Chapitre 1
Traitement automatique du langage naturel
(TALN)

1. Introduction

Un bouleversement considérable s'est apparu dans les années 90 : ordinateurs personnels standardisés, avec des capacités de stockage et de traitement en progression exponentielle, ainsi que l'apparition du Web qui a marqué l'apogée technologique en informatique. Dans tout ce changement est née « l'ingénierie linguistique ». La linguistique appelée aussi sciences du langage, est l'étude scientifique des langues naturelles de l'espèce humaine.

Les textes constituent la masse d'information la plus présente sur le Web (le son et les images sont plus récents). Ainsi toute contribution au classement, au traitement des documents textuels et l'extraction de l'information devient une préoccupation principale. C'est dans cette perspective que l'ingénierie linguistique se met ainsi au service de la "fouille de textes" où on remarque la domination des méthodes statistiques sur les méthodes symboliques.

Pour distinguer la langue humaine, on parle actuellement des "langues naturelles", contrairement aux "langues artificielles" ou "formelles" que sont les langages de programmation informatique ou la logique mathématique.

« On regroupe sous le vocable de traitement automatique du langage naturel (TALN) l'ensemble des recherches et développements visant à modéliser et à reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication » Véronis (2001) ; Tellier (2010) ; Yvon (2010)

Le traitement automatique du langage, récemment à la croisée de la linguistique, de l'informatique et de l'intelligence artificielle, voit ses applications, ses programmes et beaucoup de techniques informatiques, au service du langage humain en vue d'appréhender le sens des données en langage naturel. Une compréhension de haut niveau pour ce raisonnement humain a était longtemps recherchée et considérée comme le but extrême des premiers travaux.

Ce chapitre présente ce que peut être un traitement automatique du langage naturel TALN, son architecture, ses niveaux d'analyse du langage traité et ses différents formalismes de représentation de connaissances et du sens sont exposés. Un aperçu d'horizon sur les différents systèmes ou outils TALN, développé pour la langue arabe sera traité à la fin de ce chapitre.

2. Les différents niveaux d'analyse en TALN

2.1. L'analyse d'un système TALN

A ce niveau, deux études formelles ont été menées. L'une peu ancienne, au niveau de la morphologie et de la syntaxe, et l'autre beaucoup plus récente au niveau de la sémantique et de la pragmatique linguistique. A noter qu'on confond souvent la *sémantique lexicale*, qui explique le sens d'unités individuelles, et la *sémantique propositionnelle* qui étudie le sens d'énoncés dans son ensemble et à qui on peut lui donner une valeur de vérité.

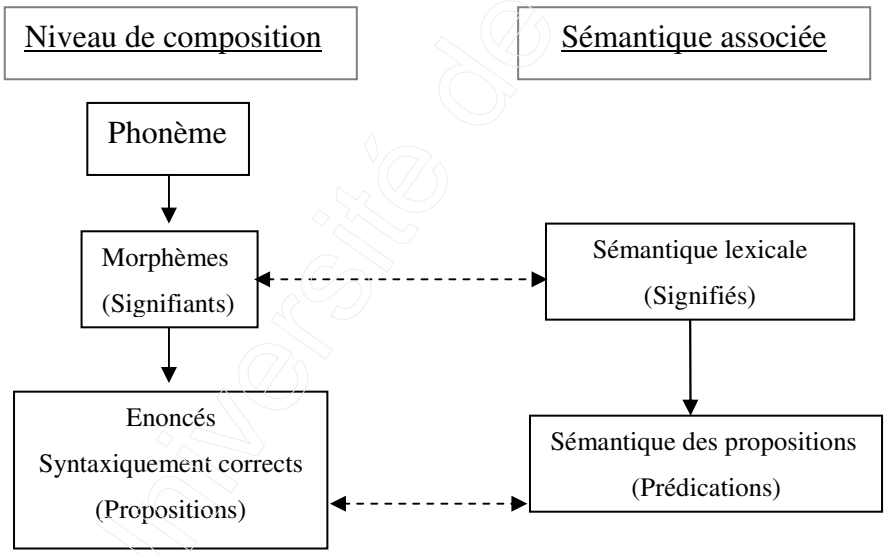

Fig. 1 : Hiérarchie des niveaux d'analyse des langues naturelles

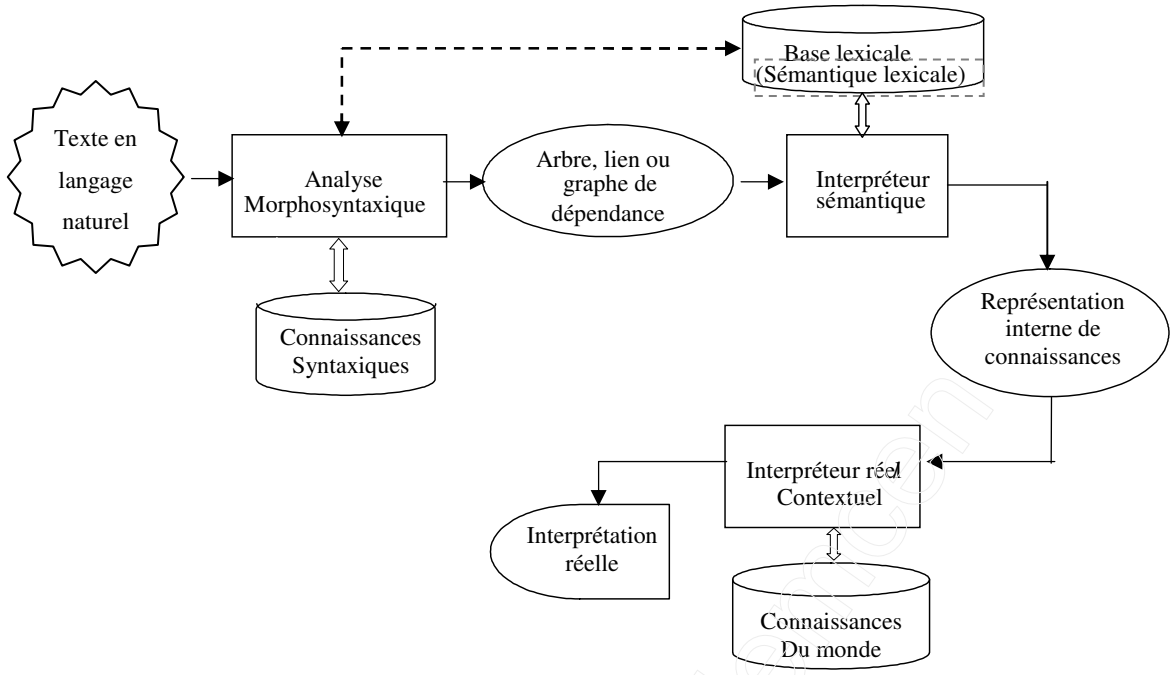

Fig. 2 : Architecture générale du TALN.

2.1.1. Analyse morphologique

La morphologie : interprète comment les mots sont structurés et quels sont leurs rôles dans la phrase. Cette analyse consiste à une segmentation du texte en unités élémentaires auxquelles sont attachées des connaissances dans le système : une fois cette segmentation effectuée, ce n'est plus le texte qui est manipulé, mais une liste ordonnée d'unités. Pour le traitement d'un texte numérique : on part d'une chaîne de caractères typographiques, et on essaie de la segmenter de manière à ce que chaque partie corresponde à une unité classée dans le système.

Exemple : soit la chaîne de caractères « **يأكل عـــمر التفاحة .** »

La segmentation se fera de la manière suivante :

U1 = يأكل

U2 = عمر

U3 = التفاحة

Maintenant, on pourra associer toutes sortes d'informations aux U_i (i = 1, 2, 3, ...), comme par exemple : U2 = عمر

Informations morpho-syntaxiques : nom propre, masculin, singulier.

Informations sémantiques : animé, humain, prénom ...

U1 = **يأكل**

Forme lemmatisée : **أكل**

Informations morpho-syntaxiques : verbe (فعل) , passé (ماضي), indicatif , 3ème personne, singulier, constructions : transitif, ...

Idem pour U3…

Remarque : il y a des phénomènes (concernant le choix et le statut des unités) qui sont répertoriés de longue date par les linguistes : qui conduisent à s'interroger sur la notion de mot : élisions[1], amalgames, flexions, dérivations, compositions, ...

2.1.2. Analyse syntaxique

C'est une partie de la grammaire qui traite la manière dont les mots peuvent se combiner pour former des propositions et de l'enchaînement des propositions entre elles. Cela consiste à associer, à la chaîne découpée en unités, une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités (voir Fig.3).

Reprenons l'exemple précédant : « يأكل عمر التفاحة . », et sa représentation morphologique:

U1 = يأكل U2 = عمر U3 = التفاحة

Le résultat de l'analyse syntaxique pourra être par exemple l'arbre suivant :

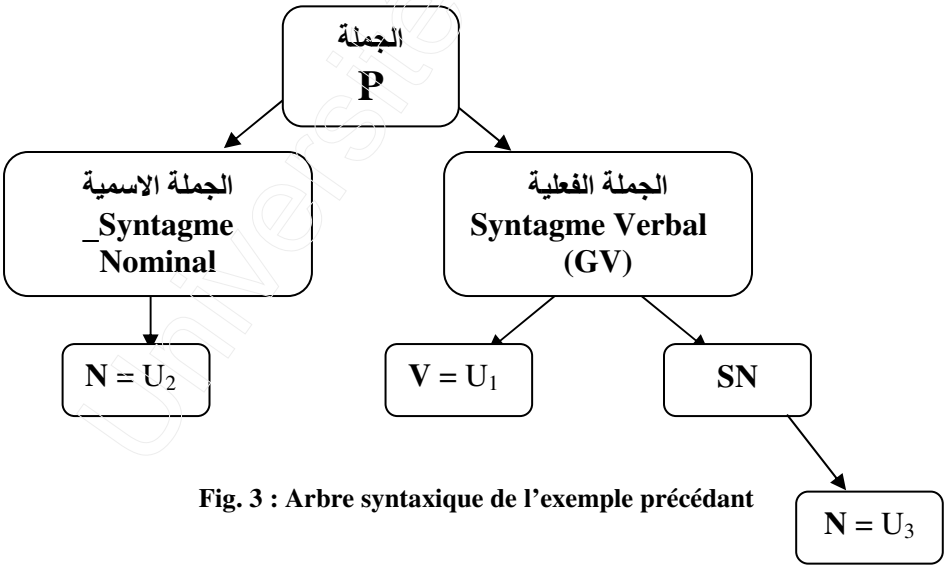

Fig. 3 : Arbre syntaxique de l'exemple précédant

1 - **élision** : nom féminin singulier (grammaire) suppression de la voyelle finale devant un mot commençant par une voyelle ou un 'h' muet, **...en arabe :**

- ترخيم : حذف الحرف الأخير أو أكثر بعد أداة النداء مثال : فاطم في فاطمة.

- حذف : إسقاط بعض أجزاء الكلمة أو الجملة أو التفعيلة لعلة.

- إدغام : إدخال الحرف في الآخر مثال : ("وَمَن يَّعْمَلْ مثقال ذرة")

P = « يأكل عمر التفاحة »

SN = عمر

SV = يأكل التفاحة

SN = التفاحة

N = عمر

V = يأكل

N = التفاحة

2.1.3 Analyse sémantique

Le niveau sémantique est encore beaucoup plus complexe à décrire et à formaliser que les niveaux précédemment énoncés. De ce fait, peu d'outils de traitement reste opérationnel ou du moins, concernent des applications très réduites où l'analyse sémantique se limite à un domaine parfaitement étroit ; par contre, il reste beaucoup à apprendre sur la manière de construire en grandeur réelle des analyseurs sémantiques généraux qui couvriraient la totalité de la langue arabe et seraient indépendants d'un domaine d'application particulier.

La phrase est l'unité d'analyse principale que prend en charge le traitement sémantique afin de représenter sa partie significative. Ces phrases, dont l'analyseur sémantique doit décrire le sens, se composent d'un certain nombre de mots identifiés par l'analyse morphologique, et regroupés en structures par l'analyse syntaxique. Ces mots et ces structures constituent autant d'indices pour le calcul du sens : *on pourrait dire, que le sens résulte de la double-donnée du sens des mots et du sens des relations entre ces mots.*

2.1.4 Analyse contextuelle

La phrase traitée hors contexte, c'est-à-dire isolé de son texte, n'a peut être pas le même sens que dans son contexte. L'analyse sémantique de la phrase isolée, nous amène à représenter la partie de la signification des mots dans cette phrase, elle n'épuise donc pas ce que l'on peut appeler la signification complète d'un texte, à savoir les relations existantes entre les phrases du texte telles que l'humain l'appréhende lors d'un processus de compréhension. C'est ainsi qu'intervient l'analyse contextuelle qui consiste à trouver la signification "réelle" des phrases liées aux conditions positionnelles et contextuelles d'utilisation des mots.

2.2. Le sens

Le sens est partout dans le traitement automatique des langues : il faudrait parler des aspects :

→ Lexicaux (quels liens existent entre les mots et leurs sens ?),

→ Syntaxiques (quel sens est porté par les structures dans lesquelles ces mots interviennent?),

→ Sémantiques bien sûr (comment sont représentées, obtenues et traitées des significations ?)

→ Contextuelles (quelles sont les influences des connaissances sur le monde et la situation pour déterminer le sens ?)…

2.3. Le problème du sens

Qu'est ce que le mot « sens » ? Tout le monde répondra à première vue que c'est « approfondir un peu », c'est-à-dire aller plus loin que "le sens d'un terme, que veut-il évoquer ?". Plusieurs interprétations du sens du mot "sens" peuvent exister. Toutes ces définitions dévoilent le flou qui couvre ce domaine, mais permettent aussi de souligner une différence entre le sens fondamental et le sens interprété, lié également à la prise en considération ou non du contexte [JPM-00]. En effet, une grande partie des travaux en intelligence artificielle et surtout en traitement automatique des langues suppose (implicitement ou non) la possibilité de calculer un sens littéral (qui relève de ce qui est alors appelé sémantique), puis de l'interpréter selon les connaissances générales sur le monde de référence, le contexte et les caractéristiques des interlocuteurs (on parle alors de contextuel).

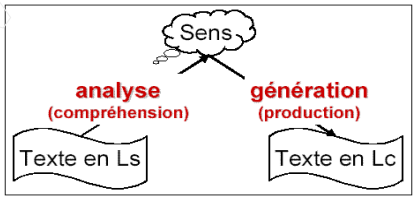

Fig. 4 : Calcul du sens

Bien entendu, cela pose la question de l'existence d'un niveau linguistique indépendant, que certains remettent en cause en arguant de l'impossibilité de séparer l'interpréteur de la chose interprétée. D'autres contestent l'existence des acceptations énumérées dans les dictionnaires pour défendre le sens littéral… Cette hypothèse est si

commode pour les traitements automatiques qu'elle est à peu près systématique même si sa validité psychologique reste incertaine. Mais, même ici, on trouvera un certain flou dans les catégories possibles ; ainsi peut-on distinguer (sans qu'il s'agisse le moins du monde d'une partition), voir figure 5 :

Fig. 5 : Différentes catégories de la sémantique

— La sémantique veri-conditionnelle précise les conditions de vérité de l'expression traitée (on parle aussi parfois de référence virtuelle…).
— La sémantique intentionnelle voit une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants.
— La sémantique extensionnelle décrit une expression comme l'ensemble des objets ou situations du monde que cette expression désigne (on parle aussi de sémantique dénotationnelle ou référentielle).
— La sémantique componentielle cherche à décomposer les mots en éléments de sens plus primitifs, puis étudie leurs possibilités de combinaison.
— La sémantique procédurale décrit le sens d'une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné.
— La sémantique argumentative (liée aux notions de supposition et de présupposition) dépasse la description d'actes de langage isolés pour étudier leurs enchaînements dans le discours et les connecteurs correspondants.

Néanmoins, une question retient l'attention des chercheurs : « est-il possible de décomposer la notion de sens ? ». Plusieurs points de vue opposés ont résulté du débat de cette question. Par exemple, W. Chafe défend l'idée que le sens est unitaire et ne peut se

décomposer. D'autres soutiennent l'idée que le sens global peut se décomposer en divers éléments, étudiés séparément. Leech, quand à lui, distingue sept formes fondamentales liées au sens (en considérant ou non le contexte et en faisant la distinction entre le sens intentionnel et le sens interprété) : conceptuel, connotatif, stylistique, affectif, réfléchi, collocatif, thématique dont il a apporté des éléments partiels.

3. Compréhension et formalismes de représentations diverses

La compréhension littérale d'un texte nécessite divers types de connaissances (modèle de la langue, modèle de la tâche, éventuellement état de la tâche, historique du dialogue et modèle utilisateur).

L'utilité de construire un module de compréhension nous donne l'avantage d'en extraire ce que nous appelons le « *sens utile* » d'un texte (informations nécessaires pour l'application). Si on situe la compréhension par rapport à un système de commandes, le *sens utile* permet de construire sa commande.

La représentation sémantique peut être vue comme la fonction de transformation d'une représentation primaire vers une autre représentation interprétable par le contrôleur de dialogue d'un système interactif.

Dans la littérature informatique, une multitude de formalismes de représentations sémantiques est proposée pour la représentation interne d'une phrase, afin d'en révéler le sens. Nous pouvons entre autre citer :

Les logiques (la logique des propositions, la logique des prédicats ou la logique modale), par exemple, le démonstrateur AGS (Audiotel Guide des Services) du CNET utilise la logique du premier ordre pour représenter le sens d'un énoncé.

Les graphes conceptuels (Sowa), appelés aussi réseaux sémantiques ou graphes de Sowa, ont été développés par Sowa. Les logiques et les graphes de Sowa sont surtout utilisés dans le domaine de l'ingénierie des connaissances linguistiques.

Les structures de traits et les ensembles d'attributs sont très courants dans les interfaces homme-machine. Le choix d'un formalisme dépend de ses propriétés et caractéristiques selon l'objectif recherché. Notons que pour des serveurs dialoguant avec des bases de données, la représentation sémantique doit permettre de générer une requête de type SQL (Structured Query Language) pour interroger la base de données. De ce fait, les types de représentations sémantiques les mieux adaptés sont les structures de traits et les ensembles d'attributs. [BOU-02]

3.1. La compréhension d'un texte

Informatiquement parlant, comprendre un texte ou un énoncé, implique sa transformation en une structure de données exploitable par la machine. C'est cette structure que nous appelons le sens du texte. Mais pour pouvoir faire cette transformation, le module de compréhension (voir figure 4) utilise de nombreuses connaissances linguistiques (lexique[1], grammaire, etc.)

A la suite de cette partie, on propose plusieurs formalismes de représentations du sens d'un texte ainsi que les connaissances utiles au processus de compréhension. Puis nous étudions quelques stratégies de compréhension typiques des systèmes de dialogue.

3.2. Le sens et sa représentation

Nous allons nous intéresser, dans cette section, aux principaux formalismes permettant la représentation interne d'une phrase, afin d'en dégager le sens. La représentation du sens d'un texte ou d'un énoncé est donc la structure obtenue en sortie du module de compréhension. Une description de la logique, des graphes de SOWA, des structures de traits et des attributs seront explicités dans cette section. Les deux premiers sont surtout utilisés dans le domaine de l'ingénierie des connaissances linguistiques, les deux autres sont très courants dans les interfaces homme-machine.

Bon nombre de ces formalismes sont presque identique à la logique des prédicats du premier ordre. Le choix d'un formalisme est donc avant tout accrédité à l'expert pour exprimer ces connaissances et aux algorithmes d'interprétation utilisés. On peut aussi associer au sein d'un même système plusieurs formalismes.

3.3. Les logiques

Plusieurs approches logiques ont vu le jour comme la logique des propositions, la logique des prédicats ou la logique modale. Notons qu'aucune logique n'a réussi à représenter une phrase de façon complète. Mais elles peuvent nous satisfaire comme dans le cas particulier des serveurs vocaux interactifs : comme le démonstrateur AGS (Audiotel Guide des Services) du CNET qui utilise d'ailleurs la logique du premier ordre pour représenter le sens d'un texte ou d'un énoncé.

1 - Lexique : Ensemble de mots constituant une langue.

3.4. Les graphes conceptuels

Les graphes conceptuels ou réseaux sémantiques, développés par Sowa, est une représentation graphique composé d'arcs orientés et de deux types de nœuds:

- Les nœuds représentant les entités (concepts) notés par des rectangles.
- Les nœuds représentant les relations notées par des ovales.

Les arcs relient deux nœuds de nature différente. Les entités sont définies par un type et un marqueur. Le marqueur peut désigner un objet en particulier (noté par le signe #, suivi d'un numéro référençant l'objet en question) ou au contraire un générique (noté par le signe *).

Un des intérêts de ce formalisme est que l'on peut très facilement ajouter des connaissances à un graphe : c'est le procédé de jointure de plusieurs graphes.

Exemple : [BOU-02], Représentons les phrases :

—" الكراس الأحمر يوجد فوق الطاولة "

—" الدفـتر يوجد فوق الطاولة "

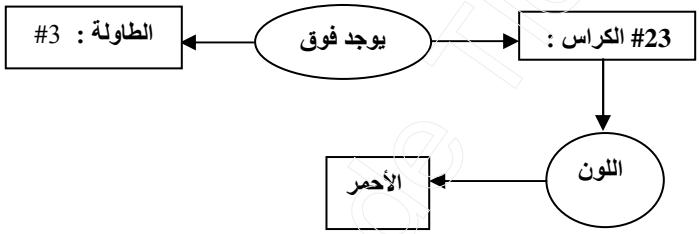

Fig. 6 : Graphe conceptuel de" يوجد الكراس الأحمر فوق الطاولة "

Fig. 7 : Graphe conceptuel de" يوجد الدفـتر فوق الطاولة "

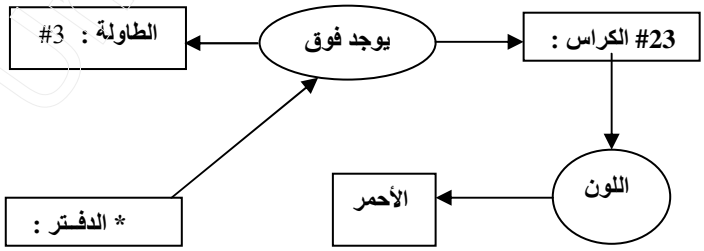

Fig. 8 : Exemple de graphe conceptuel résultant de la jointure de deux informations.

La logique classique peut représenter les graphes conceptuels. Ces dernières permettent de réaliser des inférences, déductions et autres opérations permises par la logique.

3.5. Structures de traits (SDT)

Une structure de traits (SDT) est un ensemble de couples (traits) [attribut = valeur] dont la valeur peut être un entier, un réel, une chaîne ou une autre SDT. C'est donc une structure récursive. Une SDT peut aussi être représentée par un arbre.

Exemple : [BOU-02]

Représentons la phrase :

– أريد قطارا ينطلق من الجزائر يوم الاثنين على الساعة 06 سا ويصل إلى تلمسان على الساعة 13 سا .

Fig. 9 : exemple d'une SDT

Cette SDT est équivalente avec l'arbre suivant :

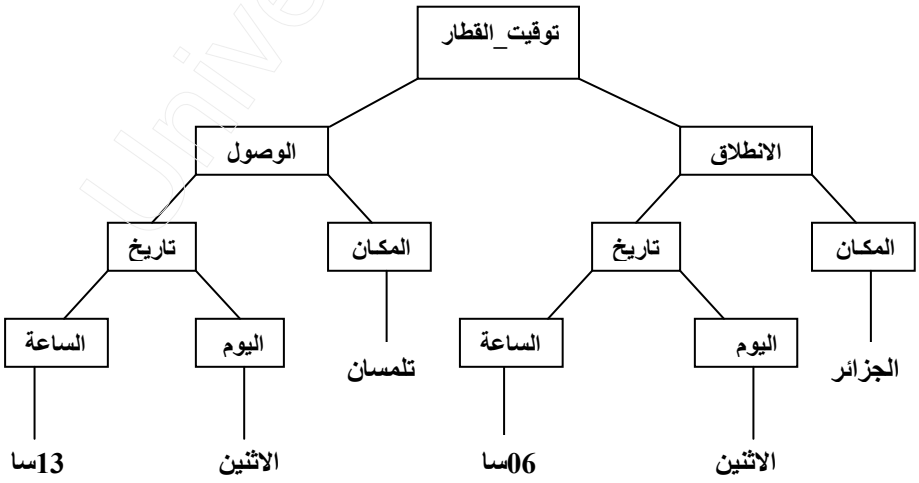

Fig. 10: Représentation d'une structure de traits par un arbre.

Ce formalisme permet de faire des opérations aussi bien que la disjonction ou l'unification (i.e. union de deux SDT). L'analyseur ALPES employé dans le système de dialogue de l'équipe GEOD (CLIPSIMAG) utilise des STD pour représenter le sens des énoncés [BOU-02].

4. Représentation des connaissances linguistiques

En clair, un système de compréhension doit disposer d'un modèle de la langue et un modèle de la tâche (c'est-à-dire un vocabulaire propre à la tâche, comme la syntaxe, la grammaire…), toutes ces connaissances sont enveloppées dans un formalisme afin que ce système puisse les identifier et comprendre un texte. Soulignons deux manières d'aborder l'analyse d'un texte :

- **soit on essaie de le comprendre en se servant des règles de syntaxe et de grammaire.**
- **soit on ne tient pas compte de sa représentation syntaxique mais uniquement des éléments porteurs de sens (appelés aussi concepts).**

4.1. Les lexiques

Un module de compréhension a besoin de connaître les mots (lexies) pour pouvoir analyser une phrase. Ainsi nous pourrons dire que les lexiques sont comme des dictionnaires permettant de décrire un vocabulaire.

Il existe plusieurs types de lexiques : certains ne contiennent que le vocabulaire, d'autres indiquent le genre, nombre et autres particularités du mot. Les mots du lexique peuvent être représentés sous la forme d'une structure de traits.

Exemple : représentons le mot '**السفن**' [BOU-02].

$$\begin{pmatrix} \text{Forme orthographique} = \textbf{السفن} \\ \text{Catégorie lexicale} = \textbf{اسم} \\ \text{Accord} = \begin{pmatrix} \text{genre} = \textbf{مؤنث} \\ \text{nombre} = \textbf{جمع} \end{pmatrix} \end{pmatrix}$$

Fig. 11 : Exemple de représentation d'un mot dans le lexique

Ce type de lexique est utilisé si l'on désire tenir compte des accords (entre l'article et le nom par exemple).

Maintenant, considérons un texte comme une suite de concept, au niveau de l'analyse le lexique peut nous renseigner dans quel(s) concept(s) se trouve le mot, afin d'éviter de donner toutes les caractéristiques du mot (citées dans Fig.11 ci-dessus).

Exemple : supposons que le mot '**السفن**' fasse partie du concept " **التـنقل** " transport [BOU-02].

السفن : وسيلة للتـنقل على البحر

L'entrée du lexique correspondant aux deux mots sera alors de la forme :

السفن : البحر

السفن : تـنقل

Le type et le contenu du lexique dépendent donc énormément de la stratégie de compréhension utilisée et du formalisme choisi pour représenter le sens d'un énoncé.

4.2. Les grammaires formelles

Les grammaires formelles permettent de décrire la syntaxe du langage.

On peut représenter une grammaire par un quadruplet (V_a, V_t, R, S) où

- V_a est le vocabulaire auxiliaire,
- V_t le vocabulaire terminal,
- R l'ensemble des règles
- et S l'axiome.

Il existe plusieurs types de grammaire rangés selon la classification de ***Chomsky***. On peut citer entre autres les grammaires régulières et les grammaires hors contexte. Exemple d'une grammaire régulière (représentable par un automate fini) :

Soient :

V_a = {P =>, SN (syntagme nominal), SV (syntagme verbale), Dét (déterminant), N (nom)},

V_t = {**ال, سفينة, أبحرت**} et S l'axiome.

Les règles sont les suivantes :

S :	P	=>	SN + SV
		SN	→ Dét + N
		SV	→ V
		Dét	→ **ال**
		N	→ **سفينة**
		V	→ **أبحرت**

Cette grammaire permet de former la phrase " **السفينة أبحرت** ".

4.3. Les mots clefs d'un texte

En lisant une phrase, Il suffit donc de comprendre quelques mots « clefs » pour pouvoir en extraire le sens. En général on ne tient pas compte de leur emplacement dans la phrase. L'analyse par mots « clefs » ne prend pas compte ni de la syntaxe de la phrase ni des mots qui ne sont pas considérés comme "clefs". Le principe des mots clefs ne peut être utilisé qu'avec des textes relativement simples. Ce sont les mots nécessaires et suffisants à la compréhension d'un texte.

5. Connaissances du monde (CM) et connaissances linguistiques (CL)

5.1 Méthodologie d'identification des connaissances encodées dans le lexique

Faisons tout d'abord le point sur la problématique d'encodage des CM afin de présenter une méthodologie permettant d'établir des CM encodées dans la langue.

5.1.1 Encodage des Connaissances du Monde (CM)

Notons que la grande partie de nos CM n'est pas encodée dans la langue. Prenons un exemple pour mieux illustrer cette notion. Soit la phrase : « **الثلج أبيض** », nous savons précisément que la neige est blanche. Nous savons aussi par exemple qu'une voiture est plus lourde – en général – qu'une bicyclette, de même si on parle d'une cuisine, on trouve une table et des ustensiles, Ces connaissances ne sont pas encodées dans la langue, n'ont pas leurs équivalents linguistiques. Bien entendu, nous pouvons les exprimer à l'aide de la langue: en formulant des énoncés ou en écrivant des textes. On peut dire que la grande partie de nos CM n'est pas encodée dans la langue. Si par contre, on arrive à trouver cette information écrite dans la langue arabe (dans un dictionnaire, dans une encyclopédie,…) alors on peut dire que c'est une connaissance linguistique (CL). Ainsi dire, une connaissance C est encodée dans une langue L signifie pour nous qu'il existe une parallèle entre X et une connaissance (règle) qui fait partie de L, en tant que système. [Ban-03].

Il est important de savoir que les connaissances lexicales (qui sont un sous ensemble des connaissances linguistiques) sont généralement des connaissances du monde, encodées dans un lexique. L'encodage linguistique de certaines connaissances est conceptuellement vraisemblable : par exemple, tout le monde sait ce que c'est qu'un restaurant et s'attendra, en fait, à ce que la définition de la lexie « **المطعم** » contienne le sens « **أَكْلٌ** », sans pour autant croire qu'on y trouvera formellement la lexie « **أكل** » : cette définition peut aussi contenir une autre lexie (par exemple « **غذاء** ») qui inclut ce même sens.

Il existe aussi des connaissances dont l'encodage linguistique est conceptuellement imprévisible : par exemple, pour parler d'un remerciement dont le degré est élevé, on peut dire des phrases comme [Ban-03] :

— **شكرا جزيلا.**

— **شكرته بحرارة.**

Mais les phrases :

— **شكرته جزيلا.**

— **شكرا بحرارة.**

seraient étranges (ou du moins incorrecte) en arabe.

Deux questions importantes que R. BANGHA a souligné, « pourquoi telle ou telle partie de nos connaissances sont encodées dans la langue et pas d'autres ? » et « Dans quelle mesure ce qu'est le monde actuellement (ou ce qu'il a été auparavant) justifie les choix d'encodage et dans quelle mesure c'est arbitraire ? »

Dans sa thèse, Robert BANGHA (2003) souligne l'idée que c'est parce que la connaissance est écrite dans une langue et qu'elle est importante et plus présente dans nos esprits. Reprenons l'exemple du restaurant « **المطعم** » : il y a des employés de cuisine, cette connaissance est également encodée dans la langue arabe à travers les lexies « **المطعم** » et «**طباخ**» et les liens lexicaux qui les unissent : notamment, « **المطعم** » est un actant[1] de « **طباخ** » – et « **طباخ** » est aussi un actant de « **المطعم** ».

D'un point de vue conceptuel, la situation est similaire par exemple : un guichet d'informations. Précisément, on y trouve un employé qui donne des renseignements, mais pas de lexie en arabe pour désigner cet employé « **المكلف بمكتب الاستعلامات** ». L'encodage lexical d'une connaissance peut fortifier les connaissances du monde à propos de ce à quoi on se réfère.

5.1.2 Comment les connaissances sont lexicalisées ?

Bien entendu, nous soutenons l'idée que la langue encode une grande quantité de nos connaissances du monde mais elle le fait de façon fortuite à partir de nos CM.

Pour commencer, il faut recenser tous les concepts et toutes les entités qui sont lexicalisées dans un certain domaine afin de relever l'ensemble des mots (qui ont un sens), « les lexies » qui marquent ce domaine. Prenons un exemple concret, celui de l'école :

1 En linguistique, le terme d'**actants** désigne les constituants syntaxiques imposés par la valence de certaines classes lexicales (comme le verbe**,** principalement, mais aussi le nom, l'adjectif, la préposition...).

« **المدرسة** ». Lorsqu'on parle d' « **المدرسة** », on peut tout de suite penser à des lexies comme « **طالب** », « **مدير** », « **ناجح** », « **الشهادة** », « **الدرس** », « **القلم** », « **القسم** », « **المرقد** », «**كشف** » [**النقاط**], « **الناظر** », « **المراقب** », « **نقاط** », « **مؤدب** », « **مربي** », « **معلم** », « مدرس », « **تدريس** », « **تلميذ** », « **الدخول** », « **دراسي** », « **غياب** », etc.

Le fait qu'il existe des lexies encodées dans la langue arabe désignant par exemple des personnes qui étudient dans des écoles : « **تلميذ** » ou « **طالب** », on parle ici de connaissances lexicalisées. Une résultante qui semble triviale, d'avoir des mots ayant du sens pour des personnes qui étudient dans une école, mais d'autres optiques[1] ne le voient pas du même angle. Si on compare les deux lexies : « **المدرسة** » et « **المرآب** » ; dans le premier, on peut trouver des écoliers « **التلاميذ** », et dans le second, on peut trouver des voitures « **السيارات** ». Ces deux connaissances sont conceptuellement comparables mais elles ne sont pas encodées de la même façon en arabe.

Le mot « **التلاميذ** » est lexicalisée et désigne des personnes qui étudient à l'école, mais la seconde « **المرآب** » ne l'est pas, car il n'existe pas de mots propres aux voitures garées dans un parking. Ici nous soulignons l'encodage imprévisible des connaissances. Bien sur, il est clair que la définition de la lexie « **المرآب** » doit contenir des éléments relatifs au stationnement des voitures. Cette comparaison nous laisse penser que la présence de la lexie « **التلاميذ** » n'est pas une exigence absolue. [Ban-03]

5.1.3. Dictionnaires et connaissances lexicalisées

Il est clair que pour définir des mots, on a toujours recours à des dictionnaires. Mais souvent les définitions ont plusieurs acceptations qui ne sont pas identiques – même si elles sont souvent semblables par rapport à la référence choisie, par exemple : « **زَيَّنَ** » est synonyme de « جَمَّلَ ». Bien qu'il existe des difficultés concernant la conception des définitions des lexies, Robert BANGHA croit qu'il est possible de surmonter ces difficultés en s'appuyant sur une lexicologie explicative et combinatoire (LEC) suivant des critères précis. Trois étapes qui ont le plus de rapport avec l'encodage des CM :

→ Déterminer le genre prochain,

→ Déterminer les différences spécifiques

→ Prendre en considération la pertinence linguistique.

1 - Optique saussurienne (**Ferdinand de Saussure**, linguiste suisse (1857-1913)), la langue est considérée comme un système relativement autonome et arbitraire – et non pas comme un simple reflet, un simple encodage du monde ou de nos CM

La définition hypéronymiques (genre prochain et différence spécifique) est souvent privilégiée en lexicographie. [Ban-03].

a/ Détermination du genre prochain (hyperonymie)

Le genre prochain ou hyperonymie « *est un concept dont les traits sémantiques sont partagés par les concepts qui lui sont immédiatement subordonnés* ». La relation fondamentale appelée IS_A, souligne l'appartenance d'une connaissance à une autre, et marque clairement cette notion en Intelligence artificielle.

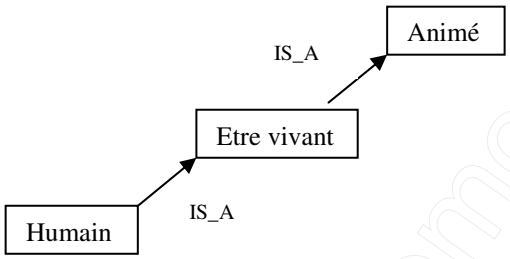

Fig. 12 : relation IS_A - extraite de l'arbre de Porphyre

Ainsi, si on remarque bien les définitions lexicographiques, on se rend compte que cette relation (IS_A) est présente et encode des connaissances similaires. Le genre prochain est la composante centrale de la définition de la lexie : il en est la paraphrase minimale, l'hyperonyme le plus proche. Par exemple, dans un dictionnaire, on trouve la définition d'un restaurant « **المطعم**» :

Is A

المَطْعَم :هو مكان تقدم فيه المأكولات والمشروبات للزبائن

Cette définition nous indique que « **مكان »** est le genre prochain de « **المطعم**». Cela signifie que nos connaissances du monde qui indiquent qu'un restaurant est un endroit sont encodées dans la langue à travers le genre prochain de « **المطعم**». [Ban-03]

b/ Détermination des différences spécifiques

Prenons un petit exemple d'oiseaux : le chardonneret et la huppe et soit les lexies suivantes « **الحسون** » ، « **الهدهد** » :

Premièrement, tous deux sont des oiseaux, de taille moyenne, le premier aux plumages bariolé et l'autre aux plumes roussâtres, etc. Parmi tout ce que nous savons à leur sujet, que faut-il encore inclure dans leur définition ? Malgré toute l'importance du genre prochain, il ne peut pas suffire à lui seul pour définir une lexie. C'est pour cette raison que l'on doit inclure également dans la définition un ensemble de composantes que l'on appelle les différences

spécifiques : elles permettent, d'une part, de faire une distinction sémantique entre les lexies qui ont le même genre prochain, d'autre part, de caractériser la combinatoire sémantique d'une lexie. [Ban-03]

5.2 Fouille des connaissances dans les liens lexicaux

Reprenons l'exemple de la figure 12, on peut dire que le concept «Animé» est l'hyperonyme du concept «Etre vivant», mais aussi on peut associer une relation d'hyponymie, en occurrence le concept « Humain » est hyponyme du concept «Etre vivant ». Remarquons par la suite, que nos connaissances du monde ne nous permettent pas d'envisager la façon dont seront encodées ces connaissances (d'un point de vue purement conceptuel) dans la langue sans connaître vraiment la langue arabe. Illustrons tout ça par un deuxième exemple, le cas de ceux qui empruntent des livres d'une bibliothèque [Ban-03]:

— القارئ للمكتبة

— المستخدم للمكتبة

Et non pas (ça serait étrange):

— المستأجر للمكتبة

6. Les outils du traitement automatique de la langue (TAL) Arabe

Le TAL est un domaine de savoir et de méthodes élaborées autour de diverses préoccupations. Beaucoup de concepts et de techniques régissent de son étude et il se trouve à l'intersection de multiples disciplines : l'informatique théorique, la logique, la linguistique, l'Intelligence Artificielle, mais aussi les neurosciences, les statistiques, etc. Pour mieux cerner le TAL considérons à juste exemple l'énoncé :

— هَتَفَ الجُمْهُورُ بِدُخُولِ اللاَّعِبِينَ المَلْعَبَ.

Remarquons l'enchainement des opérations qu'il faudra suivre pour réussir la compréhension complète et automatiquement de cette phrase. Il nous faudra :

1. Segmenter cette phrase en unités lexicales (mots) ;

— « هَتَفَ » ، « الجُمْهُور » ، « بِدُخُول » ، « اللاَّعِبِينَ »، « المَلْعَبَ » .

2. Identifier les composants lexicaux, et leurs propriétés : c'est l'étape de traitement lexical ;

— "هتف" : فعل ماضي ، منصوب بالفتحة الظاهرة على آخره ، . . .

— "الجمهور" : فاعل مرفوع بالضمة الظاهرة على آخره ،

— **بـ : حرف جرٍ .**

— **الملعبَ : مفعول به ، منصوب بالفتحة الظاهرة على آخره.**

3. Identifier des constituants (groupe) de plus haut niveau, et les relations (de dominance) qu'ils entretiennent entre eux : c'est l'étape de traitement syntaxique ;

— **دخول : اسم مجرور بالكسرة الظاهرة على آخره ، وهو مضاف**

— **اللاعبين مضاف إليه مجرور بالياء والنون.**

4. Bâtir une représentation du sens de cette phrase, en associant à chaque concept évoqué un objet ou une action dans un monde de référence (réel ou imaginaire) : c'est l'étape de traitement sémantique ;

— Le public a acclamé les joueurs dès leurs entrer dans le stade

— **يُحيِّ الجُمْهُور اللاَعِبِينَ عند دُخُولِهم المَلْعَبَ.**

5. reconnaître enfin la fonction de l'énoncé dans le contexte particulier de la situation dans lequel il a été produit : c'est l'étape de traitement pragmatique ;

— **هَتَفَ الجُمْهُور بشدّةٍ بِدُخُولِ اللاَعِبِينَ الجزائريين المَلْعَبَ يوم 12 جوان 2010.**

L'analyse morphologique, est la première étape d'un traitement linguistique de données textuelles. Le challenge à la réalisation des analyseurs morphologique et d'étiquetage grammatical non ambigu de corpus est devenu d'actualité.

6.1. Analyseurs morphologique

L'analyse morphologique est très développée pour les langues latines. Mais ce n'est pas le cas pour la langue arabe par manque de ressources linguistiques (eg. corpus, lexique de base, segmenteurs de textes en phrases,…). C'est pour cette raison, que la majorité des travaux se sont basés sur l'étiquetage morphologique s'appuyant sur des méthodes d'apprentissage et une légère analyse morphologique, par exemple khoja 2001 ; Diab et al. 2004.

L'analyseur morphologique consiste après segmentation du texte, à étudier la forme d'un mot pris isolement (sans contexte) et à déduire les informations dérivationnelles et inflectionnelles. Ainsi, l'analyseur doit générer pour le mot traité une ou plusieurs solutions morphologique décrites par les informations suivante : les suffixes, les préfixes, le radical, la forme canonique (lemme) ainsi que d'autres informations comme le genre grammatical (féminin, masculin), le nombre (singulier, pluriel) ou le temps (verbe conjugué au présent, au passé parfait,...etc.).

Le premier essai d'analyse de la langue arabe a été proposé par David Cohen dans les débuts des années soixante. (1960).

6.1.1. L'analyseur morphologique à états finis de Beesley 2001 (Xerox)

Beesley a développé un analyseur morphologique arabe utilisant les outils de Xerox de modélisation de langage à état fini. Cet analyseur donne pour chaque mot toutes ces listes de caractéristiques morphologiques possibles. Cet analyseur de Beesley trouve son utilisation comme composante d'aide à l'apprentissage dans un système de traitement de langage naturel, plus large.

6.1.2 L'analyseur morphologique de Buckwalter : Aramorph

Développé par Tim Buckwalter [Buc-04] en langage Perl pour le compte du **Linguistic Data Consortium, Université de Pennsylvanie (LDC)** ***actuellement sous Java. Cet Analyseur baptisé «*** Aramorph » est c***onsidéré comme «*** **la ressource lexicologique la plus respecté de son genre».** Le texte analysé en entrée doit être transformé en code ASCII (translitération) avant traitement et le résultat d'analyse doit retranscrit en arabe (translitération inverse) afin d'être compris.

Transliteration	Arabic Windows	Unicode Value and Unicode Name
'	C1	U+0621 ARABIC LETTER HAMZA
|	C2	U+0622 ARABIC LETTER ALEF WITH MADDA ABOVE
>	C3	U+0623 ARABIC LETTER ALEF WITH HAMZA ABOVE
&	C4	U+0624 ARABIC LETTER WAW WITH HAMZA ABOVE
<	C5	U+0625 ARABIC LETTER ALEF WITH HAMZA BELOW
}	C6	U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE
A	C7	U+0627 ARABIC LETTER ALEF
b	C8	U+0628 ARABIC LETTER BEH

Fig. 13 : Extrait du tableau de translitération arabe de Buckwalter

L'analyseur n'accepte pas du texte en arabe avec de l'alphabet romain dans le même document ; Un problème peut être rencontrée lorsque par exemple, le texte contient des étiquettes de Part of Speech ou marqueurs XML en en alphabet romain.

6.1.3 L'analyseur morphologique Sebawi de Darwish

Développé par Darwish [DAR-02] en une seule journée, Sebawi est un analyseur morphologique de la langue arabe. C'est un analyseur de surface utilisé dans des applications de recherches d'information. Il réalise seulement la recherche des racines

possible d'un mot arabe. Cet analyseur morphologique arrive dans 84% des cas à trouver avec succès la racine.

6.2. Les Part of speech Taggers:

L'étiquetage grammatical (en français) ; consiste à donner une étiquette (tag) à chaque mot analysé du corpus, décrivant sa fonction grammaticale dans une phrase donnée :

- Texte original :

Nous sommes allées à Sidi- BelAbbès faire un stage pédagogique.

Texte étiqueté :

Nous/PRO:PER sommes/VER:pres allées/VER:pper à/PRP/à Sidi- BelAbbès /NAM …

L'étiquetage grammatical peut avoir une ou plusieurs solutions possibles pour un mot analysé. C'est pour cette raison qu'il faut penser à supprimer l'ambigüité, tout en considérant le contexte dans lequel le mot apparait.

6.3. Le tagger APT de Khoja

Arabic Part-of-Speech Tagger (APT) a été développé par Shereen Khoja [KHO-01], [KHO-03]. Cette méthode se base sur une combinaison de techniques statistiques et des techniques à base de règles. Les étiquettes (Tags) du « tagset » (l'ensemble des étiquettes définies) de l'APT sont initialement dérivées des étiquettes du « tagset » du corpus BNC (British National Corpus), mais qui ont étaient modifiées en prenant en compte quelque concepts de la grammaire traditionnelle arabe.

La raison de cette modification est que la langue arabe possède ses propres systèmes syntaxique, sémantique et morphologique qui rendent difficile l'adaptation aux « tagset » des langues indo-européennes. Le « Tagset » contient 131 étiquettes qui vont être assignées aux mots analysées. Un corpus de 50.000 mots issu des articles de presse saoudienne « Al Jazzira » a était utilisé pour l'apprentissage de l'étiqueteur.

7. Conclusion

Le chapitre, ainsi exploré, donne une plate forme pour la construction d'un système de traitement du langage naturel. Ce système nécessite plusieurs ressources lors des différentes phases d'analyses (Dictionnaire, ontologie).

Notons que les outils mis à notre disposition pour le traitement de la langue arabe sont appauvrit par rapport à d'autres langues telles que le français ou l'anglais. Les outils, non disponibles gratuitement à présent, permettent d'aborder les corpus textuel arabes dans une perspective d'extraction des connaissances.