REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université Mouloud Mammeri de Tizi ouzou
Faculté de génie électrique et d'informatique

Mémoire

Présenté par : **Saada Syphax**

Pour obtenir le grade de :

MASTER EN INFORMATIQUE

Option : Système Informatique

Thème :

Construction d'une base de connaissances terminologiques

Proposé et dirigé par : Aoughlis Farida

Devant le jury suivant :

Le président de jury : Mr Oularbi Aomar.

L'examinateur 1 : Mr Habet M^{ed} Said.

L'examinateur 2 : Mr Khemliche Salem.

2010/2011

Remerciements

Grâce à Dieu vers lequel vont toutes les louanges, ce travail s'est accompli.

J'exprime ma gratitude à madame Aoughlis Farida d'avoir proposé et dirigé ce mémoire, grâce à sa disponibilité, sa patience, son immense gentillesse et ses conseils. Elle a su me faire profiter de ses nombreuses connaissances et compétences dans ce domaine, je garderai l'exemple.

J'ai l'honneur d'inscrire ici un immense remerciement à mes parents pour leur irremplaçable et inconditionnel soutien.

Je présente mes gratitudes aux membres du jury qui ont bien voulu examiner et évaluer mon travail et qui m'ont fait l'honneur de participer à la soutenance, qu'ils soient rassurés de mon plus profond respect.

Mes pensées vont à mes amis, pour tout ce qu'ils m'apportent, mais surtout pour leur amitié.

Enfin, je tiens à exprimer ma profonde affection à ma famille, qui m'a toujours soutenu.

Je remercie .également tous ceux qui ont contribué à ce travail.

Dédicaces

A mes chers parents

A mes deux chères Sarah

A mon frère Kamel

A toute ma famille

Syphax

Table des matières

Chapitre II :

L'extraction de terminologie

Chapitre III :

Acquisition automatique des connaissances

Chapitre IV :

Construction d'une base de connaissances terminologique

Liste des tableaux :

Liste des figures :

Introduction générale

Introduction générale :

L'avènement d'Internet a créé un besoin accru d'accès et de traitement de l'information.

En effet, la diversité des domaines (commercial, social, industriel, littéraire, journalistique, technique/spécialisé) et surtout les caractéristiques des textes (longs, télégraphiques) rendent le Traitement Automatique du Langage Naturel (TAL) particulièrement complexe.

Cependant, le traitement automatique des langues (TAL) est presque aussi ancien que l'outil qui l'a rendu possible.

Après l'apparition des premiers véritables ordinateurs électroniques binaires, une large gamme de programmes de recherche sur la traduction automatique ont été mis sur pied.

Et depuis de nombreuses années, certains travaux ont permis la constitution de bases terminologiques dans des domaines de référence à l'aide de certains outils utilisés qui permettent à plusieurs équipes de recherche de mener des études sur l'acquisition terminologique dans le but de construire des bases de connaissances terminologiques.

Dans le cadre de nos recherches, nous proposons une étude et des méthodes d'acquisition de connaissances et d'extraction automatique de terminologie.

Plan du travail :

Le premier chapitre de notre travail sera consacré à une étude linguistique assez complète de la langue française, qui donnera une vue synthétique sur la naissance de lettres en terme d'alphabet, puis nous donnons un bref aperçu sur la synoptique de classification de mots enfin nous achevons ce chapitre par une présentation générale des phrases.

Passant au second chapitre qui se rapportera sur l'extraction de terminologie en entamant en premier lieu les procédures d'extraction terminologique, puis en second lieu nous introduisons le concept d'outils utilisés lors de l'extraction automatique de terminologie

accompagné par la suite de diverses approches d'extraction terminologique. Nous présentons en dernier lieu les différents problèmes rencontrés lors de ces approches.

Dans le troisième chapitre, nous abordons le concept d'acquisition de connaissance et présentons les outils utilisés. Puis les différentes méthodes d'acquisition de connaissance et enfin nous concluons ce chapitre par le concept d'acquisition automatique tout en l'illustrant par un système adéquat à cette dernière.

Finalisant le travail par un dernier chapitre qui va nous permettre de mettre en pratique les recherches faites dans les chapitres précédents, et choisir une méthode retenue afin de construire une Base de Connaissances Terminologiques (BCT) qui puisse représentée un corpus.

Chapitre I

Etude linguistique

I. Introduction :

Dans le domaine de l'intelligence artificielle, le traitement des langues naturelles ainsi que la linguistique sont vues comme étant deux domaines complètement inséparables.

En effet, l'étude linguistique permet de faire une étudier sur le fonctionnement du langage redouté à travers toutes les langues parlées[1]. La linguistique s'intéresse à tous types de langues en commençant par son évolution historique puis sa construction grammaticale et par la suite ses apports, ….

Quand nous parlons d'un linguiste, cela ne voudrait pas dire automatiquement qu'il est polyglotte c'est à dire qui parle toute les langues. Cela voudrait dire qu'il va entamer une recherche sur les origines d'utilisation de ce langage, de règles ou d'erreurs commises, et comprendre par exemple que derrière une faute se cache un besoin de rationalisation (de logique) de la langue, car ce n'est ni une condition nécessaire ni suffisante de maîtriser de manière efficace cette langue pour étudier et décrire certains aspects de son fonctionnement.

En effet, La linguistique inclut un certain nombre de domaines de recherche cités comme suit:

- La phonétique qui est l'étude des sons.
- La phonologie qui se résume à l'organisation des sons dans une langue pour former des énoncés.
- La dialectologie qui se résume à l'étude des dialectes.
- La syntaxe qui se rapporte à l'étude de la combinaison des mots pour former des phrases.
- L'étymologie qui est l'étude de l'origine des mots.
- La linguistique comparée : qui est l'étude de l'histoire et d'évolution des langues ou de groupes de langues.
- La typologie des langues: qui est le classement en type et en famille de langue.

  Et l'orthographe, …

[1] http://www.studyrama.be/spip.php?article274.

Dans ce chapitre, nous allons focaliser notre recherche sur l'étude linguistique de la langue Française.

Vu la richesse de ce domaine d'étude, il nous est difficile de faire une étude complète, donc nous sommes amenés à présenter une vue globale concernant l'étude linguistique en commençant par les mots, et montrant par la suite les différentes catégories grammaticales.

II. Concepts généraux :

II.1. L'alphabet :

L'alphabet est vu comme étant un ensemble de lettres utilisées dans un système d'écriture, dans le but de reproduire des sons.

Cet alphabet comporte quatre-vingt-cinq lettres dont leur combinaison constitue le langage [Anonyme, 1829].

II.2. Le mot :

II.2.1. Définition d'un mot :

Le mot dans le langage courant est une succession de sons ou de caractères graphiques formant une unité sémantique et pouvant être distingués par un séparateur qui peut être à l'écrit un blanc typographique, et à l'oral une pause.

Cette unité la est autonome et peut être utilisée dans diverses combinaisons des énoncés [Petit Larousse, 1990].

En effet, le mot possède une fonction dans la phrase et ne peut se diviser en unités plus petites répondant à la même définition [Grevisse, 1980].

II.2.2. La Synoptique de la classification des mots selon leur nature :

II.2.2.1. Les mots grammaticaux :

Un mot grammatical (mot-outil) fait parti d'une catégorie de mots dans le quel le rôle syntaxique est plus important que le rôle sémantique.

Toutefois, les mots grammaticaux se répartissent entre, les mots grammaticaux variables (déterminants et pronoms), et les mots grammaticaux invariables (mots de liaison).

Quelques caractéristiques de base des mots grammaticaux :

- Les mots grammaticaux sont limités en nombre.
- La rareté de la création de nouveaux mots grammaticaux.
- Les mots grammaticaux sont fréquents et peu précis.
- Les mots grammaticaux sont souvent très courts.
- Ils incluent les articles, les prépositions et les adjectifs non qualitatifs.
- Ils n'ont qu'une seule fonction.

Les mots-outils regroupent :

- Les déterminants,
- Les pronoms,
- Les conjonctions de coordinations,
- Les conjonctions subordinations,
- Les propositions,
- Les adverbes.

Nous commençons par détailler chacun de ces mots :

A. Les déterminants :

Nous retrouvons différentes sortes de déterminants : les articles définis et indéfinis, et les adjectifs possessifs, démonstratifs, indéfinis, numéraux, interrogatifs, exclamatifs et relatifs[2].

A.1. L'article défini : « Le, la Les L' »

L'article défini fait parti de la sous-catégorie de déterminant défini et s'oppose ainsi à l'article indéfini et à l'article partitif

L'article défini sert à introduire un nom ou groupe nominal désignant une chose ou un être déjà identifié.

Par exemple :

Il y a eu un meurtre et **le** criminel a été arrêté.

	Singulier	Pluriel
Masculin	Le, (l')	Les
Féminin	La, (l')	Les

Tableau -1- : La représentation des articles définis.

A.2. L'article indéfini : Un Une, Des

L'article indéfini spécifie un nom en le rendant particulier sans l'identifier de manière spécifique[2].
Prenons l'exemple suivant : « J'ai rencontré un homme dont je me souviendrai toute ma vie**.**

[2]http://membres.multimania.fr/clo7/grammaire/determinant.htm.

Un détermine homme en montrant qu'il s'agit d'une personne déterminée dont l'identité ne nous est pas connue.

	Singulier	Pluriel
Masculin	Un	Des
Féminin	Une	Des

Tableau -2- : La représentation des articles indéfinis.

A.3. Les adjectifs possessifs :

Les adjectifs possessifs dits aussi déterminants possessifs spécifient le nom en terme d'appartenance.

Ils varient en genre, en nombre et en personne.

En effet, la forme de l'adjectif possessif dépend du genre et du nombre du nom déterminé par le possessif et de la personne du possesseur.

Lorsque le possesseur est la personne qui prend la parole, la forme de l'adjectif possessif s'accordera en genre et en nombre avec le nom déterminé comme par exemple:

Ma console de jeux, mon i-phone, mes CD

La forme du possessif sera remplacée dans certains emplois par un article défini devant des noms qui représente par exemple des parties corporelles

Exemple :

« J'ai mal au dos» au lieu de dire « j'ai mal à mon dos».

	Singulier		Pluriel
Personne	Masculin	Féminin	
1ére Personne	Mon	Ma	Mes
2éme Personne	Ton	Ta	Tes
3éme Personne	Son	Sa	Ses
1ére Personne	Notre		Nos
2éme Personne	Votre		Vos
3éme Personne	Leur		Leurs

Tableau -3- : Les adjectifs possessifs.

A.4. Les adjectifs démonstratifs :

Sont des adjectifs qui permettent de montrer ou désigner des êtres ou des objets.

Cet : est employé comme une forme de masculin singulier devant une voyelle ou un h muet comme :

Cet élève au lieu de ce élève ;

Cet hélicoptère car le h est muet dans ce cas là,

Mais devant un h aspiré non muet on emploie : ce hérisson au lieu de cet hérisson.

On trouve également des adjectifs démonstratifs du type ce/cette…-ci/-là :

Exemple :

Ce livre-ci, ce livre-là, cet élève-ci, cet élève-là.

« Ci » signifie la proximité dans l'espace c'est à dire qu'il signifie par exemple un objet ou un être proche de notre emplacement.

Par contre, là désigne l'éloignement, mais ce n'est pas toujours le cas.

Les adjectifs démonstratifs peuvent désigner une chose présente dans l'entourage des interlocuteurs.

Ils permettent une identification identique à celle d'un geste fait par la main comme par exemple cette phrase :

Les copies sont sur ce bureau c'est-à-dire qu'on peut s'exprimer avec la main pour monter l'emplacement des copies sur le bureau.

	Singulier	Pluriel
Masculin	Ce (Cet)	Ces
Féminin	Cette	Ces

Tableau -4- : Les adjectifs démonstratifs.

A.5. Les adjectifs indéfinis :

Les adjectifs indéfinis dits aussi déterminants indéfinis spécifient une idée de quantité.

Ils permettent d'imprimer un nom qui est comptable en le quantifiant, et sont illustrés comme suit :

« Aucun, autre, certain, chaque, différents, divers, l'un et l'autre, n'importe quel, même, nul, pas un, plus d'un, plusieurs, quel, quelconque, quelque, tel, tout » et sont employés ainsi :

Quantité indéterminée : « quelques, certains, plusieurs ».

En effet, l'adjectif **Plusieurs** est invariable, par contre **certains** varie en genre et en nombre comme par exemple: « certaines filles ».

L'adjectif **Quelque** est souvent au pluriel exemple: « **quelques** articles».

Cet adjectif peut être aussi employé au singulier avec une valeur d'indétermination :

« Il y avait au concert **quelque** cinq mille personnes ».
Les indéfinis cités juste en dessous sont généralement employés devant un nom au pluriel comme par exemple:

« Certaines images ; plusieurs voiture».

L'adjectif indéfini : tel ou telle (au féminin) détermine un nom de manière imprécise :

« Il dit l'avoir vu **tel** jour.

« Je t'ai rencontré tel jour à **telle** heure ».

Nous retrouvons également une variété de déterminants indéfinis à valeur négative comme par exemple:

« Aucun ou aucune (au féminin), nul ou nulle (au féminin) »

Mais ces adjectifs sont accompagnés d'un adverbe de négation « ne » dans des phrases négatives :

« **Aucun ordre n**'a été respecté ».

On l'a cherché mais on **ne l** 'a retrouvé **nulle** part».

A.6. Les adjectifs numéraux :

On divise les adjectifs numéraux en numéraux cardinaux et en numéraux ordinaux.

- **Numéraux cardinaux :**

Expriment tout simplement un nombre comme par exemple :

« Une table, trois minutes, deux ordinateurs ».

- **Numéraux ordinaux :**

Expriment une relation d'ordre par exemple:

« Premier, deuxième, troisième...etc ».

En effet, les adjectifs numéraux ordinaux sont placés entre un déterminant du type article défini, possessif ou démonstratif et le nom déterminé : le premier enfant ; la troisième écoute.

A.7. Les adjectifs interrogatifs, exclamatifs et relatifs :

A.7.1. Quel : est un adjectif à la fois interrogatif et exclamatif, et varie en genre et en nombre par exemple :

- Quels est un adjectifs masculin au pluriel.
- Quelle est un adjectif féminin au singulier
- Quelles est un adjectifs au féminin pluriel

Il peut être placé devant un nom dans les propositions interrogatives directes :

« Quel jour revient-il ? »

ou dans les propositions interrogatives indirectes :

« Je ne sais pas quel jour il reviendra »

ou dans les propositions exclamatives :

« Quelle bonne idée ! ».

A.7.2. L'adjectif relatif : introduit une relation entre le nom qu'il introduit et un nom placé avant. Par exemple:

« Elle s'est engagée dans une route pour laquelle il existait une interdiction locale de circuler ».

Les formes de l'adjectif relatif sont :

- Le masculin singulier : « Lequel, duquel, auquel ».
- Le féminin singulier : « Laquelle, de laquelle, à laquelle».
- Le masculin pluriel : « Lesquels, desquels, auxquels ».
- Le féminin pluriel : « Lesquelles, desquelles, auxquelles ».

A.8. Les articles partitifs :

L'article partitif permet de quantifier des noms abstraits non comptables.

Il est vu comme étant un article qui se positionne devant des choses qui ne peuvent pas se compter, pour dire qu'il s´agit d´une partie seulement ou d´une certaine quantité de ce qui est désigné par le nom.

Exemple :

« Du courage, de la force, du mépris de la tendresse ».

Il désigne également des noms impossibles à dénombrer (masse continue, qu'on ne peut fragmenter) :

Comme par exemple :

« Du lait, de l'huile, de l'eau».

Les formes de l'article partitif sont :

Au masculin singulier :

- **du** comme par exemple : « du plomb ».
- **de l'** : lorsque le nom déterminé commence par une voyelle : « de l'équipement ».

Au féminin singulier :

- **De la** comme: « de la semoule, de la farine ».
- **De l'** : devant une voyelle par exemple: « de l'électricité ».

A.9. Les conjonctions :

Les conjonctions sont des mots ou des locutions invariables qui servent à joindre, à relier, à mettre en rapport deux éléments.

On distingue deux espèces de conjonctions[3] :

- Les conjonctions de subordination.
- Les conjonctions de coordination.

[3] http://www.synapse-fr.com/manuels/CONJONC.htm.

1. les conjonctions de coordination :

Servent à joindre des éléments qui possèdent la même fonction exemple :

« La mère **et** le père sont venus » les mots mère et père possèdent la même fonction car ils sont tous les deux sujets.

Elles permettent de relier également des propositions de même nature comme :

« Il a gagné **et** il est content ».

Les principales conjonctions de coordination sont :

« Cependant, néanmoins, toutefois mais, ou, et, donc, or, ni, car, ».

2. les conjonctions de subordination :

Servent à joindre ou unir une subordonnée à une autre proposition dont elle dépend comme : « Il partira quand nous arriverons ».

Les principales conjonctions de subordination sont : « comme, lorsque, puisque, quand, que, quoique, si ».

Il faut ajouter à cette liste de très nombreuses locutions :

« Ainsi que, à mesure que, après que, à moins que, au lieu que, aussitôt que, pendant que, pourvu que, etc. ».

La subordination peut marquer soit :

- Une comparaison comme par exemple : le père comme la mère....
- Une cause comme par exemple : Il ne rentrera pas puisque...
- Exprime également le temps comme exemple : Il partira quand....

A.10. Les prépositions :

Permettent la plupart du temps, de réunir deux mots dont le deuxième complète le premier en indiquant un rapport particulier selon les circonstances.

Dans ces exemples, les prépositions sont en gras et les groupes prépositionnels sont soulignés Exemples :

« Le bijou de ma mère est un accessoire **de** la plus haute qualité ».

« **De** sept heur à midi, je serai au travail **pour** rencontrer **de** nouveaux clients ».

La préposition peut marquer également :

- Le lieu : dans, en, à, chez, sous...
- Le rang : devant, derrière, après...
- Le temps : avant, après, à, depuis, pendant...
- La cause : pour, vu...
- La manière : avec, sans, selon, de, à...
- La séparation : sauf, sans...
- Le but : pour, à, envers...
- Etc.

A.10.1. Les formes de la préposition :

A.10.1.1. Déterminants contractés :

La préposition est présente dans les déterminants contractés.

Singulier		Pluriel	
Masculin			
à + le = au	de + le = du	à + les = aux	de + les = des

Tableau -5- : La représentation des déterminants contractés

A.10.1.2. Prépositions simples :

Les prépositions simples sont celles qui ne peuvent pas se décomposées, nous les représenterons dans le tableau qui suit :

A	Durant	Pendant
Après	En	Pour
Avant	Entre	Près
Avec	Excepté	Sans
Chez	Hormis	Sauf
Concernant	Hors	Selon
Contre	Jusque	Sous
Dans	Malgré	Suivant
De	Moyennant	Sur
Depuis	Outre	Vers
Derrière	Par	Voici
Dés	Parmi	Voilà
Devant	Vu	

Tableau -6- : La représentation des Prépositions simples.

A.10.1.3. Prépositions complexes :

Par contre, les prépositions complexes peuvent se décomposer ainsi :

à cause de	au-dessous de	en face de
à condition de	avant de	en sus de
à côté de	au-dessus de	en vu de
à force de	au lieu de	face à
à la manière de	au-dessus de	en vue de
à l'abri de	au lieu de	
à l'exception de	auprès de	faute de
à l'insu de	autour de	grâce à
à l'intérieur de	avant de	hors de
à moins de	d'après	le long de
à raison de	du côté de	loin de
à travers	en comparaison de	près de
afin de	en dépit de	sauf à
au-delà de	en faveur de	vis-à-vis

Tableau -7- : la représentation des Prépositions complexes.

B. Les adverbes :

L'adverbe peut être vu comme étant une catégorie de mot qui s'ajoute à un verbe, à un adjectif, à un autre adverbe ou à un nom, dans le but de modifier le sens.

Il permet d'apporter une information supplémentaire au sens[4] :

- D'un adjectif : « Il est très précieux».
- D'un verbe : « Il comprend beaucoup».
- D'un autre adverbe : « Il boit très lentement ».
- D'une phrase tout entière : « Naturellement, il va au travail tous les jours ».

[4] http://www.etudes-litteraires.com/adverbe.php#ixzz1Mu6fWsF9.

C. Les pronoms :

Le pronom est un nom utilisé pour éviter les répétitions et certains sons peu agréables.

En effet puisque le pronom n'est qu'un nom (un être, une chose, une idée), il varie alors en genre (masculin-féminin), en nombre (singulier-pluriel) et parfois en personne (pronoms personnels et possessifs).

C.1. Les différentes catégories de pronoms :

Selon le type d'indication qu'ils portent, on classe les pronoms en différentes catégories :

- Les pronoms personnels : « je, tu, il, nous, vous, ils»
- Les pronoms possessifs : « le mien, le tien, le sien… ».
- Les pronoms démonstratifs : « ce, cela, ceci, celui, celui-ci, celui-là ».
- Les pronoms indéfinis, qui notent le caractère indéterminé : « aucun, quelqu'un, plusieurs, rien, tout… ».
- Les pronoms relatifs : « qui, que, quoi, dont, où, lequel, quiconque ».
- Les pronoms interrogatifs, qui indiquent sur quoi porte la question : « qui, que, quoi, lequel ».

II.2.2.2. Les mots lexicaux :

Sont des mots dont la longueur variable qui se distinguent par les noms, les verbes et les adjectifs qualificatifs. Ils se caractérisent des mots grammaticaux comme suit:

- Ils sont en très grand nombre.
- Ils sont de longueur variable.
- Certains mots lexicaux peuvent être remplacés par des pronoms.
- Ils peuvent être créés, s'il y a besoin.

A. Les noms :

Les noms appelés aussi substantifs sont vues comme une classe de mots servant à nommer, c'est-à-dire à désigner, les catégories d'êtres, de choses et de concepts.

Ils servent à nommer :

- Une personne : Amandine.
- Un animal : un poisson.
- Une chose : un téléphone.
- Une notion abstraite : la joie.

Nous distinguons deux sortes de noms :

- le nom commun :

Le nom commun est en général un déterminant qui commence par une lettre minuscule et désigne tous les êtres ou objets d'une même espèce.

Exemple :

L'ordinateur, le téléphone

Parmi les noms communs nous distinguons :

- Les noms abstraits (peur, faiblesse).
- Les noms concrets (table, banc).
- Les noms animés (désignant des êtres vivants - humains ou animaux) (chat; bébé).
- Les noms comptables (choses qu'il est possible de dénombrer) (une banane; deux bananes; quelques bananes).
- Les noms non comptables, (choses qu'il n'est pas possible de dénombrer) (du lait, du sable).

- Les noms propres :

Se sont des noms qui commencent par une majuscule et dénomment des entités individuelles : personnes, lieux ou événements (Amandine, la Révolution, la France).

Certains noms propres se construisent sans déterminant (Paris; Paul) et jouent le rôle d'un groupe nominal. Ils en prennent d'ailleurs la plupart des fonctions : sujet **(Paul** est gentil), objet direct (Il a vu **Hélène**), objet indirect (J'ai écrit à **Céline**), complément de nom (Je ne reçois pas le message de **Pierre**), etc.:

A.1. Le genre des noms :

Les noms prennent le genre correspond au sexe « coiffeur; coiffeuse ».

Nous pouvons trouver :

- Des noms masculins dont le genre grammatical est féminin comme : une sentinelle.
- D'autres noms féminins dont le genre grammatical est masculin :(un mannequin).
- Des noms masculins qui peuvent désigner aussi bien des hommes que des femmes (médecin, témoin; écrivain; guide; ingénieur; juge; magistrat; peintre; professeur; auteur; otage).
- Des noms féminins qui peuvent se référer à des personnes des deux sexes (victime; personne) ou encore à des animaux mâles ou femelles (souris; grenouille).

A.2. Le pluriel des noms

Le pluriel des noms se forme par l'ajout d'un S, excepté les mots qui se terminent par S, X, ou Z au singulier (cas; croix; nez).

Les noms en au, eau et eu prennent un pluriel en X (tuyaux; seaux; ennuyeux).

Les noms en ou prennent un pluriel régulier en S (trou>trous) à l'exception d'une série de sept noms pour lesquels la marque du pluriel est X (bijoux; cailloux; choux; genoux; hiboux; joujoux et poux).

Les mots en al et en ail forment leur pluriel en aux (journal>journaux; travail>travaux), à l'exception de « bal; carnaval; festival; régal; chacal; cérémonial, etc », qui forment leur pluriel selon la règle générale en als, et de détail; éventail, etc. dont le pluriel en -ails est également régulier.

Enfin, il existe un certain nombre de formes très irrégulières : œil - yeux; ciel - cieux;

aïeul - aïeux. Pour œuf - œufs et bœuf - bœufs, le pluriel n'est régulier qu'à l'écrit car à l'oral la distinction entre singulier et pluriel est nette.

B. Les verbes :

Le verbe est le noyau de la phrase qui possède des particularités qui lui sont propres. Il est variable en fonction du nombre, du temps, de l'aspect, du mode, de la voix.
Le terme Conjugaison représente l'ensemble des formes qu'un verbe peut prendre.
Celui-ci est un mot qui peut exprimer :

- L'action accomplie par le sujet.
- L'action subie par le sujet.
- L'existence du sujet.
- L'état du sujet.
- La relation entre le sujet et l'attribut.

B.1. Les traits du verbe :

Les traits du verbe sont déterminés en fonction de la constitution des compléments dans la phrase, nous citons :

B.1.1. Les verbes actifs (transitifs) et intransitifs :
En fonction du verbe qui admet un complément d'objet ou non, nous parlerons de construction transitive ou intransitive :
En effet, les verbes transitifs demandent nécessairement un complément d'objet
Exemple : accrocher, distribuer « il a accroché une photo », par contre, les verbes intransitifs ne demandent pas nécessairement de complément d'objet.

Exemple : Tomber « La pluie tombe ».

B.1.2. Les verbes pronominaux et non pronominaux :
Le verbe non pronominal se conjugue avec un seul pronom sujet.
Exemple : « Je travaille dur pour réussir ».
Par contre, le verbe pronominal se conjugue avec deux pronoms dont l'un est placé entre le sujet et le verbe.

Exemple : « Il se concentre dans ses révisions ».
Le premier pronom est sujet du verbe, le deuxième est son complément

B.1.3. Les verbes impersonnels :

Les verbes impersonnels ne se conjuguent qu'avec la troisième personne du singulier c'est à dire le « il ».

Nous distinguons deux catégories de verbes impersonnels :

- Ceux qui le sont toujours comme : neiger, pleuvoir, falloir comme :
  « Il neige, il pleut, il faut partir ».
- Ceux qui le sont dans certaines situations seulement comme :
  « Il s'est produit un grave accident ».

B.1.4. Les verbes attributifs ou d'état :

Ce sont ceux qui évoquent un lien entre le sujet et l'adjectif.

Exemple : « Il **paraît** plus jeune que son frère ».

B.1.5. Les verbes passifs :

Ce sont les verbes d'une phrase en forme passive. Exemple :

Forme active : Paul **chante** une chanson.

Une chanson **est chantée** par Paul. (est chantée : verbe passif).

C. Les adjectifs :

L'adjectif est vu comme une catégorie de mot qui s'adjoint au nom pour exprimer :

- Une qualité (adjectif qualificatif).
- Une relation (adjectif relationnel).

Ou pour permettre à celui-ci d'être actualisé dans une phrase (adjectif déterminatif).

L'adjectif se distingue notamment du déterminant par sa distribution dans la phrase.

C.1. L’adjectif qualificatif :

L’adjectif qualificatif indique la qualité, la manière d’être du nom ou du pronom auquel il est rattaché. Exemple : L’enfant est **las**.

L’adjectif qualificatif prend les marques du genre et du nombre du nom auquel il se rapporte. Il peut être attribut ou épithète.

- **Attribut** :

  L'adjectif qualificatif attribut est relié au sujet par l'intermédiaire du verbe « être » ou d'un verbe équivalent : sembler - paraître - devenir - rester - demeurer - avoir l'air - passer pour ; il exprime la qualité du sujet.

  Exemple : « : Leurs yeux étaient **larges** et **doux** ».

- **Epithète**

  Quand il est placé à côté du nom (devant ou derrière) Exemple :

  « Depuis longtemps, l'homme observe cet animal **étrange** et **fascinant**. »

C.2. L’adjectif numéral :

Est un mot qui précise le sens du mot en indiquant le nombre, l’ordre ou même le rang.

Exemple : « Un, deux, trois, quatre …… etc ».

C.3. L’adjectif interrogatif ou exclamatif :

Est un mot que l’on joint pour marquer l’interrogation. Nous citons : « quelle, quel, quels, quelles ».

Concernant l’adjectif exclamatif, il indique la surprise ou l’indignation portant sur le thème indiqué par le groupe du nom.

C.4. L’adjectif relatif :

Placé devant le nom, il lui rattache la subordonnée qu’il lui introduit.

Nous citons : « laquelle, lequel, lesquels, lesquels ».

III. Les phrases :

La phrase est une suite de mots construits selon un ordre grammatical correct et bien précis dans le but de transmettre une information[5].

Cet ensemble de mots exprime une idée complète, il commence par une majuscule et se termine par un point.

En effet, la phrase est l'unité de communication d'une langue qui exprime un jugement, une pensée sur un être, sur une chose. Cette information ne peut être transmise que si la phrase a un sens, nous distinguons :

- La phrase verbale : est construite autour d'un verbe. "Sarah mange une banane".
- La phrase nominale : est construite autour d'un nom.
  Exemple : "Rentrée des classes, le 1er septembre".

III.1. Les types de phrase :

On dénombre quatre types de phrase :

- Phrase déclarative : Pierre mange du chocolat.
- Phrase interrogative : Pierre mange-t-il du chocolat ?
- Phrase exclamative : Quel bon chocolat !
- Phrase impérative : Mange ton chocolat !

III.2. Les formes de phrase :

- Forme affirmative : Pierre mange du chocolat.

- Forme négative : Pierre ne mange pas de chocolat.

Elle peut être aussi :

[5] http://membres.multimania.fr/clo7/grammaire/phrase.htm.

- Active : Pierre mange du chocolat.

- Passive : Le chocolat est mangé par Pierre.

VI. Conclusion :

Parvenus au terme de ce chapitre introductif sur l'étude linguistique nous pouvons constater d'un point de vue général que la linguistique étudie le fonctionnement du langage appréhendé à travers toutes les langues parlées. Il en fait une description.
Celle-ci a pour but d'expliciter la nature du langage. Le linguiste va donc chercher les origines des usages de ce langage, des règles ou des erreurs qu'il décrira comme des usages particuliers, voire de comprendre que derrière une faute se cache un besoin de rationalisation de la langue.

La linguistique s'intéresse à tous les types ou groupes de langues et à toutes les facettes d'une langue (son évolution historique, sa construction grammaticale, ses apports, …), mais nous nous sommes intéressés dans ce chapitre sur une étude linguistique de la langue Française.

Chapitre II

L’extraction de terminologie

I. Introduction :

Une langue n’est pas une entité figée, fixée une fois pour toute : sans cesse des mots disparaissent, meurent, d’autre nouveaux apparaissent…le monde change, et le lexique évolue.

Pour désigner les réalités nouvelles, le français, comme toutes les autres langues, s’enrichit de nouveaux mots. Dans la langue courante, cette création est en quelque sorte spontanée, l’inventivité des jeunes, des journalistes, sans parler des écrivains et poètes… se déploie dans la plus grande liberté. Il suffit de penser à tous ces mots nouveaux que l’on entend dans les médias, que l’on voit dans les journaux. Tantôt ils passent rapidement, tantôt ils s’implantent durablement dans l’usage, et dans les dictionnaires.

Dans les domaines techniques et scientifiques, les données sont différentes et d’une toute autre ampleur : pour exprimer des notions souvent très complexes, les professionnels emploient dans leur domaine d’activité particulier des mots ou des expressions très précis, des termes, qui se dénombrent en centaines de milliers.

La terminologie consiste en l'étude du choix et de l'usage des termes faisant partie des vocabulaires de spécialité, qu'on peut trouver dans tous les domaines de connaissance : informatique, grammaire, linguistique, mathématique, philosophie, médecine, musique...

Une terminologie est donc un ensemble de termes spécialisés relevant d’un même domaine d’activité qui a son propre vocabulaire. La terminologie dénommée aussi terminographie, est loin d'être une discipline nouvelle. On peut même retracer ses origines à l'antiquité grecque. Mais on peut dater son développement du début de ce siècle, en 1906, la Commission Electrotechnique Internationale (CEI) commence le développement de son Vocabulaire Electronique International (VEI). Une autre date importante est celle de la publication, dans les années trente, du Dictionnaire de la machine outil, d'Eugen Wüster, dont les travaux assoient les bases théoriques de la terminologie moderne.

Enfin la terminologie est l'ensemble de termes, rigoureusement définis, qui sont spécifiques à une science, une technique ou à un domaine particulier de l'activité humaine.

D’après [Yamouni, 2010] « L’extraction de terminologie consiste à identifier des termes potentiels dans un texte spécifique ou un ensemble de textes (corpus) ainsi que les informations pertinentes liées à l’emploi de ces termes ou aux concepts auxquels ils renvoient (définition, contexte, etc.). ».

L’extraction de terminologie peut se faire manuellement ou bien automatiquement à l’aide d’utilisation d’outils d’extraction terminologique.

Le rôle de la terminologie

De la qualité de la terminologie dépend en grande partie la qualité d’un texte spécialisé, qu’il s’agisse d’un texte original ou d’une traduction. La terminologie employée en garantit la clarté et par conséquent la compréhension. La terminologie est essentielle comme outil de communication et de transfert des connaissances.

II. Historique :

Les premiers efforts dans la terminologie ont commencé vers les années 60, probablement en raison du rapport de la publication Automatic Language Processing Advisory Committee « ALPAC » qui a préconisé les développements d’outil logiciel à aider les traducteurs au lieu de mener des recherches de traduction automatique.

La terminologie, cependant, n'avait pas été perçue comme une discipline distincte de lexicologie et d'autres disciplines linguistiques jusqu'à la publication de l'Einführung dans la matrice terminologische Lexikographie4 d'und de Terminologielehre d'allgemeine par Eugen Wüster en 1979. Les premiers projets terminologiques étaient seulement disponibles pour de grands organismes parce que le logiciel de gestion terminologique a exigé des ordinateurs centraux dits Mainframe.

En effet, la situation a mené le développement des termbanks à grande échelle, par exemple Termium, EURODICAUTOM, Banque de terminologie de Québec (maintenant Le Grand dictionnaire terminologique).

De nos jours, les termbanks développés à ce stade sont toujours en service, bien que les systèmes aient subis des révisions générales, par exemple EURODICAUTOM fonctionne maintenant sur des plateformes entièrement nouvelles (oracle et Fulcrum).

Les années 80 ont vu les premiers dictionnaires électroniques ainsi que le logiciel de gestion terminologique qui s’est développé pour des ordinateurs individuels et qui est disponible pour différents traducteurs, après le développement du logiciel « memory translation software». Cependant, ces outils ont eu beaucoup de contraintes.

La nouvelle génération d’outils de gestion terminologique a suivi la publication de concepts du poste de travail d'un traducteur intégré à trois nivaux.

Un autre outil CAT (Center of Alternative Technology) qui est maintenant l'un des chefs du marché, était le premier à être libéré en 1993.

Depuis cette époque, beaucoup de nouveaux outils et nouvelles versions de CAT ont eu lieu, se rattrapant par rapport aux développements en technologie de langue et en informatique.

Actuellement, il y a deux tendances principales dans le développement de CAT.

D'une part, les réalisateurs de logiciel tendent à isoler les fonctionnalités qui faisaient partie des modules terminologiques dans des outils séparés, par exemple le module d'extraction de termes faisaient partie de MultiTerm 5.0 mais n’est plus dans un paquet de Multiterm IX.

D'autre part, il y a la tendance d’intégration d’outils typiques de MAHT (Machine Assisted Human Translation), y compris les outils de gestion terminologique, avec les systèmes de traduction automatique et les outils de localisation, ayant pour résultat les systèmes hybrides.

L'application des systèmes hybrides et l'environnement de traduction fortement intégré est habituellement le plus avancé dans de grands établissements, prenons comme exemple la Commission européenne.

III. L’extraction de terminologie :

Trois procédures d’extraction terminologiques sont distinctes dans ce qui suit, nous citons :

- L’extraction manuelle.
- L’extraction automatique.
- L’extraction semi-automatique.

III.1. L’extraction manuelle [Yamouni, 2010] :

Les terminologues isolent manuellement les termes à partir de textes, l’extraction manuelle de terminologie comporte trois phases fondamentales :

III.1.1 La collecte des textes :

Nous pouvons trouver des textes dans des sources différentes, telles que :

- les journaux.
- les revues de presse.
- Livres.
- Articles.

Ces textes doivent être spécifiques à une science, à une technique ou à un même domaine particulier.

III.1.2. La lecture des textes et l’extraction :

La lecture des textes doit s’effectuer d’une manière sérieusement, soigneusement et attentivement et en effectuant un repérage de termes d’un domaine particulier en s’appuyant sur les concepts de base du domaine.

III.1.3. L’acquisition :

En général, l’acquisition est l’action qui consiste à obtenir une information. Elle est vue comme étant une étape qui suit l’extraction de termes conservés et gardés comme valides.

III.2. L’extraction automatique :

L’extraction automatique de terminologie se base sur l’utilisation d’un logiciel qui fait l’extraction et la validation des candidats termes automatiquement.

Elle est effectuée aussi grâce à des outils utilisant divers méthodes d’extraction.

Ainsi, l’extraction automatique de terminologie est une méthode rapide pour acquérir des connaissances sur un domaine particulier et sur le langage spécialisé qui s’y rattache.

III.3. L’extraction semi-automatique :

L’extraction semi-automatique consiste à utiliser un logiciel pour extraire les candidats termes, Cependant, après l'extraction de ces termes potentiels par les logiciels, des spécialistes doivent décider si les résultats obtenus sont adaptés ou non.

IV. Les outils d’extraction automatique de terminologie :

Les outils d’extraction automatique de terminologie nous permettent d’extraire de nouveaux termes à partir de texte ou un ensemble de textes (corpus).

Diverses méthodes linguistiques, statistiques ou hybrides sont employées pour développer les outils d'extraction automatique.

De multiples approches de recherche terminologique ont été développées afin d’extraire les termes pertinents à partir d’un corpus.

Le logiciel TERMINO [David et al. 1990] est le premier outil d’acquisition automatique de termes à avoir vu le jour.

A cette même période c'est-à-dire 1990, plusieurs chercheurs tels que : Choueka 1988, Lebart et Salem 1988, Church et Hanks 1990 ont mis au point différents systèmes statistiques de repérage de collocations dans les corpus.

Ces techniques, bien qu’elles ne soient pas spécifiquement dédiées à l’extraction de termes, ont inspiré de nombreux travaux terminologiques.

Un recensement de divers systèmes existants [De Chalender, 2002] est fait. En effet, ces systèmes rendent le recensement possible à extraire de nouveaux termes à partir de textes ou de corpus.

Quelques critères d'évaluation de logiciel d'extraction sont examinés dans [L’Homme M.C., 2000].

Dans [Drouin, 2003], des techniques standard pour l'extraction de terme sont examinées.

Dans [Drouin, 2002] et dans [Drouin, 2003b], des pivots lexicologiques spécialisés sont employés pour l'acquisition automatique des termes.

Dans [Drouin, 2003a] un corpus non technique est employé pour l'extraction de la terminologie.

L'évaluation des outils d'acquisition pour l'extraction de l'information est faite dans [Poibeau et al., 2002],

La variation terminologique est étudiée dans [Jacquemin, 1997].

Les études basées sur une approche d’acquisition sémantique et terminologique peuvent être trouvé dans [Morin, 1999] et [De Chalendar, 2001].

Dans le cadre des corpus spécialisés dans la terminologie et l'acquisition automatique des collocations, divers travaux en marche dans [Lemay, 2003], [Orliac, 2003] et [Rochibeau, 2003] sont présentés.

Une étude au sujet de la collocation peut être trouvée dans [L'Homme M.C., 2000].

Dans [Meilland, 2003], Meilland J.C présente une terminologie automatique d'extraction des mots textuels courts.

Dans [Smadja, 1996], des méthodes statistiques sont utilisées pour traduire des collocations pour le lexique bilingue.

V. Les approches d'extraction automatique de terminologie

[Yamouni, 2010] :

Les approches d'extraction automatique de terminologie ont été regroupées en trois catégories :

- Approche linguistique appelée aussi approche syntaxique.
- Approche mixte appelée aussi approche hybride.
- Approche statistique appelée aussi approche numérique.

V .1. L'approche linguistique [Yamouni, 2010] :

Le traitement automatique des langues vise à développer les outils permettant d'analyser un ensemble de documents. Parmi les processus mis en jeu, il est généralement admis que la segmentation et l'étiquetage (affectation d'une étiquète grammaticale) de ces unités constituent les premières étapes d'une chaine de traitement linguistique.

Cette dernière s'appuie sur une analyse syntaxique des textes, qui permet, grâce à un étiquetage morpho-syntaxique préalable de repérer les syntagmes nominaux et de les analyser.

Elles sont aussi appelées méthodes symboliques, celles-ci sont syntaxiques, mais il existe aussi des méthodes basées sur la sémantique [Morin, 1999].

Des systèmes sont évalués dans [Term, 2003], nous avons XTERM développé par [Cerbah, F, 1999], FASTER [Jacquemin, 1997], LEXTER [Bourigault, 1994] et TERINE [david et al. 1990].

Nous présentons ci-dessous une description d’outils d’extraction de candidats termes des méthodes linguistique :

V.1.1. XTERM :

Le système XTERM a été développé en 1999 par [Cerbah, 1999]. Afin de repérer les parties textuelles le système XTREM effectue tout d’abord un prétraitement du corpus, ensuite réalise un étiquetage morphosyntaxique des segments textuels. Grâce à la projection de patron syntaxique propre aux termes formaliser sous forme d’automates d’état fini, un repérage de candidats termes est effectué.

Une réorganisation hiérarchique de ces candidats termes est réalisée en fonction de leurs composantes (tête, expansion). En fin les variantes sont regroupées en fonction de règles.

V .1.2. LEXTER :

La constitution, l'analyse et l'exploitation d'un lexique structuré sont réalisées par le logiciel Lexter. Développé à la Direction des Etudes et Recherches d'EDF par Didier Bourigault [Bourigault, 1994], pour répondre aux besoins de constitution de terminologies qui existent dans cette entreprise.

Lexter a été initialement conçu pour extraire automatiquement des candidats termes à partir d'un corpus de textes, sans faire appel à des ressources linguistiques extérieures, pour

constituer et mettre à jour des thésaurus utilisés par un système d’indexation automatique de textes.

LEXTER reçoit en entrée un corpus de textes techniques portant sur un domaine quelconque, et propose comme résultat un réseau terminologique, c'est-à-dire un ensemble de groupes nominaux susceptibles d'être des termes complexes du domaine, organisé en réseau à l'aide de relations grammaticales. Ces textes traités ont été soumis à une analyse morphologique c'est-à-dire que chaque mot est étiqueté avec sa catégorie grammaticale.

Ce réseau de candidats termes, avec le corpus à partir duquel il a été extrait, est soumis, à un expert ou à un terminologue sous forme d'un hypertexte c'est-à-dire des noms ou groupes de noms susceptibles de designer des concepts du domaine, pour des fins de validation.

Chacun des candidats termes trouvé lors des phases précédentes est relié aux candidats termes dont il est tête ou expansion.

Le principe de base est donc de découper le texte en repérant ces frontières potentielles entre lesquelles on isole les syntagmes nominaux maximaux en s’appuyant sur les marqueurs de frontière qui sont tous les éléments qui ne peuvent pas faire partie d’un syntagme nominal exemple : déterminant, pronom, etc. Les données d'entrée du module chargé d'effectuer le découpage sont uniquement des informations morphologiques associées à chaque mot du texte : catégorie grammaticale, traits morphologiques (en particulier genre et nombre), forme lemmatisée.

Les groupes nominaux isolés par le module de découpage sont analysés par un module de décomposition. Les syntagmes nominaux (candidats termes) sont composés d'une tête et d'une expansion.

Le module de décomposition met en œuvre une analyse syntaxique : un groupe nominal maximal est traité par une règle de décomposition, qui indique quelles sont les sous-structures qui correspondent à la tête (notée T) et à l'expansion (notée E).

Les cas d’ambigüité de rattachement sont traités par des techniques d’apprentissage endogène sur le corpus.

Evidemment, tous ces syntagmes ne sont pas des termes et un filtrage doit être effectué par un spécialiste du domaine ou un terminologue.

Deux procédures d'apprentissage endogènes se sont avérées nécessaires et ont été implémentées : l'une pour relever les participes passés se construisant avec un complément introduit par la préposition "de", et l'autre pour relever les adjectifs se construisant avec un complément introduit par la préposition "à" [Bourigault, 1994b].

Les groupes nominaux : sont composés de façon privilégiée de noms et d'adjectifs, et presque jamais de verbes à des formes conjuguées ; les prépositions sont le plus souvent "de" et "à", elles sont rarement suivies d'un déterminant, etc. [Guilbert, 1965].

Le schéma de la dénomination syntagmatique est donc le suivant : un terme complexe est constitué d'une tête, qui est le terme (représentant la classe) générique, et d'une expansion, qui est un complément (un adjectif ou un groupe nominal) mentionnant un caractère spécifique [Lethuiller, 1989].

Le mode de la dénomination syntagmatique est récurrent, un terme complexe peut lui-même devenir tête ou expansion d'un nouveau terme, etc.

Il y a des limites empiriques au nombre d'expansions [Lethuiller, 1989].

Des améliorations de LEXTER en s’appuyant sur les travaux de C. Jacquemin [Jacquemin, 1994], repris, critiqués et enrichis par B. Daille dans sa thèse [Daille, 1994].

V.1.3. FASTER :

FASTER, de Christian Jacquemin est un outil de repérage de variation de terme et un analyseur syntaxique robuste dédié à la reconnaissance en corpus de termes appartenant à une liste contrôlée fournie au système [Jacquemin, 1996]. Il a besoin de données terminologiques initiales.

Christian Jacquemin part de la constatation que les termes n’ayant pas toujours, en corpus la même forme linguistique, et pouvant apparaitre sous des formes variées.

FASTER est doté d’un ensemble élaboré de métarègles, qui lui permettent de repérer les différents types de variantes (formes linguistiques différentes) dans un corpus :

- Les variantes syntaxiques : expansion nominale remplacée par une conjonction.
- Les variantes morpho-syntaxique : la tête ou l’expansion change de partie du discours.
- Les variantes sémantico-syntaxique : la tête ou l’expansion est remplacée par un élément sémantiquement proche.

FASTER permet donc de repérer des variantes de termes, mais il peut aussi servir à acquérir de nouveaux termes simples par un processus inverse. Parmi ces termes simples, ceux qui ne font pas partie des termes d’origine. Ils peuvent être proposés comme nouveaux termes pour évaluation.

De plus, les études menées par Christian Jacquemin pour la réalisation et la validation du logiciel FASTER, ont permis d’accumuler une connaissance linguistique riche sur la variation terminologique, un phénomène massif, et longtemps sous-estimé, dans les corpus technique et scientifique.

V.1.4. TERMINAE :

C’est un outil d’aide à la construction de terminologie et d’ontologie à partir de l’étude de texte. TERMINAE [Szulman, 1999] utilise les méthodes et outil linguistique et est écrit en java (version 1.3).

C’est un prototype de recherche qui facilite la construction de terminologie et l’utilisation intégrée des logiciels d’extraction.

Dans ce système une terminologie peut être vue comme étant une constitution d’un ensemble de termes dont les caractéristiques linguistiques sont décrites dans des fiches dont les relations entre sens des termes sont représentées dans un réseau sémantique et l’ontologie comme étant un ensemble structuré de concepts définit formellement et dont certains sont étiqueté par des termes.

Une fiche terminologique est crée pour un terme et ses notions associées. Chaque notion est décrite à l’aide des informations suivantes :

Les synonymes et les termes proches, d’un ensemble de cooccurrences associées, d’une définition et éventuellement d’informations lexicales.

V.1.5. NOMINO :

Le logiciel NOMINO (à son origine TERMINO) est historiquement le premier logiciel à extraire les termes d’un corpus à partir de patrons morpho-syntaxique [David et al., 1990].

Cette application se focalise sur le repérage des syntagmes nominaux qui sont les seules structures supposées produire des termes.

D’après les travaux de Benveniste [Benveniste, 1966] les candidats termes (CTs) extraits par NOMINO sont appelés « synapsies ».

La succession de traitement de NOMINO comporte trois étapes :

a- Prétraitement du texte :

Dans cette phase, le texte est filtré et les caractères de formatage sont supprimés, cette phase est primordiale dans le cadre des outils d'acquisition de terme car les corpus traités sont de type variés.

b- Analyse et acquissions terminologique :

L'acquisition des candidats termes se fait comme suit :

Analyse morphologique à base de règles.

Analyse des syntagmes nominaux.

Génération des synapsies à partir des dépendances entre tête et compliments rencontré dans la structure de syntagme nominale retournée par l'analyseur.

c- Construction et gestion interactive d'une banque de termes :

NOMINO propose une interface graphique pour la construction et la gestion d'une base de données terminologique à partir des termes acquis à l'étape précédente.

V.2. Approche statistique [Yamouni, 2010] :

Cette méthode est très présente dans le langage naturel et afin d'identifier les termes, celle-ci met en œuvre diverses stratégies.

Elle est fondée sur le fait que les termes apparaissent souvent dans les textes spécialisés par exemple les mots simples qui apparaissent souvent ensemble sont forcément significatifs.

En effet, cette approche permet d'effectuer l'extraction de candidats termes sans analyse linguistique préalable. Elle repose sur des calculs de fréquence, des calculs statistiques qui permettent le repérage de collocation entre les unités lexicales.

Il existe plusieurs outils qui se basent sur les syntagmes nominaux, les segments répétés ainsi que les schémas productifs, on retrouve parmi eux :

V.2.1. ANA :

Le système ANA acronyme d’Acquisition Naturelle Automatique [Enguehard, 1992] :
C’est un système qui détecte à partir de deux termes connus les associations récurrentes, ou encore, repère la cooccurrence d’un mot avec un terme connu.
Il permet l’extraction automatique de concepts pour la reproduction d’un réseau sémantique.

V.2.2. MANTEX :

Le logiciel MANTEX [Oueslati, 1999], [Frath et al., 2000] est fondé selon le principe de répétition de segments qu’il repère dans un corpus.

V.2.3. LIKES :

LIKES (frath et al., 2000) acronyme de LInguistic and Knowledge Engineering Station est une station d’ingénierie linguistique destinée à traiter des corpus écrite par François Rousselot, elle permet la création de terminologie et d’ontologie, fonctionne pour l’instant sur la plupart des langues européennes : l’anglais et le français, le portugais, l’espagnol, l’allemand, le roumain et quelques langues slaves : le tchèque, le bulgare le slovaque en utilisant des ressources minimales pour chaque langue et travaille sans aucun dictionnaires pour trouver des segments répétés.

Les corpus sont constitués d’un ou plusieurs textes en ASCII (American Standard Core for Information Interchange) ou en HTML (HyperText Markup Language), l’interface donne la possibilité de constitué son corpus et d’y exécuter un certain nombre de tâches allant de simples tâches de découpage en mot, de tri ou de recherche de motifs à des tâches plus complexes d’aide à la synthèse de grammaire, d’aide au repérage de relations, d’aide à la construction d’une terminologie.

L’interface donne également la possibilité de constituer des filtres grammaticaux spécifiques à chaque utilisateur et/ou à chaque corpus qui sont utilisés dans les divers traitements de la station :

- Découpage en phrase.
- Lemmatisation
- Calcul de segments répétés.

En effet, LIKES s’adresse à des utilisateurs linguistiques, terminologues, étudiants, chercheurs ainsi que des informaticiens, et pour faciliter le travail de ce dernier toutes les tâches sont exécutables à partir d’une console.

Les traitements possibles dans LIKES vont des tâches élémentaires concernant les mots, les familles de mots, les mots composés, à celles plus complexes qui traitent des distributions et des relations.

1. Les tâches de base concernant les mots simples :

Nous distinguons quatre tâches de base citées comme suit :

- Constitution d’un corpus.
- Découpage en mots en et en phrases.
- Tri alphabétique.
- Tri par fréquence.

2. Les tâches de bases concernant les groupes de mots :

Nous distinguons deux tâches de plus haut niveau concernant les groupes de mots :

- La fusion.
- Le calcul de segment répété.

2.1. La fusion :

Sert à regrouper en une seule forme les familles de mots proches. L’usage spécifique est décidé par l’utilisateur qui peut par exemple regrouper les formes du pluriel et du singulier d’un même mot et décider par la suite le représentant de cette nouvelle famille d’occurrence qui rassemblera bien évidemment les occurrences de deux formes dans le corpus.

Exemple : « médecin » et « médecins » seront représentés par « médecin_S ».

L’utilisateur peut également rassembler deux synonymes par exemple : « AVC » et « Accidents Cardiaux Vasculaires ».

2.2. Le calcul de segment répété :

La technique de segments répétés a été utilisée dans des approches orientées statistiques d’abord pour le dépouillement d’enquêtes [Lebart et Salem, 1994], ensuite pour de l’aide à l’extraction de termes [Justeson et Katz, 1995]. Il permet de repérer des séquences de mots répétées dans le corpus.

3. Les tâches de recherches :

3.1. La recherche simple :

Ce type de recherche est réalisé grâce à un outillage d’expressions régulières qui est équivalent à des automates récursifs comparables à ceux qu’on trouve dans la station

INTEX [Ibekwe-SanJuan, 1998] et utilisant une syntaxe proche. Les entités d’expressions sont des séquences de lettres non forcément de mots. Nous pouvons utiliser des métacatactères tel que <L> pour lettre.

Nous pouvons alors manipuler des schémas morphosyntaxiques aisément comme l’expression :

« <L><L>*(ion+ment) »

Qui recherchera dans le corpus tous les mots composés d’une lettre suivie de zéro ou plusieurs lettres suivies de (ion) ou (ment).

3.2. La recherche de collocations :

Si un mot X apparaît plus fréquemment dans l’entourage d’un mot Y, alors les mots X et Y forment une collocation.

V.3. L’approche mixte [Yamouni, 2010] :

Rapidement, les chercheurs se sont orientés vers des méthodes d’extraction hybrides qui consistent à combiner les approches linguistiques avec celles statistiques.
En effet les approches linguistiques sont utilisées dans le but de filtrer les termes en fonction de leurs catégories syntaxiques quant aux approches statistiques, elles représentent la partie essentielle de la méthode d’extraction.
Un certain nombre de difficultés ont été résolues par l'utilisation conjointe de méthodes syntaxiques et statistiques.
Les méthodes syntaxiques et statistiques peuvent se compléter de deux manières : soit on procède au traitement statistique de séquences repérées au préalable grâce à des schémas syntaxique [Daille, 1994], soit à l'inverse, on procède au repérage statistique des cooccurrents, auxquels on applique ensuite des règles syntaxiques dans un second temps [Smadja, 1993].
Parmi les outils qui utilisent l’approche hybride, nous retrouvons ACABIT [Daille, 1994], XTRACT [Smadja, 1993] et ASSIUM [FAURE et Nédellec, 1998,1998a].

V.3.1. ACABIT :

ACABIT de Béatrice daille [Daille, 1994] et acronyme de Automatic Corpus based Acquisition of BInary Terms, travaille a partir d'un corpus pré-étiqueté et effectue une analyse syntaxique suivie d'un traitement statistique.

L'acquisition terminologique dans ACABIT se déroule en deux phases :

a. L'analyse linguistique et regroupement de variantes :

   Un ensemble de transducteurs analyse le corpus étiqueté pour extraire des séquences nominales et les ramener à des candidats termes.

b. Filtrage statistique :

   ACABAT repose sur l'utilisation de diverses mesures statistiques qui retiennent le mieux les candidats termes extraits dans la phase précédente.

V.3.2. ASIUM :

ASIUM de David faure [Faure et Nédellec., 1998], apprend des concepts faits de l'agrégation de mots simples. ASIUM repère dans les textes des connaissances récurrentes de verbes et de noms.

ASIUM utilise le résultat d'un analyseur syntaxique SYLEX pour détecter les associations entre un verbe et les noms têtes de ces sujets, compléments d'objet direct et autres compléments reliés aux verbes par une préposition.

Les connaissances sémantiques apprises par ASIUM sont utilisés par [Faure et Poibeau, 2000] pour l'extraction d'information à l'aide du système INTEX [Selberztein, 1993].

V.3.3. EXIT :

Le logiciel EXIT dont l'acronyme est EXtraction Itérative de la Terminologie fait partie d'une approche qui est fondée sur une méthode statistique et linguistique.

EXIT de [Roche et al 2004] est destiné à des utilisateurs experts d’un domaine et travaille sur un niveau syntaxique haut.

Il regroupe les unités adjacentes appelées aussi collocation accepté par l’expert afin de former par la suite de nouveaux termes.

L’intérêt de regrouper des unités adjacentes est d’obtenir une nouvelle unité plus riche de sens.

Le logiciel EXIT est adapté aux formalismes linguistique, il utilise des informations syntaxique, sémantique tout en étant itératifs afin d’enrichir la terminologie.

V.3.4. XTRACT :

XTRACT le logiciel de Smadja [Smadja 1993], est un extracteur de cooccurrences mais peut aussi être utilisé en acquisition de terminologie. En effet il a servi d’inspiration à certains travaux d’acquisition terminologique exploitant des techniques statistiques.

Il utilise une technique statistique pour détecter les cooccurrences et une analyse syntaxique pour les classer.

Pour une première phase XTRACT commence par repérer les cooccurrents d'un mot donné, dans une fenêtre de cinq mot à gauche et à droite du mot et note leur place.

Dans une deuxième phase, XTRACT recherche des n-grammes, c'est-à-dire des collocations de longueur supérieure à 2.

Dans une troisième phase, XTRACT utilise les étiquettes grammaticales et syntaxiques assignées au préalable par un assignateur de catégories pour valider ou rejeter des collocations.

Si une paire de cooccurrents apparaît dans une relation syntaxique d'un type spécifié par l'utilisateur, alors elle est validée, sinon elle est rejetée.

XTRACT a été conçu comme un outil d'exploration lexicographique et syntaxique de corpus pour la constitution de listes non exhaustives de collocations destinées par exemple à l'indexation, à la réalisation de glossaires, ou à l'enseignement de la langue [frath, 1997].

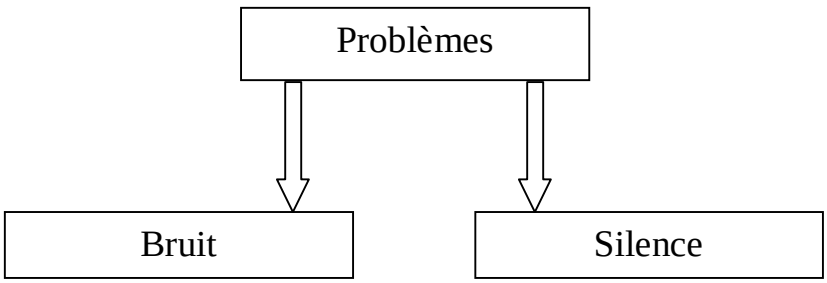
Selon [Frath, 1997] XTRACT est un outil d'exploration textuelle efficace.

Cependant, dans une optique d'acquisition automatique, il présente un inconvénient majeur :

- ce ne sont pas les chiffres qui permettent d'arriver aux conclusions.
Les raisons du rejet ne sont pas dans les chiffres, mais dans la compétence linguistique de l'utilisateur.

Les chiffres ne sont là que pour confirmer l'intuition, alors leur utilité est sans doute limitée dans une optique d'acquisition automatique.

VI. Les performances :



Les trois méthodes présentées précédemment rencontrent deux problèmes majeurs, le bruit et le silence.

VI.1. Le bruit :

C’est quand une unité est extraite par le logiciel mais elle n’est pas pertinente d’un point de vue de l’utilisateur appelé aussi « Précision ».

VI.2. le silence :

C’est quand des unités pertinentes présentées dans le texte, mais elles ne sont pas extraites par le logiciel appelé aussi « Rappel ».

VII. Récapitulatif :

Dans le tableau ci-dessous (inspiré du tableau présenté dans (Yamouni F. 2010), nous présentons un récapitulatif des outils d’extraction de terminologie.
Notons que :

- « AL » signifie Approche Linguistique.
- « AS » signifie Approche Statistique.
- « AM » signifie Approche Mixte.

Logiciels	Différentes approches			Références
	AL	AS	AM	
FASTER	X			Jacquemin 1996, 1997, 2001
TERMS			X	Justeson et Katz 1995
LIKES		X		Frath et al. , 2000
SYNTEX			X	Bourigault et Fabre 2000
EXIT			X	Roche et al. 2004
ANA		X		Enguehard 1993
NOMINO	X			David et plante 1990
TERMIGHT			X	Dagan et church 1997
INTEX	X			Savary 2000
MANTEX		X		Frath et al. 2000 Ouslati 1999
CLARIT			X	David A. Evans et Zhai 1996
ACABIT			X	Daille 1994
XTRACT			X	Smadja 1993
LEXTER	X			Bourigault 1993
TERMINAE	X			Szulman S. 1999
XTERM	X			Cerbah F. 1999
ESATEC			X	Biskri et al. 2004
FIPS			X	Nerima et al. 2003, 2006
WASPBENCH			X	Kilgariffet Tugwel 2001

Tableau -8- : Un récapitulatif des outils d'extraction de terminologie.

VIII. Conclusion :

Dans ce chapitre, nous avons commencé par présenter la terminologie, lorsque on parle de cette dernière on entend tout d’abord l’ensemble des termes appartenant à un domaine particulier par exemple : la terminologie de la médecine, du droit, de l’informatique etc.

Sur le plan scientifique et technique, la terminologie se trouve aujourd’hui au confluent de toutes les disciplines liées à la communication.

Nous avons également présenté l’extraction de terminologie, il s’agit essentiellement d’identifier des Termes Potentiels (TP) dans un texte spécifique ou un ensemble de textes (corpus). Son objectif étant l’extraction d’unités susceptibles d’être des termes, il nous faut prendre en considération les caractéristiques linguistiques de celles-ci. De nombreuses études ont permis de cerner à peu près les caractéristiques des unités terminologiques [Benveniste, 1966], [Condamines et al. 1993], [Jacquemin et al., 1994] [Bourigault, 1994], [Katz et al. 1995]. L’extraction de terminologie est un moyen rapide d'acquérir des connaissances sur un domaine et représente une étape importante lors de la création de bases de données terminologiques.

Elle s'accomplit manuellement grâce à des terminologues qui isolent manuellement les termes à partir des textes, automatiquement à l’aide d’utilisation d’outils utilisant différentes méthodes d’extraction ou bien semi-automatiquement en utilisant un logiciel pour extraire les candidats termes, Cependant, après l'extraction ,des spécialistes doivent décider si les résultats obtenus sont adaptés ou non.

L’extraction manuelle de terminologie demande beaucoup de temps, les outils d’extraction automatique sont une aide au terminologue mais la majorité des systèmes nécessitent l’intervention des spécialistes pour la validation des termes extraits.

Ce chapitre nous a permis aussi d’établir une analyse sur les différentes méthodes d’extraction qui peuvent être classé en trois catégories :

les méthodes linguistiques, les méthodes statistiques, les méthodes mixtes, ainsi que les outils d’extraction mais la plupart ne sont ni commercialisés ni d’utilisation libre.

Chapitre III

Acquisition de connaissance

I. Introduction :

Tout au long de leur scolarisation, les élèves doivent acquérir des connaissances. Cette confrontation au savoir dure aujourd'hui entre 15 et 20 ans, voir parfois plus.

Les connaissances acquises sont diverses et nombreuses. Certaines, de plus en plus spécialisées dans un champ disciplinaire donné, constitueront le fondement des futures compétences professionnelles, d'autres deviendront des connaissances générales, d'autres encore seront, selon toute apparence, oubliées.

L'école ne représente de plus qu'une partie du quotidien des élèves. S'ils y acquièrent la plupart de leurs connaissances, ils n'en continuent pas moins à apprendre chez eux, devant la télévision, devant des jeux vidéo, dans la rue, dans les bandes, dans la société, chez un professeur de musique, au fond d'une cave ou le samedi sur un stade de football. L'être humain semble se caractériser par une peur viscérale du « vide ». En effet, ne réellement rien faire paraît difficile.

Toute expérience est matière à apprentissage, la capacité à apprendre de l'être humain paraît phénoménale : se souvenir du dernier emploi du temps de ses amis, se lancer dans des discussions, poser des questions, attendre des réponses, apprendre explicitement et implicitement à conduire, à manger proprement, tout en apprenant à résoudre un problème d'équation à deux inconnues, tout en apprenant les paroles du dernier remix à la mode, diffusé en bruit de fond par la radio.

En effet L'acquisition de connaissances se définit comme « le transfert et la transformation d'une expertise d'une source de connaissances à un programme ».

"L'acquisition des connaissances est une des difficultés majeures de la conception des systèmes experts"

La conception d'un système expert exige un travail de transfert de connaissance entre des sources d'expertise (experts humains ou documents) et un outil informatique de façon à disposer ensuite d'un Système à Base de Connaissances (SBC) pouvant être consulté comme un expert.

Traditionnellement, cette construction met en jeu un ou plusieurs experts, ainsi qu'un ou plusieurs ingénieurs de la connaissance (les cogniticiens) chargés d'extraire la Connaissance des experts pour la traduire dans le formalisme de représentation des connaissances offert par un générateur de système expert cible [Hayes-Roth, 1983], [Hiarnioli, 1985], [Bonnet et al., 1986].

II. Acquisition des connaissances [Nicolas, 2000] :

Le processus d'acquisition des connaissances perme le traitement de connaissances de bas niveau (données) pour aboutir par la suite à des connaissances de haut niveau (structure). Comme l'explique l'illustration ci-dessous.

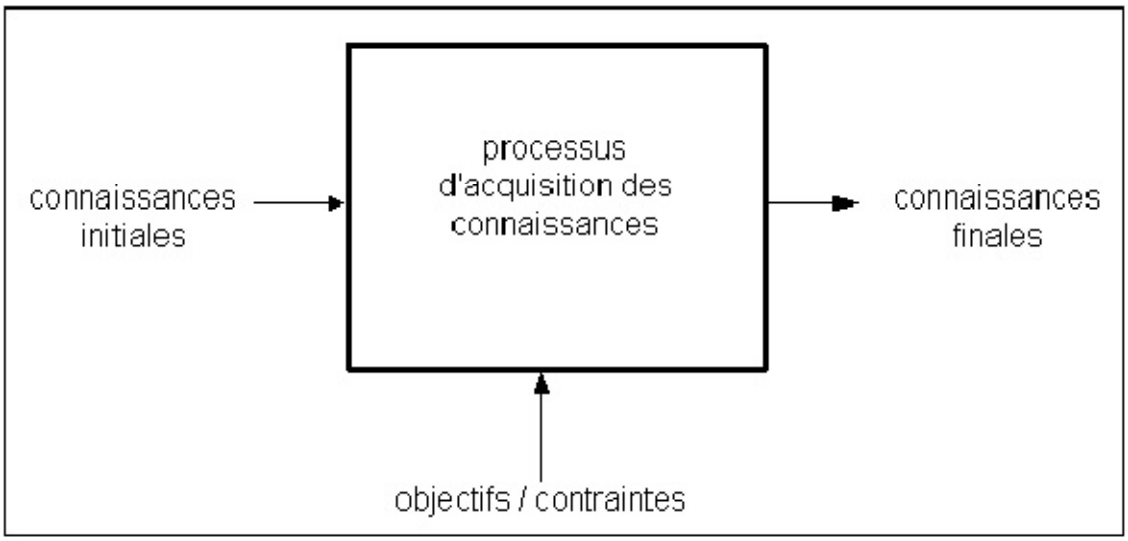


Figure 1 : processus d'acquisition des connaissances [Nicolas, 2000]

Le choix des connaissances n'est pas obligatoire mais celui-ci peut être vu comme étant le résultat d'une collecte soumise à une série de contraintes qui influent principalement sur les résultats.

La connaissance initiale rassemble l'ensemble des faits et des relations nécessaires à la résolution d'un problème et est composée deux parties :

- Les données, caractérisées par des objets et des variables :

Les variables manipulées peuvent être classées en :

- Variables qualitatives
- Variables quantitatives.

§ Les variables quantitatives :

Sont des variables dont les valeurs font parties d'un ensemble infini, complètement ordonné et assimilable à R+.

§ Les variables qualitatives :

Quant aux variables qualitatives, les valeurs font parties d'un domaine fini ou dénombrable. Elles peuvent être :

- monovaluées : chaque objet prend une seule valeur par variable (exemple: couleur="rose").

- multivaluées : chaque objet peut prendre plusieurs valeurs simultanées (exemple : avion: altitude=10 latitude=100 longitude= -15).

- Les connaissances supplémentaires :
  - L'information liée au domaine (expert…).
  - L'information liée au but (critère à optimiser...).
  - L'information sur les données : relations d'ordre (à côté, au dessus…), taxinomies (arborescences…) etc.

En effet, le but de l'Acquisition des Connaissances (AC) est de construire un Système à Base de Connaissances (SBC) qui va permettre soit :

l'exécution de tâches qui nécessitent une expertise (exemple : un système expert), soit la collecte et la représentation d'une manière explicite et structurée une expertise.

Le premier cas inclut le second, car pour les méthodologies actuelles d'acquisition des connaissances, la création d'un SBC implique tout d'abord de représenter à un niveau conceptuel l'expertise nécessaire au SBC, c'est-à-dire de créer un modèle conceptuel de cette expertise [Newell, 1981].

II.1. Acquisition de connaissances pour les systèmes à base de connaissances :

Pour [Aussenac-Gilles, Krivine et Sallantin 1992], "Le domaine de l'acquisition de connaissances pour les systèmes à base de connaissances se caractérise par l'identification et l'agencement des processus requis pour l'élaboration (conception, évaluation et évolution) d'un système à base de connaissance (SBC) à partir de sources diverses de connaissances (documentaires, humaines et expérimentales, etc).
Le résultat attendu de cette démarche est de fournir au futur système les connaissances qui seront à la base de ses compétences. Le maître d'œuvre de l'acquisition de connaissances est le cogniticien : il orchestre l'intervention des différents processus, acteurs et agents" [Aussenac-Gilles, Krivine et Sallantin 1992].

Le processus d'acquisition se fait en quatre étapes, toujours selon [Aussenac-Gilles, Krivine et Sallantin 1992] :

1. L'acquisition dirigée par les données.
2. La construction d'un schéma conceptuel du Modèle Conceptuel.
3. L'acquisition dirigée par le Modèle Conceptuel.
4. L'opérationnalisation du Modèle Conceptuel.

1. L’acquisition dirigée par les données :

Au début de la phase 1 (Acquisition dirigée par les données), aussi appelée phase d'identification du domaine, le cogniticien ne connaît pas, a priori, le domaine qu'il est chargé de structurer. Il procède donc à la collecte de données qui vont lui permettre de se faire une idée. Ses sources sont diverses :

- Entretiens avec des experts,
- Documentation technique,
- Ouvrage,
- Livre, article et journaux,
- Etc.

2. La construction d'un schéma conceptuel du Modèle Conceptuel :

Après avoir recueilli les données par le cogniticien dans la phase précédente, au niveau de celle-ci ces premières données lui donnent la possibilité de construire un schéma du modèle conceptuel, c'est-à-dire une hiérarchie de concepts qui décrit le domaine, ainsi que les concepts spécifiques permettant des raisonnements intéressants.
A ce stade, il s'agit essentiellement d'une structure vide vu que le schéma contient peu de connaissances.

3. L’acquisition dirigée par le modèle conceptuel :

Au niveau de cette phase, les experts corrigent et valident le schéma conceptuel qui sert de guide à l'acquisition de connaissances pour qu’ils se chargent de remplir le cadre avec leur connaissance du domaine.

4. L’opérationnalisation du modèle conceptuel :

Enfin, l'opérationnalisation consiste à implémenter les données dans un système informatique pour le rendre opérationnel.

II.2. Le cycle de vie d'un Système à Base de Connaissances (SBC) :

Souvent, les méthodes d'acquisition des connaissances sont associées à une certaine vision du cycle de vie du SBC. Les premières propositions [Hayes-Roth, 1983], [Hiarnioli, 1985] décrivant le cycle de vie d'un SBC étaient souvent basées sur le prototypage rapide. Celles-ci distinguaient les phases suivantes :

- L'étape d'identification permet de mettre en place des acteurs et des ressources, le choix d'un problème approprié ainsi d'objectifs précis.

- L'étape de développement en quelques mois d'une maquette : cette phase comprend la conceptualisation du problème (à partir d'entretiens cogniticien-expert. Explicitation des principaux concepts, relation et stratégies de résolution du problème). Puis la formalisation (représentation dans un formalisme). L'implantation dans un outil choisi de façon adéquate et enfin des testes pour valider cette maquette.

- L'étape de développement d'un système complet (par extension ou modification totale de la base de connaissances de la maquette).

- L'étape de validation du système complet : cette phase consiste à évaluer le système obtenu par rapport aux tests ou par d'autres experts que ceux ayant participé à son développement.

- L'étape d'intégration dans l'entreprise (accompagnée d'une éventuelle connexion à d'autres logiciels) et maintenance.

Un tel cycle de vie repose sur le prototypage rapide car le cogniticien implante dés que possible une maquette, après un nombre suffisant d'entretiens.

Il raffine ensuite cette maquette grâce à de nouvelles interviews, ce qui entraine un certain nombre de retoures-arrière. D'autres descriptions du cycle de vie [Haiker et Welz, 1989], [Klinker, 1989] sont également basées sur le prototypage rapide.
Dans une approche différente, appelée l'acquisition structurée des connaissances [Gruel et Breuker, 1985], [Wielinga et Breuker, 1985], [Wielinga et Breuker, 1986], [Johnson, 1988], l'implantation n'a lieu que plus tard, une fois la connaissance décrite dans une représentation intermédiaire.

Le cycle de vie proposé dans [Norbo et al., 1989] se rattache à cette seconde approche :

- Définition du problème.
- Modélisation de la connaissance : cette phase permet le recueil des connaissances et le développement des modèles du domaine. On obtient par la suite un ensemble de modèles sur lesquels sera basée la conception du système.
- La conception du système.
- L'implantation et tests.

Remarquons que la description du cycle de vie du SBC correspond déjà à une méthode d'acquisition des connaissances. Nous identifions ainsi deux approches :

- Le prototypage rapide.
- L'acquisition structurée basée sur une représentation intermédiaire.

III. Outils d'acquisition des connaissances :

III.1. Généralité :

On peut trouver une revue détaillée des outils d'acquisition de connaissance dans [Becker et Sehnan, 1989] et [Boom, 1989]. Comme le souligne [Boom, 1989], nous envisageons un certains nombre de critères de comparaison de tels outils présentés comme suit:

- Le niveau de généralité et le type des tâches visées :
  Même les outils d'acquisition les plus généraux ne visent pas toutes les tâches possibles. Certain outils sont spécialisés dans les tâches de diagnostic : ETS [Boose, 1985], [Boose, 1985a], MOLE [Eshelinan, 1988], [Eshelinan, 1988a], [Eshelinan et Mcdermott, 1988], MORE [Kahn et al., 1985] [Kahn, 1988], ROGET[Bennett, 1984], [Bennett, 1985] D'autres aident à des tâches d'analyse générales : AQUINAS [Boose et Bradshaw, 1987], [Boose et Bradshaw, 1988], [Boose et Kitto, 1987], KRITON [Diederich et May, 1987], [Diederich, 1987], [Linster, 1988], [Linster, 1988a] alors que quelques outils se focalise sur les problèmes de synthèse, généralement la conception comme DSPL ACQUIRER [Chiang et Brown, 1987]. SALT [Marcus, 1988], [Marcus, 1988a], [Marcus, 1988b], [Marcus, 1987] et 3DKAT [Meng, 1989], [Epp et Riera, 1989], [Faisandier, 1989].

- L'exploitation d'un domaine, d'une méthode ou d'un langage d'implantation : Certain outils exploitent la connaissance du domaine et sont adaptés à une application donnée (comme OPAL [Museu, 1987]) alors que d'autre exploitent plutôt une méthode de résolution de problème particulière pour guider l'acquisition des connaissances (AQUINAS, MOLE, SALT) ou même génèrent une base de connaissance directement utilisable (AQUINAS, ETS, KNACK [Klinker, 1988], KRITON, MOLE, MORE, OPAL, ROGET, SALT, TEREISIAS [Davis et Lento, 1982]).

- Les techniques ou méthodes d'acquisition des connaissances :
  Certain outils proposent des techniques assistées par ordinateur, interactives [Kelly, 1955] (ETS, AQUINAS, KRITON) ou sur l'analyse de protocoles (KRITON) ou sur un mode d'interview automatique (AQUINAS, KRITON, MOLE, ROGET, SALT). SIS [Kawaguchi et al., 1986], [Kawaguchi et al., 1987] aide à construire automatiquement des systèmes d'interview adaptés à des problèmes quelconque. On peut le considérer comme stockant de la connaissance sur « l'art des interview ». SIS dispose de sept stratégies primitives pour poser des questions.
  A partir de ces stratégies, on peut en définir de nouvelles, adaptées à une application particulière. KRITON, KADS ou MACAO acceptent l'utilisation de différentes techniques de recueil des connaissances.

- La phase du cycle de vie où l'outil peut être utilisé :
  Certain outils aident le cogniticien dans la phase de recueil d'expertise, avant l'entrées effective de la base de connaissance. D'autres se chargent également de la génération de la base de connaissances (AQUINAS, ETS, KNACK, KRITON, MOLE, MORE, OPAL, ROGET, SALT, TEREISIAS). D'autres proposent une méthodologie complète pour le processus d'acquisition complet (KADS, KOD, MAKAO). Dans certain, l'analyse doit être complète avant que l'implantation commence (KADS), alors que, dans d'autres l'implantation est effectués de façon incrémentale (prototypage rapide permis par AQUINAS, ETS, KITTEN, KSSO, MORE, MOLE, PLANTE et SALT).

- Le nombre de systèmes effectifs construits avec l'outil :
  De nombreux systèmes ont été construits en utilisant des outils comme AQUINAS, ETS, KNACK et MOLE, par exemple.

- L'utilisation du système :
  L'utilisateur visé (expert ou cogniticien) varie suivant l'efficacité et la vitesse d'utilisation de l'outil et le degré d'entrainement nécessaire pour l'utiliser.

Par exemple, des outils comme ETS sont facilement utilisables directement par les experts qui apprécient la transparence de l'interface [Boom et al., 1989]. De nombreux outils comme (AQUINAS, ETS, KITTEN, KNACK, MOLE, MORE, OPAL, SALT, TEREISIAS) sont présentés comme destinés à l'expert [Boom, 1989], alors que MACAO et 3DKAT visent plutôt le cogniticien et éventuellement l'expert.

III.2. Exemples :

III.2.1. Outils orientés analyse [Dieng, 1990] :

- KRITON :

KRITON [Diederich, 1987], [Diederich et May, 1987], [Linster, 1988], [Linster, 1988a], est un outil destiné aux tâches d'analyse et permet la création de bases de connaissances pour l'environnement BABYLON [Di Primio et Brewka, 1985]. Il permet le recueil des connaissances grâce à trois techniques :

- L'analyse de protocoles pour acquérir la connaissance.
- L'interview.
- L'analyse de textes pour acquérir la connaissance statique.

- L'analyse de textes :

Les textes sont lus à partir d'un fichier, les noms y sont mis en valeur. Le cogniticien peut inclure des noms dans la hiérarchie décrivant la façon dont le texte présente l'organisation des concepts du domaine.

- L'interview :

Cette technique est utilisée dans le but de compléter l'information collectée à partir de textes. Elle repose sur la technique de grilles-répertoire et permet d'acquérir les attributs qui décrivent les concepts du domaine.

- L'analyse de protocoles :

Quand l'expert résout un problème, il parle dans un microphone et mentionne ses actions, conclusion ou observation.

Le résultat de ce recueil est une représentation intermédiaire sous forme de réseaux, où les nœuds correspondent aux concepts et les liens aux relations entre eux. On distingue des relations structurelles (connaissance statique, recueillie à partir des interviews) et des relations associatives (extraites grâce à l'analyse de protocole). Le tout est ensuite traduit dans le formalisme de BABYLON (schémas et règles de production).

III.2.2. Outils orientés synthèse [Dieng, 1990] :

- 3DKAT :

3DKAT [43 [Meng, 1989], [Eshelman et Mcdermoot, 1986], [Epp et Reira, 1986], [Esterbrook, 1989] est un outil d'acquisition des connaissances dédié aux applications de conception. Il repose sur l'hypothèse que l'expert utilise en général un modèle implicite de ce qu'il veut concevoir. Au cours des entretiens, le travail du cogniticien va consister à se recréer implicitement ce modèle à partir de ce qu'il a compris des informations fournies par l'expert. 3DKAT permet donc au cogniticien de rendre explicite ce modèle.

Ce modèle est basé sur les dépendances entre les paramètres important intervenant lors de la résolution du problème. Ce modèle peut être représenté par un graphe de dépendance appelé PDOG (Parameter Dependency Oriented Graph). Un nœud du graphe correspond soit à un attribut de l'objet à concevoir ou d'un de ses composants, soit à un paramètre provenant de l'extérieure.

3DKAT propose une typologie des relations possibles entre les nœuds. Cette hiérarchie de lien est extensible par exemple par des relations spécifiques à une application.

Une relation particulière permet d'exprimer comment un paramètre influe sur un autre. Elle permettra de visualiser dynamiquement l'influence de la modification d'un paramètre donné et donc de valider l'aspect dynamique de la connaissance extraite.

Au cours de la phase d'acquisition des connaissances. Le cogniticien modélise sous forme d'un PDOG le modèle de dépendance à l'expert et le complète ou le corrige avec lui.

IV. Méthodes d'acquisition des connaissances [Dieng, 1990] :

Dans Cette section, nous examinons quelques méthodes d'acquisition :

IV.1. KADS :

- ### Présentation générale :

  KADS acronyme de Knowledge Analysis and Design System est une méthodologie de développement pour les systèmes à base de connaissances développé à Amsterdam au début des années 1990.

  La méthode KADS [Breukcer and Wielinga, 1985], [Anjewierden, 1987], [Bredeweg et al., 1988], [Breuker et Wielinga, 1989], [Brunet et Breuker, 1990] traite tout le processus d'acquisition des connaissances, depuis l'organisation du domaine jusqu'au développement d'un système complet. Elle permet de décomposer le problème et de séparer l'analyse de la connaissance et l'implantation.

- ### Le modèle de cycle de vie :

  KADS propose un modèle de cycle de vie, basé sur les méthodes de développement de logiciels. Une analyse complète des données précède la conception et l'implantation du système expert, ce qui diffère donc du prototypage rapide. Les résultats de la phase d'analyse servent d'entrée à la phase de conception.

KADS repose sur l'hypothèse que les phases d'analyse et de conception doivent être séparées : il n'y a pas de retour sur la phase d'analyse, à partir de celle de conception ou d'implantation [Breuker et Wielinga, 1989].

- Les langages de modélisation :

KADS comprend :

- Un langage de modélisation conceptuelle (un langage pour le modèle conceptuel KCML).
- Un module de conception.
- Un module de modalités

- Le langage de modélisation conceptuelle KCLM (Kerridge-C-Macro-Language) :
  Ce langage peut être utilisé comme une représentation intermédiaire entre les données de l'expertise et la conception / implantation d'un système expert. Indépendant du formalisme d'implantation du future SBC. KCLM permet une conceptualisation de haut niveau de l'expertise du domaine.

- Un module de conception :
  Le module de conception décrit comment le modèle conceptuel peut être traduit dans une architecture adéquate.

- Un module de modalité :
  Ce module spécifie la coopération et la communication du SBC avec l'utilisateur ou avec d'autre système.

- Technique et outils :

  KADS dispose de diverses techniques de recueil des connaissances et de différentes méthodes d'analyse des données recueillies. KADS Power Tools fournit des outils facilitant les phases d'analyse et de conception. L'outil SHELLEY [Bouchet et Brunet, 1989] permet d'analyser les données ainsi que de construire un modèle conceptuel.

Comme le souligne [Bouchet et Brunet, 1989], la méthode KADS peut être utilisée seule ou en complément d'autre méthode. Elle se base sur la séparation des phases d'analyse et de conception.

IV.2. MACAO :

MACAO s'adressent plutôt au cogniticien mais permettant aussi à l'expert d'exprimer directement ses connaissance. MACAO [Aussenac, et al., 1989], [Aussenac, 1989], [Aussenac et Souhie, 1990] est une méthode complète de développement d'un système à base de connaissance.

- Le cycle de vie :

La méthode MACAO se déroule en quatre étapes :

- Identification de l'expertise.
- Recueil des protocoles de résolution de problème.
- Conceptualisation, analyse et formalisation de la connaissance.
- Validation interactive des connaissances par l'expert.

- Identification de l'expertise :

  Cette étape est basée sur une succession de dialogues expert-cogniticien et sur des observations de l'expert. On peut ainsi spécifier l'environnement de l'utilisateur, ainsi que les principales caractéristiques de l'expertise.

- Recueil des protocoles de résolution :

  L'expert est invité à résoudre divers problème des catégories identifiées dans la première étape, et ce, dans des conditions aussi proches que possible de la réalité.

  Il explique ce qu'il fait en raisonnant à haute voix ou en fournissant toute information supplémentaire demandée par le cogniticien. Ces dialogues sont enregistrés.

- Analyse et formalisation de la connaissance :

  Le cogniticien, guidé par le logiciel MACAO, analyse les informations précédentes en se basant sur le modèle cognitif, grâce à cette analyse, on obtient plusieurs graphes de connaissance correspondant à chaque problème (graphe instanciés) ou décrivant chaque catégorie problème (graphe génériques).

- Validation interactive des connaissances par l'expert :

  La validation a lieu directement par l'expert qui, lors d'entretiens centrés, peut corriger le graphe général des connaissances. MACAO lui offre un éditeur interactif pour afficher ou mettre à jour la connaissance. La validation est achevée quand l'expert est satisfait des différents graphes obtenus.

V. Acquisition automatique :

Selon [Frath, 1997] l'acquisition de connaissances est donc dirigée par les données dans un premier temps, puis par le modèle un second temps. Dans [Bourigault et Lépine, 1994] et [Bourigault, 1994], les auteurs décrivent comment est utilisé LEXTER lors de ces deux phases.

Une phase communément dénommée dépouillement permet au cogniticien de se familiariser avec le domaine et de repérer la terminologie par élimination des candidats termes non pertinents.

Lors de la phase descendante, appelée fouille, le cogniticien possède un schéma de modèle conceptuel incomplet, qu'il s'agit maintenant de compléter.

Dans ce cas, le logiciel LEXTER, fournit simplement une aide au cogniticien pour la construction manuelle du modèle conceptuel.

Cependant certains travaux, constatant que les textes contiennent des connaissances, tentent à les utiliser soit pour produire automatiquement ou semi-automatiquement une représentation du domaine à partir des textes (phase ascendante), soit pour extraire des connaissances sur le texte par rapport à un modèle existant (phase descendante), soit lors des deux phases. Bourigault appelle ces logiciels des outils de transfert.

A titre d'exemple, nous allons étudier ci-dessous le projet PLINIUS, dont l'objectif est de développer un système capable d'acquérir de manière semi-automatique les connaissances à partir de résumés de communications scientifiques [van der Vet, de Jong, Mars, Speel et Stal, 1994].

V.1. PLINIUS : un système d'acquisition semi-automatique de connaissances à partir de textes [Frath, 1997].

L'ambition du projet PLINIUS est de construire un système d'acquisition semi-automatique.

PLINIUS est destiné à traiter des résumés de publications scientifiques, ce qui présente l'avantage de réduire les difficultés linguistiques : le corpus ne contient pas de phrases interrogatives ni impératives, le nombre d'ambiguïtés sémantiques est réduit puisque les mots sont toujours utilisés dans un contexte connu.

V.1.1. Architecture générale :

Dans une première étape, l'objectif général est de transformer les phrases du corpus en un langage de représentation de connaissance, puis de les sauvegarder dans une Base de Connaissances Intermédiaires (BCI).

Dans une seconde étape, le système procède à une intégration des données de la BCI dans une base de données intégrée.

A chaque étape, les résultats sont contrôlés et éventuellement modifiées par un intervenant humain. Une interface utilisateur devra permettre l'interrogation du système.

V.1.2. Traitement linguistique du corpus :

La phase du traitement linguistique du corpus consiste à transformer le texte en un formalisme de représentation de connaissances.

En effet, les auteurs déclarent dans cette section que la majorité de leurs procédures de traitement linguistique sont soit en cours de développement, soit en projet.

V.1.3. Traitement non-linguistique :

La phase du traitement non-linguistique permet d'utiliser les résultats obtenus dans la phase précédente, qui stockés dans une base de connaissances intermédiaire (BCI), dans le but de répondre aux différentes requêtes de l'utilisateur.

- Exemple de déroulement d'une requête :

  Par exemple, Si l'utilisateur veut connaître la force de torsion d'un matériau dénommée « hydroxyapatite » alors le système collectera en premier temps les assertions dans la BCI concernant cette requête.

  Ensuite, il tentera en second lieu la résolution de divers conflits qu'il ya entre les différentes assertions par exemple « les différentes valeurs de torsion » en ayant ainsi éliminé le conflits ou déclarant être incapable de le résoudre.

En effet, les auteurs présentent PLINIUS comme étant un projet et les problèmes de la capacité du système à répondre à des requêtes autres que numériques, sont loin d'être réglés.

VI. Conclusion :

L'acquisition de connaissances est une méthodologie pour acquérir et représenter des connaissances à partir de sources différentes, essentiellement des entrevues avec des experts et des textes.

La recherche sur l'acquisition des connaissances est extrêmement vivante, comme en témoignent les séminaires de plus en plus nombreux consacrés à ce thème : Knowledge Acquisition for Knowledge Based Systems Workshop (KAW), European Workshop ou Knowledge acquisition (EKAW). Journées française sur l'Acquisition des Connaissances (JAC).

En France, quelques outils commerciaux commencent à apparaitre : KATE (acquisition des connaissances par apprentissage), KOD-STATION, NEXTRA, SHELLEY alors qu'au niveau recherche, divers projets sont en cours sur des méthodes ou outils : MACAO, SMAC, 3DKAT.

D'autre part, une bibliographie extrêmement fournie est présentée dans [Sigart, 1989a], [Sigart, 1989], [Aussenac, 1989], [Boom, 1989] : le lecteur y trouvera les références d'un nombre considérable d'article. Un résumé de nombreux outils d'acquisition est offert dans [Boom, 1989].

Chapitre IV

Construction d’une base de connaissance terminologique

I. Introduction :

Ce quatrième et dernier chapitre va nous permettre de mettre en pratique les recherches faites dans les chapitres précédents, et de proposer des méthodes afin de construire une Base de Connaissances Terminologiques (BCT) qui puisse représenter un corpus.

Depuis quelque temps, au niveau de l’université de Toulouse L’Equipe de Recherche en Syntaxe et Sémantique (ERSS) s’est engagée dans la constitution de Bases de Connaissances Terminologiques (BCT) à partir de corpus d'entreprises, le plus souvent dans le domaine spatial.

Dans les chapitres précédents, nous avons plusieurs fois mentionné la notion de Base de Connaissances Terminologiques (BCT), il s'agit maintenant d'approfondir cette notion puis de voir comment une Base de connaissances terminologiques pourrait être construite aussi automatiquement que possible à partir de textes.

2. Notion de Base de Connaissances Terminologiques :

La notion de base de connaissances terminologiques (BCT) est encore très mouvante, et selon [Séguéla et Aussenac, 1997], les Base de connaissances terminologiques développées sont très peu nombreuses.

Afin de l’étudier nous nous basons dans cette section sur les travaux de [Condamines & Rebeyrolle 1998] et [Séguéla & Aussenac 1997].

La notion de base de connaissances terminologique doit être analysée par rapport aux méthodes d'acquisition de connaissances que nous avons examinées dans le chapitre précédent.

Pour [Condamines et Rebeyrolle, 1998], une base de connaissances terminologique est organisée autour de quatre champs :

- Le terme.
- Le concept.
- Le lien terme-concept.
- Le texte.

La description linguistique du « terme » comprend les données proprement linguistiques : nature et genre, variantes de formes, etc.

Dans le système d'A. Condamines et J. Rebeyrolle, le « concept » n'est pas décrit de manière formelle, il est simplement associé à une définition en langue naturelle. Il comporte les données qui concernent le concept dénommé par le terme, sous forme d'une définition et de relations explicites.

Le champ « lien terme-concept » comporte des informations sur la validité du terme pour dénommer tel concept.

Enfin, en ce qui concerne le « texte », celui-ci permet de rendre compte des liens entre un terme et ses occurrences dans le corpus dont il est extrait.

En ce qui concerne les auteurs, « une BCT est ... avant tout un inventaire des termes du domaine enrichi d'informations conceptuelles permettant de donner un sens à ce dernier, de définir les notions qu'ils désignent et de justifier leur place dans la terminologie ».

Elle doit fournir aux utilisateurs un accès au corpus, et à des listes de termes et de concepts présentés par ordre alphabétique.

3. Quels outils utiliser ?

Il n’est pas envisageable de travailler sur des corpus sans la présence d'outils. En effet, les corpus sont trop volumineux pour pouvoir effectuer un traitement manuel.

Deux catégories d’outils sont envisageables pour construire une Base de Connaissance de Terminologie :

- Les outils terminologiques (outils d'extraction de candidats termes).
- Les outils d’analyse de corpus.

3.1 Outils à vocation terminologique :

Parmi les outils construits pour l’extraction des candidats termes. Deux classes ont été distinguées en fonction de la méthode d’analyse.

Certains outils mettent en œuvre une « méthode descendante » qui pose que le texte est l'actualisation d'un système notionnel qui lui préexiste. C'est le cas des outils Seek [Jouis, 1995], Coatis [Jackiewicz, 1996], Text Analyser [Davidson et al. 1998]. C'est le cas aussi de Termino [David & Plante 1990].

D'autres outils mettent en œuvre une « méthode ascendante » qui part du texte pour extraire des connaissances sur le domaine, essentiellement des termes.

Les outils d'extraction de terminologies qui fonctionnent par repérage de segments répétés, comme ANA [Enguehard, 1994], utilisent cette approche, de même que les outils qui utilisent les méthodes statistiques, comme Syclade [Habert et al. 1995].

3.2. Outils d'analyse de corpus :

La majorité des outils destinés à analyser les corpus pour le traitement automatique peuvent être également utilisés dans le contexte de l'extraction de données terminologiques.

Nombreux sont les outils de ce type (Sato, Tact, Hyperbase, etc.), et tous se caractérisent par la souplesse de l'accès au texte qu'ils proposent. En effet, ils laissent beaucoup de liberté à l'utilisateur en lui proposant de combiner des critères afin de se constituer une interrogation spécifique, adaptée à ses propres objectifs.

4. Méthode de construction de BCT à partir de textes :

Fondée sur l'analyse linguistique de corpus de textes d'un domaine, la méthode que l'équipe de recherche ERSS ont adopté consiste en l'intégration et en l'interprétation des résultats fournis par différents outils (Lexter, Sato).

Précisons que cette équipe de recherche s'est guidée, non seulement par son objectif de constitution d'une Base de Connaissance Terminologique (BCT), mais également par les connaissances qu'elle a du fonctionnement de la langue.

Après avoir rappelé le contexte dans lequel s'inscrit le travail de cette équipe, nous allons présenter dans ce qui suit la méthode proprement dite.

4.1. Le contexte de travail de l’équipe de l’ERSS :

La réflexion de l’ERSS concernant l’élaboration d'une Base de Connaissance Terminologique a trouvé un nouvel élan dans un projet financé par le GIS (Geographisches Informations System) Sciences de la Cognition, « Terminologie, modélisation des connaissances et systèmes hypertextuels de consultation de documentation technique », qui l’a amené à collaborer avec l'IRIT (Institut de Recherche en Informatique de Toulouse) et EDF.

En effet, la constitution d’une Base de Connaissances Terminologiques « BCT » doit s’appuyer, selon cette équipe, non seulement sur l’analyse linguistique de corpus de textes caractéristiques du domaine qu’elle veut modéliser, mais aussi sur les résultats fournis par les outils.

Précisons que le projet final consiste à évaluer les possibilités de réutilisation d'une Base de Connaissances Terminologiques pour construire une Base de Connaissances, qui doit être intégrée à un outil d'aide au repérage de données dans des documents.

Nous allons donner une brève présentation du corpus sur lequel travaille l’équipe ERSS ainsi que les outils utilisés.

4.1.1. Le corpus :

Il s’agit du guide MOUGLIS (Méthodes et OUtils de Génie Logiciel pour l’Informatique Scientifique) qui est constitué d’un ensemble de documents méthodologiques sur l'organisation de projets et les techniques de génie logiciel appliquées au développement de logiciels scientifiques.

MOUGLIS est à la fois un guide qui précise comment réaliser les différentes phases de la conception d’un logiciel, et également un modèle de rédaction de documents qui accompagnent chacune de ces phases.

4.1.2. Les outils utilisés :

Deux types d’outils ont été utilisés dans ce projet :

- Un outil d'extraction de candidats termes : Lexter
- Un logiciel d'analyse de textes : Sato.

Le logiciel LEXTER (Logiciel d’EXtraction de TERminologie) a été conçu par Didier Bourigault [Bourigault et Lépine, 1994] au sein de la Direction des études et Recherches d’EDF pour l’extraction de données terminologiques.

Le logiciel SATO (Système d'Analyse de Textes par Ordinateur) a été conçu par Jean-Guy Meunier et développé par François Daoust [Daoust, 1992]. L’accès au texte s'effectue au moyen de concordances, c'est-à-dire la recherche de l'ensemble d’occurrences d'un mot dans chacun de ses environnements contextuels. Sato permet l'ajout de propriétés aux mots ou aux segments textuels qui permettent de procéder à des catégorisations grammaticales, sémantiques, etc.

4.2. Les étapes de la méthode :

Deux étapes principales ont été distinguées :

- Etape 1 :

  Cette étape se base généralement sur l’utilisation de LEXTER pour le repérage de termes ainsi que ces variantes.

- Etape 2 :

Celle-ci correspond à l'utilisation de Sato et permet un travail sur les concepts et les relations conceptuelles.

A partir de la liste des « termes candidats », l'équipe de recherche ERSS travaille sur l'interprétation des contextes dans lesquels ils cooccurrent ce qui permet d'identifier les relations sémantiques (considérées comme équivalentes aux relations conceptuelles).

A la fin de cette deuxième étape, seuls les termes candidats qui appartiennent à la terminologie du domaine seront définitivement considérés comme des termes.

Etape 1 :

4.2.1. Traitement des résultats de Lexter : des « candidats termes » aux « termes candidats ».

L'analyse faite par l'équipe de recherche ERSS de la liste des « candidats termes » proposés par Lexter engendre beaucoup de bruit, c'est pourquoi, cette équipe à décrit un ensemble de critères linguistiques lui permettant de trier cette liste afin d'obtenir un ensemble de « termes candidats ».

Nous présenterons rapidement ces principaux critères de tri en les illustrant par des exemples :

4.2.1.1. Les résultats de Lexter : quelques chiffres

Pour le corpus MOUGLIS, Lexter propose 5878 « candidats termes » d'environ 50 000 mots. C'est cet ensemble d'unités qu'il s'agit de trier sur la base des critères qui seront présentés prochainement. L'application de ces critères linguistiques permet d'éliminer 74% des unités proposées par Lexter. L'équipe ERSS a obtenu ainsi une liste de 1516 « termes candidats» dont le fonctionnement en contexte peut ensuite être évalué grâce à Sato.

4.2.1.2. Critères de rejet des « candidats termes » de Lexter :

Face à un chiffre remarquable d'environ 6000 propositions de termes (4.2.1.1), il s'agit de spécifier des critères linguistiques stables permettant d'éliminer les unités qui constituent incontestablement du bruit et qui multiplient inutilement les données à analyser en contexte.

L'équipe de recherche ERSS a proposé d'éliminer toutes les unités qui répondent à l'un des critères suivants :

- critères syntaxiques.
- critères sémantiques.

- Critères syntaxiques de filtrage de la liste Lexter :

Pour les raisons syntaxiques suivantes, l'équipe de recherche ERSS a éliminé les candidats qui ne peuvent pas être des termes:

- erreur de découpage due à une confusion entre forme verbale et forme nominale (la forme verbale utilisée est considérée par Lexter comme une forme nominale) : offre, demande. Seul un examen en contexte permet d'identifier cette erreur.

- formes non-terminologiques : structures qui sont syntaxiquement correctes mais qui ne sont pas terminologiques, c'est le cas de : est-il, a-t-il, choix entre.

- erreur de découpage syntaxique effectuée par Lexter, par exemple : présume pas de la méthodologie de développement utilisée, au lieu de « ne présume pas de la méthodologie de développement utilisée.

- Critères sémantiques de filtrage de la liste Lexter :

Certaines formes sont considérées comme étant trop générales :

- certains groupes nominaux exemple : Synthèse du projet,
- certaines locutions prépositionnelles exemple : à la suite de, à l'issue de,
- certaines locutions adverbiales exemple : coup par coup,
- certains candidats termes dont la tête joue un rôle de déterminant,
- des formes contenant un anaphorique ou un cataphorique, dont l'interprétation est directement liée au contexte exemple : phase suivante de développement,
- des formes contenant un déictique, dont l'interprétation est liée à la situation d'énonciation exemple : mise à jour du présent document,
- des formes contenant un adjectif qualificatif trop vague ou trop général exemple : classique.

4.2.1.3. Critères de conservation des « candidats termes » :

Parallèlement, l'équipe de recherche ERSS a distingué deux critères de conservation de certains items :

- Critères morpho-syntaxiques
- Sémantiques qui suivent.

- Critères morpho-syntaxiques de conservation des candidats Lexter

  Il s'agit d'un cas où il existe une équivalence entre candidats termes qui manifeste la présence d'un concept unique.

  L'équipe de recherche ERSS a proposé des équivalences entre des candidats morphologiquement proches, comme : outil de GL du projet et outil génie logiciel du projet. Le logiciel Lexter propose d'ailleurs des hypothèses d'équivalence entre les candidats termes, en repérant des variations de détermination : Terme1 + déterminant Terme2 vs Terme1 + Terme2, comme dans l'exemple suivant : état de configuration logiciel vs état de configuration du logiciel.

De même, cette équipe rapproche des équivalences entre les formes siglées et les formes développées correspondantes : PDL = Plan de Développement Logiciel, DCP = Dossier de Conception Produit.
Cependant, il s'agit d'autant d'hypothèses d'équivalence qu'il est nécessaire de valider avec les experts du domaine.

- Critères sémantiques de conservation de candidats Lexter

  Cette équipe conserve les candidats termes qui sont en relation sémantique avec d'autres candidats termes. Nous distinguons deux types de relations :

  - des relations sémantiques, appelés de « bas niveau » : existence de paradigmes d'équivalence ou d'opposés parmi les expansions. Des structures du type : Terme1 + Adjectif 1, Terme 1 + Adjectif 2, dans lesquelles Adjectif 1 et Adjectif 2 sont équivalents (ou opposés) en langue. Exemple : conception générale vs conception détaillée.

  - Des relations sémantiques communément appelés de « haut niveau » : existence d'un paradigme taxinomique, les paradigmes proposés par Lexter sous forme de tête (T) / expansion (E) permettent de faire l'hypothèse qu'il existe une relation de hiérarchie entre T et T + E (ex : test, test d'acceptation, test de qualification, test de validation, test de recette, etc.).

A la fin de cette première étape, l'analyse des résultats de Lexter permet d'acquérir ou d'ordonner des connaissances sur les données suivantes :

- définition d'une liste de termes candidats,
- identification de variantes de termes (sigles, ellipses),
- identification de termes équivalents (validés par un expert),
- identification de relations conceptuelles de hiérarchie.

L'étape 2 :

4.2.2. Utilisation de Sato :

Ce qui guide l'équipe de recherche ERSS est que les données conceptuelles sont exprimées de manière implicite au niveau linguistique par des marqueurs, qui sont soit :

- des marqueurs indépendants d'un corpus particulier.
- des marqueurs dépendants d'un corpus.

L'étude de différents corpus et spécialement, l'analyse des marqueurs de relations définitoires [Péry-Woodley et Rebeyrolle, 1998], en particulier l'hyperonymie [Borillo 1997], a permis à l'équipe de recherche ERSS de spécifier le fonctionnement des marqueurs dans les corpus spécialisés.

Le travail de L 'ERSS sur les corpus consiste à utiliser les connaissances qu'ils ont, sur les marqueurs de relation, soit pour les adapter au corpus, soit pour en découvrir de nouveaux.

Pour cela, l'interaction permanente avec le logiciel LEXTER permet de prendre en compte les résultats immédiatement et d'ajuster leurs patrons de fouille.

À partir de marqueurs considérés comme généraux (indépendants du domaine), l'ERSS a définit une première série de patrons de fouille qu'ils ont testé sur le corpus. A partir de ces résultats produits, ils adaptent ces patrons de fouille de façons à les rendre plus pertinents, afin de limiter les bruits.

Suite aux couples de termes obtenus, ils recherchent de nouveaux patrons de fouilles de plus en plus spécifiques au domaine, et identifient les contextes dans lesquels ces couples cooccurrent.

Remarque :

Le patron de fouille est représenté de la manière suivante :

Il s'agit de rechercher tous les mots qui sont des termes identifiés par Lexter suivis (strictement adjacent) :

- du verbe être (au présent ou au futur),
- d'un déterminant (défini ou indéfini),
- d'un Nom repéré comme marqueur de la méronymie,

- éventuellement d'une préposition,
- et d'un autre terme de Lexter.

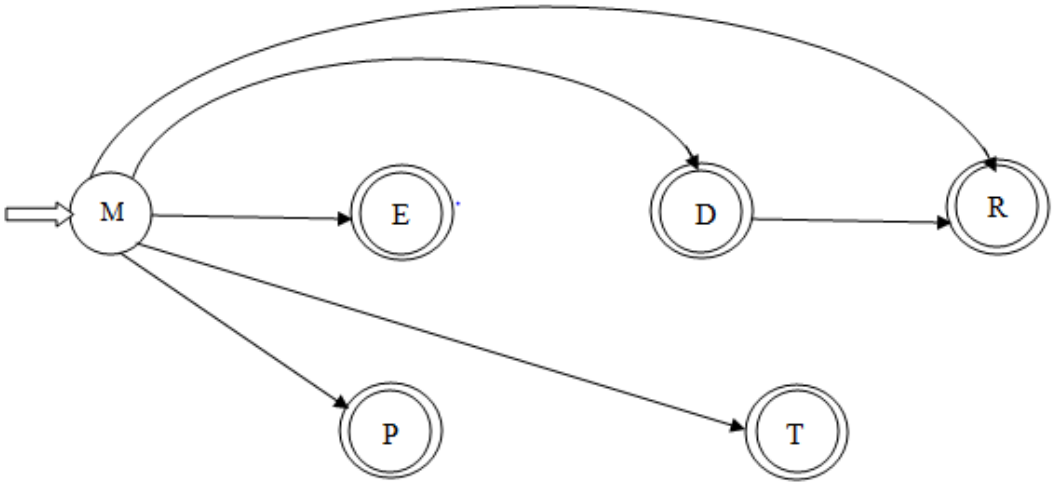


Figure 2 : Le patron de fouille.

Notons que :

M signifie : tous les mots qui sont des termes identifiés par LEXTER.

E signifie : verbe être au présent ou au future.

D signifie : déterminant (défini ou indéfini).

R signifie : marqueur de la méronymie.

P signifie : préposition.

T signifie : terme de LEXTER.

4.2.2.1. Un exemple de l'utilisation de Sato : repérage des concepts reliés par la relation partie-de dans le corpus.

Nous montrons étape par étape sur un exemple, partant d'une structure générale, comment on peut affiner une requête et mettre en évidence des phénomènes propre à un corpus.

- Utilisation de marqueurs :

  En s'appuyant sur les différents travaux sur la méronymie[1] [Otman, 1996] et [Jackiewicz, 1996], l'équipe ERSS a identifié un ensemble de marqueurs lexicaux de la relation de partie-tout en français. Il s'agit de certains noms, comme, partie, élément, phase, étape, membre, composant, constituant, famille, groupe, lot, ensemble, morceau, section, fragment, unité, pièce, portion, quartier, tranche, parcelle, particule, échantillon et de certains verbes : faire partie de, être membre de, être composé de, se composer de, constituer, inclure, comprendre, comporter, consister en, contenir, renfermer, englober, embrasser, appartenir à, être rattaché à.

  Appliquer dans Sato un patron de fouille qui fait intervenir l'un de ces marqueurs, sans autre précision, permet d'obtenir une variété de résultats qui contiennent beaucoup d'autres choses que l'expression d'une relation de méronymie. En effet, l'application de ce patron, dans Sato, donne 863 unités. Ceci peut en particulier s'expliquer par le fait que les mots « composant », « phase » et « unité » sont aussi des termes du domaine et qu'ils sont très fréquents : donc ne jouent pas toujours le rôle de marqueurs.

  Pour limiter le bruit, en filtrant les contextes les plus pertinents, cette équipe a proposé de définir des structures lexico-syntaxiques, plutôt que de se limiter à de simples marqueurs lexicaux très polysémiques.

[1] La méronymie est une relation sémantique et partitive entre mots d'une même langue exemple, un méronyme A d'un mot B dont le signifié A désigne une sous partie du signifié B. Ex : Le bras est un méronyme du corps.

- Recours à des structures lexico-syntaxiques : détermination de couples entretenant une relation partie-de :

Pour limiter le bruit, cette équipe a appliqué, dans Sato, des patrons de fouilles qui font intervenir non seulement les marqueurs lexicaux identifiés mais aussi les termes proposés par Lexter ayant la configuration syntaxique du type suivant : T1 est_une partie de T2.
Ce patron de fouille fournit un ensemble de seulement 21 occurrences parmi lesquelles :

Un état de configuration est un ensemble d'unités de configuration
La documentation de réalisation est une composante essentielle du produit-logiciel
La phase d'architecture est une phase du cycle de vie

Ces résultats obtenus permettent d'identifier certains couples de termes qui dénomment des concepts reliés par la relation partie-de (Etat de Configuration / Unités de Configuration, documentation de réalisation / produit-logiciel, phase d'Architecture / cycle de vie, etc.).
Ce sont ces résultats qui seront exploités dans la phase suivante de cette analyse.

- Retour au corpus : repérage de nouveaux marqueurs

L'ERSS a élaboré de nouveaux patrons de fouille en recherchant cette fois les contextes dans lesquels interviennent ensemble les termes identifiés comme étant effectivement reliés pour trouver d'autres marqueurs de méronymie et dont certains sont probablement spécifiques au domaine.
Avec cette méthode nous obtenons d'autres marqueurs, des noms comme : « enchaînement », « prolongement » et des verbes comme : « regrouper », « recouvrir », « être accompagné de », « comporter ».

- Repérage de nouveaux couples de termes reliés par un type de partie-de

  En utilisant les marqueurs ainsi repérés dans l’étape précédente, cette équipe de recherche a confectionné de nouveaux patrons dans le but d’identifier de nouveaux couples de termes.

L’intégration méthodique des résultats fournis par Sato dans la définition de nouveaux patrons de fouilles, permet la définition d'un réseau conceptuel qui est soumis, dans une dernière étape à un expert du domaine qui le valide. Alors seule, la liste des termes et des relations retenus est validé.

Sur le corpus MOUGLIS, l'application de cette méthode a engendré les résultats suivants :

- 10 relations de méronymie ont été identifiées nous les citons comme suit : est composé de, a lieu dans, a lieu pendant, débute pendant, se termine pendant, précède, conditionne le début de, conditionne la fin de, est le résultats de, est mis à jour pendant. Parmi eux, certaines relations sont spécifiques au corpus comme conditionne le début de et conditionne la fin de, d'autres sont indépendantes du domaine.

- Pour la relation, « est composée de », 41 marqueurs ont été définis et pour chacune des autres, 1 seul marqueur a été défini.

- Grâce aux marqueurs définis pour ces relations, 269 paires de concepts ont été repérées et réparties de la manière suivante : est composé de : 133, a lieu dans : 8, a lieu pendant : 59, débute pendant : 5, se termine pendant : 8, précède : 16, conditionne le début de : 11, conditionne la fin de : 14, est le résultat de : 8, est mis à jour pendant : 7.

4.3. Tableau récapitulatif :

Le tableau ci-dessous synthétise le rôle des outils à chaque étape lors de la construction d'une base de connaissance terminologique.

	OUTILS	
	LEXTER	SATO
Repérage des termes	- sélection des termes candidats sur la base de critères linguistiques.	- filtrage sur marqueurs lexico-syntaxiques.
Repérage des termes équivalents	- identification des variations morphologiques d'un terme.	- identification des synonymes - identification des formes développées, des formes sigles.
Repérage des relations entre concepts	- hypothèses sur les relations Hiérarchiques.	- identification de marqueurs spécifiques. - sélection de couples de termes en relation.

Tableau -9- : Les rôles des outils à chaque étape lors de la construction d'une BCT.

5. Conclusion :

Parvenu au terme de ce chapitre dont le rôle est de voir les étapes de construction d’une base de connaissances terminologiques, nous pouvons constater que la BCT se veut simplement une structuration de données aussi fidèles que possible au corpus d'origine pour permettre des utilisations applicatives.

En effet, s’il est incontestable ou évident de construire une BCT à partir de corpus qui ne peut être faite sans avoir recours à des outils, il est utile alors de s’interroger sur les résultats qu’ils fournissent au niveau d’une analyse linguistique.

A partir de cette interrogation, la méthode proposée dans cette section tente d’apporter plus de clarté c'est-à-dire une réponse en précisant comment l’analyse linguistique de corpus peut être assistée par les résultats de différents outils à chaque étape de la constitution d’une BCT.

Conclusion générale :

Le projet réalisé s'inscrit dans le cadre du traitement automatique des langues naturelles, il peut être utile dans plusieurs domaines tels que les correcteurs orthographiques ou grammaticaux, les dictionnaires électroniques, l'interrogation de bases de données en langue naturelle, la traduction automatique….etc.

Arrivés au terme de ce travail dont le but de l'étude actuelle est de représenter les recherches acquises concernant l'extraction automatique de la terminologie et l'acquisition de connaissances, il s'agit maintenant de l'achever de manière générale et de cerner ce qui nous semble en être l'intérêt essentiel.

En effet, le chapitre introductif nous a donné l'opportunité d'assimiler de manière générale certaines connaissances concernant l'étude linguistique de la langue française.

Puis dans le second chapitre, nous avons abordé le concept de l'extraction de terminologie qui est une application du traitement automatique du langage naturel consistant à extraire une liste de termes à partir de textes spécifiques ou un ensemble de textes appelé corpus.

Quant au troisième chapitre, nous nous sommes focalisés sur l'acquisition de connaissances à partir de textes qui consiste, idéalement, à générer une représentation structurée de textes fournis en entrée à un système informatique.

Le dernier chapitre nous a permis de proposer les étapes ainsi que les méthodes nécessaires à la construction d'une base de connaissances terminologiques.

Les perspectives :

Du point de vue pratique, il s'agit de poursuivre les recherches actuelles afin d'implémenter et construire une véritable Base de Connaissances Terminologiques (BCT) capable de représenter le contenu d'un corpus.

Conclusion générale

En effet, nous pouvons constater que la demande des entreprises pour la constitution de données terminologiques est forte, c'est sans doute cette demande qui a fait la multiplication de prototypes visant à automatiser le processus de constitution de données terminologiques à partir de corpus.

Liste des acronymes

ACABIT: Automatic Corpus based Acquisition of Binary Terms.

ALPAC : Automatic Language Processing Advisory Committee.

AL: Approche Linguistique.

AM: Approche Mixte.

ANA: Acquisition Naturelle Automatique ou Apprentissage Naturel Automatique.

AS: Approche Statistique.

ASCII: American Standard Core for Information Interchange.

BCI: Base de Connaissance Intermédiaires.

CAT : Center of Alternative Technology.

CEI : Commission Electrotechnique Internationale.

CT: Candiats Termes.

DCP: Dossier de Conception Produit.

EKAW: European Workshop.

ERSS: Equipe de Recherche en Sytaxe et Sémantique.

EXIT: EXtraction Itérative de la Terminologie.

FASTER : Filtrage et Acquisition Syntaxique de TERmes.

GL: Génie Logiciel.

HTML: HyperText Markup Language.

IRIT: Institut de Recherche en Informatique de Toulouse.

JAC: Journée française sur l'Acquisition de Connaissances**.**

KADS: Knowledge Analysis and Design System.

KAW: Knowledge Acquisition for knowledge based systems Workshop.

KCLM : Kerridge-C-Macro-Language.

KOD : Knowledge Oriented Design.

LEXTER: Logiciel d'EXtraction TERminologique.

LIKES: Linguistic and Knowledge Engineering Station.

MAHT : Machine Assisted Human Translation.

MOUGLIS: Méthodes et Outils de Génie Logiciel pour l'Informatique Scientifique.

PDL : Plan de Développement Logiciel.

PDOG : Parameter Dependency Oriented Graph.

SALT : Standards-Based Access to Multilingual Lexicons and Terminologies.

SATO: Système d'Analyse de Textes par Ordinateur.

TAL : Le traitement automatique des langues.

TP: Terme Potentiel.

VEI : Vocabulaire Electrotechnique Internationale.

XTERM : Logiciel d'Extraction TERminologique.

X.M.L. : Xtensible Markup Language.

Bibliographie :

[Anjewierden, 1987] A. Anjewierden : The KADS System. In Proceedings of the 131. European. Workshop on Knowledge Acquisition for Knowledge-Based Systems (EKAW 87). Reading University. September 1987.

[Anonyme, 1829] : (Anonyme, "Civilisation des Indiens Chérokées", Revue des Deux Mondes, 1829, tome1).

[Aussenac, 1989] N. Aussenac : Conception d'une méthodologie et d'un outil d'acquisition des connaissances expertes. Octobre 1989. Thèse de Doctorat en Informatique, Université Paul Sabatier de Toulouse.

[Aussenac, et al., 1989] N. Aussenac, L. Soubie and Frontin : A mediating representation to assist knowledge with MACAO, Acquisition for Knowledge-Based Systems (EKAW 89); pages 516-529, Paris. July 1989. Also in Proceedings of the 4th Workshop on. Knowledge Acquisition for Knowledge-Based Systems (KAW 89), Banff. Canada, October 1989.

[Aussenac et Souhie, 1990] N. Aussenac and L. Souhie : Place d'un outil d'acquisition des connaissances dans la conception des systèmes intelligents. Avril 1990.

[Aussenac-Gilles, Krivine et Sallantin 1992] AUSSENAC-GILLES Nathalie, KRIVINE Jean-Paul et SALLANTIN Jean : « L'acquisition de connaissances pour les systèmes à base de connaissances ». In Revue d'intelligence artificielle, Vol. 6, n°1-2/1992, pp.7-18.

[Becker et Sehnan, 1989] S. Becker and B. Sehnan : An Overview of Knowledge Acquisition Methods for Expert Systems. Technical Report. CSR1-184, University of Toronto. Canada, 1989.

[Bennett, 1984] S. Bennett : Acquiring the Conceptual Structure of a Diagnostic Expert System. In IEEE Proceedings Workshop on. Principles of Knowledge-Based Systems. pages 83-88, December 1984.

[Bennett, 1985] S. Bennett : A Knowledge-based System for acquiring the Conceptual Structure of a Diagnostic Expert System. Journal for Automated Reasoning, 1:49-74, 1985.

[Benveniste, 1966] Benveniste E : Fondements syntaxique de la composition nominale et formes nouvelles de la composition nominale, Problèmes de la linguistique générale, Vol.2, Gallimard, Paris, 145-176

[Biskri et al. 2004] Biskri Ismail, Jean-Guy Meunier, Sylvain Joyal : L'extraction des termes complexes : une approche modulaire semi-automatique, JADT 2004 : 7es journées internationales d'analyse statistique des Données Textuelles.

[Bonnet et al., 1986] A. Bonnet, J. P. Raton. and M. Truong-Ngoc. Systèmes experts : Vers la maîtrise technique. 1986

[Boom, 1989] H. Boom : A Survey of Knowledge Acquisition Techniques. Knowledge Acquisition., 1(1):3-37, 1989.

[Boom et al., 1989] H. Boom, M. Bradslia, M. Kitto, and B. Schenut : Front ETS to AQUINAS: Six years of Knowledge Acquisition Tool Development. ln Proceedings of the 3rd European Workshop on Knowledge Acquisition. for Knowledge-Based Systems (EKAW 89). Pages 502-515. Paris, July 1989.

[Boose, 1985] H. Boose : A Knowledge Acquisition Program for Expert. Systems based on Personal Construct Psychology, 1985.

[Boose, 1985a] H. Boose : Personal Construct. Theory and the Transfer of Human Expertise. In Proc. of the 4th. National Conference of Artificial Intelligence. Pages 27-33. Austin. Texas, USA. August 1984.

[Boose et Bradshaw, 1987] H. Boose and M. Bradshaw : Expertise transfer and complex problems: using AQUINAS as a knowledge-acquisition workbench for knowledge-based systems. International Journal of Man-Machine Studies, 26:3 28. 1987.

[Boose et Bradshaw, 1988] H. Boose, M. Bradshaw and D. B. Schema. Recent progressiv AQUINAS: a Knowledge Acquisition Workbench. 1988.

[Boose et Kitto, 1987] H. Bouse and C. M. Kitto. Choosing Knowledge Acquisition Strategies for Application Tasks. IEEE Journal. 96-103. 1987.

[Borillo 1997] Borillo A : Repérage automatique et identification de la relation lexicale d'hyperonymie, LINX, n°34-35, p 113-121.

[Bouchet et Brunet, 1989] C. Bouchet and E. Brunet : SHELLEY an integrated workbench for KBS development. In Actes des 9éme Journées Internationales sur les Système Experts et leurs Applications, pages 317- 328. Avignon. France. Mai-Juin 1989.

[Bourigault, 1993] Bourigault D : Analyse syntaxique locale pour le repérage de termes complexes dans un texte, TAL 34(2), pp. 105-118, 1993.

[Bourigault, 1994] Bourigault D : « LEXTER, un logiciel d'extraction de terminologie. Application à l'extraction de connaissances à partir de textes ». Thèse de doctorat en mathématique, informatique appliquée aux sciences de l'homme, Paris, Ecole des hautes Etudes en Sciences Sociales.

[Bourigault, 1994b] Bourigault D : LEXTER, un logiciel d'Extraction de TERminologie. Application à l'acquisition des connaissances à partir de textes, Thèse de l'Ecole des Hautes Etudes en Sciences Sociales, Paris.

[Bourigault et Fabre 2000] Bourigault D et Fabre C : Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de grammaire 25 : 131-151 (EACL), P. 131-141.

[Bourigault et Lépine, 1994] Bourigault D et Lépine P : Méthodologie d'utilisation de Lexter pour l'acquisition des connaissances à partir de textes, In Actes de JAVA.

[Bredeweg et al., 1988] B. Bredeweg G. Sehreiber, J. Breuker and B. Wieliuga : Modeling in KBS Development. In Proceedings of the 2nd European. Workshop on Knowledge Acquisition. For Knowledge-Based Systems (EKAW 88), pages 7.1-7.15, Bonn, RFA, 1988.

[Breukcer and Wielinga, 1985] Breukcer and Wielinga. RADS: Structured Knowledge. Acquisition for Expert. Systems. In Proc. of the 5th Internetionel Worshop on. Expert Systems and their Applications. Avignon. France. 1985.

[Breuker et Wielinga, 1989] Breuker and B. Wielinga : Models of Expertise in Knowledge Acquisition. North-Holland. Elsevier Science Publishers B.V. 1989.

[Brunet et Breuker, 1990] E. Brunet and Breuker : La méthode KADS pour le développement des systèmes à base de connaissances. Avignon. Mai-Juin 1990.

[Cerbah, F, 1999] Cerbah F : Acquisition de ressources terminologiques, description technique des composants d'ingénierie linguistique, Rapport Terminologique, aviation Dassault.

[Chiang et Brown, 1987] L. Chiang and C. Brown : DSPL ACQUIR.ER - A System of Acquisition of Routine Design Knowledge. In Sriram and Ailey, editors. Proc. of the 2nd Conference on Applications of Artificial Intelligence. Mech. Comp. Publ.. Cambridge. MA. USA. August 1987.

[Condamines et al. 1993] Condamines A et Amsili P : Terminology between language and knowledge. An example of terminological base. Actes TKE'93, Terminology and Knowledge engineering. Frankfurt, Indeks-Verlag, 316-323.

[Condamines et Rebeyrolle, 1998] Condamines A et Rebeyrolle J : a corpus-based approach to a Terminological Knowledge Base, In Proceedings of the First Workshop on Computational Terminology, D.

[Dagan et church, 1997] Dagan Ido et church Ken : Termight coordination humans and machines in bilingual terminology acquisition, Machine Translation, 12: 89-107.

[David et al., 1990] David., P. Plante. 1990 : De la nécessité d'une approche morpho-syntaxique en analyse de textes, in OICO, Vol. 2(3), Québec, (1990) 140-155.

[Daille, 1994] Daille B : "Extraction automatique de terminologie monolingue", Actes du colloque Informatique et Langues Naturelles, Nantes.

[David & Plante 1990] David S et Plante P : Termino version 1.0, Rapport du Centre d'Analyse de textes par Ordinateur, Université du Québec à Montréal.

[David et Zhai, 1996] David A et Zhai C : Noun phrase Analysis in unrestricted Text for Ionformation retrieval, dans Proceeding of the ACL-96, 34th annual meeting of the association for Computational Linguistics, p. 17-24, Santa Cruz, USA,1996.

[Davidson et al. 1998] Davidson L, Kavanagh J, Mackintosh K, Meyer I et Skuce D: Semi-automatic extraction of knowledge-rich contexts from corpora, In Proceedings of the First Workshop on Computational Terminology, D. Bourigault, C. Jacquemin & M.C. L'Homme Eds, COLING-ACL'98, Montréal, Quebec, Canada, p 50-56.

[Davis et Lento, 1982] R. Davis and D. Lento : Knowledge-based System. Artificial Intelligence. McGraw-Hill. New-York. 1982.

[Daoust, 1992] Daoust F : SATO (Système d'Analyse de Textes par Ordinateur) version 3.6, Manuel de référence, Centre ATO Université du Québec à Montréal, 170p.

[De Chalendar, 2001] De Chalendar, G : STEVLAN: un système de structuration du lexique guidé par la détermination automatique du contexte thématique ».Thèse de docteur de l'université de Paris XI, Orsay, (2001).

[De Chalender G. 2002] Les systèmes d'acquisition de connaissances à partir de textes. http:// www.limsi.fr/individu/gael/manuscritThèse/html/node64.html

[Diederich, 1987] Diederich : Knowledge-based Knowledge Elicitation. In Proc. of the 10th _TICAL pages 201.204. Milan, Italy. August, 1987.

[Diederich et May, 1987] Diederich Ruhlmann. and M. May : CRITON a Knowledge-Acquisition Tool fur Expert Systems. International Journal of Aften-Machine Studies. 26:29 40, 1987.

[Dieng, 1990] Dieng R, Méthodes et outils d'acquisition de connaissances. Rapports de Recherche, Institut National de Recherche en Informatique en Automatique, France, 1990.

[Di Primio et Brewka, 1985] F. Di Primio and G. Brewka : BABYLON Kernel System of an Integrated Environment for Expert System Development. And Operation. In Proc of the 5th International Workshop on Expert. System and their Applications, pages 573-.583. Avignon, France. 1985.

[Drouin, 2002] Drouin P. 2002 : Acquisition automatique de termes : l'utilisation des pivots lexicaux spécialisés. Thèse de doctorat présenté à l'Université de Montréal.

[Drouin, 2003] Drouin P. 2003 : Extraction de termes : techniques courantes, in Ateliers sur les corpus spécialisés en terminologie. http://www.ling.umontreal.ca/l'homme/cgi-bin/pD.zip

[Drouin 2003a] Drouin p : « Term extraction using non technical corpora as a point of leverage ». in Terminology, vol 9, N° 1, pp. 99-117.

[Drouin 2003b] Drouin P : Acquisition des termes simples fondée sur les pivots lexicaux spécialisés, dans Actes des cinquièmes rencontres terminologie et intelligence artificielle (TIA 2003), Strasbourg, pp.183-186.

[Enguehard, 1992] Enguehard C. Acquisition naturelle automatique d'un réseau sémantique, Thèse de doctorat de l'UTC, Compiègne.

[Enguehard 1993] Enguehard C : Acquisition de terminologie à partir de gros corpus. ILN'93, Nantes, Information et langue naturelle, pp. 373-384.

[Enguehard, 1994] Enguehard C : Acquisition of a terminology from colloquial texts, Computational Linguistics for Speech and Handwriting Recognition, *CLSHR*, Leeds, England.

[Epp et Riera, 1989] P. Epp and N. Riera : 3DKAT Un outil d'aide à l'acquisition des connaissances. 1989.

[Eshelman et Mcdermoot, 1986] L. Eshelman and J. McDermott : MOLE A knowledge acquisition tool that uses ifs head. In Proc. of the National Conference on. Artificial Intelligence. pages 950-955, Philadelphia, USA, August 1986.

[Eshelinan, 1988] L. Eshelinan : A knowledge acquisition tool. International Journal of Man-Machine Studies. 29:563-577. 1988.

[Eshelinan, 1988a] L. Eshelman : A Knowledge-Acquisition Tool for Cover-and-Differentiate Systems. In Marcus editor, Automating Knowledge Acquisition for Expert Systems. Pages 37--80. Kluwer Academie Publishers, 1988.

[Eshelinan et Mcdermott, 1988] L. Eshelman and J. McDermott : A knowledge acquisition tool that uses ifs head. Artificial Intelligence. pages 950-955, Philadelphia, USA, August 1986.

[Esterbrook, 1989] S. M. Easterbrook. Distributed Knowledge Acquisition a Model for Requirements Elicitation. In Proc of the 3rd European Workshop on Knowledge Acquisition for Knowledge-based Systems (EXAW 89). Pages 530-- 543. Paris. July 1989.

[Faisandier, 1989] E. Faisandier : Introduction des conditions dans l'outil d'acquisition des connaissances 3DKAT. Novembre 1989. Rapport de stage. ENSIMAG.

[Faure et Nédellec, 1998] Faure D et Nédellec C : " ASIUM : Learning subcategorization frames and restriction of selection." in Proceeding of the Text Mining workshop, 10th European Conference on Machine Learning (ECML 98), Kodratoff Y. (Ed.), Chemnittz, Allemagne, avril 1998.

[Faure et Nédellec, 1998a] Faure D et Nédellec C : "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition" in Proceedings of Adapting lexical and corpus resources to sublanguages and application, workshop of the 1st International Conference on language resources and Evaluation (LREC), p. 1-8 Velaedi p. (Ed.), Grenade, Espagne, Mai 1998.

[Faure et Poibeau, 2000] Faure D, Poibeau T : Extraction d'information utilisant Intex et des connaissances sémantique apprises par Asium première expérimentation. Congrès Francophone AFRIF-AFIA de Reconnaissances des Formes et Intelligence Artificielle (RFIA' 2000), pp. 91-100.

[Frath, 1997] FRATH Pierre : Sémantique, Référence et Acquisition automatique de connaissances à partir de textes. Thèse de doctorat. L'université Des Sciences Humaines De Strasbourg, 1997.

[Frath et al., 2000] Frath P, Ouaslati R et Rousselot F : Genectic programing applied to model identification, p. 291-304, Eyrolles, Paris.

[Grevisse, 1980] : M.GREVISSE, « le bon usage, Grammaire Française ». Edition Duculot, 1980.

[Guilbert, 1965] Guilbert L : La formation du vocabulaire de l'aviation, Larousse, Paris.

[Habert et al. 1995] Habert B, Barbaud P, Dupuis F et Jacquemin C : Simplifier des arbres d'analyse pour dégager les comportements syntaxique-sémantiques des formes d'un corpus, Cahiers de grammaire, 20, 1-32.

[Haiker et Welz, 1989] M. Haiker and U. Welz : Software life-cycle for knowledge-based systems. Journées Internationales sur les Systèmes Experts et leurs Applications. pages 329-342. Avignon. France, Mai-Juin 1989.

[Hayes-Roth, 1983] F. Hayes-Roth, D. B. Lentt and D.A. Waterman : Building Expert Systems. Reading MA: Addison-Wesley Pubiishing Compagny. 1983.

[Hiarnioli, 1985] P. Hiarnioli and D. King Expert Speerte3: Artificial intelligence in Business, 1985.

[Ibekwe-SanJuan, 1998] Ibekwe-SanJuan F. Terminological variation, a means of identifying research topics from texts, in Joint International Conference on Computational Linguistics (COLINGACL' 98), Montréal Québec, 10-14, août 1998, 564-570, 1998.

[Jackiewicz, 1996] Jackiewicz A : L'expression lexicale de la relation d'ingrédience (partie-tout), Faits de Langues, 7, Paris, Ophrys, 53-62.

[Jacquemin, 1994] Jacquemin C : "Quelques mécanismes spécifiques d'une grammaire d'unification adaptée à l'extraction terminologique", Actes 9ème Congrès Reconnaissance des Formes et Intelligence Artificielle, Paris, p. 385-396.

[Jacquemin, 1996] Jacquemin C : A symbolic and surgical acquisition of terms through variation, dans Springer, S. Wermeter, E. Rotoff, G. Scheedler editors, connectionist, Statistic Approaches to learning for Natural Language Processing, p. 425-438.

[Jacquemin, 1997]. Jacquemin C : Variation terminologique et acquisition automatique de termes et leurs variantes en corpus. Habilitation à diriger des recherches en informatique, IRIN, Université de Nantes, 1997.

[Jacquemin, 2001] Jacquemin C : Spotting and disovering Terms through NLP, MIT Press, Cambrige Massachusetts.

[Jacquemin et al., 1994] Jacquemin C et Royauté J : Retrieving terms and their variants in a lexicalized unification-based framework, Actes ACM-SIGIR 94, Dublin, juillet, 132-141.

[Johnson, 1988] N. E. Johnson : Knowledge Acquisition for Knowledge-Based Systems (EKAW 88), pages 23.1-23.10, 13011.11, Julie 1988.

[Jouis, 1995] Jouis C : SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe, Actes des Journées d'Acquisition de Connaissances du PRC-GDR-IA du CNRS, Grenoble, 5-7 avril.

[Justeson et Katz, 1995] Justeson John et Kats Slava : « Technical Terminology : Some Linguistic Properties and an Algorithm for Identification in Text ». In Natural Language Engineering 1(1), pp. 9-27, Cambridge University Press.

[Kahn, 1988] G. Kahn : Front Observing Knowledge Engineers to Autotnating Knowledge Acquisition. In Marcus. editor. Automating Knowledge Acquisition for Expert Systems. Pages 7-35, Kluwer Academie Publimhers, 1988.

[Kahn et al., 1985] G. Kahn, S. Nowlau, and .1. McDermott. MORE: An Intelligent Knowledge Acquisition Tool. In Proc. of the 9tle LICA, Los Angeles, 1985.

[Kawaguchi et al., 1986] A. Kawaguchi et R Mizoguchi. T. Yamaguchi. and O. Kakusho : An Intelligent Interview System for Conceptual Design of Database. In Proc. of ECAI-86, 1986.

[Kawaguchi et al., 1987] A. Kawaguchi. R. Mizoguchi. T. Yamaguchi. and O. Kakusho : A Shell for Interview Systems. Page 359-361. Vol. 1, Milan, Italy, 23-28 August 1987.

[Kelly, 1955] Kelly A : The Psychology of Personal Constructs. Norton. 1955.

[Kilgariffet et Tugwel, 2001] Kilgariffet Adam, Tugwel David : Wasp-bench : an MT lexicographer's workstation supporting state-of-the-art lexical disambiguation, in proc. MT Summit VIII, Santiago de Compostela, Spain, September, pp. 187-190.

[Klinker, 1988] Klinker : Sample-Driven Knowledge Acquisition for Reporting Systems. In Marcus editor. Automating Knowledge Acquisition for Expert Systems. Pages 125-174. Klnwer Academie Publishers, 1988.

[Klinker, 1989] Klinker : A Framework for Knowledge Acquisition. In Proceedings of the 3rd European. Workshop on Knowledge Acquisition for Knowledge-Based Systems (EKAW 89), pages 102-116, Paris, July 1989.

[Lebart et Salem, 1994] LEBART L. et SALEM A : Statistique textuelle. Dunod.

[Lemay, 2003] Lemay C : Utilisation de corpus de référence pour dégager la terminologie d'un corpus. In Ateliers sur les corpus spécialisés en terminologie. Présentation téléchargeable. http://www.ling.umontreal.ca/l'homme/cgi-bin/CL2.zip

[Lethuiller, 1989] Lethuiller J : "La synonymie en langue de spécialité", Meta, XXXIV, 3, p. 443-449.

[L'Homme M.C., 2000] Evaluation de logiciels d'extraction de terminologie : examen de quelques critères. Communication donnée à la Réunion Inter institutions sur la terminologie et la traduction assistées par ordinateur (JIAMCATT), Office des Nations unies, Vienne (Autriche)

[Linster, 1988] M. Linster : KRITON it Knowledge Elicitation Tool for Expert Systems. In Proceedings of European Workshop on Knowledge Acquisition for knowledge-Based Systems (EKAW 88). pages 4.14.9. Bonn. RFA. Julie 1988.

[Linster, 1988a] M. Linster : Towards a second generation knowledge. Acquisition tool. Knowledge Acquisition. 1(2):163183. 1989.

[Marcus, 1987] S. Marcus. Taking backtracking with acquisition grain of SALT. International Journal of Man-machine Studies, 1987.

[Marcus, 1988] S. Marcus : A knowledge representation scheme for acquiring design knowledge. In Tong and Sriram. Artificial Intelligence Approach's to Engineering Design. Addison-Wesley, Reading, MA. 1988.

[Marcus, 1988a] S. Marcus : SALT a knowledge acquisition language for propose and revise systems. In Sandra Marcus editor. Automating Knowledge Acquisition for Expert Systems. Pages 81 123, 1988.

[Marcus, 1988b] S. Marcus and .1. McDermott. SALT: a knowledge acquisition language for propose-and-revise systems. Artificial Intelligence., 1988**.**

[Meilland, 2003] Meilland, J : Extraction de terminologie à partir de libellés textuels courts. In: P.U.R., (Presses Universitaires de Rennes), Linguistique de corpus (2003).

[Meng, 1989] B. Meng : Utilisation de connaissances graduelles en intelligence artificielle. 9éme Journée Internationale sur les Systèmes Expert et leurs Applications.. Avignon. France, 1989.

[Morin, 1999] Morin E : Extraction de liens sémantiques entres termes à partir de corpus de textes techniques, Thèse de doctorat, IRIN, Nantes.

[Museu, 1987] A. Museu : Use of a dornain model to drive an interactive kuowledge-editing tools. International Journal of Man-machins Studies, 26:105-121, 1987.

[Nerima et al. 2003] Nerima L, Sieretan V, Wehrli E : Creating a multilingual collocations dictionary from large text copora, dans Proceeding of Conference of the European Chapter of the Association for Computational Linguistics.

[Nerima et al. 2006] Nerima L, Sieretan V, Wehrli E : Le problème des collocations en TAL, Nouveux cahiers de linguistique francaise 27(2006), 95-115.

[Newell, 1981] Newell, Simon (1972) A. Newell et H. Simon. *Human Problem Solving*, Englewood Cliffs, N.J., Prentice-Hall, 1972.

[Nicolas, 2000] Nicolas T, Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles. Thèse de doctorat d'état, université LOUIS-PASTER, STRASBOURG, 2000.

[Norbo et al., 1989] L. Nordbo, M. Vestli. and I. Solvberg : METAMETH. Methodology for Knowledge Engineering. In Proceedings of the 3rd European Workshop on Knowledge Acquisition. for Knowledge-Based Systems (EKAW 89). Pages 226-238, Paris, July 1989.

[Orliac, 2003] Orliac B : Acquisition de collocation verbe + nom à partir de représentations syntaxiques, dans Ateliers sur les corpus spécialisés en terminologie. Version téléchargeable : http://www.ling/umontreal.ca/lhome/cgi-bin/BO.zip

[Otman, 1996] Otman G : Le traitement automatique de la relation partie-tout en terminologie, Faits de langue, 7, Paris, Ophrys, p 43-52.

[Ouaslati, 1999] Ouaslati R : Aide à l'acquisition de connaissances à partir de corpus. Thèse de doctorat, Université Luis Pasteur Strasbourg.

[Petit Larousse, 1990] : Petit LAROUSSE illustré, Librairie LAROUSSE, PARIS 90.

[Péry-Woodley et Rebeyrolle, 1998] Péry-Woodley M.P et Rebeyrolle : Domain and genre in sublanguage text: definitional microtexts in three corpora. Proceedings of First International Conference on Language Resources and Evaluation, Granada 28th-30th May, Spain, pp. 987-992.

[Poibeau T. et al. 2002] Poibeau T , Dutoit D, Bizourad S : Evaluation resource acquisition tools for information extraction. Proceeding of the International language resource and Evaluation Conference (LRERC 2002), Las Palmas, Les canaries.

[Roche et al 2004] Roche M, Heitz Thomas, Matte-Tailliez Oriane et Kodratof Yves : EXIT un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés, Exposé dans le cadre de la conférence JADT'04 (7th Intenational Conference on the Statistical Analysis of Textual Data), Louvain-La-Neuve, Belgique.

[Rochibeau, 2003] Rochibeau B : Extraction de collocation fondée sur les méthodes statistique, in Ateliers sur les corpus spécialisés en terminologie, http://www.ling-umontreal.ca/lhomme/cgi-bin/BR.zip

[Savary, 2000] Savary A : Recensement et description des mots composés, méthodes et application. Thèse de doctorat en informatique fondamentale, université de paris7.

[Selberztein, 1993] Selberztein M : Dictionnaires électronique et analyse automatique de textes : le système INTEX, Edition Masson, Paris.

[Séguéla et Aussenac, 1997] Séguéla P et Aussenac N : Un modèle de base de connaissance terminologique, In Actes de TIA'97, p 47-68.

[Sigart, 1989] Sigart Newsletter : Knowledge Acquisition. October 1989.

[Sigart, 1989a] Sigart Newsletter : Knowledge Acquisition Specified, April 1989.

[Smadja, 1993] SMADJA Franck (1993) : « XTRACT : An Overview ». In Computers and the Humanities 26, pp. 399-413, 1993. Kluwer Academic Publishing.

[Smadja, 1996] Smadja, F, McKeown, K, Hatzivassiloglou, V : Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics 22(1), 1–38 (1996).

[Szulman, 1999] Szulman S : Alinguitic-Based Tools for the building of Domain Ontology. EKAW 1999: 49-66.

[Term, 2003] : Constitution de corpus à partir du web pour l'acquisition automatique de terminologie : une expérience, http://www.poleia.lip6.fr/~slodzian/sberland/chapitre1.html.

[van der Vet, de Jong, Mars, Speel et Stal, 1994] Vander Vet Paul, De Jong Hilde, Mars Nicolaas, Speel Piet-Hein, Ter Stal Wilco (1994) : « PLINIUS Intermediate Report ». Memoranda Informatica 94-35, University of Twente, Netherlands.

[Wielinga et Breuker, 1985] D.Wielinga and J. A. Breuker : Interpretation of verbal data for knowledge acquisition.

[Wielinga et Breuker, 1986] L. Wielinga and J. A. Breuker : Models of expertise. Proceedings of the 7th Conference on Artificial intelligence. Pages 30G-318. Elsevier Science Publishers, North Holland, Brighton, July 1986.

[Yamouni, 2010] Yamouni Farida : Construction d'un dictionnaire électronique de terminologie informatique et analyse automatique de textes par grammaires locales. Thèse de doctorat d'état, Université Mouloud Mammeri de Tizi Ouzou, 2010.

.