uncover%2:39:00:: !! lemma=كشف#v
d001 d001.s001.t014 charge%1:21:02:: !! lemma=شحنة#n
d001 d001.s001.t015 practice%1:04:00:: !! lemma=ممارسة#n
d001 d001.s001.t016 development%1:22:02:: !! lemma=تطور#n
d001 d001.s001.t017 comfort%1:26:00:: !! lemma=راحة#n
d001 d001.s001.t018 circulate%2:38:03:: !! lemma=أذاع#v
d001 d001.s001.t019 interstice%1:08:00:: !! lemma=ثغرة#n
d001 d001.s001.t020 contract%1:10:00:: !! lemma=عقد#n
d001 d001.s001.t021 three%1:23:00:: !! lemma=ثلاثة#n
d001 d001.s001.t022 past%1:28:01:: !! lemma=الماضي#n
d001 d001.s001.t001 augmentation%1:04:00:: !! lemma=زيادة#n
d001 d001.s001.t002 need%1:26:00:: !! lemma=حاجة#n
d001 d001.s001.t003 adulthood%1:28:00:: !! lemma=رشد#n
d001 d001.s001.t004 idealization%1:04:00:: !! lemma=تعظيم#n
d001 d001.s001.t005 benefit%1:21:00:: !! lemma=استفادة#n
d001 d001.s001.t006 resource%1:21:00:: !! lemma=مورد#n
d001 d001.s001.t007 leave%2:42:01:: !! lemma=أجاز#v
d001 d001.s001.t008 orphan%2:40:00:: !! lemma=يتم#v
d001 d001.s001.t009 proceeding%1:04:00:: !! lemma=إجراء#n
d001 d001.s001.t010 many%3:00:00:: !! lemma=عديد#a
d001 d001.s001.t011 check%1:04:05:: !! lemma=مراجعة#n
```

Figure 7-7 Annotation d'un texte dans le format requis pour le script d'évaluation

Dans la plupart des cas les informations qui permettent de faire le choix du meilleur sens d'un mot ne sont pas de nature thématique, d'autres critères prévalent. Les fonctions sémantiques, comme la synonymie et l'antonymie jouent très souvent un rôle important. La fréquence de distribution des sens peut également être pertinente. Pour prendre en compte cette information, nous pourrions faire bénéficier les sens les plus fréquents d'un avantage au détriment des sens les plus rares. A cet effet, il serait possible de biaiser la distribution de ressources initiales de manière à ce que l'émergence des sens les plus rares n'ait lieu que lorsque le contexte les favorise fortement.

7.8.4 Comparaison de notre travail avec d'autres travaux connexes

Plusieurs approches qui s'intéressent au problème de la désambiguïsation lexicale de la langue arabe ont été présentés au cours des dernières années. Le résultat de ces méthodes ne peut être comparé directement à ceux de notre approche parce que ces méthodes ont été utilisées pour d'autres tâches et leurs résultats ont été obtenus par l'utilisation des différents ensembles des informations.

Pour cette raison, on a opté pour une comparaison dans laquelle on a utilisé les mêmes textes déjà utilisé pour ces approches. Les résultats sont illustrés dans le tableau suivant (tableau 7.8).

Tableau 7-8 Comparaison de nos résultats avec ceux des autres méthodes

Méthodes	**Score (%)**
ACARWSD : Notre approche proposée	80
algorithme Génétique (Menai, 2014a, 2014b; Menai & Alsaeedan, 2012)	78.9
Classifieur naïf bayes (Zouaghi, Merhbene, & Zrigui, 2011)	76.6
Version modifiée du Lesk (Zouaghi, Merhbene, & Zrigui, 2011)	67
Le travail de Mona Diab (Diab, 2004)	56,9

Nous remarquons que notre approche donne un taux de 80% qui se rapproche de taux manuel 100%. Ce résultat est le meilleur que les autres approches, nous pouvons dire que le fait de combiner l'algorithme local LESK et l'algorithme global Colonie de fourmi peut être à la base de ce taux, car les autres travaux sont juste basés soit sur l'algorithme local ou un classifieur.

7.9 Conclusion

Dans ce chapitre, une présentation détaillé de ce que nous avons réalisé, en commençant par la conception puis l'implémentation de système.

L'objectif de cette implémentation est basé sur la performance et l'efficacité d'algorithme de désambiguïsation sémantique non-supervisé (qui est l'algorithme de fourmi) et l'algorithme Lesk qui a été utilisé comme une mesure locale.

Les critères de performance sont basés sur deux facteurs : le temps et la qualité qui donnent un meilleure résultat.

Enfin, du point de vue de l'évaluation, le résultat obtenu par notre système par apport à la désambiguïsation sémantique manuelle est environt de 80%. Concernant les mots non étiquetés, la structure de notre base lexicale sera améliorée.

Conclusion générale

« Ce n'est pas la fin. Ce n'est même pas le commencement de la fin. Mais, c'est peut-être la fin du commencement »
Winston Churchill.

Il était bénéfique pour nous d'avoir vécu l'expérience de la mise en place d'un système de désambiguïsation lexicale pour la langue arabe, malgré les difficultés que nous avons été confrontés, dont les plus remarquables est la rareté des études et des approches qui proposent des solutions intégrées (théoriquement et pratiquement) dans le domaine de sémantique et le manque de ressources lexicales, en plus des problèmes techniques liés à la programmation, imposées par la nature de la langue arabe.

Pour ce faire, nous avons fixé nos objectifs au début et nous avons organisé notre travail selon trois étapes principales. D'abord, étude théorique sur le domaine, nous avons récolté et préparé toutes les données linguistiques nécessaires : corpus de travail, base lexicale.

Ensuite, nous avons développé notre système en utilisant la méthode de Lesk comme une mesure locale et l'algorithme de fourmi comme un algorithme global pour lever l'ambiguïté. Enfin, nous avons terminé par une évaluation quantitative et qualitative de notre système.

Il nous apparut que l'approche de désambigüisation suivi, est une étape cruciale pour obtenir par la suite un étiquetage de bonne qualité. Les expériences que nous avons effectuées ont montré que les variantes et le choix des méthodes de désambiguïsation assez simple, comme l'algorithme de Lesk et l'algorithme de fourmi, pourrait produire des résultats comparables à d'autres techniques, plus compliquées ou nécessitant des ressources couteuses ou difficiles à construire.

Il est bien entendu que le type idéal de ressource capable de fournir une information suffisante, mais la discrimination correcte des sens reste encore questionnable, douteux et limité en pratique.

Enfin, nous avons obtenu des résultats encourageants au niveau de l'étiquetage sémantique qui augmentent notre désir de continuer à travailler, comme l'a montré l'évaluation qualitative et quantitative effectuée sur notre système.

D'un point de vue général, le traitement automatique de la langue arabe et en particulier la désambiguïsation sémantique reste un domaine très ouvert et présente des marges de progression importantes, du fait des caractéristiques de l'arabe, la complexité et la richesse morphologique. Il reste encore beaucoup de travail dans l'avenir afin d'introduire la langue arabe dans le monde du TALN de portes plus larges, comme c'est le cas avec autres langues.

Afin d'améliorer la performance de notre approche, nous proposons dans nos futurs travaux de recherches ce qui suit :

- L'enrichissement automatique des bases lexicales peut aider à construire un dictionnaire pour couvrir un maximum des définitions, ce qui mène à améliorer considérablement la performance du système de désambiguïsation ;
- L'architecture globale de l'approche que nous proposons est adaptée pour être étendue à d'autres castes de fourmis. Dans le prototype que nous avons développé jusqu'à présent, une seule caste de fourmi est présente, s'occupant d'informations thématiques par le biais des vecteurs.
- L'utilisation d'un autre algorithme local pour améliorer les performances de notre système.

Bibliographie

Abeillé, A. (1998). Grammaires génératives et grammaires d'unification. *Revue Langages, No. 129*, 24-36.

Abouenour, L., Bouzoubaa, K., & Rosso, P. (2008). Construction de l'ontologie Amine Arabic WordNet dans le cadre des systèmes Q/R. *Proc. 2nd JOurnées Scientifiques en Technologies de l'Information et de la Communication JOSTIC-2008, Rabat, Marroco*.

Agirre, E., & Martinez, D. (2000). *Exploring automatic word sense disambiguation with decision lists and the Web.* Paper presented at the Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content.

Agirre, E., & Rigau, G. (1996). *Word sense disambiguation using conceptual density.* Paper presented at the Proceedings of the 16th conference on Computational linguistics-Volume 1.

Apidianaki, M. (2008). *Acquisition automatique de sens pour la désambiguïsation et la sélection lexicale en traduction.* thèse de doctorat, Université Paris-Diderot-Paris VII.

Atserias, J., Climent, S., Farreres, X., Rigau, G., & Rodriguez, H. (2000). Combining multiple methods for the automatic construction of multilingual WordNets. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 327-340.

Audibert, L. (2003a). *Étude des critères de désambiguïsation sémantique automatique: résultats sur les cooccurrences.* Paper presented at the 10ème conférence sur le Traitement Automatique des Langues Naturelles (TALN-2003).

Audibert, L. (2003b). *Outils d'exploration de corpus et désambiguïsation lexicale automatique.* Doctoral dissertation, Université de Provence-Aix-Marseille I.

Azzoug, W. (2014). *Contribution à la définition d'une approche d'indexation sémantique de documents textuels.* Doctoral dissertation.

Baker, C. F. (2009). La sémantique des cadres et le projet FrameNet: une approche différente de la notion de «valence». *Langages*(4), 32-49.

Baker, C. F., Fillmore, C. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography, 16*(3), 281-296.

Bakhouche, A., & Yamina Tlili-Guiassa. (2013). Extraction des relations sémantiques à partir du Wiktionnaire Arabe. *Revue RIST| Vol, 20*(2), 47.

Bakhouche, A., & Yamina, T.-G. (2012). Meaning representation for automatic indexing of Arabic texts. [Article]. *International Journal of Computer Science Issues, 9*(6), 173-178.

Bakhouche, A., & Yamina, T. (2011). التمثيل الدلالي للغة العربية. Paper presented at the le contenu numérique en langue arabe dans le système d´administration électronique, Algérie

Bakhouche, A., & Yamina, T. (2012). الفهرسة الآلية للنصوص العربية. Paper presented at the Proceedings of the 4th International Conference on Arabic Language Processing, Rabat, Maroc.

Bakhouche, A., Yamina, T., Schwab, D., & Tchechmedjiev, A. (2015). Ant colony algorithm for Arabic word sense disambiguation through English lexical information. *International Journal of Metadata, Semantics and Ontologies, 10*(3), 202-211.

Banerjee, S., & Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet *Computational linguistics and intelligent text processing* (pp. 136-145): Springer.

Barzilay, R., & Elhadad, M. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, 111-121.

Baschung, K. (1990). *Grammaires d'unification à traits et contrôle des infinitives.* Doctoral dissertation, Clermont Ferrand 2.

Belfadel, D., & Diaf, M. (2014). De la fourmi réelle à la fourmi artificielle. *Revue CAMPUS, No 2*, 22-33.

Berthiau, G., & Siarry, P. (2001). État de l'art des méthodes "d'optimisation globale". *RAIRO-Operations Research, 35*(03), 329-365.

Black, E. (1988). An experiment in computational discrimination of English word senses. *IBM Journal of research and development, 32*(2), 185-194.

Bouillon, P. (1998). *Traitement automatique des langues naturelles*: De Boeck Supérieur, Book.

Bruce, R., & Wiebe, J. (1994a). *A new approach to word sense disambiguation.* Paper presented at the Proceedings of the workshop on Human Language Technology.

Bruce, R., & Wiebe, J. (1994b). *Word-sense disambiguation using decomposable models.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Buckwalter, T. (2002). Arabic Morphological Analyzer (AraMorph), software.

Buscaldi, D., & Rosso, P. (2008). A conceptual density-based approach for the disambiguation of toponyms. *International Journal of Geographical Information Science, 22*(3), 301-313.

Cailliau, F. (2010). *Des ressources aux traitements linguistiques: le rôle d 'une architecture linguistique.* Doctoral dissertation, Université Paris-Nord-Paris XIII.

Candito, M.-H. (1996). *A principle-based hierarchical representation of LTAGs.* Paper presented at the Proceedings of the 16th conference on Computational linguistics-Volume 1.

Candito, M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées.* Doctoral dissertation, Université Paris 7.

Carpuat, M., & Wu, D. (2007). *Improving Statistical Machine Translation Using Word Sense Disambiguation.* Paper presented at the Proceedings of the EMNLP-CoNLL.

Chan, Y. S., Ng, H. T., & Chiang, D. (2007). *Word sense disambiguation improves statistical machine translation.* Paper presented at the Annual Meeting-Association for Computational Linguistics.

Chaumartin, F.-R. (2007). *WordNet et son écosysteme: un ensemble de ressources linguistiques de large couverture*. Paper presented at the In Colloque BD lexicales.

Cowie, J., Guthrie, J., & Guthrie, L. (1992). *Lexical disambiguation using simulated annealing.* Paper presented at the Proceedings of the 14th conference on Computational linguistics-Volume 1.

Cunningham, H., & Scott, D. (2004). Software architecture for language engineering. *Natural Language Engineering, 10*(3-4), 205-209.

de Loupy, C., El-Beze, M., & Marteau, P.-F. (1998). *Word sense disambiguation using HMM tagger.* Paper presented at the Proceedings of LREC.

de Loupy, C., El-Bèze, M., & Marteau, P. (1998). *WSD based on three short context methods.* Paper presented at the SENSEVAL Workshop, Herstmontceux.

Desmarais, L. (1994). *Proposition d'une didactique de l'orthographe ayant recours au correcteur orthographique (Proposal for a Teaching Methodology for Spelling Using a Spell-Checker)*: ERIC, Book.

Diab, M. T. (2004). *An unsupervised approach for bootstrapping Arabic sense tagging.* Paper presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.

Dréo, J., Pétrowski, A., Taillard, É. D., & Siarry, P. (2003). Métaheuristiques pour l\'optimisation difficile.

Dubois, J. (1969). Grammaire générative et transformationnelle. *Langue française*(1), 49-57.

Eaton, H. S. (1940). *Semantic frequency list for English, French, German, and Spanish: a correlation of the first six thousand words in four single-language frequency lists*: The University of Chicago Press, Book.

Ehrig, M., Haase, P., Hefke, M., & Stojanovic, N. (2005). Similarity for ontologies-a comprehensive framework. *ECIS 2005 Proceedings*, 127.

El-Beze, M., Michelon, P., & Pernaud, R. (1999). An integer programming approach to word sense disambiguation. *Submitted to European Journal of Operational Research (EJOR-1999)*.

Elmougy, S., Taher, H., & Noaman, H. (2008). *Naïve Bayes classifier for Arabic word sense disambiguation.* Paper presented at the proceeding of the 6th International Conference on Informatics and Systems.

Escudero, G., Màrquez, L., & Rigau, G. (2000a). *A comparison between supervised learning algorithms for word sense disambiguation.* Paper presented at the Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7.

Escudero, G., Màrquez, L., & Rigau, G. (2000b). Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. *arXiv preprint cs/0007011*.

Fellbaum, C. (1998). *WordNet*: Wiley Online Library, Book.

Firer-Blaess, S. (2007). Wikipédia: le refus du pouvoir *Mémoire de l'IEP, Université Lyon, 2.*

Francoise Forest, L. N., Gerard Sabah & Anne Vilnat, , . (1988). Une etude sur l'integration de l'utilisateur dans un dialogue homme-machine en langage naturel, notes et documents *LIMSI, LIMSI-CNRS*(12).

Gardent, C., Guillaume, B., Falk, I., & Perrier, G. (2005). Le lexique-grammaire de M. Gross et le traitement automatique des langues. *LORIA & ATILF, 55*.

Gazdar, G. (1985). *Generalized phrase structure grammar*: Harvard University Press, Book.

Gazdar, G., & Pullum, G. K. (1982). *Generalized phrase structure grammar: a theoretical synopsis* (Vol. 7): Indiana University Linguistics Club Bloomington, Book ,Indiana, .

Gelbukh, A., Sidorov, G., & Han, S.-Y. (2003). Evolutionary approach to natural language word sense disambiguation through global coherence optimization. *WSEAS Transactions on Computers, 2*(1), 257-265.

Gibbons, F. X., Gerrard, M., Blanton, H., & Russell, D. W. (1998). Reasoned action and social reaction: willingness and intention as independent predictors of health risk. *Journal of personality and social psychology, 74*(5), 1164.

Halliday, M. A., & Hasan, R. (1976). Cohesion in. *English, Longman, London*.

Hammersley, J. (2013). *Monte carlo methods*: Springer Science & Business Media, Book.

Harabagiu, S., Miller, G., & Moldovan, D. (1999). *Wordnet 2-a morphologically and semantically enhanced resource.* Paper presented at the Proceedings of SIGLEX.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database, 305*, 305-332.

Holland. (1975). Adaptation in Natural and Artificial Systems. *University of Michigan Press :*

Ann Arbor.

Ide, N., & Véronis, J. (1995). Knowledge extraction from machine-readable dictionaries: An evaluation *Machine translation and the lexicon* (pp. 17-34): Springer.

Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics, 24*(1), 2-40.

Imbs, P., & QUEMADA, B. (1971). Trésor de la langue française. Dictionnaire de la langue du XIX e et du XX e siècles: Paris, Éditions du Centre National de la Recherche Scientifique, 1296-1297.

Jacquet, G. (2005). *Polysémie verbale et calcul du sens.* Doctoral dissertation, Paris, EHESS.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Johansson, S. (1980). The LOB corpus of British English texts: presentation and comments. *ALLC journal, 1*(1), 25-36.

Joshi, A. K. (1987). An introduction to tree adjoining grammars. *Mathematics of language, 1*, 87-115.

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars *Handbook of formal languages* (pp. 69-123): Springer.

Jousse, A.-L. (2010). Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales.

Kaplan, R. M., & Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. *Formal Issues in Lexical-Functional Grammar*, 29-130.

Kelly, E. F., & Stone, P. J. (1975). *Computer recognition of English word senses* (Vol. 13): Book, North-Holland.

Kenter, T., & Maynard, D. (2005). Using gate as an annotation tool. *University of Sheffield, Natural language processing group*.

Kilgarriff, A., & Rosenzweig, J. (2000). *English Senseval: Report and Results.* Paper presented at the LREC.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *science, 220*(4598), 671-680.

Kolhatkar, V. (2009). *An extended analysis of a method of all words sense disambiguation.* Doctoral dissertation, Citeseer.

Kray-Baschung, K. (1992). Grammaires d'unification à traits et contrôle des infinitives en français.

Ku, H., & Francis, W. N. (1967). Computational Analysis of Present-Day {A} merican {E} nglish.

Lapata, M., & Brew, C. (1999). *Using subcategorization to resolve verb class ambiguity.* Paper presented at the Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Lapata, M., & Brew, C. (2004). Verb class disambiguation using informative priors. *Computational Linguistics, 30*(1), 45-73.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database, 49*(2), 265-283.

Lecomte, A. (2007). Sémantique. [Cours de Licence de Sciences du Langage].

Lesk, M. (1986). *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.* Paper presented at the Proceedings of the 5th annual international conference on Systems documentation.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*: University of Chicago press, Book.

Lin, D. (1998). *An information-theoretic definition of similarity.* Paper presented at the ICML.

Loiseau, S., Gréa, P., & Magué, J.-P. (2011). Dictionnaires, théorie des graphes et structures lexicales. *Revue de Sémantique et Pragmatique*(27), 51--78.

Lönneker-Rodman, B., & Baker, C. F. (2009). The FrameNet model and its applications. *Natural Language Engineering, 15*(03), 415-453.

Magnini, B., Strapparava, C., Ciravegna, F., & Pianta, E. (1994). *A Project for the Constraction of an Italian Lexical Knowledge Base in the Framework of Wordnet*: Istituto per la Ricerca Scientifica e Tecnologica, Book.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*: MIT press, Book.

Menai, M. E. B. (2014a). Word sense disambiguation using an evolutionary approach. *Informatica, 38*(2), 155.

Menai, M. E. B. (2014b). Word sense disambiguation using evolutionary algorithms–Application to Arabic language. *Computers in Human Behavior, 41*, 92-103.

Menai, M. E. B., & Alsaeedan, W. (2012). *Genetic algorithm for Arabic word sense disambiguation.* Paper presented at the Software Engineering, Artificial Intelligence,

Networking and Parallel & Distributed Computing (SNPD), 2012 13th ACIS International Conference on.

Meyer, C. M., & Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. *Electronic Lexicography*, 259-291.

Mihalcea, R., & Chklovski, T. (2004). Building sense tagged corpora with volunteer contributions over the web. *Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, 260*, 357.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Mooney, R. J. (1996). Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *arXiv preprint cmp-lg/9612001*.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics, 17*(1), 21-48.

Mouelhi, Z. (2008). AraSeg.*.: un segmenteur semi-automatique des textes arabes. *Actes des JADT 2008*, 867-878.

Nameh, M., Fakhrahmad, S., & Jahromi, M. Z. (2011). *A New Approach to Word Sense Disambiguation Based on Context Similarity.* Paper presented at the Proceedings of the World Congress on Engineering.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., ShuKai, H., Tzu-Yi, K., . . . Chu-Ren, H. (2009). *Wiktionary and NLP: Improving synonymy networks.* Paper presented at the Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR), 41*(2), 10.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review, 45*(2), 167-256.

Ng, H. T. (1997). Exemplar-based word sense disambiguation: Some recent improvements. *arXiv preprint cmp-lg/9706010*.

Ng, H. T., & Lee, H. B. (1996). *Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach.* Paper presented at the Proceedings of the 34th annual meeting on Association for Computational Linguistics.

Nguyen, T. M. H. (2006). *Outils et ressources linguistiques pour l'alignement de textes multilingues français-vietnamiens.* Doctoral dissertation, Université Henri Poincaré-Nancy I.

Nichols, E., Bond, F., & Flickinger, D. (2005). *Robust ontology acquisition from machine-readable dictionaries.* Paper presented at the IJCAI.

Palmer, H. (1933). 2nd interim report on english collocations. *Institute for Research in English Teaching. Tokyo*.

Patwardham, S. (2003). Incorporating dictionary and corpus information in a measure of semantic relatedness. *MS Ph. D Thesis, University of Minnesota, Duluth, MN*.

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003). Using measures of semantic relatedness for word sense disambiguation *Computational linguistics and intelligent text processing* (pp. 241-257): Springer.

Pedersen, T. (2001). *Machine learning with lexical features: The duluth approach to senseval-2.* Paper presented at the The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems.

Pedersen, T. (2002). A baseline methodology for word sense disambiguation *Computational Linguistics and Intelligent Text Processing* (pp. 126-135): Springer.

Pedersen, T., & Bruce, R. (1997). *A new supervised learning algorithm for word sense disambiguation.* Paper presented at the AAAI/IAAI.

Pedersen, T., Bruce, R., & Wiebe, J. (1997). *Sequential model selection for word sense disambiguation.* Paper presented at the Proceedings of the fifth conference on Applied natural language processing.

Pereira, F., Tishby, N., & Lee, L. (1993). *Distributional clustering of English words.* Paper presented at the Proceedings of the 31st annual meeting on Association for Computational Linguistics.

Pérennou, G., & De Calmes, M. (2000). *MHATLex: Lexical Resources for Modelling the French Pronunciation.* Paper presented at the LREC.

Pérez, L. A., Oliveira, H. G., & Gomes, P. (2011). *Extracting lexical-semantic knowledge from the portuguese wiktionary.* Paper presented at the Proceedings of the 15th Portuguese Conference on Artificial Intelligence, EPIA.

Pierce, J. R., & Carroll, J. B. (1966). Language and machines: Computers in translation and linguistics.

Pincemin, B. (2004). *Lexicométrie sur corpus étiquetés.* Paper presented at the 7es Journées internationales d'analyse statistique des données textuelles (JADT 2004).

Piron, S. (2004). Contraintes syntaxiques et préférences sélectionnelles du verbe entendre. *Actes des Journées d'Analyse statistique de Données Textuelles*.

Polguère, A. (1998). La théorie sens-texte. *1998a, Dialangue, 8*, 9.

Polguère, A. (2003). *Lexicologie et sémantique lexicale: notions fondamentales*: Pum, Book.

Polguère, A., & Tremblay, O. (2003). Qu'y a-t-il à l'intérieur de NOIX? Ou comment décortiquer les unités lexicales. *La Lettre de l'AIRDF, 33*, 2003-2003.

Pollard, C., & Sag, I. (1987). Information-based Syntax end Semantics, CSLI Series: University of Chicago Press, Book.

Prolo, C. A. (2002). *Generating the XTAG English grammar using metarules.* Paper presented at the Proceedings of the 19th international conference on Computational linguistics-Volume 1.

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

Rakho, M., Pitel, G., & Mouton, C. (2008). Désambiguïsation automatique à partir d'espaces vectoriels multiples clutérisés. *rapport intermédiaire université paris*.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Rosso, P., Masulli, F., Buscaldi, D., Pla, F., & Molina, A. (2003). Automatic noun sense disambiguation *Computational Linguistics and Intelligent Text Processing* (pp. 273-276): Springer.

Roussarie, L. (2000). *Un modèle théorique d'inférence de structures sémantiques et discursives dans le cadre de la génération automatique de textes.* Doctoral dissertation, Paris 7.

Roussel, F. (2007). Comprendre un blog chinois, lire un site arabe, tchater en allemand... Internet fait rêver d'un monde sans barrière de langues et relance la recherche sur les systèmes de traduction automatique. *Babel Web*.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). FrameNet II: Extended theory and practice.

Sadat, F., & Terrasa, A. (2010). *Exploitation de wikipédia pour l'enrichissement et la construction des ressources linguistiques.* Paper presented at the Proceedings of TALN.

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Sanderson, M. (1994). *Word sense disambiguation and information retrieval.* Paper presented at the Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval.

Schütze, H. (1992). *Dimensions of meaning.* Paper presented at the Supercomputing'92., Proceedings.

Schütze, H. (1998). Automatic word sense discrimination. *Computational linguistics, 24*(1), 97-123.

Schwab, D. (2005). *Approche hybride-lexicale et thématique-pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* Doctoral dissertation, Université Montpellier II-Sciences et Techniques du Languedoc.

Schwab, D., Goulian, J., & Guillaume, N. (2011). Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis. *Actes des conférences TALN 2011 et Recital 2011*, 185.

Schwab, D., Tchechmedjiev, A., Goulian, J., Nasiruddin, M., Sérasset, G., & Blanchon, H. (2013). GETALP: Propagation of a Lesk Measure through an Ant Colony Algorithm. *Atlanta, Georgia, USA*, 232.

Seco, N., Veale, T., & Hayes, J. (2004). *An intrinsic information content metric for semantic similarity in WordNet.* Paper presented at the ECAI.

Sells, P. (1989). Lectures on contemporary syntactic theories: an introduction to government-binding theory, generalized phrase structure grammar, and lexical-functional grammar.

Shah, R., Dhillon, P. S., Liberman, M., Foster, D., Maamouri, M., & Ungar, L. (2010). *A new approach to lexical disambiguation of Arabic text.* Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

Silber, H. G., & McCoy, K. F. (2000). *Efficient text summarization using lexical chains.* Paper presented at the Proceedings of the 5th international conference on Intelligent user interfaces.

Sinclair, J. M. (1996). The empty lexicon. *International journal of corpus linguistics, 1*(1), 99-119.

Sproat, R., Hirschberg, J., & Yarowsky, D. (1992). *A corpus-based synthesizer.* Paper presented at the ICSLP.

Stokoe, C., Oakes, M. P., & Tait, J. (2003). *Word sense disambiguation in information retrieval revisited.* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Tateisi, Y., Torisawa, K., Miyao, Y., & Tsujii, J. i. (1998). *Translating the XTAG English grammar to HPSG.* Paper presented at the Proc. of TAG.

Tchechmedjiev, A. (2012). État de l'art: mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances. *JEP-TALN-RECITAL 2012*, 295.

Tchechmedjiev, D. S. J. G. A. (2013). Désambiguïsation lexicale de textes: efficacité qualitative et temporelle d'un algorithme à colonies de fourmis.

Thorndike, E. L. (1948). On the frequency of semantic changes in modern English. *The Journal of general psychology, 39*(1), 23-27.

Torres-Moreno, J.-M. (2011). *Résumé automatique de documents*: Lavoisier, Book.

Torres-Moreno, J. M. (2011). Résumé automatique de documents: une approche statistique. *Hermes-Lavoisier*.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review, 89*(2), 123.

Uzuner, O., Katz, B., & Yuret, D. (1999). Word sense disambiguation for information retrieval. *AAAI/IAAI, 985*.

Van Rijsbergen, C. (1979). Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*.

Vasilescu, F. (2003). Désambiguïsation de corpus monolingues par des approches de type Lesk.

Vasilescu, F., & Langlais, P. (2003). *Désambiguïsation de corpus monolingues par des approches de type Lesk.* Doctoral dissertation, Université de Montréal.

Véronis, J. (2004a). Hyperlex: lexical cartography for information retrieval. *Computer Speech & Language, 18*(3), 223-252.

Véronis, J. (2004b). Quels dictionnaires pour l'étiquetage sémantique? *Français moderne, 72*(1), 27-38.

Veronis, J., & Ide, N. M. (1990). *Word sense disambiguation with very large neural networks extracted from machine readable dictionaries.* Paper presented at the Proceedings of the 13th conference on Computational linguistics-Volume 2.

Vickrey, D., Biewald, L., Teyssier, M., & Koller, D. (2005). *Word-sense disambiguation for machine translation.* Paper presented at the Proceedings of the conference on

Human Language Technology and Empirical Methods in Natural Language Processing.

Voorhees, E. M. (1993). *Using WordNet to disambiguate word senses for text retrieval.* Paper presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Wagner, W., Schmid, H., & Im Walde, S. S. (2009). *Verb sense disambiguation using a predicate-argument-clustering model.* Paper presented at the Proceedings of the CogSci Workshop on Distributional Semantics beyond Concrete Concepts.

Walker, D. E. (1986). *Knowledge resource tools for accessing large text files*: Citeseer, Book.

Weale, T., Brew, C., & Fosler-Lussier, E. (2009). *Using the wiktionary graph structure for synonym detection.* Paper presented at the Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources.

Weaver, W. (1949). {Translation}.

Weiss, S. F. (1973). Learning to disambiguate. *Information Storage and Retrieval, 9*(1), 33-41.

Wilks, Y. (1993). Providing machine tractable dictionary tools *Semantics and the Lexicon* (pp. 341-401): Springer.

Wilks, Y., Fass, D., Guo, C., MacDonald, J., & Plate, T. Slator. B.(1990). Providing Machine Tractable Dictionary Tools. *Machine Translation, 5*, 99-154.

Wilks, Y., & Stevenson, M. (1998). *Word sense disambiguation using optimised combinations of knowledge sources.* Paper presented at the Proceedings of the 17th international conference on Computational linguistics-Volume 2.

Wilks, Y., & Stevenson, M. (2000). Combining independent knowledge sources for word sense disambiguation. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 117-130.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Yarowsky, D. (1992). *Word-sense disambiguation using statistical models of Roget's categories trained on large corpora.* Paper presented at the Proceedings of the 14th conference on Computational linguistics-Volume 2.

Yarowsky, D. (1995). *Unsupervised word sense disambiguation rivaling supervised methods.* Paper presented at the Proceedings of the 33rd annual meeting on Association for Computational Linguistics.

Yarowsky, D. (1997). Homograph disambiguation in text-to-speech synthesis *Progress in speech synthesis* (pp. 157-172): Springer.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. *Computers and the Humanities, 34*(1-2), 179-186.

Yarowsky, D., Cucerzan, S., Florian, R., Schafer, C., & Wicentowski, R. (2001). *The Johns Hopkins senseval2 system descriptions.* Paper presented at the The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems.

Zaghouani, W. (2015). *Le développement de corpus annotés pour la langue arabe.* Doctoral dissertation, Université Paris 7.

Zaidi–Ayad, S. (2013). *Une plateforme pour la construction d'ontologie en arabe: Extraction des termes et des relations à partir de textes (Application sur le Saint Coran).* Doctorat Doctoral dissertation, Université Badji Mokhtar de Annaba.

Zargayouna, H., & Salotti, S. (2004). Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML. *IC 2004*, 249-260.

Zesch, T., Müller, C., & Gurevych, I. (2008). *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.* Paper presented at the LREC.

Zhong, Z., & Ng, H. T. (2012). *Word sense disambiguation improves information retrieval.* Paper presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1.

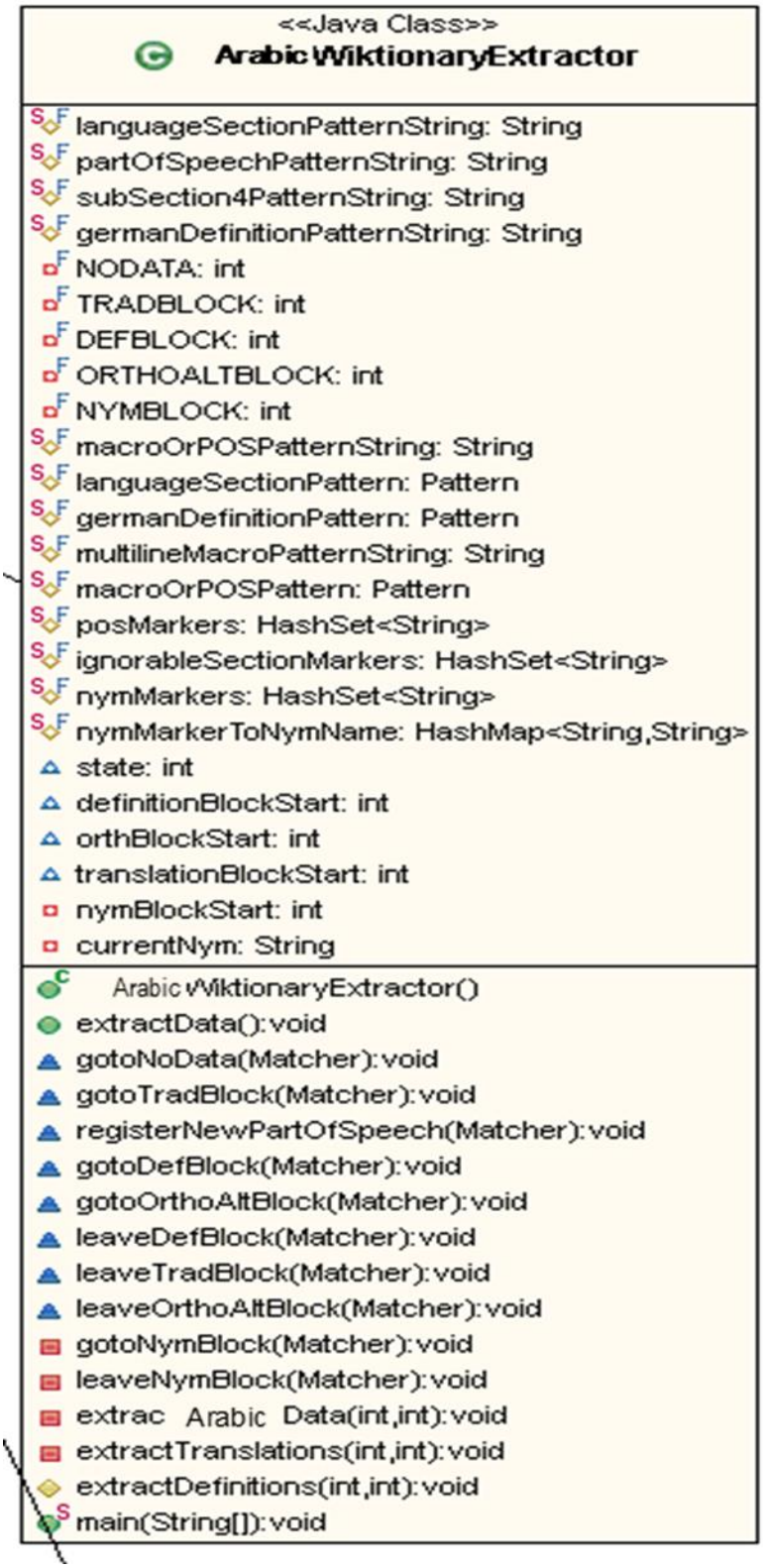
Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of general psychology, 33*(2), 251-256.

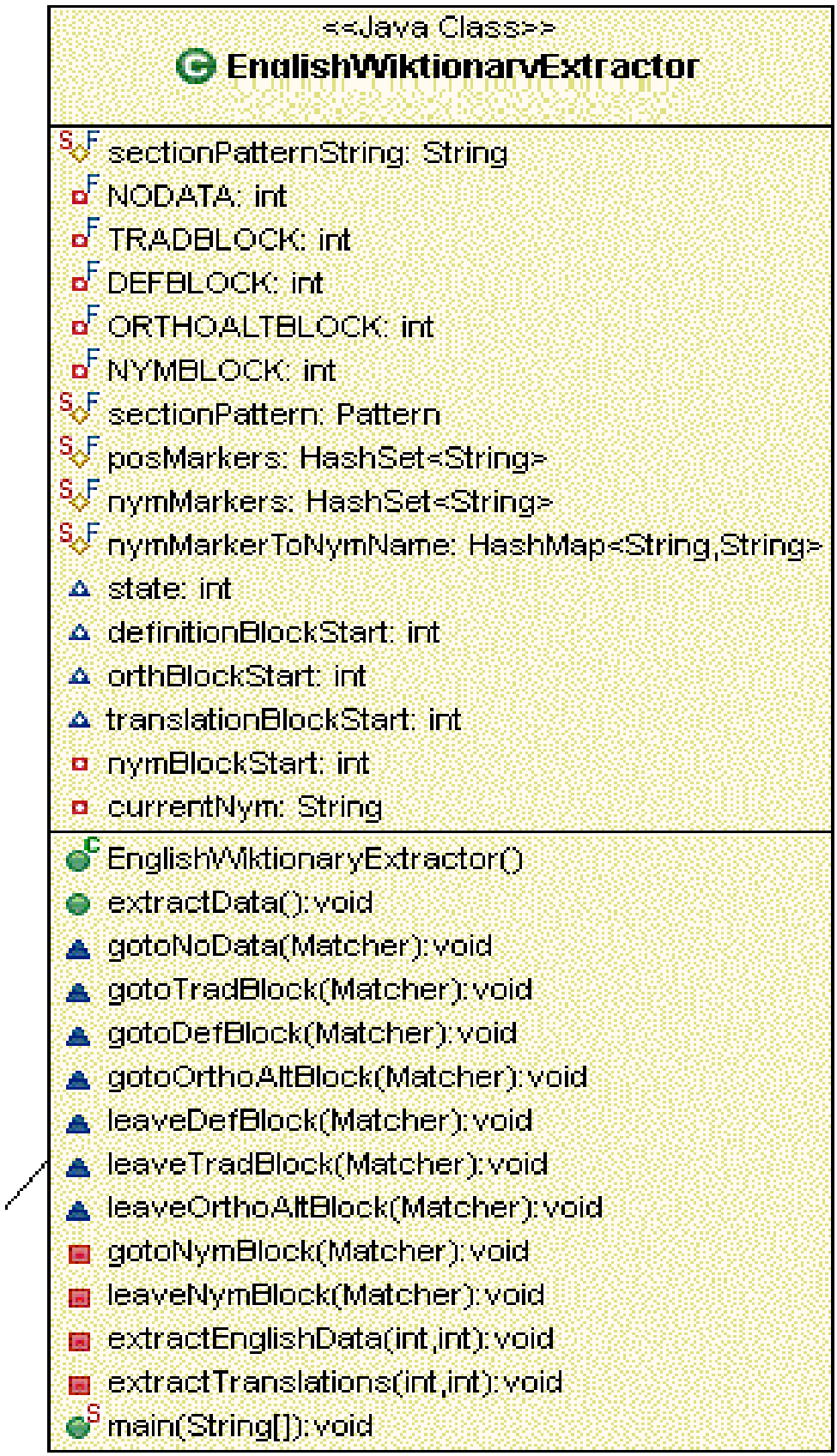
Zouaghi, A., Merhbene, L., & Zrigui, M. (2011). Word Sense disambiguation for Arabic language using the variants of the Lesk algorithm. *WORLDCOMP, 11*, 561-567.

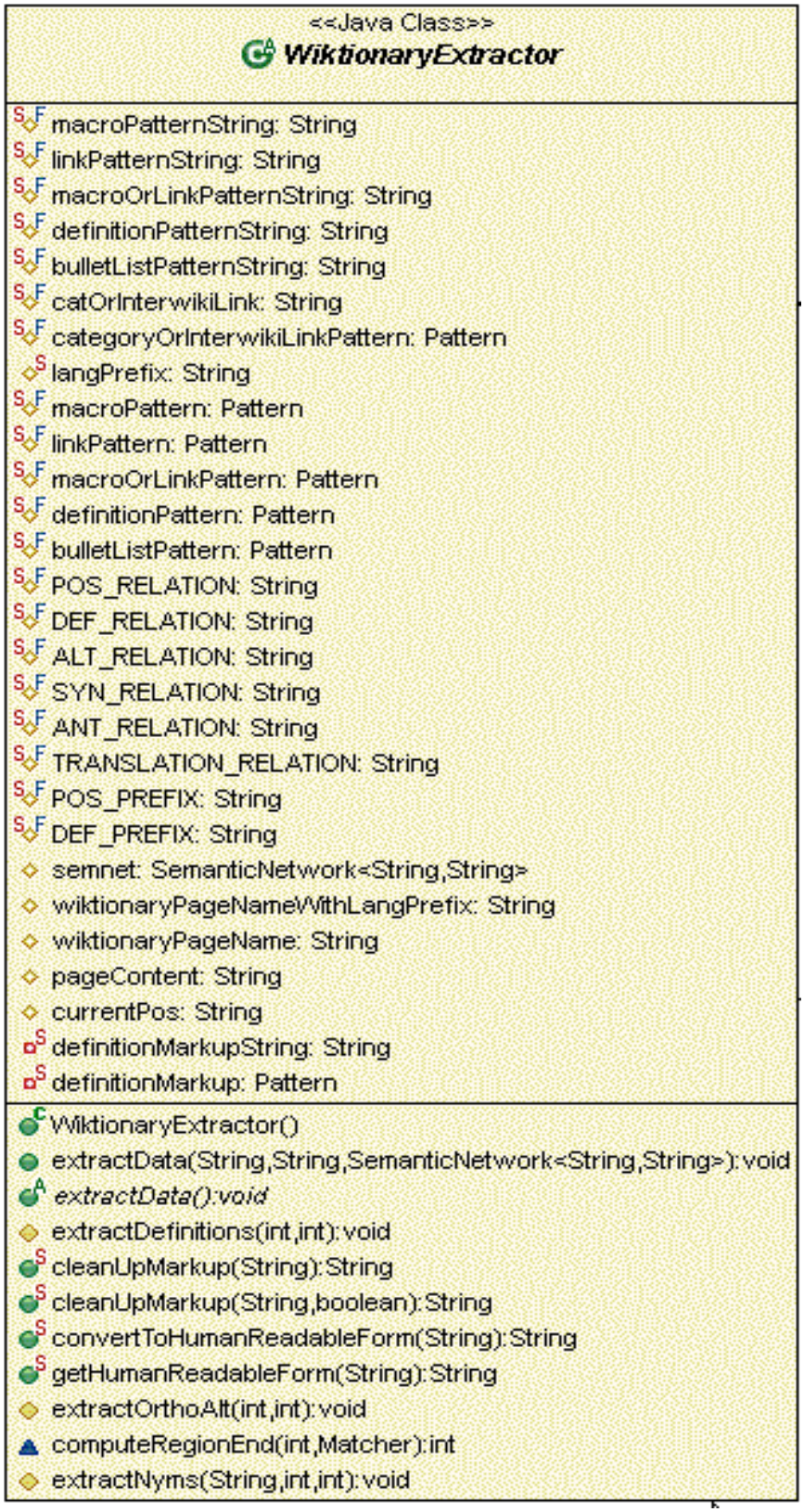
Annexe

Le diagramme de classe est le point central dans un développement orienté objet. En analyse, il a pour objectif de décrire la structure des entités manipulées par les utilisateurs. En conception, le diagramme de classes représente la structure d'un code orienté et met en évidence d'éventuelles relation entre ces classes .

On montre dans cette partie les détails de quelques classes principales du code source d'extraction des relations sémantiques de Wiktionnaire.


<<Java Class>>
ArabicWiktionaryExtractor
languageSectionPatternString: String
partOfSpeechPatternString: String
subSection4PatternString: String
germanDefinitionPatternString: String
NODATA: int
TRADBLOCK: int
DEFBLOCK: int
ORTHOALTBLOCK: int
NYMBLOCK: int
macroOrPOSPatternString: String
languageSectionPattern: Pattern
germanDefinitionPattern: Pattern
multilineMacroPatternString: String
macroOrPOSPattern: Pattern
posMarkers: HashSet<String>
ignorableSectionMarkers: HashSet<String>
nymMarkers: HashSet<String>
nymMarkerToNymName: HashMap<String,String>
state: int
definitionBlockStart: int
orthBlockStart: int
translationBlockStart: int
nymBlockStart: int
currentNym: String
ArabicWiktionaryExtractor()
extractData():void
gotoNoData(Matcher):void
gotoTradBlock(Matcher):void
registerNewPartOfSpeech(Matcher):void
gotoDefBlock(Matcher):void
gotoOrthoAltBlock(Matcher):void
leaveDefBlock(Matcher):void
leaveTradBlock(Matcher):void
leaveOrthoAltBlock(Matcher):void
gotoNymBlock(Matcher):void
leaveNymBlock(Matcher):void
extrac Arabic Data(int,int):void
extractTranslations(int,int):void
extractDefinitions(int,int):void
main(String[]):void


<<Java Class>>
EnglishWiktionaryExtractor
sectionPatternString: String
NODATA: int
TRADBLOCK: int
DEFBLOCK: int
ORTHOALTBLOCK: int
NYMBLOCK: int
sectionPattern: Pattern
posMarkers: HashSet<String>
nymMarkers: HashSet<String>
nymMarkerToNymName: HashMap<String,String>
state: int
definitionBlockStart: int
orthBlockStart: int
translationBlockStart: int
nymBlockStart: int
currentNym: String
EnglishWiktionaryExtractor()
extractData():void
gotoNoData(Matcher):void
gotoTradBlock(Matcher):void
gotoDefBlock(Matcher):void
gotoOrthoAltBlock(Matcher):void
leaveDefBlock(Matcher):void
leaveTradBlock(Matcher):void
leaveOrthoAltBlock(Matcher):void
gotoNymBlock(Matcher):void
leaveNymBlock(Matcher):void
extractEnglishData(int,int):void
extractTranslations(int,int):void
main(String[]):void


<<Java Class>>
WiktionaryExtractor
macroPatternString: String
linkPatternString: String
macroOrLinkPatternString: String
definitionPatternString: String
bulletListPatternString: String
catOrInterwikiLink: String
categoryOrInterwikiLinkPattern: Pattern
langPrefix: String
macroPattern: Pattern
linkPattern: Pattern
macroOrLinkPattern: Pattern
definitionPattern: Pattern
bulletListPattern: Pattern
POS_RELATION: String
DEF_RELATION: String
ALT_RELATION: String
SYN_RELATION: String
ANT_RELATION: String
TRANSLATION_RELATION: String
POS_PREFIX: String
DEF_PREFIX: String
semnet: SemanticNetwork<String,String>
wiktionaryPageNameWithLangPrefix: String
wiktionaryPageName: String
pageContent: String
currentPos: String
definitionMarkupString: String
definitionMarkup: Pattern
WiktionaryExtractor()
extractData(String,String,SemanticNetwork<String,String>):void
extractData():void
extractDefinitions(int,int):void
cleanUpMarkup(String):String
cleanUpMarkup(String,boolean):String
convertToHumanReadableForm(String):String
getHumanReadableForm(String):String
extractOrthoAlt(int,int):void
computeRegionEnd(int,Matcher):int
extractNyms(String,int,int):void