République Algérienne Démocratique et Populaire
الجـمهوريـــــة الجـــزائريـــة الديمـــقراطيـــة الشـــعبية
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
وزارة التـعليـم العـــالي و البـــحث العلـــمي
Université Abou Bekr BELKAID TLEMCEN
Faculté des Sciences de l'ingénieur
Département d'informatique

THESE

Présentée par

Mohammed El Amine ABDERRAHIM

Pour obtenir le titre de doctorat
Université **Abou Bekr BELKAID TLEMCEN**

Spécialité : **Informatique**

Titre :

Reconnaissance des unités linguistiques signifiantes

Soutenue le 08 Juillet 2008 devant le jury composé de :

Mr Mohamed Amine CHIKH (MC Univ. Tlemcen)	Président
Mr Fethi BEREKSI REGUIG (Prof. Univ. Tlemcen)	Encadreur
Mme Laurence BALICCO (Prof. GRENOBLE 3)	Encadreur
Mr Abdelkader BENYETTOU (Prof. Univ. USTO)	Examinateur
Mr Sidi Mohamed RETERI (Prof. Univ. Tlemcen)	Examinateur

Année universitaire: 2007/2008

A Djamila
Je dédie cette thèse.

Chapitre 3 : Les analyseurs existants et le choix d'une organisation

Chapitre 4 : Définition du modèle linguistique Et description de l'analyseur morphologique

Bibliographie & Annexes

Introduction

Introduction

Introduction

Le traitement automatique des langues naturelles (TALN) est un domaine à la frontière de la linguistique et l'informatique, il a pour objectif de développer des logiciels capables de traiter de façon automatique des données linguistiques exprimées dans une langue naturelle donnée et pour une application bien définie. Cet objectif passe nécessairement par l'explicitation des règles de la langue puis les représenter dans un formalisme calculable et enfin les implémenter à l'aide des programmes informatiques.

Parmi les applications les plus connues du TALN, on peut citer :

- la traduction automatique ;
- la correction orthographique ;
- la recherche d'information et la fouille de textes ;
- le résumé automatique ;
- la génération automatique de textes ;
- la synthèse de la parole ;
- la reconnaissance vocale ;
- la reconnaissance de l'écriture manuscrite ;

Le travail que nous présentons dans cette thèse constitue une contribution modeste à l'effort qui vise à doter la langue arabe par des outils performants de traitement automatique. L'enjeu est double :

- d'un point de vue culturel, les langues qui ne seront pas informatisées, risquent d'être exclues des médias modernes de production et de diffusion de l'information ;
- d'un point de vue économique, le marché des applications pour le TALN arabe connaît une forte croissance et les gains en productivité dans de nombreux secteurs ne sont pas à démontrer.

On distingue deux aspects différents pour le TALN écrite : l'analyse et la génération. L'analyse se compose d'une suite de traitements (morphologique, syntaxique,

sémantique,…), elle consiste à construire une représentation formelle du texte en entrée, cette représentation doit être facile à manipuler par la machine. Par ailleurs la génération consiste à générer des textes à partir d'une représentation interne. Il s'agit de la fonction inverse de celle de l'analyse, mais elle n'est pas forcément obtenue en inversant le processus. La génération de textes apparaît dans des applications comme la traduction automatique de texte, le résumé automatique de texte, la génération de comptes-rendus boursiers ou météorologiques, etc.

Dans le cadre de cette thèse, nous nous positionnons dans le cadre de l'analyse, nous nous intéressons plus particulièrement à un aspect fondamental du TALN arabe écrite à savoir : la reconnaissance des unités linguistiques signifiantes. Autrement dit nous nous intéressons à la formation des formes (mots) au travers des processus de flexion (marques de genre, nombre, de conjugaison...) et d'agglutination. Étant donné une forme, il s'agit de déterminer quelles sont les unités minimales de sens qui le composent (Ces unités minimales de sens sont appelées morphèmes). En terme de niveau d'analyse d'une langue nous nous situons au niveau morphologique qui est un premier pas vers l'analyse syntaxique.

Reconnaître les unités linguistiques signifiantes constitue donc un enjeu fondamental dans le cadre des applications pour le TALN arabe. En effet, la qualité de ces applications dépend largement de la qualité de reconnaissance de ces unités. Il s'agit donc, dans cette thèse, de développer un modèle et par conséquent un outil pour la reconnaissance des unités linguistiques signifiantes pour la langue arabe écrite, facilement réutilisable par les applications de TALN arabe et non lié à un domaine particulier.

La reconnaissance des unités linguistiques signifiantes ne constitut pas une fin en elle-même, mais plutôt un préalable nécessaire à diverses applications.

Ce travail s'inscrit dans le cadre général d'un projet scientifique développé au Laboratoire de Traitement Automatique de la Langue Arabe (désormais LTALA) à l'université Abou Bekr Balkaid Tlemcen Algérie que dirige Mr Sidi Mohamed RETERI (professeur dans la même université et enseignant-chercheur à LTALA).

Le LTALA s'est constitué autour d'un projet principal qui est la mise au point de logiciels pour le traitement automatique de la langue arabe. Dans cette optique, les activités portent essentiellement sur deux axes :

- d'une part, la modélisation linguistique propre à la langue arabe,

- d'autres part, la conception et la réalisation de logiciels pour divers domaines d'applications.

Dans ce contexte, on doit disposer d'un ensemble d'outils permettant de faciliter la mise au point des analyseurs et des dictionnaires. Ces outils peuvent nous servir pour :

- la mise au point des grammaires,
- la construction et la mise à jour des dictionnaires,
- la validation des données linguistiques.

L'objectif de ce travail est de construire un modèle de base pour le traitement automatique de l'arabe. Ce modèle repose sur une modélisation des unités linguistiques signifiantes. Pour la validation nous avons réalisé concrètement un analyseur morphologique de l'arabe écrit voyellé ou non. Cet objectif comprend donc deux phases :

- l'élaboration d'un modèle linguistique pour le traitement (on parle souvent de modélisation linguistique),
- la validation du modèle construit par la réalisation et l'expérimentation d'un analyseur morphologique pour les textes arabe.

L'analyseur ainsi construit sera utilisé et réutilisable pour de nombreuses applications de l'équipe. La polyvalence et le caractère réutilisable de cet analyseur justifié l'effort nécessaire à fournir pour son développement.

Le présent document décrit notre travail, ce dernier étant effectué au sein :

- du LTALA sous la direction du Mr le professeur F. BEREKSI,
- du Groupe de recherche sur les enjeux de la communication GRESEC équipe CRISTAL à l'université STENDHAL GRENOBLE 3 sous la direction de Mme le professeur L. BALICCO.

Il s'organise en cinq chapitres et se termine par une conclusion.

Pour situer notre étude dans la chaîne du traitement automatique d'une langue, nous passons en revu les différents niveaux d'analyse d'une langue naturelle dans le premier chapitre.

Dans le deuxième chapitre, nous présentons un modèle pour le mot (ou forme) graphique arabe. Ce modèle va nous servir de support pour l'élaboration de notre modèle conceptuel pour le traitement automatique de l'arabe. Différents problèmes du traitement automatique des mots arabes seront ainsi discutés.

Différentes approches et organisations pour l'analyse sont présentées dans le chapitre trois. La solution adoptée est motivée suite à une analyse détaillée de tous les systèmes existants.

Avant toute analyse automatique d'une langue naturelle, il est indispensable de définir un modèle linguistique à priori. La première partie du chapitre quatre présente notre modèle linguistique pour l'analyse en terme de classes et d'objets, par ailleurs, la seconde partie décrit le détail de la procédure d'analyse.

La réalisation et l'expérimentation de notre analyseur sur différents textes a fait l'objet du dernier chapitre. Le détail de l'implantation en terme de structures de données et les résultats obtenus par l'analyseur seront ainsi abordés.

Chapitre 1

Le traitement automatique des langues naturelles

Chapitre 1

Le traitement automatique des langues naturelles

1. Introduction

Le Traitement Automatique des Langues Naturelles (TALN) a fait l'objet de développement important ces dernières années.

L'analyse et la génération sont deux aspects différents pour le TALN écrite. L'analyse est une suite de traitements (morphologique, syntaxique, sémantique,…), elle consiste à construire une représentation formelle du texte en entrée, cette représentation doit être facile à manipuler par la machine. Par ailleurs la génération consiste à générer des textes à partir d'une représentation interne. Autrement dit, elle consiste à faire produire des textes par un ordinateur, de façon à exprimer automatiquement un contenu formel en langue naturelle. La génération comprend deux partie : le " quoi dire ? " qui consiste à déterminer le contenu du texte à engendrer (on parle de génération profonde), et le " comment le dire ? " qui est son expression en langue naturelle (on parle de génération de surface). [BALI-1993], [BALI-2000]

On ne peut considérer la génération comme l'inverse pur de l'analyse, il suffit de comparer l'effort demandé pour écrire un article avec celui de l'effort requis pour le lire. La comparaison entre l'analyse et la génération n'est pas évidente, car, « *Les difficultés de l'analyse et de la génération se placent à des niveaux différents et il ne faut pas chercher à comparer ces traitements en se demandant lequel est le plus simple. Il est vrai qu'il est compliqué de comprendre quelqu'un, mais il est également très délicat de se faire comprendre, et surtout se faire bien comprendre, sans ambiguïté.* » [BALI-1993 ; pp.18-19]

Dans le cadre de cette thèse, nous nous positionnons dans le cadre de l'analyse, nous nous intéressons plus particulièrement au niveau morphologique. Il convient donc, dans cette optique, de proposer dans ce qui suit un rapide survol des différents niveaux d'analyse.

Formellement, la langue écrite est un ensemble de chaînes de caractères. On convient d'appeler MOT toute chaîne de caractères délimitée par des blancs. Par conséquent on admet qu'une phrase est constituée par une succession de MOTS. Sur le plan linguistique, ces définitions posent des problèmes, d'ailleurs, [JAYE-1985] écrit :

« *Sur le plan linguistique, de telles définitions soulèvent de nombreuses difficultés : sur le mot : au fur et à mesure, … recouvrent un sens mais apparaissent comme plusieurs segments graphiques ; … Nous conservons cependant des définitions aussi imprécises pour la bonne raison que la complexité même du problème ne permet pas aux écoles linguistiques actuelles d'aboutir à une définition du mot qui soit réellement opératoire sur machine ; dans ces conditions, il nous semble préférable de fonder la notion de mot sur le seul découpage de texte que l'on sait reproduire par machine, à savoir celui que l'on obtient par reconnaissance des blancs ou caractères spéciaux comme la virgule, le point…* » [JAYE-1985 : p.65]

2. Différents niveaux d'analyses

A partir des séquences de chaînes de caractères, différents niveaux d'analyses (traitement) peuvent être envisagés. On parle dans la littérature d'analyse morphologique, d'analyse syntaxique, d'analyse sémantique, et d'analyse pragmatique.

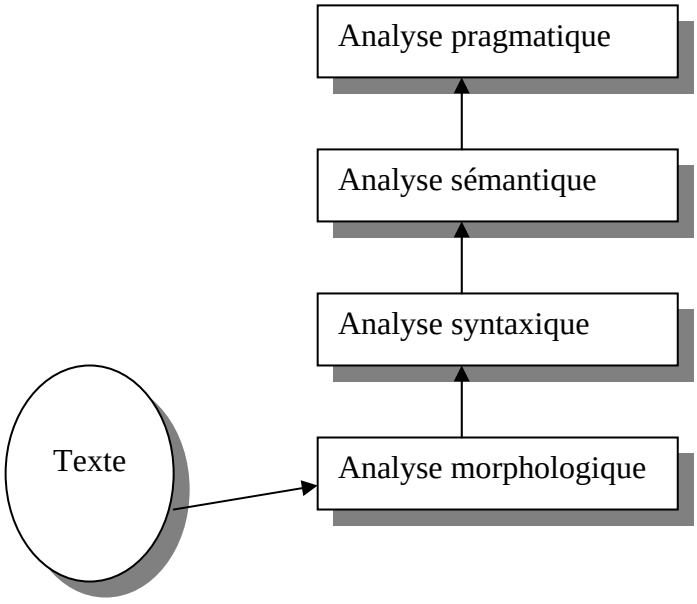


Figure 1 : Les différents niveaux d'analyse d'un texte

Le traitement automatique des langues se heurte à deux difficultés :

- L'ambiguïté de la langue : elle concerne les différents types d'ambiguïté propres à chaque niveau d'analyse. On parle souvent d'ambiguïté morphologique, d'ambiguïté syntaxique, d'ambiguïté sémantique, et d'ambiguïté pragmatique.
- La complexité des connaissances qui doivent être mises en œuvre à tous les niveaux d'analyse.

Dans ce qui suit nous allons décrire brièvement les différents niveaux d'analyse d'un texte en langue naturelle.

2.1. Analyse morphologique

L'analyse morphologique est indispensable pour tout système de traitement automatique de la langue naturelle, cette analyse permet de regrouper les mots en classes utilisables par les autres niveaux d'analyse. La définition de ces classes varie en fonction des traitements envisagés. A chaque classe on associe une étiquette appelée catégorie grammaticale ou catégorie lexicale. Il arrive qu'un même mot peut avoir différentes catégories grammaticales, on dit qu'il y a ambiguïté grammaticale ou une homographie.

L'analyse morphologique des langues comme le français ou l'anglais ne pose plus un problème et bons nombres d'analyseurs efficaces sont réalisés. L'analyseur morphologique du français proposé par [LALL-1990] en est un bon exemple. Ce dernier va nous servir comme base pour l'élaboration de notre analyseur pour la langue arabe, il comprend trois étapes :

- Préparer le texte en entrée à l'analyse. L'objectif de cette étape est de simplifier les phase ultérieures de l'analyse par la normalisation des caractères (par exemple le codage du texte à l'aide uniquement du code ASCII standard, la substitution d'une chaîne de caractères par une autre…) et le découpage du texte en formes. Ces prétraitements font partie du modèle linguistique, ils sont regroupés dans deux grandes classes : les prétraitements morpho-graphiques (par exemple le traitement des majuscule, le traitement des ponctuations…) et les prétraitements morpho-syntaxiques. Ces derniers sont basés sur l'application d'un ensemble de règles pour régulariser la surface du texte tout en amorçant l'analyse. Parmi ces règles on trouve par exemple celles de l'éclatement d'amalgames orthographiques, ou la suppression de formes.

- Chaque forme est traitée isolement par l'analyseur. Une ou plusieurs interprétation possibles en terme de couple (entrée lexicale, catégorie) sont associées à la forme dans cette étape.
- Levée des ambiguïtés par l'utilisation du contexte, ce qui a pour conséquence la réduction des interprétations multiples.

L'analyse d'une forme dans [LALL-1990] revient à trouver tous les découpages base + (flexions) attestés. Les bases sont données par un dictionnaire, par contre les flexions sont données par une liste de flexions qui sont particularisées.

Pour réaliser sa tâche, l'analyseur morphologique de [LALL-1990] a besoin d'un ensemble de données :

- la chaîne à analyser (issue du prétraitement),
- le dictionnaire,
- la liste des modèles des noms-adjecifs et des verbes,
- les régularisations de formes et de bases,
- la liste des flexions et leur compatibilité.

A l'heure actuelle, la conception et la réalisation d'un analyseur morphologique pour une langue comme le français ou l'anglais sont très bien maîtrisées et les analyseurs produits sont jugés efficaces, d'ailleurs [PITR-1985] confirme tout cela en écrivant :

« *L'analyseur comporte un programme d'analyse morphologique des mots.* ***Il s'agit d'un problème bien connu pour lequel nous savons réaliser des programmes efficaces.*** » [PITR-1985 : p.77]

Malheureusement, pour la langue arabe les choses ne font que commencer et ce domaine pose encore des problèmes. [BEN-1998] résume bien la situation en écrivant dans la page numéro huit de sa thèse :

« ***12.problématiques bien actuelles*** *… En effet, s'il est vrai que les travaux portant sur la langue arabe et sur son traitement par des moyens automatiques foisonnent, il reste aussi vrai que des résultats à la fois éprouvés et disponibles sont toujours attendus. Au moment où les dictionnaires électroniques pour l'anglais ou le français existent et sont même disponibles dans le domaine public, pour la langue arabe, un dictionnaire électronique de formes fléchies plutôt exhaustif fait encore tout simplement défaut. En la matière, les annonces faciles faites ici et là ne doivent nullement induire en erreur...* » [BEN-1998 : p8]

Une expérience (réalisée par [HASS-1987]) consistant à adapter un analyseur du français (AMEDE[1]) pour analyser des textes arabes a montré que :

- L'analyseur AMEDE peut être utilisé comme un outil de traitement automatique de la langue arabe.
- L'analyse des textes vocalisés donne un excellent résultat (presque sans ambiguïté).
- L'analyse des textes non vocalisés produit toujours des ambiguïtés. [HASS-1987]

A propos de l'adaptation de modèles utilisés pour le français [HASS-1987] conclu :

« *Les particularités de la langue arabe, et notamment l'importance de la structure à base de consonnes et le système de dérivation qui en découle, nous ont montré que le modèle linguistique utilisé pour le français ne permettait la prise en compte de l'arabe qu'au prix de traitement ad hoc et artificiels.* » [HASS-87 : p.25]

L'analyse morphologique permet de reconnaître une chaîne de caractère comme étant un mot de la langue. A chaque chaîne sont associés :

- une classe lexicale (grammaticale) décrivant la fonction syntaxique du mot. Par exemple verbe, substantif, adjectif, etc.…
- certaines variables. Ce sont des compléments de la description de la fonction syntaxique du mot. Par exemple le genre, le nombre pouvant prendre les valeurs singulier, pluriel, etc.… [MERL-1982]. Ces variables sont à définir sur des bases linguistiques et en fonction des besoins des traitements envisagés. Par exemple, dans le cadre de l'analyseur de [LALL-1990], ces variables sont réparties en trois classes suivant leur définition : syntaxique, flexionnelles et lexicales. On retrouve[2] :
    - o des variables de sous-catégorisation syntaxique qui caractérisent le type nominal, le temps des participes, les préverbaux…
    - o des variables de sous-catégorisation flexionnelle qui caractérisent le genre, le nombre, la personne, le mode, le temps…
    - o et éventuellement des variables de sous-catégorisation lexicales qui caractérisent le sous-type nominal, l'auxiliaire…

En d'autres termes un analyseur morphologique assure :

1 Analyseur morphologique pour le français réalisé par le laboratoire d'informatique documentaire de l'université LYON1.

[2] Pour une étude détaillée de ces variables et leurs valeurs voir [LALL-1990 ; pp 93-99]

- la segmentation du texte en "mots",
- le contrôle de validité de ces "mots",
- le calcul des variables morphologiques,
- la recherche de toutes les solutions possibles.

L'analyse morphologique suppose construire : un ensemble de règles (que nous appellerons une grammaire) et un dictionnaire. La grammaire contient des règles qui contrôlent la composition des formes à partir des éléments contenus dans le dictionnaire. Cette démarche permet de séparer la grammaire de la lexicographie d'où la possibilité d'enrichir le dictionnaire au fur et à mesure de son utilisation.

Le dictionnaire, élément commun à la plupart des systèmes de traitement automatique des langues naturelles, conditionne dans une large mesure la stratégie et par conséquent la qualité du système. Il contient généralement le « vocabulaire » de la langue naturelle traitée, permettant ainsi d'identifier les différents éléments constituant un texte. En d'autres termes, « *Le dictionnaire est le complément de l'analyseur morphologique. L'analyseur y accède pour y prendre toutes les informations linguistiques dont il a besoin et qu'il ne sait pas calculer.* » [LALL-1990 ; p.42]. Donc l'analyseur morphologique et le modèle linguistique définissent dans une large mesure le contenu du dictionnaire. Par exemple le dictionnaire utilisé par [LALL-1990] comprend environ 70000 enregistrements (soit 50000 entrées lexicales) et chaque enregistrement contient :

- une base prétraitée qui correspond à la partie fixe d'une forme,
- la forme canonique non prétraitée associée à la base (verbes à l'infinitif, noms au singulier et les adjectifs au masculin singulier),
- la catégorie syntaxique de la base,
- un indicateur I suivi des valeurs de variables flexionnelles associées à la base si cette dernière est invariable, sinon, un indicateur M suivi du nom du modèle morphologique,
- les valeurs de variables associées à la base, à l'exclusion des variables de genre et de nombre [LALL-1990 ; p.42].

En constatant que le « vocabulaire » d'une langue est très vaste, il sera donc tout à fait inadapté de construire un dictionnaire a priori. Dans ce cadre une construction progressive de ce dernier s'impose, ce qui implique que nous devons disposer d'un ensemble d'outils permettant sa mise à jour.

Espace et temps sont à la base de l'évaluation de chaque stratégie, fonder un choix revient donc à résoudre un problème déjà très connu :

« *Pour une technologie donnée, pour une certaine configuration machine et pour une application souhaitée : un compromis doit souvent être trouvé entre stocker des informations dans un espace mémoire minimal mais au prix d'une efficacité de traitement faible et implanter des informations dans un volume mémoire important mais permettant un accès très rapide.* » [GRAN-1975 : p.2]

Deux solutions se dégagent selon que l'on essaie de minimiser la taille du dictionnaire (réduire la place mémoire) ou de simplifier la grammaire (réduire la complexité des traitements).

Dans sa thèse, [VERG-1989] montre qu'il est possible d'analyser morphologiquement et syntaxiquement des textes (français) en n'utilisant qu'un dictionnaire réduit comportant des déterminants, des adjectifs indéfinis et numéraux, des conjonctions et prépositions et quelques adverbes. La taille de ce dictionnaire est de quatre-vingts formes.
L'intérêt d'une telle approche est de se passer du dictionnaire et donc, éviter tout un processus très complexe de construction et de mise à jour de ce dernier. [CLAV-1995] reproche à cette démarche d'être insuffisante et non généralisable :

« *De notre point de vue, il est illusoire de pouvoir jamais se passer d'un dictionnaire fût-il réduit, ne serait-ce que parce que ce dernier comporte des informations catégorielles et sémantiques qu'il faut avoir préalablement définies.* » [CLAV-1995: p.121]

La deuxième solution consiste à construire un dictionnaire comportant tout le vocabulaire de la langue. L'analyse se résume donc à un simple accès au dictionnaire, ce qui présente l'avantage de réduire l'effort déployé pour l'élaboration d'une grammaire d'analyse. Cette solution n'est pas envisageable, [CLAV-1995] a qualifié cette démarche d'être passéiste :

« *Nous nous inscrivons donc en faux vis-à-vis des positions contraires qui mettent en avant la mémoire prodigieuse des ordinateurs et leur rapidité de traitement pour justifier l'existence de volumineux dictionnaires. Cette dernière solution est passéiste même si elle présente certains avantages dans le cadre des applications actuelles.* » [CLAV-1995: p.123]

Pour conclure, on peut dire que les avantages de l'une sont les inconvénients de l'autre, et inversement. Toutefois, face à la première solution et à la deuxième solution, il semble et il parait plus logique de trouver une solution entre les deux, c'est-à-dire élaborer un dictionnaire et une grammaire.

2.2. Analyse syntaxique

« Personne ne connaît la syntaxe d'une langue en termes d'algorithmes, programmables sur ordinateur ; les ouvrages de grammaires traditionnels formulent des règles destinées à être appliquées par des êtres sensés et déjà pourvus de la connaissance de la langue. » [JAYE-1985 : p.177]

Un langage formel est défini par sa grammaire, alors que la langue naturelle ne l'est pas. En effet une langue n'est pas définie par sa syntaxe, car cette dernière (la syntaxe) est écrite postérieurement et ne présente qu'une approximation, d'où on parle de modèle syntaxique. C'est cette approximation qui fait que l'analyse syntaxique n'est pas précise et pose des difficultés. D'ailleurs, [MERL-1982] souligne ce fait et écrit :

« *L'impossibilité d'effectuer une analyse syntaxique complète et fiable et d'obtenir une "compréhension" suffisante de textes libres (en vue, par exemple, d'une interprétation des textes par un logiciel), provoque une situation de blocage.* » [MERL-1982 : p4]

Plusieurs méthodes d'analyse syntaxique se sont développées, mais la plus célèbre est sans doute la notion de grammaire formelle. Une grammaire formelle est présentée sous la forme d'un ensemble de règles de dérivation, ces règles exprimant la structure des entités syntaxiques telles que la phrase (PH), le groupe nominal (GN), le groupe verbal (GV) etc. Pour exprimer par exemple qu'une phrase est composée d'un groupe nominal et d'un groupe verbal, on utilise la règle PH → GN + GV. Et qu'un groupe nominal est composé d'un déterminant et d'un nom, on utilise la règle GN → Det Nom.
A l'aide de cet ensemble de règles, il est donc possible d'analyser un certain nombre de phrases.

Les constituants regroupant plusieurs mots (syntagme) possèdent donc leurs propres catégories (syntagme nominal, syntagme verbal, syntagme prépositionnel, etc.). La structure de ces constituants peut être représentée sous la forme d'un arbre (appelé arbre syntaxique) ou par d'autres systèmes de représentation comme par exemple le système de parenthèses. Toutefois la représentation la plus utilisée reste la représentation sous la forme d'arbre ou représentation arborescente. Dans cette représentation les

branches successives décrivent la décomposition en constituants, chaque noeud a une étiquette qui correspond à un syntagme ou à une catégorie et à chaque noeud terminal est associé un mot (ou item lexical). La figure suivante (voir figure 2) montre un exemple de représentation arborescente de la phrase « شرب الولد الماء ».

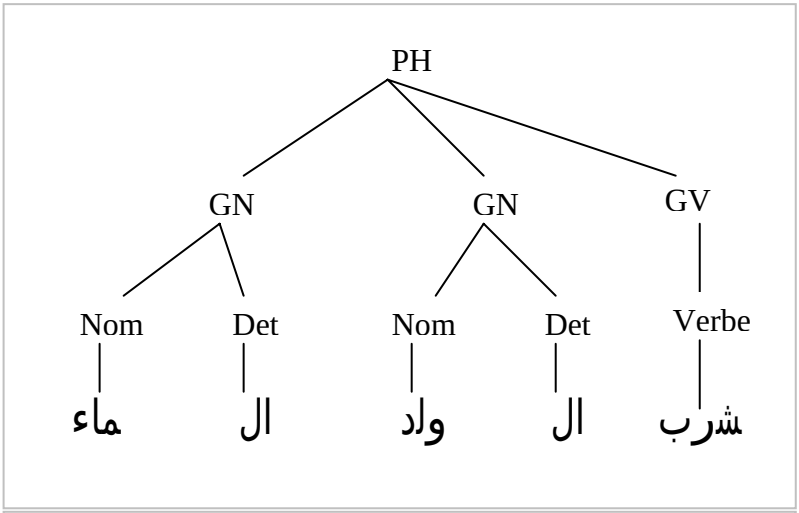


Figure 2 : Arbre syntaxique

Ce modèle est appelé dans la littérature, le modèle en constituants immédiats, ce dernier peut être formalisé par les grammaires indépendantes du contexte (formalisme appelé aussi système de réécriture : on réécrit les symboles de la partie gauche d'une règle à l'aide de sa partie droite).

Etant donnée une phrase, on peut donc l'analyser à l'aide des règles de la grammaire. Deux techniques d'analyse sont possibles :

- l'analyse descendante consiste à partir de la racine (dans l'exemple précédent le nœud « PH ») et essayer toutes les dérivations pour aboutir à la phrase à analyser. Si après avoir essayé toutes les dérivations et on est toujours en situation de blocage, alors la phrase ne fait pas partie du langage engendré par la grammaire.
- L'analyse ascendante consiste à partir de la phrase à analyser et appliquer les règles de dérivation à l'envers pour remonter à la racine. Si après avoir essayé toutes les possibilités et que l'élément racine n'a pas été retrouvé, alors la phrase ne fait pas partie du langage engendré par la grammaire.

Les grammaires indépendantes du contexte souffrent d'un certain nombre de limitations (l'étude de ces limitations sort du cadre de cette présentation), qui ne leur permettent pas de représenter certains aspects importants des langues. [VERO-2001]

Pour pallier ces déficiences (limitations), de nombreux formalismes ont été introduits. Ces modèles utilisent le mécanisme d'unification d'où leur nom de grammaire d'unification. Parmi ces grammaires on trouve :

- les grammaires lexicales fonctionnelles,
- les grammaires syntagmatiques généralisées,
- les grammaires syntagmatiques guidées par la tête,
- grammaires d'arbres adjoints.

Pour une étude détaillée de ces grammaires on pourra consulter [ABEI-1993].

2.3. Analyse sémantique

« *Dans un système de traitement automatique, l'analyse du sens des phrases consiste généralement à en extraire une représentation simplifiée, stylisée, de type logico-mathématique, qui va permettre des calculs et raisonnements ultérieurs.* » [VERO-2001 : p.10]

L'analyse sémantique des énoncés s'appuie sur une analyse syntaxique préalable. Elle cherche à construire une représentation formelle permettant des raisonnements et donc d'inférer de nouvelles informations à partir des informations présentes dans l'énoncé.

Parmi les représentations, on trouve la logique des propositions.

« Une proposition est un énoncé auquel on peut attribuer sans ambiguïté l'une des valeurs vrai ou faux. » [VERO-2001 : p.5]

A l'aide de connecteurs logiques (comme la conjonction « $\wedge$ », la disjonction « $\vee$ », la négation « $\neg$ » etc.), on peut former à partir des propositions de nouvelles propositions complexes.

Exemple

Proposition	Négation
شرب الولد الماء	لم يشرب الولد الماء

La logique des propositions ne s'intéresse pas au contenu des propositions mais seulement à leurs valeurs de vérité.

La logique des prédicats est une autre forme de représentation des énoncés. Par prédicat on entend une propriété telle que 'Homme', 'Mortel', etc.

Dans cette logique, pour créer des propositions, il faut combiner un prédicat avec un argument.

Exemple

On peut former la proposition « Socrate est un homme » : *Homme (Socrate)* où Homme représente le prédicat et Socrate représente l'argument

Au lieu de Socrat (individu particulier) on peut utiliser des variables d'individus (ne correspondent à aucun individu en particulier), par exemple *Homme (X)*. Et pour transformer des expressions (contenant des variables) en proposition on utilise des quantificateurs (comme la quantificateur existentiel « $\exists$ » ou bien le quantificateur universel « $\forall$ ») et des connecteurs.

Exemple

A partir de :

$\forall X$ *Homme (X)* $\Rightarrow$ *Mortel (X)* [Tout homme est mortel]

Homme (Socrate) [Socrate est un homme]

On peut déduire : *Mortel (Socrate)* [Socrate est mortel]

Malheureusement, de nombreux phénomènes ne peuvent pas être représentés en logique des prédicats et pour pallier cette limitation, plusieurs extensions ont été proposées :

- logiques multivaluées,
- logiques modales,
- logiques temporelles,
- logiques déontiques,
- logiques épistémiques.

Pour une étude plus approfondie on pourra consulter [SABA-1988].

Une autre forme de représentation appelée « réseaux sémantiques » a été proposée, son principe consiste à représenter la connaissance sous la forme d'un graphe (ou réseau) de concepts. Les nœuds représentent les concepts et les arcs les relations entre ces concepts. Plusieurs types de relations entre concepts existent comme : EST-UN, SORTE-DE, EST-PARTIE-DE, etc.

L'utilisation de ces concepts passe par des outils de navigation dans le graphe, afin de comprendre le sens de la phrase et les relations entre les différents mots qui la constituent.

Les réseaux sémantiques ont été étendus pour améliorer la représentation et l'inférence des connaissances. Parmi ces extensions, on trouve par exemple les graphes conceptuels de Sowa dont la présentation déborderait du cadre de cette thèse.

2.4. Analyse pragmatique

« La pragmatique concerne l'étude de l'environnement d'une phrase, au moment où elle est émise ; elle découle de l'idée qu'une phrase (un énoncé) ne peut prendre tout son sens que si on la (le) replace dans son milieu d'origine ; c'est la prise en compte de toutes les conditions de production d'une phrase, tant il est vrai qu'un acte linguistique effectif ne peut avoir lieu qu'à l'intérieur d'une certaine situation de communication. » [JAYE-1985 : p.69]

Ce niveau d'analyse recouvre tout ce qui est lié à l'implicite dans la communication. C'est donc le niveau qui pose le plus de problème à concevoir et par conséquent il est beaucoup plus complexe à établir, ce qui explique qu'il n'existe que peu de réalisation opérationnelle, et qui ne concerne que des applications limites. On est donc encore loin de savoir construire des analyseurs pragmatiques pour le TALN.

A propos des connaissances manipulées par ce niveau d'analyse, [PITR-1985] écrit :

« Les connaissances pragmatiques portent sur le domaine traité dans les textes proposés au programme. Il est bien connu qu'il ne suffit pas de connaître la grammaire et le sens des mots pour comprendre un texte. » [PITR-1985 : p.9]

3. Construction d'un système de TALN

La construction d'un système pour le TALN fait intervenir les trois étapes suivantes :

a) explicitation des connaissances externes (ou des connaissances du domaine). L'objet de l'étude étant la langue naturelle, donc les connaissances relatives à la langue doivent être explicitées, on parle de connaissances morphologiques, syntaxiques, sémantiques et pragmatiques. On fait remarquer que ces connaissances externes sont directement liées aux domaines d'applications considérés.

b) explicitation des représentations internes (système symbolique qui permettra de représenter les connaissances externes à l'intérieur de la machine). Il s'agit de déterminer la représentation la plus appropriée (grammaire formelle, logique des prédicats, graphe conceptuel…) pour la représentation et par conséquent la manipulations des connaissances externes.

c) explicitation des programmes (algorithmes) permettant d'exploiter la représentation interne en vu du traitement automatique de la langue. [SABA-1989]

Dans cette optique, pour élaborer avec succès un système de TALN, on doit tenir compte des points suivants :

- ✓ Tenir compte de la modélisation linguistique : Il faut veiller à ce que le modèle soit général pour couvrir la totalité des phénomènes de la langue. Quelques violations du modèle sont inévitables (vu la complexité de la langue), elles seront compensées par des procédures ad hoc.
- ✓ Choisir une organisation permettant un meilleur équilibre entre grammaire et lexique.
- ✓ Structurer le ou les lexiques (le contenu doit être cohérent, non redondant et homogène).
- ✓ Gérer en amont les ambiguïtés lexicales (si possible). Autrement dit, ne pas laisser le traitement de l'ambiguïté à un niveau supérieur d'analyse (syntaxique, sémantique) si on peut le faire au niveau morphologique.
- ✓ Utiliser l'approche mixte[3] (méthodes statiques et méthodes linguistiques) pour résoudre les ambiguïtés (c'est l'approche la plus prometteuse pour le TALN).
- ✓ Optimiser les programmes d'analyse morphologique.

Les attentes les plus prégnantes que l'on puisse formuler à l'égard de tout système de TALN sont:

- ➢ Grandeur réelle : les ressources linguistiques (lexiques et grammaires) sont à large couverture de la langue.

- ➢ Performance : Donner des analyses en nombre limité (minimiser le nombre de solutions possibles) tout en conservant des durées de traitement raisonnables (le temps de réponse du système doit être convenable pour ne pas pénaliser l'activité de l'utilisateur).

- ➢ Robustesse : le système est capable de traiter des textes non voyellés[4], partiellement voyellés et complètement voyellés ou des formes non-attendues

[3] Voir chapitre suivant, la section : Approches pour résoudre l'ambiguïté
[4] Voir chapitre suivant, la section : les différents types de textes.

4. Conclusion

Après avoir situé notre étude dans la chaîne du traitement automatique d'une langue (à savoir le niveau morphologique), nous avons dégagé les principales étapes et recommandations pour élaborer avec succès un système pour le TALN. Les attentes les plus prégnantes d'un tel système étant : la grandeur réelle, la performance et la robustesse. Dans le chapitre suivant nous commençons par l'explicitation des connaissances externes (connaissances liées à la langue arabe). Dans cette perspective nous présentons un modèle pour le mot graphique arabe et nous discutons les problèmes du traitement automatique de ce dernier.

Chapitre 2

Mot graphique et traitement automatique de l'arabe

Chapitre 2

Mot graphique et traitement automatique de l'arabe

1. Introduction[1]

La langue arabe est une langue sémitique, elle s'écrit et se lit de droite à gauche. L'alphabet arabe comporte :

- 29 consonnes
- 11 voyelles dont
    - Trois voyelles brèves (voyelles courtes)
    - Trois semi-voyelles (voyelles longues)
    - Le 'sukun' (signe de quiescence ou absence de désinence): Se place au-dessus de la consonne et marque l'absence de voyelle. Ce signe n'affecte jamais la première consonne d'un mot.
    - Le 'tanwin' (trois signes): Se place toujours sur la dernière lettre, il est associé à une voyelle brève. La voyelle se prononce avec un n à la fin du mot. Le 'tanwin' présente le signe d'indétermination.
    - Le 'shadda' (signe de gémination de la consonne) : Se place sur la lettre et marque le redoublement de la consonne. Le 'shadda' est toujours associé à une voyelle brève ou un signe de 'tanwin'. Ce signe n'affecte jamais la première consonne d'un mot. Il correspond au dédoublement de lettres en français.

NB : Si on considère que la co-présence du signe de gémination 'shadda' avec une voyelle brève ou un signe de 'tanwin' comme une voyelle combinant les deux, alors le nombre de voyelles sera 16.

[1] Nous adoptons la norme AFNOR NF ISO 233-2 pour la translittération des caractères arabes en caractères latins ; Z46-002 AFNOR ; Association Française de Normalisation Décembre 1993. [AFNO-1993]

Les lettres changent de formes de présentation selon leur position dans le mot (i.e. : début, milieu, fin), et toutes les lettres se lient entre elles sauf quelques lettres qui font exception (ex. : waw, rae, dal) et ne se joignent pas à gauche. Un mot est une suite de consonnes et de voyelles. Les voyelles brèves presque inexistantes dans les textes (sauf dans certains documents scolaires et le coran) sont écrites en dessus ou en dessous des consonnes. La lecture d'un texte arabe non voyellé exige une bonne connaissance de la langue de la part du lecteur, car sans voyellation plusieurs analyses peuvent être attribuées à un même mot. Concernant ce problème de voyellation [HASS-1987] écrit :

« *… En fait, lire de l'arabe non voyellé exige du lecteur qu'il comprenne les mots avant de les lire. Car il faut comprendre le mot et le connaître pour lui attribuer les voyelles manquantes et il faut définir sa place grammaticale dans la phrase pour lui attribuer les voyelles finales flexionnelles convenables. A la limite, un mot non connu ne peut être lu correctement s'il n'est pas voyellé.* » [HASS-1987 : p.11]

[ALAA-1989] considère le processus de lecture d'un texte arabe non voyellé comme étant un processus non déterministe.

« Sans les voyelles brèves, on peut lire un texte en saisissant le sens d'après le contexte. Cette manière de compréhension est simulée par le mécanisme non-déterministe dans le sens du retour en arrière, et/ou le regard en avant… » [ALAA-1989]

Donc raisonner sur un texte non voyellé implique une perte d'informations très importante, [GAUB-2001] décrit cette situation en écrivant

« *… tout se passe comme si l'on projetait un signifiant non (ou peu) ambigu, les mots voyellés, dans un système intrinsèquement ambigu, l'alphabet privé des voyelles brèves.* »[GAUB-2001 : p50]

Dans sa thèse, [ACHO-1998] compare le problème de voyellation de l'arabe avec celui de l'accentuation du français, il conclut :

« Si l'on estimait la complexité de la tâche de voyellation par le nombre de voyellations qu'il est possible d'associer en moyenne à tout mot d'un texte non voyellé (11,5) et celle de l'accentuation par le nombre d'accentuation qu'il est possible d'associer en moyenne à tout mot d'un texte non accentué (1.3), alors on pourrait dire que la tâche de voyellation est à peu près neuf fois plus complexe que la tâche d'accentuation. » [ACHO-1998 : p37]

En plus de l'absence totale de voyellation, l'écriture arabe a la particularité de l'agglutination des mots à l'intérieur d'une phrase. En effet les particules (articles,

pronoms, prépositions…) s'écrivent attachées sous forme de préfixes et/ou suffixes aux noms et aux verbes.

2. Les constituants du mot graphique arabe

2.1 Le mot graphique[2]

« *Le mot graphique est facile à identifier : c'est ce qui s'écrit en un seul bloc entre deux blancs.* » [KOUL-94 : P.55]

Rappelons ici que nous ne partons pas de rien, mais la description du mot graphique Arabe déjà existante dans la littérature (voir : [COHE-70], [DICH-90], [HASS-87]) va nous servir comme point de départ. Par mot graphique (MG) on entend toute séquence graphique de caractères, séparée par des délimiteurs comme le blanc ou les signes de ponctuation. A l'inverse de la structure simple du MG d'une langue comme le français ou l'anglais, le MG arabe (MGA) possède une structure complexe et demande une modélisation plus riche pour sa prise en charge par un système de TALN. Un MGA peut être soit simple, soit complexe. Un MGA simple est un mot attesté de la langue, il est formé par la concaténation d'une base avec d'éventuels affixes (préfixes et suffixes). Il ne constitue pas un mot attestable de la langue sans les affixes.

MGA simple[3] = Préfixes + Base + Suffixes

Un MGA complexe est formé par la concaténation d'un mot simple et un ensemble de clitiques (proclitiques et enclitiques) [DICH-90], [HASS-87].

MGA complexe[4] = Proclitiques # mot simple # Enclitiques

MGA complexe = Proclitiques # Préfixes + Base + Suffixes # Enclitiques

Ou : MGA complexe = Prébases + Base + Postbases

Avec : Prébases=Proclitiques # Préfixes ; Postbases=Suffixes # Enclitiques

[2] C'est toute séquence graphique de caractères séparée par des signes séparateurs. En traitement automatique de langue naturelle on parle de « forme graphique », en documentation on parle « d'uniterme ».

[3] Dans la littérature un mot simple est appelé aussi RADICAL.

[4] Dans la littérature un mot complexe est appelé aussi HYPERMOT.

La concaténation notée par le symbole « # » dans l'expression précédente, exprime une liaison faible, autrement dit, un MGA complexe attestable de la langue peut se réaliser sans les proclitiques (respectivement les enclitiques).

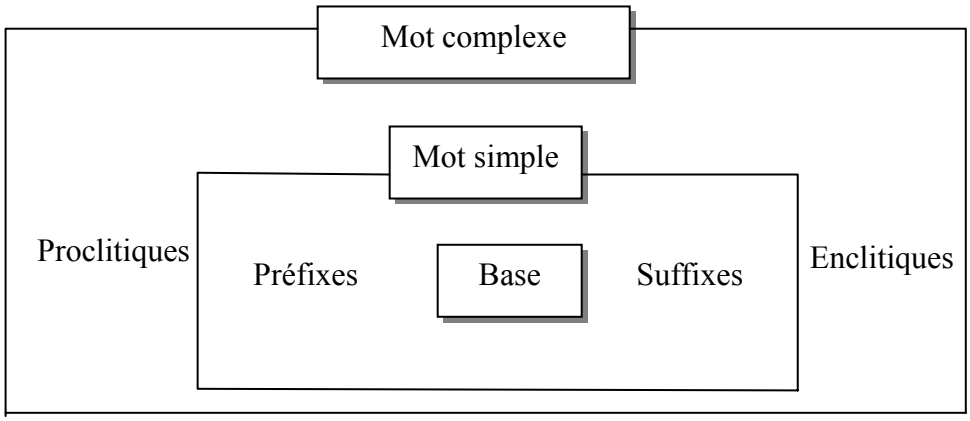


Figure 2 : les constituants du mot graphique arabe

Exemple : La forme graphique **'فسيشكرونهم'**

Postbases			*Prébases*	
Enclitiques	*Suffixes*	*Base*	*Préfixe*	*Proclitiques*
هم	ون	شكر	ي	فَس

2.2. Les clitiques

Les unités lexicales qui ne comportent qu'une consonne et une voyelle brève, comme par exemple fa (ف), wa (و), ka (ك), li (ل), ne peuvent pas être écrites isolément entre deux blancs et sont donc rattachées à la séquence qui suit, pour ne former avec elle qu'un seul mot. Ces composants attachés en début ou en fin de mot sont appelés des clitiques ou des enclinomènes.

En arabe, les conjonctions, les prépositions et l'article collent aux mots en position avant (proclitique), les pronoms collent à la fin du mot (enclitique).

Exemple : وَكتابهُ → | هُ | + | كتاب | + | وَ |

Enclitique = هُ

Proclitique = وَ

2.3 La base

Un mot simple sans ses affixes ne constitue pas un mot attestable de la langue mais un composant stable du mot que l'on appelle sa base.

Exemple

maktab → ktab مكتب → كتب

Une base présente une propriété essentielle qui est celle d'être analysable en deux composantes distinctes. La première constituée uniquement de consonnes, ordonnées de façon stricte, s'appelle la racine du mot. La seconde, constituée de voyelles longues ou brèves et parfois aussi de consonnes, s'appelle le schème. [COHE-70] confirme cela en écrivant : « Toute base d'une forme linguistique, c'est-à-dire toute forme linguistique dépourvue de ses éléments flexionnels, peut s'analyser fondamentalement en une racine et un schème. » [COHE-70 : p.50]
Par forme linguistique [COHE-70] désigne l'unité significative minimale autonome. Cette unité ne coïncide pas parfaitement avec le mot dans le sens de « mot graphique ». Par mot graphique [COHE-70] entend tout segment graphique délimité par deux blancs successifs.
Les mots qui ont même racine présentent en général une parenté sémantique plus ou moins marquée, et ceux qui ont même schème appartiennent en principe à la même classe des mots et subissent normalement les mêmes règles morphologiques (déclinaisons, conjonctions, etc.).

Il faut noter toutefois que la majorité des mots outils grammaticaux (conjonctions, prépositions, pronoms, démonstratifs) et la plupart des noms d'origine étrangère ne sont pas analysables en racine et schème. [KOUL-94]

2.3.1 La racine

« *La racine est une entité abstraite, que l'on ne rencontre jamais comme telle, mais que l'on déduit de l'analyse morphologique de familles de mots apparentés. Elle est toujours constituée exclusivement de consonnes, et est donc, en soi, imprononçable tant qu'elle n'est pas "coulée" dans un schème.* » [KOUL-94 : P.60]

Dans les dictionnaires arabes, tous les mots ayant la même racine sont groupés ensemble et la recherche se fait dans l'ordre alphabétique des racines.

D'après [KOUL-94], une analyse statique d'un dictionnaire de la langue moderne de 50000 mots montre qu'il y a 6500 racines et que 45% de ces dernières ne produisent que des notions nominales, ce qui implique qu'en moyenne chaque racine est capable de générer plus de six mots environ sachant que les racines exclusivement nominales ne génèrent qu'un ou deux mots.
Par ailleurs ces mêmes statistiques montrent que l'essentiel du vocabulaire fréquent et utile de la langue moderne est généré à partir d'un nombre réduit de racines (un millier au maximum).

En pratique, deux ou trois cents racines fournissent l'essentiel du vocabulaire indispensable.

Les racines sont généralement triconsonantique (trilitère), c'est-à-dire formée d'une suite ordonnée de trois consonnes, théoriquement leur nombre est de 21952 (si on peut prendre plus d'une fois la même consonne, le nombre de groupements possibles de 28 consonnes est : 28*28*28 soit 21952).

Un comptage des racines effectué dans [BEN-98] conclu que sur un total de 4906 racines, il y a 4260 racines trilitères (86,8%) et 646 racines quadrilitères (13,2%) ;

En pratique le nombre effectif des racines triconsonantiques attesté est très inférieur. La raison à cela est que certaines racines théoriquement possibles ne sont pas actualisées dans la langue d'une part, et que certains groupements ne sont pas tolérés par la langue d'autre part.
On convient généralement de nommer R1 la première consonne d'une racine (on dit première radicale), R2 la seconde, et R3 la troisième.

On convient aussi d'appeler :

- Racine "redoublée" ou "sourde" toute racine triconsonantique dont R2 et R3 sont identiques.
- Racine "faible" ou "malade" toute racine triconsonantique qui contient une ou plusieurs radicales glides[5] ou hamza.

[5]Les glides sont les trois voyelles longues : wa (و) précédé de u (ضمة), ya (ي) précédé de i (كسرة), et ā (ا) précédé de a (فتحة).

- Racine "forte" ou saines" toute racine triconsonantique qui ne contient pas de radical faible

D'une façon générale il existe trois types de racines trilitères, les racines dites normales, les racines dites semi-normales et les racines dites anormales. Le tableau 1 (repris de [SARO-89]) résume ces différents types.

Classe	**Racine R1, R2, R3**	**Condition**	**Description**	**Exemple**
Semi-normale	Redouble	R2=R3	La seconde et la troisième consonne sont identiques	جر
	HAMZE	$\exists i \in \{1,2,3\}$/ Ri=E	L'un des trois consonnes est un HAMZA	قرأ
Anormale	Assimilée	$R1 \in \{W,Y\}$	La première consonne est W ou Y	وصل
	Concave	$R2 \in \{W,Y\}$	La seconde consonne est W ou Y	كون
	Défectueuse	$R3 \in \{W,Y\}$	La troisième consonne est W ou Y	دعو
	Repliée Groupée	Concave et défectueuse	Racine à la fois concave et défectueuse	شوى
	Repliée séparée	Assimilée et défectueuse	Racine à la fois assimilée et défectueuse	وقى
Normale	Saine	$\forall i \in \{1,2,3\}$/ Ri ≠W, Ri ≠Y, Ri ≠E, et R2≠R3	Racine n'appartenant ni à la classe semi-normale, ni à la classe anormale.	كتب

Tableau 1 : Classification des racines trilitères [SARO-89]

Tous les dictionnaires de la langue arabe sont classés par ordre alphabétique (de la première ou de la dernière consonne) des racines et non des mots, comme c'est le cas pour plusieurs langues tel que le français par exemple. Donc pour retrouver un mot dans le dictionnaire, il faut tout d'abord rechercher la racine de ce dernier.
Par ailleurs pour former un mot arabe, il suffit de couler sa racine dans le schème approprié.

2.3.2 Le schème

Tous les noms et tous les verbes arabes peuvent se regrouper en un certain nombre de schèmes, séquences formelles de voyelles et de consonnes caractérisant des classes de mots.

Les consonnes que l'on peut utiliser pour la formation des schèmes sont au nombre de dix, et sont regroupées dans le mot " سألتمونيها " (vous me l'avez demandée). En dehors de ces dix, toute consonne qui apparaît dans un mot minimal est nécessairement radicale.

On distingue deux types de schèmes :

- Les schèmes verbaux : environ une soixantaine, à partir desquels par préfixation et / ou suffixation sont formés tous les paradigmes de conjugaison.
- Les schèmes nominaux : sont de plusieurs centaines.

Exemple de schème : darasa (درس) → R1aR2aR3a

Une autre façon de représenter un schème consiste à remplacer R1 par f, R2 par c et R3 par l. Cette façon de faire a été adaptée par les grammairiens arabes.

Exemple : darasa (درس) → facala

La relation entre racine et schème ne reflète pas fidèlement la réalité de la langue arabe, d'ailleurs [AMMA-1999] souligne ce fait « … *cette vision idéaliste de la relation entre racine et schèmes, très utile pour apprendre l'arabe et comprendre –tout particulièrement – son système verbales et déverbales, ne correspond que partiellement à la réalité…* » [AMMA-1999 ; p10].

Par ailleurs, la notion de base et de schème sont relativement identiques. Les bases sont des entités concrètes, issues du découpage d'un mot réel, par contre, les schèmes sont des entités abstraites issues de la généralisation de l'analyse en bases. Cette généralisation consiste à remplacer toutes les radicales concrètes par des symboles généraux (R1, R2, R3 ou f, c, l فعل . [KOUL-94 : P.68])

2.4. Flexion

Une flexion est une marque purement casuelle sans relation avec les classes de genre et de nombre, comme dans les langues indo-européennes.

Les flexions du masculin et du féminin, du singulier et du pluriel sont identiques.

Les mots en arabe sont soit à flexion معرب (la désinence varie selon la position du mot dans la phrase, exemple : verbes à l'inaccompli), soit construits مبني (la désinence est inchangée quelque soit la position du mot dans la phrase, exemple : verbes à l'accompli et les mots outils tel que les prépositions et conjonctions). [SARO-89]

On distingue trois cas de flexion pour la conjugaison du verbe et pour la déclinaison du nom. Pour les verbes, ce sont l'indicatif, le subjonctif et l'apocopé auxquels on associe respectivement les voyelles "u" ضمة, "a" فتحة, "o" سكون.

Pour les noms, ce sont le nominatif, l'accusatif et le génitif auxquels on associe respectivement les voyelles "u" ضمة, "a" فتحة, "i" كسرة. [SARO-89]

La conjugaison des verbes comprend l'accompli avec un seul mode, qui exprime une action déjà réalisée et l'inaccompli avec trois modes : l'indicatif (présent et futur), le subjonctif pour les subordinations et l'apocopé pour les phrases impératives.

Il existe trois nombres : singulier, duel et pluriel et deux genres : masculin et féminin ce qui donne dix huit formes possibles pour les trois personnes, mais seulement treize formes sont distinctes. Le tableau suivant montre ces formes.

		3ème personne	2ème personne	1ère personne
Singulier	Masculin	8	3	1
	Féminin	9	4	
Duel	Masculin	10	5	2
	Féminin	11		
Pluriel	Masculin	12	6	
	Féminin	13	7	

On remarque qu'il n'y a pas de distinction entre :

- le masculin et le féminin à la 1ère personne,
- le duel et le pluriel de la 1ère personne,
- le masculin et féminin à la 2ème personne au duel.

Par ailleurs, un verbe est désigné par la troisième personne du masculin singulier à l'accompli.

3. Les différents types de textes arabes

On distingue quatre types de textes :

1. Texte non voyellé
2. Texte non voyellé avec 'Shadda'
3. Texte partiellement voyellé
4. Texte complètement voyellé

3.1 Texte non voyellé

C'est un texte formé seulement de consonnes et de voyelles longues[6] (autrement dit, c'est un texte qui ne comporte pas de : voyelles brèves, 'Sukun', 'tanwin', et 'Shadda'). Ce type de texte concerne la quasi-totalité des documents produits en langue arabe (documents administratifs, journaux, romans…).

Exemple : Texte non voyellé

> اعلم أن الكلام هو الذي يعطي العلوم منازلها، ويبين مراتبها، ويكشف عن صورها، ويجني صنوف ثمرها، ويدل على سرائرها، ويبرز مكنون ضمائرها.
>
> أسرار البلاغة للجرجانى

3.2 Texte non voyellé avec 'Shadda'

C'est un texte non voyellé mais qui incorpore aux mots seulement le signe de gémination 'Shadda' (autrement dit, c'est un texte qui ne comporte pas de : voyelles brèves, 'Sukun', et 'tanwin'). Ce type texte se trouve dans certains livres scolaires.

Exemple : Texte non voyellé avec 'Shadda'

> اعلم أنّ الكلام هو الّذي يعطي العلوم منازلها، ويبيّن مراتبها، ويكشف عن صورها، ويجني صنوف ثمرها، ويدلّ على سرائرها، ويبرز مكنون ضمائرها.
>
> أسرار البلاغة للجرجاني

3.3 Texte partiellement voyellé (semi-voyellé)

C'est un texte qui ne comporte que quelques signes vocaliques. En général, ces signes permettent d'éviter des confusions de lecture. Parmi ces signes on trouve le signe de gémination ('shadda') et la voyelle finale (pour préciser la flexion casuelle). Ce type texte se trouve dans certains livres scolaires.

[6] Les voyelles longues sont considérées comme des consonnes, donc ils appartiennent à tous les types de texte.

Exemple : Texte partiellement voyellé

اعلم أنّ الكلام هو الّذي يعطي العلوم منازلها، ويبيّن مراتبها، ويكشِفُ عن صورها، ويجني صنوفَ ثَمرها، ويدل على سرائرها، ويبرز مكنون ضمائرها.

أسرار البلاغة للجرجاني

3.4 Texte complètement voyellé

C'est un texte qui incorpore aux mots leurs signes vocaliques complets (voyelles brèves, 'sukun', 'tanwin', et 'shadda'). Ce type de texte ne concerne en général que les livres destinés à l'apprentissage de la langue arabe ou le Coran.

Exemple : Texte complètement voyellé

اِعلَم أن الكَلاَم هو الذِي يعطِي العلوم منَازِلَها، ويبين مراتِبها، ويكْشِفُ عن صورها، ويجنِي صنُوفَ ثَمرها، ويدل علَى سرائِرها، ويبرز مكْنُون ضمائِرها.

أسرار البلاغة للجرجاني

4. Le problème de la voyellation

Par convention la voyellation d'un mot représente l'ensemble des voyelles associées aux consonnes de ce mot. Un mot est dit ambigu s'il admet plusieurs voyellations potentielles hors-contexte. Par exemple la forme graphique : (أهل) admet cinq interprétations possibles (voir le tableau 2). Un texte arabe complètement voyellé est donc un texte non ambigu, par contre, un texte non voyellé est un texte ambigu.

Forme 1 (forme nominale : les proches, digne)	أَهل
Forme 2 (forme verbale : marier)	أَهل
Forme 3 (forme verbale : qualifier)	أَهّل
Forme 4 (forme verbale : apparaître)	أَهلّ
Forme 5 (mot outil : est-ce que est-ce que)	أَهلْ

Tableau 2 : Voyellations potentielles pour la forme : (أهل)

Pour mettre en évidence le problème de voyellation de la langue arabe, [ACHO-1998] a introduit deux types d'ambiguïtés (ambiguïté en définition/ambiguïté en usage) et a procédé à leur évaluation.

4.1 Ambiguïté en définition

C'est le taux de mots ambigus calculé à partir d'un dictionnaire de la langue (un dictionnaire de la langue contient plusieurs entrées, et une entrée correspond à un mot attesté de la langue. Pour l'arabe, une entrée est un mot simple[7] non voyellé).

4.2 Ambiguïté en usage

C'est le taux de mots ambigus calculé à partir d'un corpus[8].

L'étude de l'ambiguïté en définition effectuée par [ACHO-1998] sur un dictionnaire[9] de formes fléchies (mots simples) conclue que 56% des formes sont ambiguës. Cette même étude montre que l'ambiguïté en usage, calculée à partir d'un corpus (un texte arabe non voyellé de taille 22375 mots) est très importante (environ 90% des formes sont ambiguës). Ceci s'explique par le fait que l'agglutination des clitiques aux formes simples augmente considérablement l'ambiguïté d'un texte arabe. Il suffit pour s'en convaincre de remarquer que le nombre de voyellation hors-contexte d'un mot simple est toujours inférieur ou égal au nombre de voyellation hors-contexte de ce même mot simple agglutiné à un enclinomène. Ce qui confirme que l'agglutination des enclinomènes aux mots simples augmente l'ambiguïté d'un texte arabe.
La figure 2 montre l'ambiguïté vocalique en fonction du type de texte.

[7] Mot simple = préfixes + base + suffixes.

[8] « *Un corpus est un ensemble de documents, artistiques ou non (textes, images, vidéos, etc.), regroupés dans une optique précise. On peut utiliser des corpus dans plusieurs domaines : études littéraires, linguistiques, scientifiques, etc.* » [http://fr.wikipedia.org/wiki/Corpus] Pour la création d'un corpus bien formé il faut prendre en compte plusieurs caractéristiques : la taille, le langage du corpus, le temps couvert par les textes du corpus et le registre.

[9] Voir la description détaillée de ce dictionnaire dans le chapitre suivant.

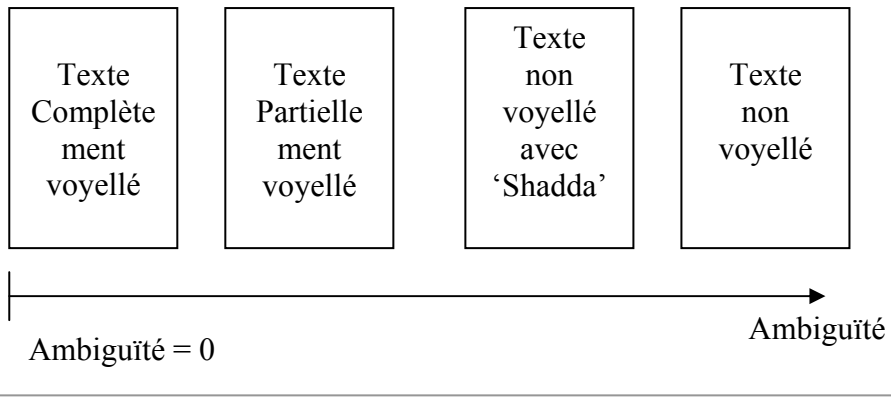


Figure 3 : L'ambiguïté vocalique d'un texte arabe

4.3 Approches pour résoudre l'ambiguïté

Pour résoudre l'ambiguïté morphologique lors de l'analyse d'un texte on utilise généralement trois approches :

1. Approche linguistique : Consiste à utiliser des critères et des propriétés linguistiques sous la forme des règles exprimant le fonctionnement de la langue naturelle utilisée. Par exemple l'utilisation des règles contextuelles comme la règle suivante : Y + […,Y,…] → Y + Y qui signifie : si nous avons une succession de deux formes f1(de catégorie Y) et f2 (qui a une catégorie parmi […,Y,…]), alors, la catégorie de f2 sera Y. Cette démarche est efficace, mais insuffisante car elle ne peut résoudre qu'un nombre limité de cas d'ambiguïté. [KALL-1987]
2. Approche statique : On calcule la fréquence d'occurrence pour chaque mot, et la fréquence d'occurrence des séquences de mots à partir d'un large corpus (estimé représentatif de la langue). Ces fréquences sont utilisées pour sélectionner l'analyse la plus probable pour un mot. Le modèle des chaînes de Markov est utilisé par exemple par [KALL-1987] pour la levée de l'ambiguïté dans le cadre de l'analyseur du français proposé par [LALL-1990]. Toujours d'après [LALL-1990 ; p.47] ce modèle a donné « d'excellents » résultats.
3. Approche mixte : Cette approche combine les deux autres approches, certainement elle est la meilleure, mais c'est la plus difficile à réaliser. La première approche (et par conséquent la dernière approche) est très difficile à réaliser car elle se base sur une description syntaxique de la langue que

personne ne prétend la faire exactement. S'il en est ainsi pour la syntaxe, alors le problème est plus grave et persiste encore pour la sémantique !

Dans le cadre de cette thèse nous ne abordons pas cette problématique, toutefois il nous semble que la réalisation des outils pour la désambiguïsation est nécessaire vu les résultats obtenus par notre analyseur morphologique pour la langue arabe (voir chapitre 5 de cette thèse).

5. La ponctuation

Etant d'un usage récent en arabe, la ponctuation ne s'applique qu'aux textes modernes. Il s'agit du groupe des poins, marquant la fin de phrase (point, point d'exclamation, point d'interrogation) et du groupe des virgules, marquant une pause dans la phrase (virgule, point-virgule, deux points).

6. Problèmes du traitement automatique de l'arabe

Nous pouvons résumer les problèmes liés au traitement automatique de la langue arabe dans les points suivants :

- la voyellation multiple,
- la structure complexe du mot graphique arabe (phénomènes d'agglutinations qui caractérisent la langue ; un mot graphique arabe peut correspondre à une phrase française).
- L'ordre des mots est relativement libre dans une phrase arabe (Verbe + Sujet + Complément ; Verbe + Complément + Sujet ; Complément + Verbe + Sujet).
- Traitement des cas particuliers : traitement de la 'hamza'', traitement de la 'Shedda', traitement de l'altération de la forme du mot.
- Traitement des racines analogues ne donnant pas lieu à des dérivations analogues.
- Traitement de la racine d'un mot issu d'une racine anormale.
- Traitement des mots homographes (une même chaîne de caractère qui suivant le contexte recouvre deux notions différentes) Exemple : le mot **في** : verbe impératif ou préposition.

7. Conclusion

Dans ce chapitre, nous avons explicité les différentes connaissances liées au traitement automatique à la langue arabe. Comme nous l'avons déjà laissé entendre précédemment, le mot graphique arabe représente l'entité de base sur lequel nous allons construire notre système pour le TALN. Dans ce contexte, nous avons essayé de dégager les problèmes qui peuvent entraver la construction d'un système de TALN arabe. Dans le chapitre suivant nous présentons les différentes approches et organisations pour le TALN arabe. La solution adoptée est motivée suite à une analyse détaillée de tous les systèmes existants.

Chapitre 3

Les analyseurs existants et le choix d'une organisation

Chapitre 3

Les analyseurs existants et le choix d'une organisation

1. Introduction

L'analyse d'un texte écrit en langue naturelle se construit sur la base d'une architecture à deux composantes : le dictionnaire et la grammaire. Deux solutions (approches) se dégagent selon que l'on essaie de simplifier la grammaire (réduire la complexité des traitements) ou de minimiser la taille du dictionnaire (réduire la place mémoire) (voir tableau : 3). La première solution consiste à construire un dictionnaire comportant toutes les formes fléchies des mots. L'analyse se résume donc à un simple accès au dictionnaire, ce qui présente l'avantage de réduire l'effort déployé pour l'élaboration d'une grammaire d'analyse.

La deuxième approche présente l'avantage du gain en espace mémoire lié à la construction d'un dictionnaire ne contenant que le strict minimum de formes (une liste des préfixes, des suffixes, des bases, et d'autres éléments). Par ailleurs l'élaboration d'une grammaire (assez importante) permettant de valider ou d'invalider les décompositions possibles d'une forme est inévitable, ce qui présente un inconvénient que l'on reproche à cette démarche.

Pour conclure, on peut dire que les avantages de l'une sont les inconvénients de l'autre, et inversement.

	Approche I	**Approche II**
Minimiser la place mémoire	Non	Oui
Complexité de l'algorithme d'analyse	Non	Oui
Rapidité des traitements	Oui	Non
Rapidité d'accès au lexique	Non	Oui

Tableau 3 : Comparaison des deux approches pour l'analyse d'un texte

2. Les analyseurs existants

Dans cette partie nous passons en revue les principaux analyseurs pour la langue arabe proposés par différents auteurs. L'analyse de ces analyseurs va nous permettre de justifier notre choix concernant l'approche à adopter pour la construction de notre système pour le TALN arabe.

2.1 L'analyseur de [HASS-1987]

Il s'agit là disons-le tout de suite d'un analyseur morphologique pour des textes vocalisés. Pour ce faire [HASS-1987] utilise trois lexiques :

- Le lexique des exceptions : il regroupe les prépositions, les conjonctions, les pronoms, les noms propres, etc.
- Le lexique des racines : il regroupe les racines de l'ensemble des mots.
- Le lexique des vecteurs booléens : Il concerne seulement les formes verbales et nominales.

Les étapes essentielles de la méthode d'analyse se résument ainsi :

a) Partant d'un mot, l'analyseur cherche ce dernier dans le lexique des exceptions. En cas de succès, l'analyseur fournira à ce mot sa catégorie morphologique.

b) En cas d'échec, l'analyseur détermine pour ce mot une chaîne de voyelles, et une chaîne de consonnes. De la chaîne de consonne, l'analyseur extrait la racine du mot et cherche s'il se trouve dans le lexique des racines. Si la racine est trouvée, l'analyseur va calculer le vecteur booléen du mot et cherche s'il se trouve dans le lexique des vecteurs booléens. Si le vecteur est trouvé, l'analyseur fournira au mot sa catégorie morphologique.

Dans le cadre de cette expérience effectuée sur plusieurs textes, [HASS-1987] conclu que les résultats obtenus sont satisfaisants ! Mais l'expérience n'a pas été quantifiée en terme de chiffres (taux de reconnaissance taux d'ambiguïté, etc.). Il faut noter toutefois que [HASS-1987] n'a pas donné la taille des différents lexiques utilisés dans l'expérimentation d'une part, d'autre part il n'a pas montré comment l'analyseur extrait la racine du mot à analyser, car c'est sur la base de cette procédure que dépend toute la méthode d'analyse à notre avis.

2.2 L'analyseur du projet SAMIA [HASS-1987]

SAMIA : **S**ynthèse et **A**nalyse **M**orphologiques **I**nformatisées de l'**A**rabe en vue d'une application en Enseignement Assisté par ordinateur (E.A.O.) : est un programme de recherche dirigé par J.E. BENCHEIKH à l'université de Paris8.

Le modèle linguistique du projet SAMIA, repose sur un vecteur descriptif du mot qui est composé :

- d'un ensemble d'éléments cités ci-dessous :

1- PCL3 : pour proclitique de 3ème ordre (par ordre on entend la position de l'élément dans la chaîne graphique. Les proclitiques sont agencées selon un ordre bien défini. Cet ordre de bonne formation structurelle du mot repose sur les propriétés distributionnelles des différents formants du mot, par exemple : un article n'est jamais suivi d'une base verbale)
2- PCL2 : pour proclitique de 2ème ordre
3- PCL1 : pour proclitique de 1er ordre
4- PRF : pour préfixe
5- ART : pour article
6- BVP : pour base verbale perfective
7- BVI : pour base verbale imperfective
8- BN : pour base nominale
9- SUF : pour suffixe
10- ECL : pour enclitique

- d'une matrice de compatibilité entre les formants ; appelée aussi règles de formation structurelle du mot. Elle a pour objet d'écarter les incompatibilités entre les formants d'un mot.

2.2.1 Le modèle d'analyse du projet SAMIA

Le modèle d'analyse proposé dans le cadre du projet SAMIA consiste à analyser des mots graphiques non vocalisés. La première étape de l'analyse consiste à faire une projection du mot sur un vecteur d'analyse composé de 32[1] cases (voir figure 5), ce qui va permettre de retrouver la place exacte de chaque lettre du mot dans ce vecteur. Cette opération va permettre d'identifier les différents formants du mot en question :

- ses proclitiques,
- son préfixe,
- ses lettres radicales,

[1] Le nombre 32 représente la taille maximale d'une forme graphique arabe voyellé.

- les éléments non-radicaux de la base,
- ses suffixes,
- ses enclitiques.

Une fois ces formants identifiés reste maintenant à construire la base du mot, et reconnaître une racine attestée.

Mot maximal (1-32)																						
PCLg						PRF	Base graphique							Suffixes				ENCL1			ENCL2	
1	2	3	4	5	6	7	8	9	10	11	..	17	..	22	23	..	26	27	28	29	30	31	32
										R1	..	R3	..										
						Mot minimal (7-26)																	

Figure 5 : Vecteur d'analyse du mot graphique [HASS-1987 : p.81]

La réalisation de la projection du mot graphique minimal sur le vecteur d'analyse ne pose pas de problème, par contre, il semble qu'il y ait des difficultés pour le mot graphique maximal, d'ailleurs [HASS-1987] écrit :

« *Pour un mot maximal cela nous a paru très difficile à réaliser compte tenu des nombreuses possibilités pour chaque lettre du mot, surtout si cette dernière est un élément non-radical de la base ou un clitique…* ». [HASS-1987 : p.81]

C'est dans ce contexte, qu'une nouvelle proposition s'est dégagée. Cette dernière consiste à décomposer le mot en 11 cases au lieu de 32 cases (voir figure 6). On ne parle plus de projection dans ce cas mais plutôt de segmentation du mot au maximum, en 11 éléments.

Une grammaire d'analyse dans ce cas doit s'assurer de:

1- la bonne succession des constituants,
2- la bonne compatibilité entre les constituants.

Parallèlement à cela, il ne faut pas oublier au moins 11 consultations du dictionnaire chaque fois pour attester les différents constituants et assurer leur compatibilité. Tout cela conduit à une analyse complexe.

PCL1	PCL2	PCL3	PCL4	PREF	BASE	SUF1	SUF2	SUF3	ECL1	ECL2

Figure 6 : Vecteur d'analyse du mot graphique [HASS-1987] : p.82

Pour palier cette complexité d'analyse, l'équipe SAMIA a proposé un nouveau vecteur d'analyse, ce dernier est composé de trois éléments au lieu de 11 éléments (voir figure 7).

1- pré-base : composée de proclitique et de préfixe,
2- Base,
3- Post-base : composée de suffixes et des enclitiques.

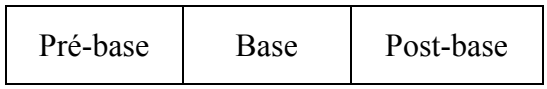


Figure 7 : Le nouveau vecteur d'analyse [HASS-1987 : p.84]

Ce nouveau vecteur d'analyse conduit à une segmentation du mot en trois formants et suppose donc l'existence de ces trois éléments dans le dictionnaire. De toute évidence, la grammaire d'analyse dans ce cas sera réduite au profit du dictionnaire, car beaucoup de règles de compatibilités et de succession des constituants du vecteur d'analyse seront éliminées (les compositions attestées des constituants sont enregistrées dans le dictionnaire).

Donc pour construire ce système d'analyse on doit :

1- Elaborer trois lexiques (le lexique des pré-bases, le lexique des bases, et le lexique des post-bases).
2- Elaborer une grammaire de compatibilité entre pré-base, base et post-base.

Par conséquent, l'analyse d'un mot graphique dans ce modèle consiste à :

1- Segmenter le mot en trois sous-chaînes.
2- Consulter le dictionnaire pour valider la segmentation.
3- Attester la compatibilité des sous-chaînes de la segmentation.

2.2.2 L'analyse dans le projet SAMIA

L'analyse du mot graphique consiste à associer à chaque mot graphique non-vocalisé, l'ensemble des segmentations possibles pour ce mot.

Les différents constituants du mot sont : pré-base, base, et poste-base.

Dans ce contexte, il a été question de construire des listes contenant ces éléments, leurs combinaisons et leurs compatibilités. La liste de pré-bases est formée à partir de la combinaison proclitiques-préfixe, et la liste de post-bases est formée à partir de la combinaison suffixes-enclitiques.

La construction automatique du vecteur-mot comporte deux étapes :

- Etape 1 : Extraction de la base du mot.

Elle consiste à :

- Identifier la partie pré-base, à partir du début du mot (la plus longue chaîne formée de proclitiques et de préfixe).
- Identifier la partie post-base, à partir de la fin du mot (la plus longue chaîne formée de suffixe et d'enclitiques).
- Comparer les résultats pré-base et post-base de sorte à éliminer les décompositions incompatibles.

- Etape 2 : Identifier la base.

  Elle consiste à rechercher la base trouvée dans l'étape 1 à partir d'une liste des bases (avec leurs racines) enregistrées dans un dictionnaire. Si la base est trouvée, il y a identification de la racine.

Le modèle d'analyse proposé par le projet SAMIA fournit, pour chaque mot graphique non vocalisé arabe, l'ensemble des informations :

1- Les affixes (proclitiques, préfixe, suffixes et enclitiques) accompagnés de leurs valeurs grammaticales ;
2- La racine ou le représentant du mot :
   - la racine du mot (si la base de celui-ci peut être rapportée à une racine arabe attestée),
   - le représentant graphique du mot dans le cas contraire ;
3- Le ou les formes graphiques vocalisées du mot ;
4- Pour chaque forme graphique vocalisée :
   - la catégorie morphologique de la forme minimale (la forme sans ses proclitiques et ses enclitiques),
   - des variables de type grammatical pour chaque forme, telles que le genre et le nombre, l'aspect et le mode, etc.
5- Des informations de type micro-syntaxique pour les mots outils en particulier. [HASS-1987 : p.86]

Le lexique qui a fait l'objet de l'étude de [HASS-87] est conçu sous la forme d'une base de données relationnelle. Les informations linguistiques sont structurées pour constituer un ensemble de traits linguistiques qui peuvent être attribués à toute forme rencontrée dans un texte en langue arabe. Ce lexique comprend :

1- les bases (sous forme graphique vocalisée et non vocalisé, et sous forme phonologique),

2- les proclitiques, les préfixes, les suffixes, et les enclitiques (sous forme graphique vocalisée et non vocalisé),
3- les combinaisons des proclitiques, des préfixes, des suffixes, et des enclitiques (sous forme graphique vocalisée et non vocalisé).
4- les autres constituants (constituants abstraits tel que la racine et le schème) et les représentants de certains mots non assimilés par le système dérivationnel (exemple : le représentant de ليبيا, ليبي est ليبي.

Les documents que nous possédons ne donnent aucune idée sur :

- l'implantation sur machine du projet (surtout l'analyseur morphologique) ;
- l'évaluation des performances de l'analyseur.

Le modèle d'analyse proposé par le projet SAMIA nous semble intéressant vu la simplicité de la procédure d'analyse. Cette simplicité est due principalement à l'élaboration d'un lexique très riche en informations permettant ainsi d'éviter beaucoup de calcul. Toutefois, toute la difficulté et par conséquent l'enjeu de ce modèle réside dans l'élaboration de ce lexique.

2.3 L'analyseur de [SARO-1989]

L'objectif des travaux de la thèse de [SARO-1989] est de mettre au point un modèle lexical de l'arabe écrit, qui comportera un lexique aussi exhaustif que possible, afin d'augmenter le nombre de phrases analysables. Cet objectif global se subdivise en deux sous objectifs intermédiaires :

1- Développer une base de données lexicale de l'arabe écrit pour l'analyse morpho-syntaxique. Cette base rassemble le maximum d'information sur chaque racine tel que le schème, le préfixe, le suffixe et les valeurs grammaticales. « *Cette base doit être suffisamment complète pour générer la majorité des mots arabes. Elle doit être complétée par les règles d'adaptation qui interviennent au cours de l'analyse et la synthèse des mots à morphologie irrégulière...* » [SARO-1989 ; p.33] .
2- Développer sous forme d'un système expert deux composantes à savoir : une composante morphologique et une composante syntaxique. La base de données représente la base des faits et la grammaire arabe représente la base de règles.

La base de données lexicale de [SARO-1989] présente une analogie avec BDLEX, une base de données développée pour le français dans le même laboratoire (IRIT-CERFIA[2]). Le modèle de données choisi étant le modèle relationnel.

L'approche préconisée dans l'analyseur de [SARO-1989] consiste à concevoir un lexique réduit (petits dictionnaires) et une grammaire importante. Le lexique est composé d'un :

- lexique des racines (2000 racines du dictionnaire AL-SABIL),
- lexique des schèmes,
- lexique des suffixes,
- lexique des préfixes,
- ensemble de règles dérivationnelles et flexionnelles permettant de générer environ 200000 formes fléchies.

Chaque racine pointe sur l'ensemble des règles de dérivation qui lui correspondent (on obtient l'ensemble des parentés sémantique de la racine), et chaque règle de dérivation pointe à son tour sur les règles de flexion (on obtient l'ensemble des formes fléchies d'un mot de base). Les liens de chaînage entre la racine d'une part, et les règles de dérivation et de flexion d'autre part, sont spécifiés par un expert.

2.3.1 L'Analyse morphologique

L'analyse morphologique commence par une segmentation préalable de chaque mot, ce qui permet de trouver les différentes unités morphologiques qui composent le mot (i.e: les préfixes, les préfixes, les suffixes, la base ou radical : analysable en racine et schème, et la désinence)

Exemple :

Soit le mot :

MAKOTABATUN مكتبة

MA	KOTAB	AT	UN
Préfixe	Base	Suffixe	Flexion

La base KOTAB est analysable en racine et schème :

Racine (KOTAB) = KTB

Schème (KOTAB)= R1OR2AR3 avec R1=K, R2=T, et R3=B

Par la suite de cette décomposition, on associe à chaque unité trouvée un ensemble de valeurs grammaticales hors contexte.

[2] Les activités de recherches du laboratoire IRIT-CERFIA sont focalisées sur l'étude de la syntaxe d'une phrase (simple ou complexe) arabe.

Exemple (analyse morphologique) [SARO-1989 : p.63]

Soit le mot : WAEALOMADOPAOATI والمدرسة

Ensemble des préfixes du mot = {WA, EALO, MA}

Ensemble des suffixes du mot = {AT, I}

Les valeurs grammaticales associées sont :

- WA : conjonction
- EALO : Déterminant
- MA : Préfixe faisant partie du schème
- AT : Marque du féminin (désinence)
- I : Flexion (génitif)
- (DRS, R1OR2AR3) : Base du mot (analysable en racine et schème)

Le modèle d'analyse morphologique tel qu'il a été décrit ne permet pas de comparer directement un mot à analyser avec les éléments de la base de données lexicale. Dans un premier temps, les mots doivent être décomposés en différentes unités lexicales (préfixe, suffixe, racine, schème). En fonction des unités lexicales qui composent le mot à analyser et la base de données lexicale, le générateur engendre un sous lexique par synthèse, qui va servir pour l'identification du mot à analyser par un processus de comparaison. La démarche d'analyse se fait selon le schéma de la figure 8.

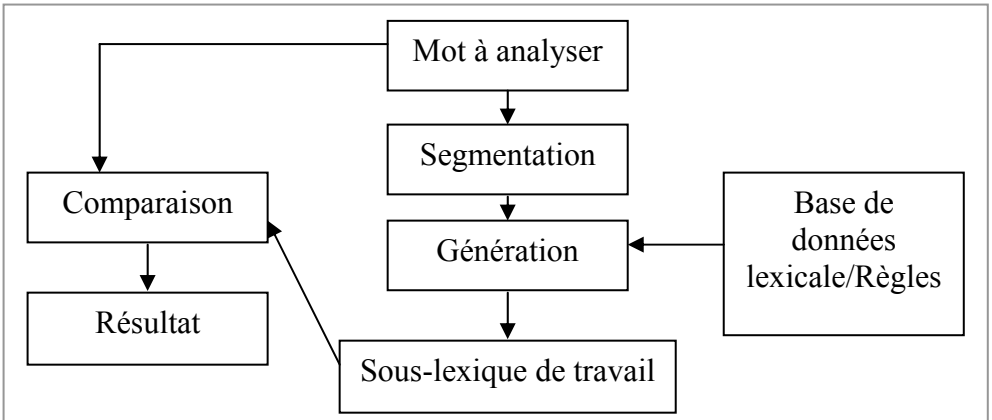


Figure 8 : Analyse par génération du dictionnaire [SARO-1989 : p.64]

Les différentes tâches à exécuter pour l'analyse d'un mot dans [SARO-1989] sont (voir figure 9) :

1- Extraction des préfixes du mot.
2- Extraction des suffixes du mot.
3- Synthèse des composants du mot (contrôle de compatibilité des différents segments).

4- Extraction de la racine et du schème de la base.
5- Localisation de la racine et du schème dans la base de données lexicale.
6- Calcul des valeurs grammaticales hors contexte des différentes unités lexicales du mot.

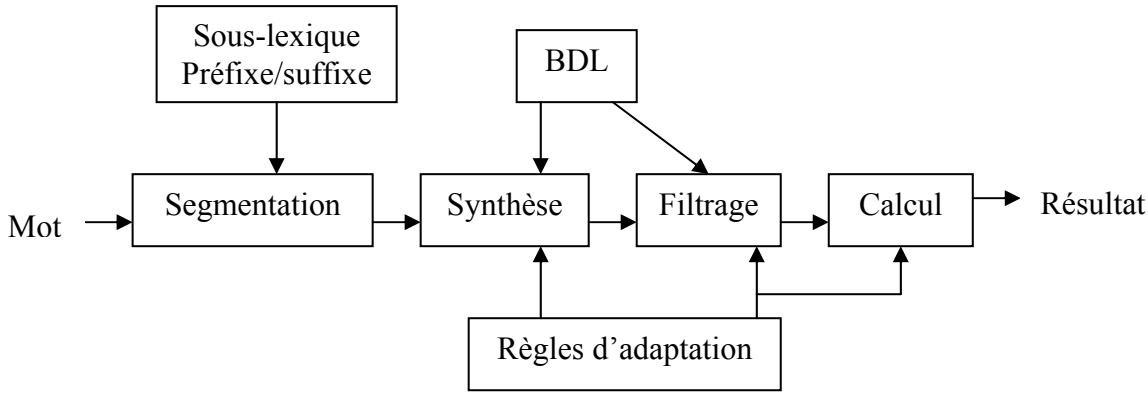


Figure 9 : Modèle d'analyse morphologique [SARO-1989 : p.66]

Le générateur morphologique de [SARO-1989] (voir figure 10) est défini par le triplet <LEX, RD, RFLX>, où LEX est l'ensemble des entrées lexicales (racines), RD représente l'ensemble de règles de dérivation et RFLX l'ensemble de règles de flexion. On parle de deux types de génération :

1- La génération lexicale minimale à partir d'une racine, qui produit n mots : M1, M2, …, Mn avec n ≥ 1, où chaque mot Mi est le résultat de l'opération de dérivation de la racine (r) par une règle de dérivation (Rdi). On peut noter cette opération de génération par le couple <r, RDi>.
2- La génération lexicale maximale à partir d'un mot Mi qui produit m mots : Mi1, Mi2,…, Mim avec m ≥ 2, où chaque mot fléchi Mij est le résultat de l'opération de flexion du mot Mi par une règle de flexion RFLXj. On peut noter cette génération par le couple <Mi, RFLXj>.

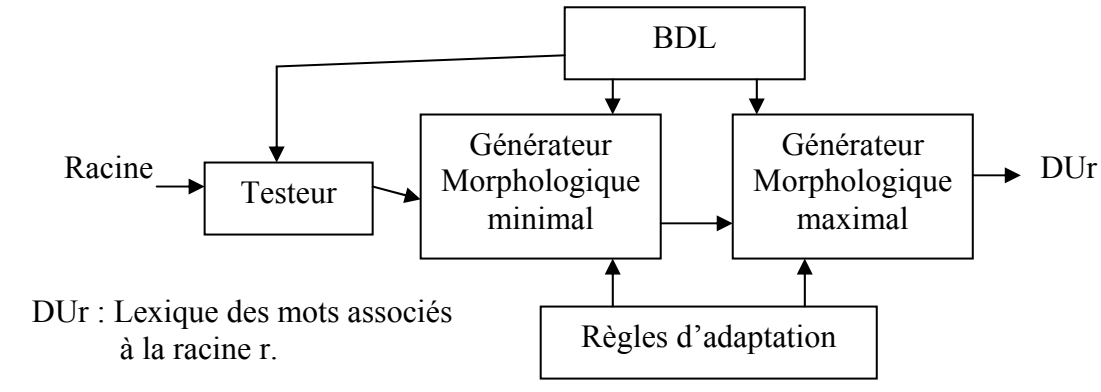


Figure 10 : Schéma fonctionnel du générateur [SARO-1989 : p.90]

Le testeur recherche la racine candidate dans le lexique « racine ». En cas de succès, il retourne, avec la racine, l'ensemble de ses attributs avec les références des règles de dérivation qui s'appliquent à elle. Par ailleurs, si le résultat de la génération (dérivation ou flexion) est un mot non attesté et en fonction de la catégorie de la racine, le générateur applique des règles d'adaptation (suppression/remplacement d'une chaîne de caractères) afin d'obtenir un mot attesté.

2.3.2 Le dictionnaire

Le rôle du générateur est donc la construction d'un dictionnaire de la langue arabe à partir de la base de données lexicale. Pour chaque mot non vocalisé, le dictionnaire contient l'ensemble des formes voyellées hors contexte ainsi que l'ensemble des valeurs grammaticales qui leur sont attachées.

Le dictionnaire de [SARO-1989] comporte les attributs suivants :

- MOT : représentation graphique du mot généré, exemple : كتب كتب
- CAT : catégorie syntaxique,
- PRE : préfixe du mot, exemple مكتب → م
- SUF : suffixe du mot
- V_MOR : valeurs morphologiques du mot (la personne, le nombre, le genre, etc.),
- CAS : cas casuel du mot (nominatif,…)

Le dictionnaire des racines contient 2000 entrées. A partir de ces racines et d'un ensemble de règles (dérivationnelles et flexionnelles), le système de génération morphologique de [SARO-1989] peut générer jusqu'à 200 000 formes fléchies (formes conjuguées pour les verbes, déclinaisons pour les substantifs et les adjectifs). Pour une application réelle, la taille de ce dictionnaire (200 000 formes) me semble insuffisante pour la simple raison que la langue arabe compte beaucoup plus de formes ([ATTI-2000] compte environ 6.10^{10}, et [OUER-2002] l'estime à environ 6 millions de formes). Par ailleurs, il est très important de souligner que le fait de générer à chaque analyse d'une forme un lexique spécifique cela entraîne une perte de temps considérable ce qui présente donc un inconvénient dans cette démarche d'analyse. A cause de ces deux derniers obstacles il me semble que la démarche de [SARO-1989] pour l'analyse n'est pas appropriée. Il faut noter ici que les documents que nous possédons ne donnent aucune idée sur l'implantation et l'évaluation des performances de cet analyseur !

2.4 L'analyseur morphologique de [ZOUA-1989]

Le lexique de [ZOUA-1989] constitué d'un ensemble de clitiques (proclitiques et enclitiques) et de formes fléchies réduit considérablement l'algorithme de segmentation d'une part, et la grammaire qui doit valider les segments décomposés d'autre part (l'auteur n'effectue pas la décomposition en préfixe, radical et suffixe).

L'analyseur procède en deux étapes (voir figure 11) :

Etape 1

- Segmentation du texte d'entrée en unités morphologiques.
- Analyse de ces unités morphologiques pour découvrir les entrées lexicales qui les composent.

Etape 2

Attribution à chacune des unités lexicales ainsi reconnues l'ensemble des informations linguistiques qui leur sont attachées.

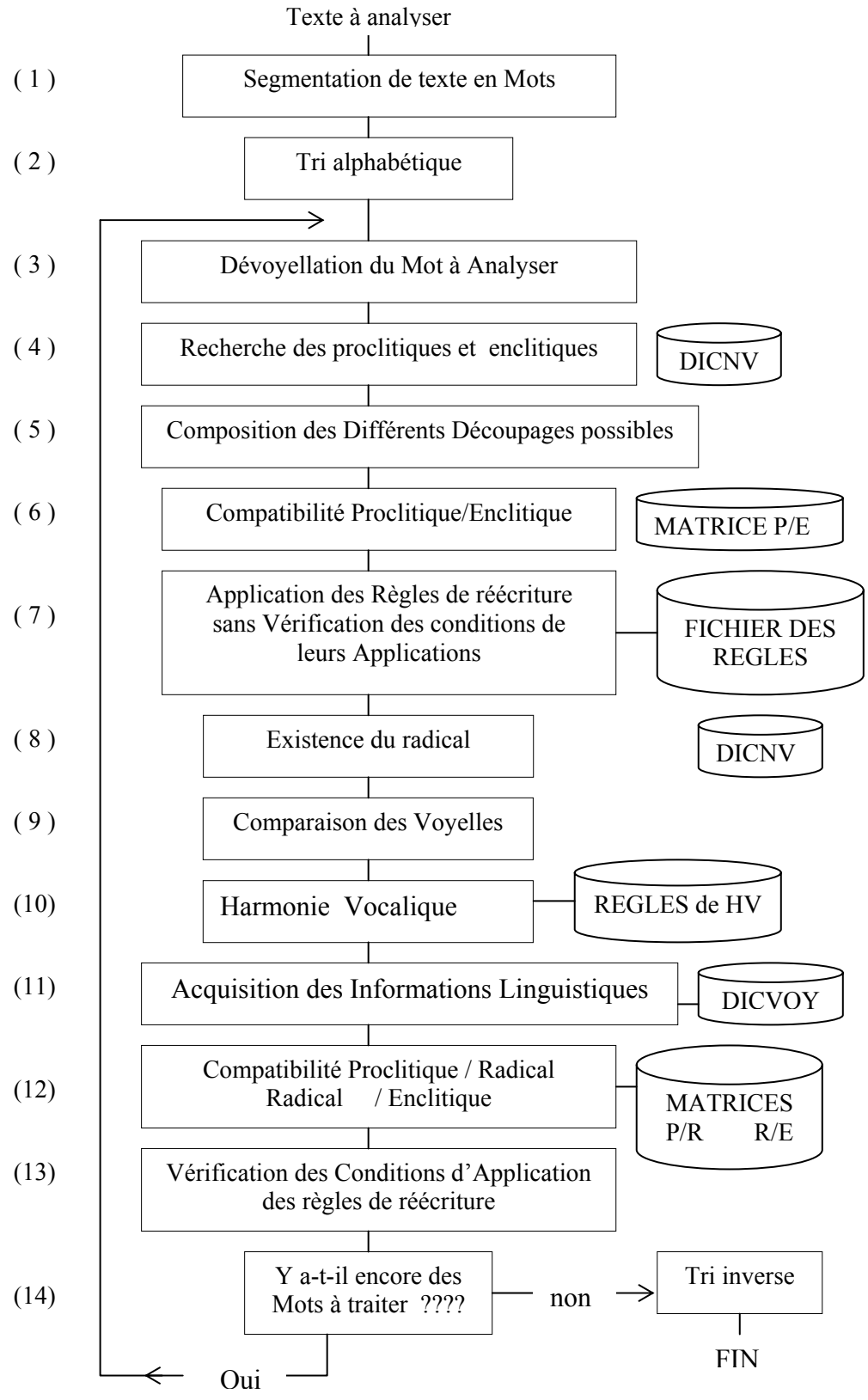


Figure 11 : Algorithme d'analyse morphologique [ZOUA-1989 : p66]

Avec : **DICNV** : Dictionnaire non voyellé ; **DICVOY** : Dictionnaire voyellé

MATRICE P/E : Matrice de compatibilité entre proclitiques et enclitiques.
REGLES de HV : Règles d'harmonie vocalique.
MATRICES P/R R/E : Il s'agit de deux matrices. Matrice de compatibilité entre proclitiques et radicaux / Matrice de compatibilité entre les radicaux et les enclitiques.

Le tri permet d'éviter de refaire plusieurs fois l'analyse d'un même mot, l'analyseur vérifie si le mot à analyser est identique ou précédent, si oui alors il recopie le résultat de l'analyse précédent. Cette procédure permet donc d'économiser les temps d'analyse.

2.4.1 Construction du dictionnaire

La méthode de construction du dictionnaire dans [ZOUA-1989] consiste à générer automatiquement toutes les formes dérivées à partir d'un dictionnaire de racines. Le dictionnaire ainsi construit devra fournir pour chaque mot non voyellé l'ensemble des formes voyellées qui lui correspond, et pour chaque forme voyellée l'ensemble des informations grammaticales hors-contexte (genre, nombre… etc). Ce dictionnaire présente un double intérêt, d'une part il permet de simplifier la procédure d'analyse (éviter de découper la forme analysée en préfixe, base et suffixe), de l'autre part il présente l'avantage d'être généré automatiquement.

Il faut noter toutefois qu'aucune indication de taille n'a été donnée pour ce dictionnaire.

2.4.2 Les résultats de l'analyseur

Dans le cadre de cette expérience effectuée sur *plusieurs mots* (vingt mots comme exemple d'analyse), [ZOUA-1989] conclu que les résultats obtenus sont satisfaisants ! Mais l'expérience n'a pas été évaluée.

2.5 L'analyseur morphologique de [BEN-1998]

Cet analyseur n'est qu'une version actualisée de l'analyseur de [ZOUA-1989]. Pour analyser une forme, l'analyseur morphologique de [BEN-1998] utilise deux dictionnaires (un dictionnaire des formes fléchies et un dictionnaire des enclinomènes) et un ensemble de règles. Cet analyseur peut analyser un texte complètement voyellé,

partiellement voyellé[3] ou non voyellé. La figure 12 présente le schéma général de cet analyseur.

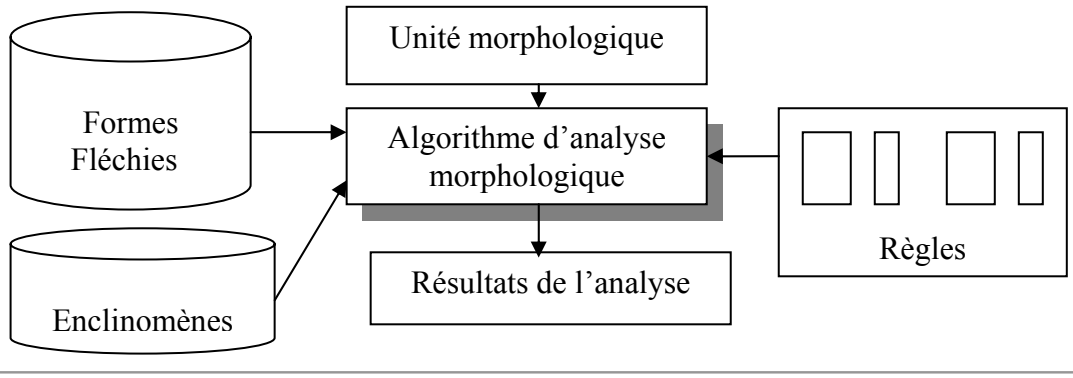


Figure 12: Schéma général de l'analyseur morphologique [BEN-1998 : p 79]

2.5.1 Le lexique de [BEN-1998]

Ce lexique comprend deux dictionnaires :

a) Un dictionnaire des formes fléchies (ce dictionnaire est généré d'une manière semi-automatique): Il comprend les formes fléchies voyellées avec un ensemble d'information linguistique qu'elles peuvent avoir. Une entrée du dictionnaire comprend deux zones. La première zone contient la forme fléchie non voyellée, par contre la deuxième zone contient un ensemble d'information :
   1) le schéma vocalique[4] : c'est l'ensemble des voyelles d'une forme.
   2) les catégories grammaticales (183 catégories)
   3) le genre (masculin/féminin) et nombre (singulier/duel/pluriel)
   4) la transitivité (ne vaut que pour les formes verbales, elle comprend les valeurs : transitif direct/transitif indirect/intransitif)
   5) le lemme : le lemme d'une forme verbale est composé des deux formes résultant de la conjugaison à l'accompli et à l'inaccompli du verbe au troisième personne du singulier. Pour les formes nominales, c'est la forme au masculin singulier ou la forme au féminin singulier. Les lemmes des particules sont les formes elles-mêmes.

   Ces champs figureront au compte de chaque forme non voyellée autant de fois que celle-ci admet de voyellation.

[3] Voir le chapitre 2 page 31 pour la définition de ce type de texte
[4] Chaque forme possède un schéma consonantique et un schéma vocalique.
Par exemple si la forme : F=kataba Alors le schéma consonantique (les consonnes de la forme F) = ktb, et le schéma vocalique (les voyelles de la forme F)= aaa.

Le nombre d'entrées de ce dictionnaire est de 599 361 mots non voyellés. Par contre le nombre de mots complètement voyellés de ce dictionnaire est de 1 622 802. En moyenne une forme non voyellée a 2.8 voyellations différentes possibles.

b) Un dictionnaire des enclinomènes : Il s'agit d'une liste de 89 particules pouvant s'agglutiner au début ou en fin des formes fléchies. Les enclinomènes ont été groupés en classes. Douze classes de proclitiques et onze classes d'enclitiques ont été établies.

2.5.2 Les règles

Les règles sont de trois types :

a- Les règles de compatibilité du triplet (proclitique, radical, enclitique) : elles permettent de valider les juxtapositions de proclitiques, radicaux et enclitiques.

b- Les règles de réécriture : Elles permettent la prise en compte des phénomènes d'altération qui surviennent lors de l'agglutination de certaines enclinomènes à certaines formes.

c- Les règles d'harmonie vocalique : Elles permettent de vérifier que les voyelles entre le radicale et les enclitiques sont compatibles.

2.5.3 Les étapes de l'analyse

Sept étapes sont proposées pour l'analyse. La figure 13 reprise de [BEN-98 : p83] montre ces étapes.

2.5.4 Les résultats de l'analyseur de [BEN-1998]

Pour mesurer les performances de son analyseur, [BEN-1998] a utilisé la notion de bruit et de silence.

Bruit : « *On parle de bruit chaque fois qu'un découpage, une voyellation, une catégorie grammaticale ou toute autre information linguistique reste indûment associée à la forme textuelle analysée. On distinguera le bruit total où toutes les informations qui sont associées à une forme textuelle sont inadéquates.* » [BEN-1998 : p88]

Silence : « *On dit qu'il y'a silence quand un découpage, une voyellation, une catégorie grammaticale ou toute autre information linguistique est censée être associée à la forme textuelle analysée et qu'elle ne l'est pas. Le silence est dit total quand aucune des informations censées être associées à la forme textuelle ne l'est.* » [BEN-1998 : p89]

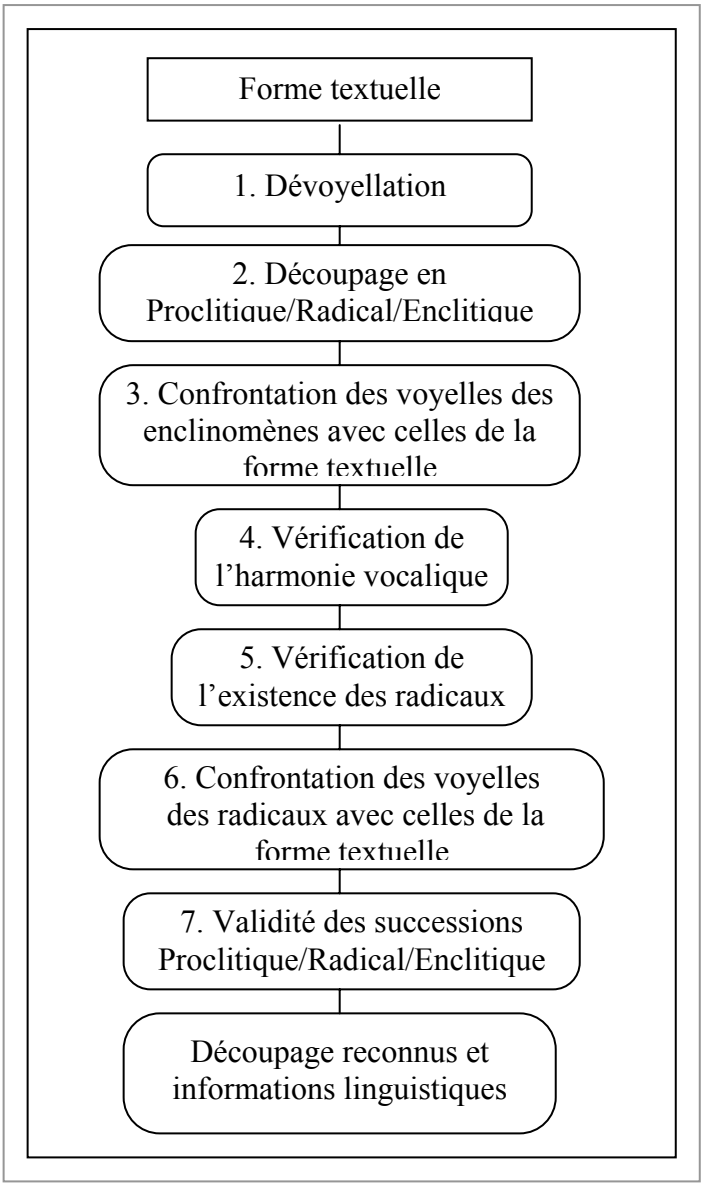


Figure 13: Schéma général des étapes de l'analyseur morphologique [BEN-98 : p83]

Les corpus utilisés pour l'expérimentation de l'analyseur sont de 21 903 mots non voyellés. Le tableau 4 donne respectivement pour chacun des quatre types d'informations ses taux de bons résultats, de bruit et silence.

	Découpage	Voyellations	Lemmes	Catégories grammaticales
Analyse correcte	74,46%	10,55%	43,44%	5,24%
Bruit	25,25%	89,07%	54,67%	93,54%
Silence	0,20%	0,21%	0,20%	0,20%

Tableau 4 : Résultats de l'analyseur morphologique de [BEN-1998]

Une première constatation de ce tableau implique que les cas de silence sont rares (environ 0,20%). On peut dire que l'analyseur présente des bonnes performances pour la tâche de découpage (74,46%), par contre il présente de mauvais résultats pour les autres tâches (voyellations : 10,55%, lemmatisation[5] : 43,44%, étiquetage grammaticale : 5,24%), ce qui démontre bien la difficulté de la tâche des autres niveaux d'analyse (notamment l'analyse syntaxique).

[BEN-1998] estime que les résultats de son analyseur sont assez encourageants.

2.6 L'analyseur morphologique de [ACHO-1998]

L'objectif de la réalisation de cet analyseur est l'étude de la voyellation automatique de l'arabe. La démarche préconisée consiste à utiliser un lexique de toutes les formes de type mot minimal (Un mot minimal = Préfixes + (Base ou Pro-base) + Suffixes) avec une liste des Proclitiques et Enclitiques). Ce choix a été justifié par les raisons suivantes :

- Le dictionnaire des formes fléchies qui présente l'inconvénient d'être volumineux a l'avantage d'être construit automatiquement. La construction automatique de formes fléchies ne constitue pas un avantage à mon avis, car :
  1- Elle ne supprime pas l'effort déployé pour construire le dictionnaire.
  2- La procédure de génération automatique n'est pas et ne peut être généralisable à la totalité des formes de la langue arabe.
  3- Dans sa thèse consacrée à la construction automatique d'un dictionnaire de formes fléchies, [BEN-1998] a démontré l'impossibilité de réaliser un système de synthèse lexicographique intégralement automatique. D'ailleurs, il conclut : « *Le système de synthèse lexicographique que nous avons développé est-il plutôt manuel ou plutôt automatique ? Est-ce un système de synthèse assisté par ordinateur ou un système de synthèse automatique assisté par l'Homme ? Nous préférons plus simplement dire qu'il s'agit d'un système de synthèse mixte ou semi-automatique.* » [BEN-1998 : p74]
- La grammaire est relativement simple.
- La taille du dictionnaire a moins d'impact vue l'évolution technologique dans le domaine des mémoires.

Il faut noter que cet analyseur est une version modifiée d'un analyseur déjà existant qui a été réalisé par [ZOUA-1989].

[5] La procédure qui permet de représenter chaque mot du texte par une forme canonique s'appelle lemmatisation. (ex. le nom au singulier, le verbe à l'infinitif, l'adjectif au masculin singulier).

2.6.1 Description du lexique utilisé

Ce lexique se compose de deux parties : Lexique de formes simple et le lexique des enclinomènes.

a) Lexique de formes simples

Il contient pour les noms, toutes les formes nominales fléchies (pouvant être au singulier, au duel ou au pluriel) ; et pour les verbes, toutes les formes verbales conjuguées (à l'accompli, l'inaccompli et l'impératif avec toutes les personnes). Ce lexique de formes simples compte 502 924 entrées dont 214 805 formes nominales, 304 014 formes verbales et 410 outils ou particules.

A chaque entrée de ces formes est associé un ensemble de traits linguistiques (genre, nombre…).

b) Lexique des enclinomènes

C'est une liste des proclitiques et des enclitiques (elles peuvent être simples ou composées). A chaque élément non voyellé de la liste sont associés :

- un type (proclitique / enclitique),
- les voyellations potentielles hors contexte,
- un indicateur : élément de locution (oui/non), cet indicateur permet de savoir si l'enclinomène est simple ou composé,
- une classe,
- des catégories grammaticales hors contexte.

2.6.2 Description de l'analyseur

La figure 14 reprise de [ACHO-1998] présente le principe de fonctionnement de l'analyseur.

La partie grammaire représente un ensemble de règles permettant d'agencer correctement les différents éléments contenus dans les deux lexiques (lexique des formes fléchies et le lexique des enclinomènes).

On distingue cinq types de règles :

1- Les règles d'harmonie vocalique : permettent de gérer la compatibilité entre la voyellation du radical et celle de l'enclitique qui lui colle.
2- Les règles de compatibilité ternaire : permettent de gérer la compatibilité entre classes de proclitiques, classes de radicaux, et classes d'enclitiques.
3- Les règles de réécriture : permettent de retrouver la bonne graphie d'une forme à partir de sa forme altérée à la suite de son agglutination à un enclinomène.

4- Les règles de soudure : permettent de souder correctement des enclinomènes à une forme simple (c'est le traitement inverse des règles de réécriture).
5- Les règles de compatibilité entre catégories grammaticales et flexions casuelles : permettent de faire correspondre à chaque catégorie grammaticale la flexion casuelle qui lui correspond.

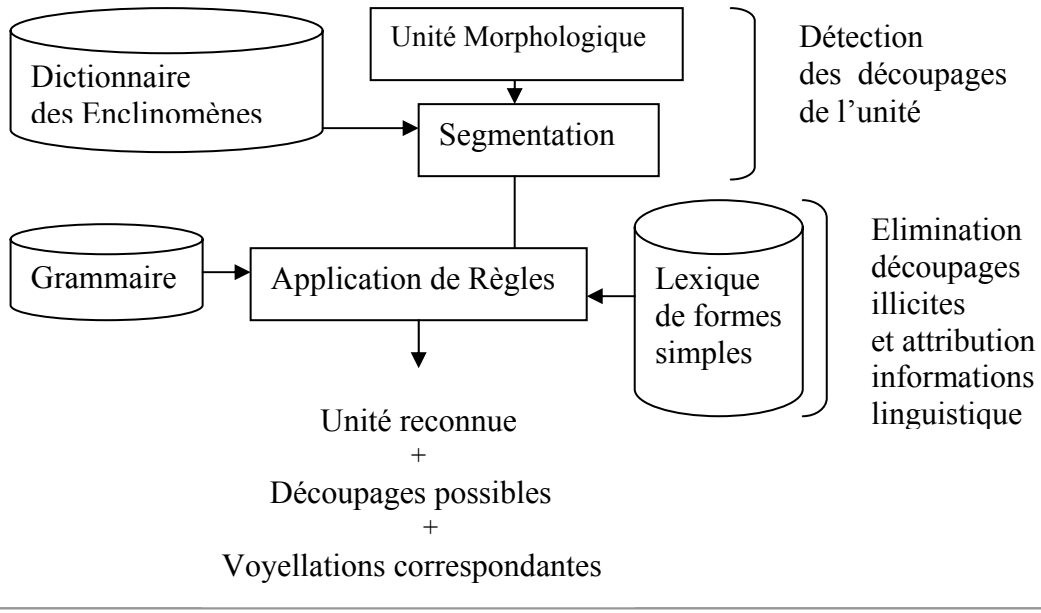


Figure 14 : Principe général de l'analyse morphologique [ACHO-1998 : p55]

Le principe de l'analyse morphologique se résume ainsi :

1- Réaliser un découpage de l'unité morphologique de toutes les façons possibles. Le résultat étant un ensemble de triplets (proclitique, radical, enclitique).
2- Valider chaque découpage obtenu. La validation passe par :
    a) L'accès au dictionnaire de formes simples pour vérifier l'existence du radical. S'il existe, l'analyseur lui associe l'ensemble de ses informations linguistiques (voyellations…). Il faut noter ici que l'analyseur applique sur le radical des règles de réécriture pour retrouver sa graphie initiale si ce dernier a été altéré suite à une agglutination à un enclinomène.
    b) L'accès au dictionnaire des enclinomènes pour vérifier l'existence du proclitique et de l'enclitique.
    c) L'application des règles d'harmonie vocalique pour valider la compatibilité des voyellations associées au radical avec celles de l'enclitique.
    d) L'application des règles de compatibilité ternaire pour vérifier l'existence de triplet (proclitique, radical, enclitique) compatible.
    e) L'application des règles de soudure pour :

i. Ressouder les composants non voyellés afin de comparer le résultat avec le mot initial du texte.
ii. Ressouder les composants voyellés pour obtenir le mot initial complètement revoyellé.

2.6.3 Les résultats de l'analyseur

Dans le cadre de la thèse de [ACHO-1998], les résultats de l'analyseur sont donnés en terme de voyellations, autrement dit ces résultats permettent d'évaluer l'apport de l'analyseur morphologique dans le processus de voyellations (aucun résultat n'est donné concernant l'analyseur en terme de reconnaissance des mots).

L'analyseur est expérimenté sur trois corpus (au hasard) respectivement d'une taille de 264 mots, 416 mots et 469 mots. L'évaluation est faite en terme de bruit et silence.

« *L'analyse morphologique d'un mot est bruyante lorsque, parmi les voyellations qu'elle lui a associées, il existe une ou plusieurs qui ne lui correspondent pas.* » [ACHO-1998 ; p.66]

« *L'analyse morphologique d'un mot est silencieuse lorsqu'il existe une ou plusieurs voyellations qui correspondent au mot et qui pourtant ne lui ont pas été associées.* » [ACHO-1998 ; p.66]

Les taux de bruit, de silence et de mots non reconnus sont donnés en se rapportant au nombre total de mots dans le texte :

Taux (Bruit) = (nombre de mots où il y a bruit / nombre total de mots) *100

Taux (Silence) = (nombre de mots où il y a silence / nombre total de mots) *100

Taux (Bruit) = (nombre de mots non reconnus / nombre total de mots) *100

Les résultats observés sur les corpus non voyellés sont :

	Bruit	Silence	Mots non reconnus
Corpus 1	0%	0,37%	0,37%
Corpus 2	0,24%	0%	0%
Corpus 3	0%	0,42%	0,42%

Tableau 5 : Résultats de l'analyseur morphologique de [ACHO-1998]

D'après [ACHO–1998] :

- Pour le corpus 1 : 0,37 % de mots non reconnus représente un mot dans le texte qui n'a pas été reconnu par l'analyseur (ce mot n'existe pas dans le dictionnaire). Le silence de 0,37% est dû à ce mot non reconnu.
- Pour le corpus 2 : Le bruit évalué par 0,24% (1 mot sur les 416 pour lequel l'analyse a été bruyante) est provoqué par une voyellation qui a été associée à tort au mot. [ACHO–1998] remonte la cause à un défaut dans l'algorithme d'analyse (étape de l'application des règles d'harmonie vocalique) qui peut être résolu sans difficulté !
- Pour le corpus 3 : Le silence de 0,42% est provoqué par deux mots (absents du dictionnaire) qui n'ont pas été reconnus.

[ACHO-1998] conclut que le silence est dû au manque de couverture du lexique (mots absents du dictionnaire), et le bruit est généré par l'algorithme d'analyse. Autrement dit si le dictionnaire contient toutes les formes rencontrées dans le corpus analysé (mots non reconnus égale à zéro) alors le silence sera toujours égal à zéro. Donc le fait que les deux taux (silence et bruit) soient égaux n'est pas un simple hasard mais montre bien que l'algorithme de [ACHO-1998] ne produit pas de silence.
A mon avis la taille des trois corpus de l'expérimentation n'est pas représentative, et pour s'en convaincre il suffit de consulter les résultats (voir le tableau 4 page 55 de cette thèse) obtenus par cet même analyseur et réalisés par [BEN-1998] (voir [BEN-1998 : p 27]). Etant donné que [ACHO-1998] et [BEN-1998] ont travaillé avec le même analyseur (du laboratoire de l'équipe de DEBILI), il nous semble que les résultats avancés par les deux sont en parfaite contradiction ! (si ce n'est pas une erreur de frappe!)

2.7 L'analyseur morphologique de [ATTI-2000]

Cet analyseur représente une version améliorée de l'analyseur morphologique développé par Nagy Fatehy Mohammad dans les laboratoires de RDI[6]. D'après son auteur cet analyseur n'est pas spécifique à une application particulière (non lié à un domaine particulier). [ATTI-2000] organise le mot autour des composants : préfixes, schèmes, racines, et suffixes.

Un mot = préfixe + radical + suffixe

Avec : radical = racine * schème ; Le symbole '*' dénote l'opération d'intrication (ou combinaison).

[6] RDI : Une société privée qui produit des logiciels à titre commercial (http: //www.rdi.com)

Si le préfixe et/ou le suffixe est vide alors le radical ne peut pas être vide.

Par analogie au modèle du mot graphique du projet SAMIA (déjà décrit dans la partie précédente), le radical représente la partie base, le préfixe représente la partie pré-base, et le suffixe représente la partie post-base.

Un mot = pré-base + base + post-base

Avec : base = racine * schème ; Le symbole '*' dénote l'opération d'intrication.

Une conséquence immédiate de cette organisation est que la base ne fera plus partie du lexique qui sera utilisé par l'analyseur, mais elle sera remplacée par un lexique des racines, un autre pour les schèmes et éventuellement une procédure de calcul, qui, à partir d'une racine et un schème engendre la base appropriée.

2.7.1 Description du lexique utilisé

Pour construire son lexique, [ATTI-2000] a divisé les mots arabes en deux classes :

- les mots d'origine arabe : cette classe se divise en deux sous classes :
  - Les mots dérivables : des mots qui possèdent une base qui s'analyse en racine et schème. Ces mots représentent la majorité du lexique de la langue arabe. Les bases de ces mots se divisent en deux sous classes :
    - ❖ les bases régulières : une base de cette classe est issue directement de l'intrication d'une racine et d'un schème.
    - ❖ les bases irrégulières : une base de cette classe représente une forme altérée de la forme issue de l'intrication d'une racine et d'un schème (voir exemple en bas).
  - Les mots outils (ou fixes) : C'est une liste fermée (environ 260 mots). A l'inverse des mots dérivables, les mots outils ne s'analysent pas en racine et schème.
- les mots étrangers : se sont des mots issus d'autres langues.

Racine	Base régulière non utilisée	Base Irrégulière
و ه م	وهم	تهم
و ق ى	وقوى	تقوى

Exemples de bases irrégulières

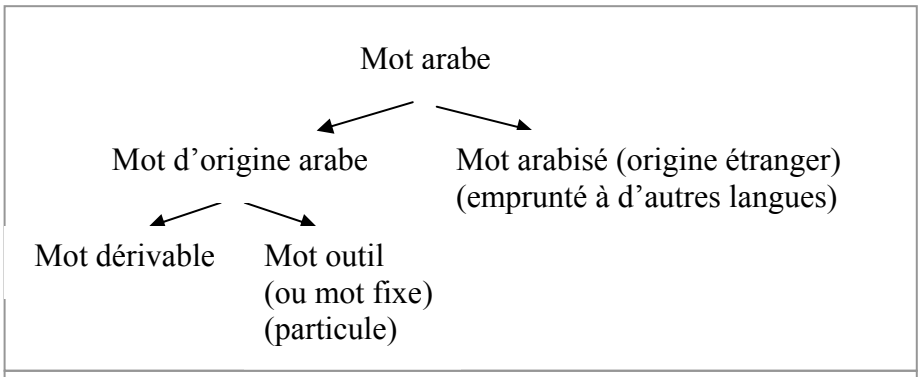


Figure 15 : Les différentes classes de mots arabes

Le modèle du mot graphique arabe utilisé par [ATTI-2000] repose sur les éléments suivants :

- Base (body) : **b** (cette base est analysable en racine et schème ***b = r*f*** avec ***r*** représente la racine et ***f*** représente le schème).
- Pré-base (préfixe) : ***p***
- Post-base (suffixe) : ***s***

Donc un mot arabe ***W*** est :

$$W = p + f*r + s \;\; = \;\; p + b + s$$

Le symbole '+' dénote l'opération de concaténation, pour réaliser cette concaténation, il faut que les couples ***(p,b), (b,s), et (p,s)*** soient compatibles.

En d'autres termes un mot arabe peut être représenté par le quadruplet :

$$Q = (t : p, r, f, s)$$

Avec t : représente le type de la base et par conséquent la classe du mot en question. On distingue quatre types :

a) Dérivé régulier.
b) Dérivé irrégulier.
c) Fixe (outils).
d) Arabisé

En utilisant ce modèle la synthèse d'un mot graphique arabe peut être vue comme un processus à trois étapes :

a) Construire la base : ***b = r*f***
b) Concaténer le préfixe à la base : ***p + b***
c) Concaténer le suffixe au résultat précédent : ***W = p + b + s = p + r*f + s***

Par contre l'analyse morphologique d'un mot graphique arabe W consiste à trouver le quadruplet ***Q = (t : p, r, f, s)***. En d'autres termes faire la segmentation du mot W en trois éléments : la base, le préfixe, et le suffixe. Par la suite essayer d'analyser la base ***b*** pour extraire le schème ***f*** et la racine ***r***.

En se basant sur le modèle décrit précédemment, [ATTI-2000] propose une base de connaissance morphologique construite autour de neuf entités différentes, chaque entité est composée de deux parties :

- Une partie description de l'entité en contexte libre, c'est la forme graphique de l'entité (isolated part).
- Une partie interactive qui décrit l'interaction de l'entité avec d'autres entités (interactive part).

Les neuf entités sont :

- Préfixes (P) (au nombre de 250)
- Racines des mots dérivables (R_d) (au nombre de 4500)
- Schèmes des mots réguliers (F_{rd}) (au nombre de 1000)
- Schèmes des mots irréguliers (F_{id}) (au nombre de 300)
- Racines des mots fixes (R_f) (au nombre de 120)
- Schèmes des mots fixes (F_f) (au nombre de 250)
- Racines des mots arabisés (R_a) (au nombre de 300)
- Schèmes des mots arabisés (F_a) (au nombre de 300)
- Suffixes (S) (au nombre de 550)

[ATTI-2000] estime que l'utilisation de 250 préfixes, 4500 racines des mots dérivables, 1000 schèmes des mots réguliers, 550 suffixes, et en prenant en compte les contraintes de combinaisons de ces entités, permet de couvrir un lexique de taille :

$$(250*4500*1000*550)/10 \cong 6x10^{10} \text{ mots arabes}$$

La figure 16 reprise de [ATTI-2000] présente la classification de ces entités.

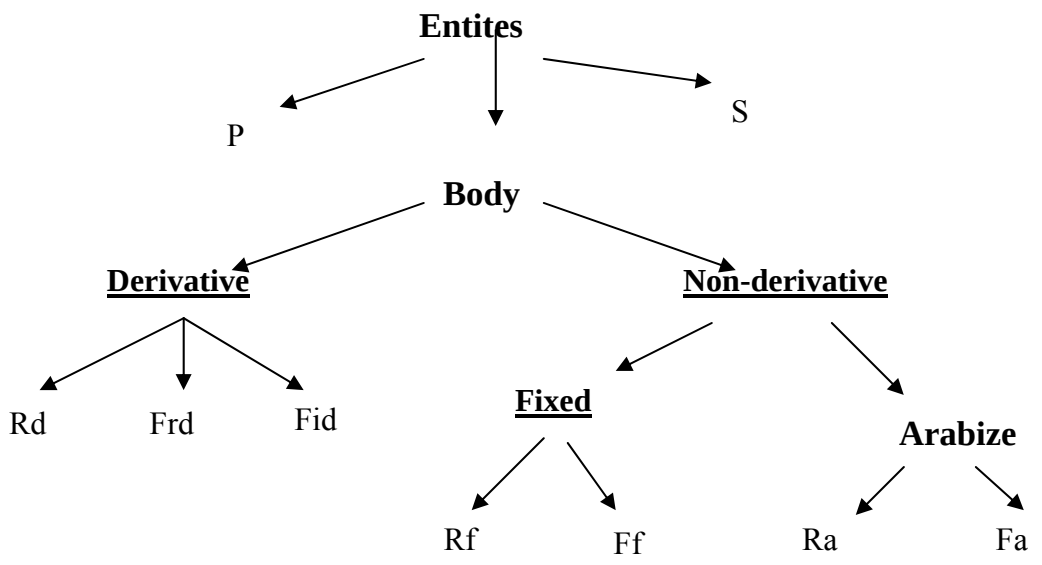


Figure 16 : Classification des entités [ATTI-2000 : p34]

2.7.2 Description de l'analyseur

La procédure d'analyse pour une forme (mot) s'annonce ainsi :

1- Extraction de tous les préfixes possibles.
2- Extraction de tous les suffixes possibles.
3- Vérification de la compatibilité des couples (préfixes, suffixes) en utilisant une matrice de compatibilité prédéfinie.
4- Extraction de toutes les bases possibles respectant les couples (préfixes, suffixes) compatibles.
5- Unification de toutes les bases similaires trouvées dans l'étape 4. Il s'agit de regrouper les triplets (préfixe, base, suffixe) qui présentent une même base.
6- Découpage de chaque base en racine et schème (extraire la racine et le schème de chaque base).
7- Vérification de la compatibilité des paires (schèmes, préfixes-suffixes)

Cette analyse va permettre donc de trouver tous les quadruplets (racine, schème, préfixe, suffixe) valides.

2.7.3 Les résultats de l'analyseur

Pour mesurer les performances de son analyseur [ATTI-2000] a utilisé la notion de couverture. Elle est définie par :

$$C = M / M_T \leq 1$$

Avec : MT : le nombre de mots analysé du texte.

M : le nombre de mots correctement analysé du texte

Un analyseur avec C < 98% a au moins un mot non analysé correctement pour un texte de 50 mots (M= 1, $M_T = 50$).

Pour comparer deux analyseurs en utilisant le paramètre de couverture on calcule le rapport de non couverture (uncoverage ratio) :

$$U_{12} = (1 - C_2) / (1 - C_1)$$

Si $U_{12} > 1$ alors la couverture du premier analyseur est U_{12} fois meilleure que le deuxième analyseur, sinon si $U_{12} < 1$, alors la couverture du deuxième analyseur est $1/U_{12}$ fois meilleure que le premier.

Après l'expérimentation de cet analyseur sur des corpus de tailles 15000, 50000, 250000 et 500000 mots (malheureusement l'auteur n'a pas spécifié le type de ces corpus : un facteur très important à prendre en compte dans l'évaluation des résultats de l'analyseur) il a affiché les résultats suivants (voir tableau 6) :

Taille du corpus	**M_T**	**M**	**$C = M / M_T$**
15 000	5 000	4 900	98 %
50 000	10 000	9 873	98,73 %
250 000	10 000	9 890	98,9 %
500 000	10 000	9 905	99,05 %

Tableau 6 : Résultats de l'analyseur morphologique de [ATTI-2000]

Les résultats présentés dans le tableau 6 suggèrent que l'analyseur donne de bonnes performances (un taux de couverture entre 98% et 99,05%). Toutefois il semble qu'il y ait une relation étroite entre taille du corpus et le taux de couverture. Plus le corpus est grand, plus la couverture est grande. Mais sachant que la procédure d'analyse n'a pas de relation avec la taille du corpus analysé, alors comment explique t-on ces résultats ?

Par ailleurs, l'auteur ne donne aucune indication sur la nature des mots qui ne sont pas correctement analysés ainsi que, la raison de l'échec de son analyseur dans le traitement de ces mots, est-ce que :

- Un défaut dans l'algorithme d'analyse ?
- Mots absents du lexique ?

2.8 L'analyseur morphologique de [GAUB-2001]

Cet analyseur entre dans le cadre de la réalisation d'un logiciel (dénommé Sarfiyya) pour l'expérimentation d'une approche linguistique (théorie de la minimalité) pour le traitement automatique de l'arabe développée par Claude Audebert et André Jaccarini. Ce logiciel comporte un ensemble de modules dédiés à la conception des grammaires morphologiques minimales, la variation de ces grammaires et la mesure de leur impact sur un texte réel.

L'objectif principal de la minimalité s'annonce :

« Parvenir à une analyse morpho-syntaxique de la phrase arabe, sans toutefois en explorer la sémantique, par une exploitation optimale et une description algorithmique pertinente de toutes les régularités morphologiques et syntaxiques de la langue, sans recours au lexique ni à une liste exhaustive de règles de syntaxe. » [GAUB-2001 : p7]

Les principes fondamentaux de cette théorie sont la régularité de la morphologie (permettant de s'affranchir du lexique) et la primauté de la syntaxe appuyée par le rôle central des mot-outils. Le formalisme retenu pour la mise en œuvre pratique de ces deux principes est celui des AFND (Automate Fini Non Déterministe) et des ATN (Réseau de Transition Augmenté).

D'après [GAUB-2001] l'étude des éléments du lexique arabe se décompose en deux parties :

- La morphologie externe : Etude de la définition des éléments du lexique arabe.
- La morphologie interne : Etude de la structure interne des éléments du lexique arabe.

Autrement dit l'étude du mot graphique revient à étudier le triplet (préfixe, radical, suffixe), la morphologie externe c'est l'étude des couples (préfixe, suffixe) et la morphologie interne concerne l'élément central qui est le radical. Le radical est le résultat de l'intrication d'une racine et d'un schème.

A partir d'une liste de racines et de schèmes, et en utilisant les AFND et les ATN, [GAUB-2001] propose donc de construire son analyseur morphologique.

2.8.1 Description du lexique utilisé

Le lexique de [GAUB-2001] se divise en deux composantes : les racines et les schèmes.

a) Les racines (issue du dictionnaire 'AL-SIHAH')

C'est une liste de 4800 racines trilitères, 774 racines quadrilitères et 35 racines quinquilitères. Cette liste est enregistrée sous forme de R1R2R3 pour les racines trilitères (R1, R2, R3 sont les consonnes de l'alphabet arabe) avec :

- dimension de R1 et R2 : 28 éléments
- dimension de R3 : 27 éléments (le 'WAW' و et le 'YA' ى ne sont pas distingués)

Il est à noter que seules les racines trilitères sont prises en compte dans la version actuelle de cet analyseur.

b) Les schèmes

C'est une liste (de 100 schèmes les plus courants nominaux et verbaux dévoyellés) enregistrée dans un fichier sous la forme (fa'ala) فعل pour les trilitères et (fa'lala) فعلل pour les quadrilitères avec un ensemble d'informations du genre : Ce schème est-il celui d'un substantif (actif ou passif) ? Ce schème est-il celui d'un verbe à l'accompli ou à l'inaccompli ? …

2.8.2 Description de l'analyseur

Trois catégories sont proposées : Nom (N), Verbe (V) et Atomes ou Tokens (T).
L'analyseur de [GAUB-2001] fonctionne en deux passes, la première passe de l'analyse restitue, pour un mot donné et selon un automate donné, une réponse sous forme de descriptions morphologiques « lettre à lettre ». (On parle de descriptions morphologiques élémentaires. Une description morphologique élémentaire représente la catégorie lexicale de chaque caractère : voir les différentes catégories lexicales dans [GAUB-2001 : pp 62-64]). Chaque réponse est composée d'interprétations explicitant la catégorie de chaque lettre. La deuxième passe effectue les tâches d'étiquetage, de recomposition et de contrôle de morphèmes ou informations lexicales propres à l'arabe.

L'analyse morphologique permet donc de déterminer la description morphologique élémentaire de chaque mot analysé, ce qui va permettre par la suite de décider quelles étiquettes employer pour les interprétations de l'analyse.

Dans l'état actuel de Sarfiyya seuls la détermination pour les noms et l'aspect/mode pour les verbes sont utilisés pour l'étiquetage.

2.8.3 Les résultats de l'analyseur

D'après [GAUB-2001] cette nouvelle approche (minimalité) conduit à une couverture presque totale de la morphologie, soit 99 %, avec peu de règles, l'automate nominal comporte 350 transitions pour 51 états et l'automate verbal plus de 500 transitions pour 54 états. Toujours selon [GAUB-2001], le taux de non couverture, soit 1 %, est dû principalement aux mots étrangers. Parallèlement à cette large couverture, l'ambiguïté reste toujours inévitable, [GAUB-2001] enregistre environ 2,1 interprétations par mot. A mon avis, bien que cette approche donne un taux de reconnaissance remarquable, son inconvénient reste toujours la complexité croissante de la maintenance des réseaux construits à partir de 350 (voire 500) transitions. Par ailleurs, cette approche ne permet pas d'aller vers la sémantique, car, le recours à un lexique est inévitable.

2.9 L'analyseur morphologique de [OUER-2002]

La construction de cet analyseur morphologique entre dans le cadre global qui est celui de la réalisation d'un analyseur morpho-syntaxique pour la détection et le diagnostic des fautes d'accord.

Dans cette étude, l'organisation du lexique retenue consiste à utiliser une liste de bases avec une liste de proclitiques, préfixes, suffixes et enclitiques.

Le choix de cette organisation a été justifié par les raisons suivantes :

- Les traitements sont beaucoup moins coûteux en terme de temps que les accès et la recherche dans des lexiques volumineux.
- L'utilisation d'un lexique de formes canonique implique un gain important en terme de taille.
- Le système de [OUER-2002] est destiné à des applications qui nécessitent des analyses fines. A mon avis cet argument dépend plutôt du contenu du lexique et non pas de l'organisation choisie, car on peut toujours réaliser des analyses fines en utilisant une autre organisation.

Il faut noter que cet analyseur est une version adaptée d'un analyseur existant de l'équipe dans laquelle travail [OUER-2002].

2.9.1 Description du lexique utilisé

Pour réaliser cet analyseur, [OUER-2002] a construit un lexique regroupant quatre parties (ce lexique a été généré à partir de la base de données lexicale DIINAR1):

- Liste des mots outils : Cette liste compte 350 éléments répartis en deux listes, une liste contenant les propositions, conjonctions, démonstratifs, relatifs et une liste contenant les adverbes, les cardinaux, etc.
- Listes des composants d'un mot : comptent 2365 éléments répartis comme suit :
    - Liste des préfixes (au nombre de 9),
    - liste des suffixes (au nombre de 61),
    - liste des couples préfixe-suffixe compatibles (au nombre de 179),
    - liste des proclitiques, elle contient des proclitiques et des combinaisons de proclitiques (au nombre de 68),
    - liste des enclitiques, elle contient des enclitiques et des combinaisons d'enclitiques (au nombre de 49),
    - liste des prébases (au nombre de 279),
    - liste des Post-bases (au nombre de 1720).
- Lexique des données morpho-syntaxiques : Il contient les schémas vocaliques et les informations morpho-syntaxiques qui peuvent être associés aux bases non-voyellées, comme par exemple la racine, la catégorie grammatical, le genre,…
- lexique des bases : Ce lexique contient 196818 bases non-voyellées de DIINAR1 réparties en 39000 bases nominales, 79818 bases verbales et 78000 bases déverbales.

2.9.2 Description de l'analyseur

L'analyseur découpe de toutes les façons possibles un mot graphique arabe pour obtenir toutes les combinaisons acceptables (valides) en clitiques, préfixes, suffixes et bases. Cette décomposition est faite selon le modèle du projet SAMIA décrit précédemment dans de ce chapitre.

L'analyse se déroule en sept phases :

a) Régularisation pré-morphologique : cette étape comprend

- Le traitement des articles contractés (lil → li + al , لل→ل + ال).
- Le traitement de la 'Hamza' en remplaçant les formes 'ى' ، ' أ ' ، ' ؤ ' par la forme canonique de la 'Hamza' .' ء '
- Le traitement de la 'shadda' : Elimine la 'shadda' si la première lettre de la base est une consonne solaire[7].

[7] Lors du soudage du proclitique 'al' avec la première lettre de la base, on ajoute la 'shadda' sur cette première lettre (si cette lettre est solaire). Une lettre solaire appartient à la liste des consonnes suivante : ل، ز، د، ذ، ت، س، ش، ض، ظ، ص، ط، ث، ر، ن

- Dévoyellation du mot en entrée : le schéma vocalique du mot est conservé pour une comparaison ultérieure.

b) Consultation des mots outils : Accès à la liste des mots outils pour vérifier si le mot appartient à cette liste.

c) Décomposition du mot : Consiste à segmenter le mot de toutes les façons possibles en prébases, bases et post-bases.

d) Validation des décompositions du mot en utilisant la matrice de compatibilité entre prébases et post-bases, cette étape consiste à rejeter les segmentations non valides.

e) Affinage des décompositions : consiste à calculer pour chaque prébase (respectivement postebase) trouvée précédemment, ses couples de proclitiques, préfixe (respectivement suffixe, enclitique) qui le composent.

f) Consultation des lexiques et récupération des informations : l'analyseur doit valider chaque segmentation trouvée en utilisant les différents lexiques. L'analyseur associe à chaque décomposition valide l'ensemble de ses informations morpho-syntaxiques.

g) Régularisation post-morphologique : A l'inverse de la segmentation, cette étape consiste à souder (concaténer) les différents composants trouvés (proclitique, préfixe, base, suffixe et enclitique) lors de l'étape précédente. Pour réaliser cette tâche, un ensemble de règles dites morpho-phonologique propre à la langue arabe sont nécessaires comme par exemple les règles du support de la 'Hamza' , la règle de la 'tâ marbûta' تاء المربوطة,…

Cet analyseur est implémenté avec le langage Visual C++ sous Windows.
La figure 17 reprise de [OUER-2002] représente le schéma général de l'analyse.

2.9.3 Les résultats de l'analyseur

Le découpage d'un mot en ses formants n'est généralement pas unique, d'où les sources d'ambiguïtés, [OUER-2002] compte un nombre moyen de 1.25 découpages possibles pour un mot non voyellé arabe et ce sur la base d'un comptage statistique fait dans un corpus de 37952 mots non-voyellé. Le tableau 7 résume les résultats de l'analyseur.

Les résultats obtenus par l'analyseur de [OUER-2002] montrent que 81% des mots sont ambigus et que le nombre d'analyses par mot est de six catégories.

[OUER-2002] souligne que pour le français 20% et pour l'anglais 11% des mots sont ambigus et que le nombre d'analyses par mot est de deux catégories pour le français.

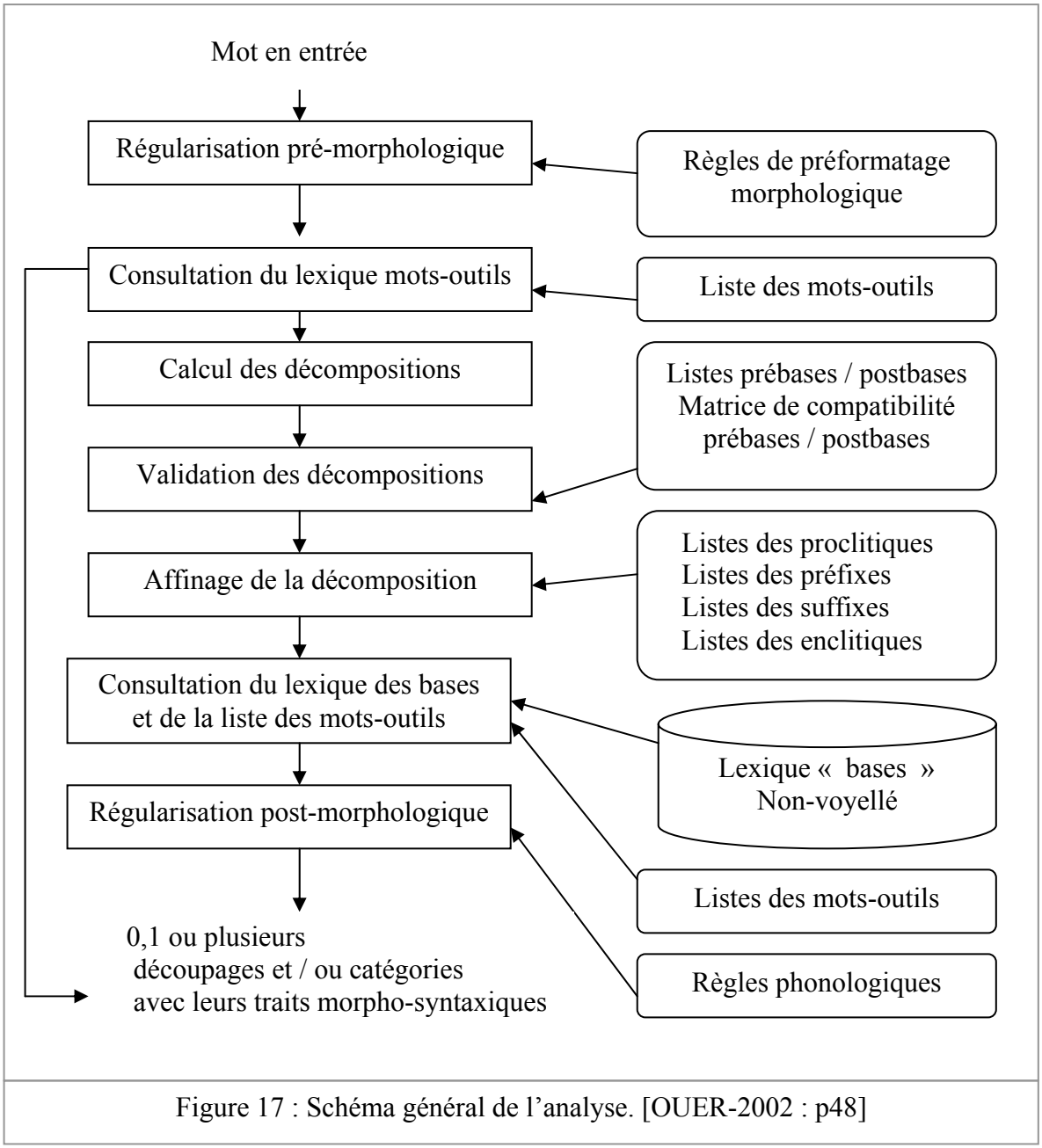


Figure 17 : Schéma général de l'analyse. [OUER-2002 : p48]

Taille du corpus	% reconnaissance	Nombre d'analyse Par mot	% mot non ambigu	Nombre de découpage
37 952 mots	89 %	5,99	19 %	1,25

Tableau 7 : Résultats de l'analyseur morphologique de [OUER-2002]

L'analyseur de [OUER-2002] affiche un taux de 89% de mots reconnus, les 11% non reconnus sont dus à des erreurs de numérisation (reconnaissance optique) dans le corpus utilisé. Pour un texte saisi à la main (de 7802 mots non voyellés) le taux de reconnaissance de l'analyseur est de 97%, les 3% de mots non reconnus sont dus essentiellement aux fautes d'orthographe, les noms propres et les mots étrangers.
Ces résultats nous laissent penser que le programme d'analyse, aussi complexe qu'il soit, ne produit pas d'erreurs (de segmentation, de validation…)!

2.10 L'analyseur morphologique de [ZAAF-2002]

L'objectif de la réalisation de cet analyseur est l'apprentissage avec ordinateur de l'arabe langue étrangère. L'organisation retenue dans cette étude consiste à utiliser un lexique de toutes les formes de type mot minimal (Un mot minimal = préfixes + (Base ou Pro-base) + suffixes) avec une liste des proclitiques et enclitiques). Ce choix a été justifié par les raisons suivantes :

- L'arabe est une langue fortement agglutinée.
- Les textes peuvent être non voyellés, partiellement voyellés ou complètement voyellés.

Autrement dit si on opte pour l'organisation qui utilise le mot maximal, on doit prévoir un lexique d'environ une centaine de millions de mots, chose qui est non raisonnable. Donc [ZAAF-2002] met en avant le problème de la taille du lexique pour justifier le rejet d'organiser son lexique autour du mot maximal (malgré l'existence des ordinateurs ayant une mémoire de taille très importante !). A mon avis, même le lexique des formes fléchies présente l'inconvénient d'être volumineux (estimé à environ six millions de formes dans [OUER-2002]). Toutefois il présente l'avantage d'être généré automatiquement à partir de DIINAR[8].

Il faut noter que cet analyseur est une version adaptée d'un analyseur existant de l'équipe dans laquelle travail [ZAAF-2002].

2.10.1 Description du lexique utilisé

Ce lexique est généré à partir de DIINAR, il se compose de sept parties [ZAAF-2002 ; p.80]:

1. Lexique des formes conjuguées : il comprend l'ensemble des formes conjuguées à partir des verbes de DIINAR. Chaque entrée est composée de la forme conjuguée

[8] DIINAR est une base de données lexicale.

non vocalisée, un schéma vocalique, le verbe, la racine, le pronom de conjugaison et l'aspect de conjugaison.

2. Lexique des formes nominales : comprend l'ensemble des bases nominales obtenues à partir de DIINAR. Chaque entrée est composée de la base nominale non vocalisée, un schéma vocalique, le genre, le nombre, ses différents suffixes et déclinaisons possibles, et si oui (ou non) elle accepte l'article de détermination ('al').
3. Lexique des déverbaux : comprend l'ensemble des déverbaux de DIINAR. Chaque entrée (au singulier masculin) est composée de la base non vocalisée du déverbal, le schéma vocalique, la catégorie du déverbal, le verbe, la racine du verbe et les différentes déclinaisons.
4. Lexique des mots outils : il regroupe toutes les formes fléchies des mots outils. Chaque entrée de ce lexique comprend la forme non vocalisée du mot outil et son schéma vocalique.
5. Lexique des noms propres : comprend l'ensemble des bases nominales obtenues à partir de DIINAR. Chaque entrée est composée de la base nominale non vocalisée, le schéma vocalique, la catégorie sémantique, le genre, le nombre et les différents suffixes et déclinaisons possibles.
6. Liste des enclitiques : cette liste comprend tous les enclitiques. Chaque élément de cette liste se compose de l'enclitique non voyellé et son schéma vocalique.
7. Liste des proclitiques : cette liste comprend tous les proclitiques. Chaque élément de cette liste se compose du proclitique non voyellé et son schéma vocalique.

Aucune indication sur la taille de ces différents lexiques n'a été donnée.

2.10.2 Description de l'analyseur

La figure 18 reprise de [ZAAF-2002] présente le principe de fonctionnement de l'analyseur.

L'analyse morphologique se divise en plusieurs étapes :

1- Dévoyellation de l'entrée : permet de traiter les textes partiellement vocalisés et complètement vocalisés.
2- Consultation de la liste des mots outils : d'après [ZAAF-2002] la fréquence des mots outils dans un texte arabe est très importante. C'est pour cette raison que le lexique des mots outils est consulté au début du processus d'analyse.
3- Identification des enclitiques : consiste à trouver les différents enclitiques de la forme à analyser.

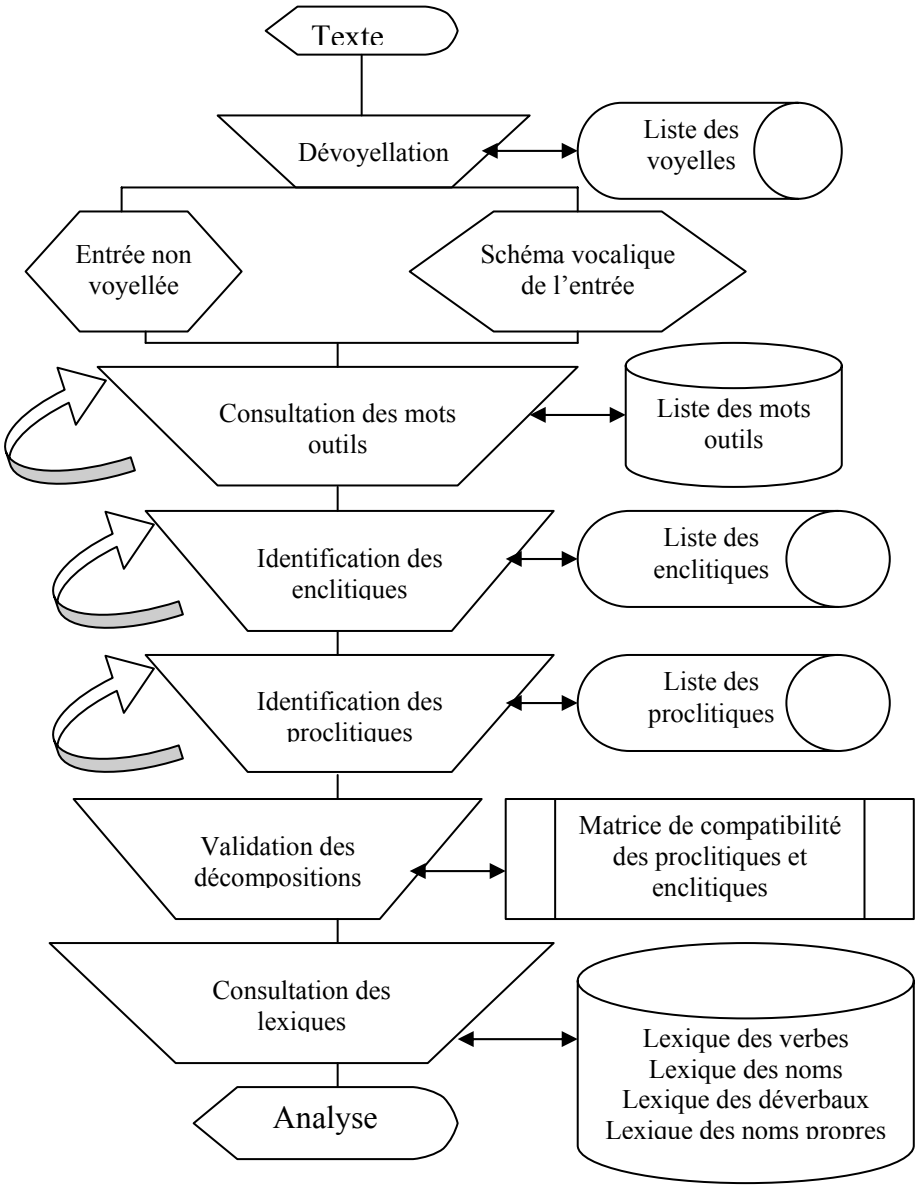


Figure 18 : Principe général de l'analyse morpho. [ZAAF-2002 : p71]

4- Identification des proclitiques : consiste à trouver les différents proclitiques de la forme à analyser.

5- Validation des décompositions : en utilisant une matrice de compatibilité entre proclitique et enclitique, il s'agit de ne retenir que les couples (proclitique, enclitique) valides (ou compatibles). La matrice de compatibilité entre proclitique et enclitique permet de dire si oui ou non le couple (enclitique, proclitique) est

compatible, si c'est le cas alors elle renseigne sur le type présumé du mot (un verbe, un nom ou un déverbal, un nom ou un déverbal ou un verbe).

6- Consultation des lexiques : consiste à attester l'existence du mot minimal dans l'un des différents lexiques et de récupérer les diverses informations relatives à ce mot minimal.

2.10.3 Les résultats de l'analyseur

Les documents que nous possédons ne donnent aucune idée sur :

- l'implantation sur machine de l'analyseur morphologique ;
- l'évaluation des performances de l'analyseur.

3. Résumé des analyseurs existants

Le tableau 8 résume les différents analyseurs existant étudiés.

N°	Analyseurs existants	Labo / Equipe	Organ-isation **	Taille du lexique	Domaine d'application	Observations
1	[HASS-1987]	SAMIA	**X**	Pas d'indication	Enseignement Assisté par ordinateur (E.A.O.) de l'arabe.	Analyseur morphologique pour des textes vocalisés seulement.
2	[HASS-1987]	SAMIA	4	Pas d'indication	Enseignement Assisté par ordinateur (E.A.O.) de l'arabe.	Pas d'implantation sur machine
3	[SARO-1989]	IRIT-CERFIA	2	200000 formes fléchis	Pas de domaine particulier	Prototype
4	[ZOUA-1989]	F. DEBILI	2	Pas d'indication	Pas de domaine particulier	Prototype analysant un mot à la fois
5	[BEN-1998]	F. DEBILI	2	599361 formes fléchies+ 89 enclinomènes	Détection et correction des erreurs orthographiques dans les textes arabes	version actualisée de l'analyseur de [ZOUA-1989]
6	[ACHO-1998]	F. DEBILI	2	502924 (formes nominales +formes verbales +outils)	Voyellation automatique de l'arabe	version modifiée d'un analyseur déjà existant qui a été réalisé par F. DEBILI et son équipe
7	[ATTI-2000]		4	7570 (préfixes+suffixes + schèmes+racines)	Pas de domaine particulier	Prototype

8	[GAUB-2001]		5	5709 (racines+schèmes)	Expérimentation de la minimalité pour le traitement de la morphologie et la syntaxe arabe	Prototype
9	[OUER-2002]	SAMIA	3	199254 (bases+outils+préfixes+suffixes+enclitiques+proclitiques+prébases+Post-bases)	Détection et diagnostique des fautes d'accord de l'arabe	Prototype
10	[ZAAF-2002]		2	Lexique généré à partir de DIINAR projet SAMIA	Apprentissage avec ordinateur de l'arabe langue étrangère	Prototype

X : Cette organisation ne se base pas complètement sur le modèle du mot graphique arabe (Voir la description du lexique utilisé par [HASS-1987]).

****** : Les cinq organisations possibles :

1- Liste de tous les mots de la langue (mot maximal).

2- Liste des formes fléchies (mot minimal), des proclitiques et des enclitiques.

3- Liste des bases (BAS/PBA), des préfixes, des suffixes, des enclitiques et des proclitiques.

4- Liste des prébases, des racines, des schèmes et des Post-bases.

5- Liste des racines et schèmes.

Tableau 8 : Résumé des analyseurs existants

Nous concluons l'étude des analyseurs existants par les remarques suivantes :

1. Tous les analyseurs procèdent par une approche séquentielle (c'est à dire faire l'analyse morphologique ensuite l'analyse syntaxique et non pas fondre l'analyse morphologique dans l'analyse syntaxique.
2. Pour mesurer les performances d'un analyseur, il y a toujours recours à des métriques type maison. Certaines études mesurent les performances en terme de silence/bruit, d'autres en termes de mot ambigus/non ambigus. Cette situation rend impossible toute tentative de comparaison entre ces analyseurs. Donc il n'y a pas de métriques unifiées pour mesurer la performance d'un analyseur et par conséquent, il n'y a pas d'outils standard pour l'évaluation des performances d'un analyseur.
3. L'expérimentation est faite d'une manière aléatoire. Certaines le font sur des mots, d'autres sur des corpus (plus ou moins sélectionnés, plus ou moins représentatifs…). La taille des corpus utilisés est très différente ([OUER-2002] a utilisé un corpus de 37952 mots, [ACHO-1998] a utilisé un corpus de 469 mots environ…). Le type de corpus utilisé pour l'évaluation n'a pas été clairement défini, sauf le corpus utilisé par [OUER-2002] qui comprend des textes littéraires, journalistique (politique, culture) et scientifique (géographie, astrologie, biologie).
4. Tous ces analyseurs sont des prototypes!
5. Le dictionnaire (élément très important dans toute application en TALN) est très peu abordé dans ces études. Par dictionnaire on entend : la conception, la réalisation et la mise à jour du lexique utilisé. Certaines études supposent l'existence de cette ressource (le dictionnaire), ce qui laisse entendre qu'il s'agit d'un problème bien résolu !
6. Une plus grande importance est donnée à la stratégie d'analyse (procédure d'analyse) au détriment bien sûr du lexique.
7. Certaines études exposent des résultats difficilement interprétables (rend très difficile une interprétation scientifique nous permettant de remonter aux origines des phénomènes observés).
8. Si on estime la taille objective d'un dictionnaire (tel qu'estimée par [ATTI-2000], environ 6.10^{10}) alors la plupart des dictionnaires réalisés jusqu'à présent ne sont pas complets ! (donc on ne peut pas prétendre à une large couverture des analyseurs).

9. Il n'y a pas un modèle linguistique standard sur lequel on peut construire des systèmes de TALN.
10. Il n'existe pas encore d'analyseur général de la langue arabe, c'est à dire dont la couverture soit suffisamment large pour satisfaire aux exigences d'application à large échelle comme la correction ou la traduction automatique. [OUER-2002].

4. Le choix d'une organisation

Sur la base du modèle du mot graphique arabe décrit dans le chapitre précédent nous distinguons cinq organisations possibles du lexique :

1- Liste de tous les mots de la langue (mot maximal).
2- Liste des formes fléchies (mot minimal), des proclitiques et des enclitiques.
3- Liste des bases (BAS/PBA), des préfixes, des suffixes, des enclitiques et des proclitiques.
4- Liste des prébases, des racines, des schèmes et des Post-bases.
5- Liste des racines et schèmes.

On peut classer ces organisations selon l'approche[9] d'analyse comme dans le tableau 9.

Approches	Organisations
I	1
II	2, 3, 4, 5

Tableau 9 : Approches et organisations

Essayons d'analyser ces différentes organisations.

- La première consiste à utiliser un lexique de toutes les formes (objets) de type mot maximal (Un mot maximal = PCL # PRE + (BAS ou PBA) + SUF # ECL). Cette organisation présente l'avantage de la simplification de la stratégie d'analyse (une simple consultation du lexique), mais pose le problème de la taille du lexique et du temps d'accès à ce dernier; pour la langue arabe on compte des centaines de millions de mots ([ATTI-2000] compte environ 6.10^{10}). Cette organisation n'est donc pratiquement pas envisageable malgré sa faisabilité technique qui est liée à la présence actuelle

[9] Voir la page 39 de ce chapitre pour l'explication des deux approches (I et II).

d'un matériel informatique sophistiqué (mémoire de taille importante, vitesse des traitements plus élevée)!

- La deuxième consiste à utiliser un lexique de toutes les formes de type mot minimal (Un mot minimal = PRE + (BAS ou PBA) + SUF) avec une liste des PCL et ECL. La liste des formes de type mot minimal représente la liste de toutes les formes fléchies obtenues à partir de la liste des bases, des préfixes et des suffixes. La taille du lexique sera moins considérable (502924 formes dans [ACHO-1998] et estimé à environ 6 millions[10] de formes dans [OUER-2002]), par conséquent la procédure d'analyse doit tenir compte des règles de construction d'un mot maximal en fonction du triplet (mot minimal, PCL, ECL). Généralement ces règles de construction sont de deux types, soit des règles de concaténation soit des règles de compatibilités entre composants. L'algorithme d'analyse évite tous les phénomènes de flexion et de dérivation. Il ne traite que les phénomènes d'agglutination des clitiques aux formes fléchies.

- La troisième consiste à utiliser une liste de bases (BAS/PBA) avec une liste de PCL, PRE, SUF et ECL. La taille du lexique est raisonnable (199254 entités dans [OUER-2002]), mais la procédure d'analyse sera plus complexe comparée à celle de l'organisation précédente du moment où elle doit tenir compte des règles de construction du 5-uplet (bases, PCL, PRE, SUF, ECL). A partir du lexique de cette organisation on peut toujours générer le lexique de l'organisation précédente par une opération de dérivation et de conjugaison.

- La quatrième organisation consiste à utiliser une liste des racines, une liste des schèmes, une liste des prébases et une liste des Post-bases. Dans cette organisation un mot se présente comme un quadruplet (prébase, racine, schème, Post-base). L'intrication du schème avec la racine forme la partie radicale. La taille du lexique est minimale (7570 entités dans [ATTI-2000]), mais la procédure d'analyse sera plus complexe que les organisations précédentes, car en plus de la complexité de la construction d'un mot à partir des formants du mot en question, une autre complexité vient s'ajouter qui est

[10] Ce lexique de mot fléchis (préfixe-base-suffixe) est généré à partir de la base de données lexicale DIINAR.1 du projet SAMIA.

celle de l'opération d'intrication[11] (l'analyse et/ou génération des bases à partir des schèmes et des racines).

- La cinquième consiste à utiliser une liste très réduite (5709 éléments) de racines et de schèmes. Dans cette organisation, il s'agit donc, d'utiliser un lexique réduit au minimum. L'importance est focalisée entièrement sur la stratégie d'analyse. L'avantage de cette démarche est évident (pas de lexique), mais les inconvénients a mon avis seront la complexité de la stratégie d'analyse, et la complexité de la maintenance (l'évolutivité) des analyseurs construits autour de cette démarche ; pour cela, il faut remarquer l'effort produit pour la mise à jour d'un lexique (organisation 1, 2, 3 ou 4) comparé à celui de la mise à jour d'un analyseur. Enfin la construction d'un analyseur moyennant cette organisation ne favorise en aucun cas la séparation entre objets linguistiques et les traitements qui leurs sont associés. Cette organisation ne me semble pas appropriée pour la construction d'un analyseur morphologique surtout si on envisage d'aller vers la sémantique, car le recours à un lexique devient inévitable!

5. La solution adoptée

Pour choisir une organisation on doit prendre en compte trois paramètres : la taille du lexique, la complexité de la procédure d'analyse et l'application ciblée. Une organisation qui minimise la taille et la complexité sera la plus favorable, mais malheureusement on remarque que lorsque la taille diminue en allant de la première organisation vers la dernière organisation, la complexité augmente.

Si la première organisation ne peut faire l'objet d'un choix raisonnable, et on remarque que l'organisation 2 est une variante de l'organisation 3, car on peut toujours générer le lexique de l'organisation 2 à partir de l'organisation 3 par un dériveur et un conjugueur, alors le choix sera fait parmi trois organisations possibles (à savoir 3, 4 ou 5). Envisager un peu de la sémantique après l'analyse morphologique, c'est écarter l'organisation 5 et de se fait réduire le choix à deux cas (organisation 3 ou 4).

Minimiser la taille du lexique revient à choisir l'organisation 4, mais en dehors d'une valeur concrète de la complexité de l'algorithme d'analyse, base d'une comparaison objective entre les deux organisations (3 et 4), le choix reste difficile à

[11] Opération qui consiste à combiner une racine avec un schème pour construire une base (voir chapitre 2).

faire. Pour toutes ces raisons, il nous paraît justifié d'adopter la troisième attitude pour représenter le lexique du système. Ceci nous permet, d'une part, d'éviter de gérer un lexique trop volumineux et difficile à mettre à jour et d'autre part, de ne garder dans la base que les éléments essentiels nous permettant de générer les mots dont on a besoin.

6. Conclusion

L'étude des analyseurs existants nous a permis de construire une base sur laquelle nous avons justifié notre choix pour l'organisation. Ce choix étant engagé nous pouvons maintenant proposer dans le chapitre suivant une modélisation conséquente pour le traitement automatique de la langue arabe. Le modèle ainsi construit sera considéré comme une plate forme commune et réutilisable par toutes les applications de TALN arabe.

Chapitre 4

Définition du modèle linguistique Et description de l'analyseur morphologique

Chapitre 4

Définition du modèle linguistique Et description de l'analyseur morphologique

1. Introduction

Avant toute analyse automatique d'une langue naturelle, il est indispensable de définir un modèle linguistique à priori, ce dernier doit répondre à certaines caractéristiques globales : La généralisation, l'optimisation et la calculabilité. [LALL-1990]

- La généralisation : il faut que le modèle linguistique soit le plus général pour couvrir une plus grande partie de la langue (maximum de phénomènes linguistiques, un dictionnaire complet…). On parle souvent de couverture de la langue. Notre analyseur devra donc permettre de traiter n'importe quel texte écrit en arabe, facilement réutilisable par les applications de TAL arabe et non lié à un domaine particulier.
- L'optimisation : par optimisation on entend le processus de réduction au maximum de l'ambiguïté intrinsèque à la langue. Pour l'arabe, on ne voit qu'une seule issue « affiner le modèle », une modélisation fine est donc nécessaire pour répondre à cet objectif.
- La calculabilité : en effet, un modèle calculable est un modèle implémentable sur une machine. Notre objectif de départ étant la réalisation concrète d'un analyseur pour la langue arabe.

La modélisation linguistique consiste à classer les mots de la langue selon deux aspects, le premier est purement syntaxique en revanche le second est lexical. Chaque classe représente une catégorie grammaticale ou une classe syntaxique. Un ensemble de variables grammaticales est associé à chaque classe syntaxique. Ces variables représentent les traits linguistiques associés à ces classes.

Dans ce qui suit nous allons proposer notre modèle linguistique pour le traitement automatique de l'arabe. Le modèle construit sera considéré comme une plate forme commune et réutilisable par toutes les applications de TALN arabe. La démarche de

construction de ce modèle est relativement inspirée des travaux de modélisation réalisés par A. BERRENDONNER [BERR-1990], [LALL-1990] pour le français. En effet, sur la base de l'organisation que nous avons adoptée (voir chapitre précédent), nous allons construire un modèle en classes (classe dans le paradigme objet) pour le TALN arabe. Et pour valider ce dernier nous allons réaliser un analyseur morphologique pour la langue arabe. Nous utiliserons le langage UML[1] (en anglais **U**nified **M**odeling **L**anguage, « langage de modélisation unifié ») pour décrire nos différents modèles. UML est un langage graphique de modélisation des données et des traitements. « *C'est une formalisation très aboutie et non-propriétaire de la modélisation objet utilisée en génie logiciel.*» [http://fr.wikipedia.org]

Plusieurs raisons conduisent à préconiser l'utilisation d'UML:

a) sa normalisation par l'OMG (www.omg.org) (les spécifications sont accessibles gratuitement), c'est un langage universel pouvant servir de support pour tout langage orienté objet,
b) faciliter le dialogue entre concepteurs (linguistes et informaticiens),
c) possibilité d'utiliser le même atelier de génie logiciel (de l'expression des besoins à la génération de tout ou partie de l'application),
d) utilisation des principes et concepts objet (enrichir la démarche de conception, richesse, modularité, cohérence et rigueur…),
e) utilisation d'une représentation visuelle permettant la communication entre les concepteurs d'un même projet (linguistes et informaticiens),
f) utilisation d'une notation graphique simple, compréhensible même par des non informaticiens. Elle permet d'exprimer visuellement une solution objet, ce qui facilite la comparaison et l'évaluation de solutions,
g) son indépendance par rapport aux langages de programmation, aux domaines d'application et aux processus, en font un langage universel.

2. Les catégories grammaticales

Dans la langue arabe, il y a trois catégories de mots (formes) : le nom, le verbe et la particule.

Le premier niveau de catégorisation comprend donc:

[1] Langage de référence, retenu comme norme de modélisation par l'OMG (Object Management Group) et permettant d'augmenter la rigueur et la qualité des travaux quand on construit une application informatique. (pour des références à UML voir par exemple : [CHAR-2005], [BOOC-2000], [FOWL-2004], [BLAH-2005], [LAI-2004]…)

CAT = {Nom (N), Verbe (V), Particule (P)}

A cette réalité on peut associer un modèle conceptuel en classes composé de quatre classes avec des relations de spécialisation (ou généralisation). La figure 19 présente ce modèle.

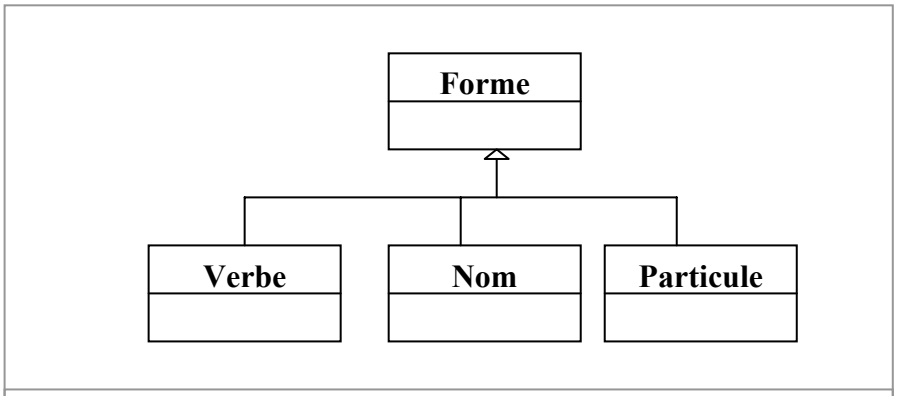


Figure 19 : Modèle objet d'une forme (mot) arabe

Cette catégorisation est trop grossière pour être utilisée par le système de traitement automatique de la langue arabe que nous envisageons de construire. Une catégorisation plus fine est donc nécessaire.

En se basant sur la documentation des grammairiens arabes et surtout sur [FOUA-1973], nous avons établi la classification suivante :

2.1 La catégorie nom (N)

Cette catégorie comprend un deuxième niveau de catégorisation. Elle se compose de neuf sous catégories (SCATN) :

- **M**asdar (NM),
- Nom **C**ommun (NC),
- Nom **P**ropre (NP),
- Nom de **T**emps et de lieu (NT),
- **A**djectif assimilé (NA),
- **P**articipe **A**ctif (NPA),
- **P**articipe **P**assif (NPP),
- Nom d'**I**nstrument (NI),
- **Q**ualificatif de supériorité (NQ)}.

SCATN = {NM, NC, NP, NT, NA, NPA, NPP, NI, NQ}

2.2 La catégorie particule (P)

Les particules sont des mots invariables, ils s'écrivent soit isolés, soit agglutinés[2] à d'autres formes dans une phrase. Les particules agglutinées sont de quatre types : les proclitiques, les préfixes, les suffixes et les enclitiques.

Les particules s'emploient avec le nom et avec le verbe et forment une liste finie (voir annexe C), on peut donc classer ces dernières en fonction de leurs emplois dans une phrase, il y a :

- celles qui s'emploient toujours avec un nom (PN),
- celles qui s'emploient toujours avec un verbe (PV),
- et éventuellement celles qui s'emploient indifféremment avec un nom ou un verbe (PNV).

SCATP = {PN, PV, PNV}

De cette description nous avons retenu deux propriétés pour les particules : un type et un emploi.

La propriété « Type » divise les particules en deux classes : celles qui sont agglutinées, elles se composent aussi de quatre classes (classe des proclitiques, classe des préfixes, classe des suffixes et éventuellement la classe des enclitiques) et celles qui sont isolées. La figure 20 montre ces différentes classes.

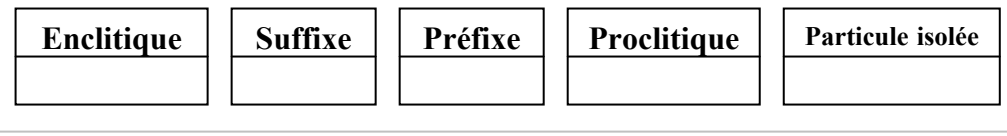


Figure 20: Modèle objet des particules

La propriété « Emploi » pose problème, en effet elle ne peut être ni une classe, ni une association. Donc, elle donne lieu nécessairement à un attribut d'une classe. Une première solution consiste à mettre cet attribut dans chaque classe de particules, mais sachant que plusieurs particules peuvent partager les mêmes propriétés, alors il est plus intéressant de regrouper les propriétés dans une classe à part et éviter ainsi la redondance

[2] L'écriture arabe a la particularité de l'agglutination des mots à l'intérieur d'une phrase. En effet les particules (articles, pronoms, prépositions…) s'écrivent attachées sous forme d'affixes et/ou clitiques aux noms et aux verbes.

des informations dans le modèle, ceci dit, la deuxième solution, que nous avons adoptée, consiste à introduire une nouvelle classe « Propriétés Particule», cette dernière regroupe toutes les propriétés concernant une particule, on trouve entre autres l'attribut « Emploi », l'attribut « Type » et bien d'autres. Une association de type agrégation (tous/partie) existe entre une particule et ses propriétés, La figure 21 montre le nouveau modèle conceptuel objet des particules.

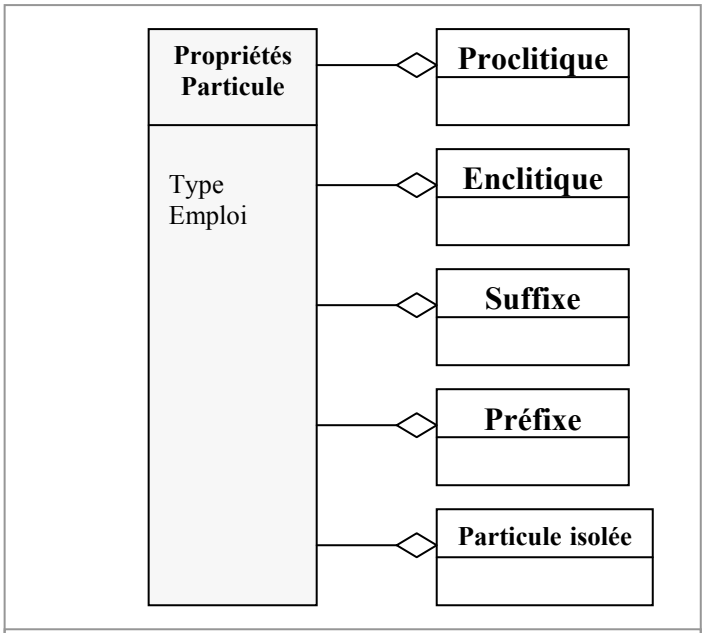


Figure 21: Nouveau modèle objet des particules

La reconnaissance de PN ou de PV dans une phrase peut déterminer la catégorie de la forme avec laquelle elle s'emploie.

Cette propriété, très intéressante, peut être exploitée par l'analyseur morphologique pour orienter ses investigations. En effet, les observations opérées sur les corpus de texte arabe attestent que la fréquence des particules est très élevée, d'où l'idée de faire précéder l'analyse morphologique d'une forme par une simple analyse de son contexte droit (voir la figure 22). Cette fonction d'aiguillage, que nous appelons aussi moniteur permettra donc de décider du type d'analyse que va subir la forme :

- analyser la forme comme un nom si la particule qui la précède s'emploie seulement avec un nom (PN),
- analyser la forme comme un verbe si la particule qui la précède s'emploie seulement avec un verbe (PV),
- effectuer les deux analyses autrement.

L'intérêt de ce moniteur est double : dans un premier temps il écarte en amont certains découpages « invalides » de la forme (plus de détails sur cette question seront abordés par la suite dans la partie description de l'analyseur) ; dans un deuxième temps il évite la consultation des deux lexiques[3] à la fois (dans le cas ou la particule qui précède la forme s'emploie seulement avec un nom ou un verbe), ce qui a pour conséquence un gain non négligeable en terme de temps d'analyse.

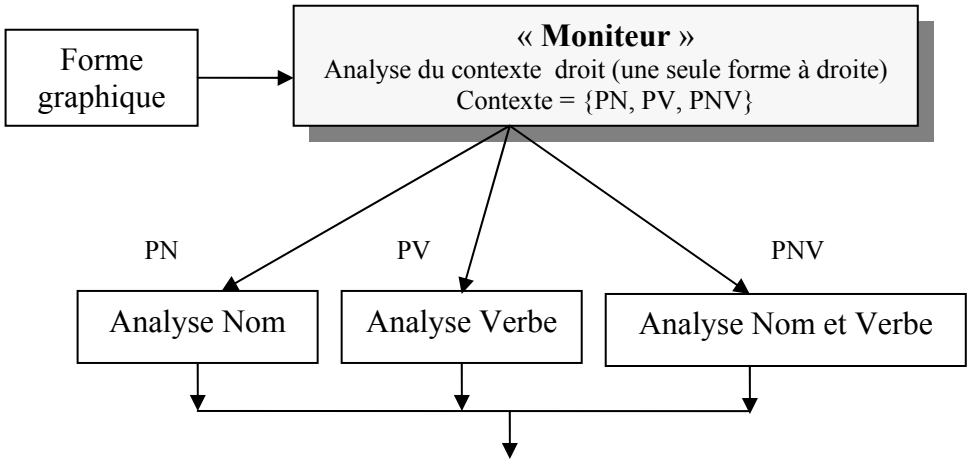


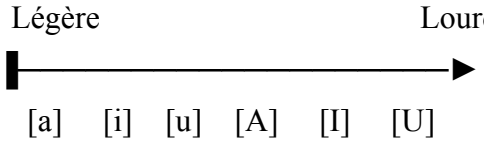
Figure 22 : Analyse de la forme selon son contexte droit

2.2.1 Catégories des particules pré et postfixés

Nous avons recensé 204 catégories de particules pré et postfixées. Cependant nous en avons retenu que 201 suite à la remarque de [BENH-1993] qui affirme que les catégories suivantes ne sont pas usitées en raison de leur lourdeur (ce mot est repris de [YAVA-1988], voir définition de la lourdeur par la suite) : Inaccompli énergique II : 11, 13 ; l'impératif énergique II : 7, 11 ; l'inaccompli énergique II : 7, 8 (voir l'annexe D pour la liste complète de ces catégories de particules).

Selon les grammairiens arabes, les voyelles sont classées en fonction de leurs légèretés ou de leurs lourdeurs. Par exemple, la voyelle « fatha : [a] » est la plus légère, par contre la voyelle longue « [U] » est la plus lourde (voir figure suivante). [YAVA-1988]

[3] Pour effectuer sa tâche l'analyseur consulte deux lexiques : un lexique pour les bases Nominales, et un autre pour les bases verbales. La description de ces deux lexiques se trouve dans la partie spécification de l'analyseur de ce chapitre.



Les voyelles selon le degré léger/lourd

Si la phonologie arabe a comme vocation la construction des séquences légères, elle tâche surtout d'éviter des séquences lourdes. En effet la majorité des transformations que subissent les formes graphiques sont dues principalement aux phénomènes de lourdeur. Autrement dit les grammairiens arabes considèrent la lourdeur comme étant la cause principale de nombreuses transformations.

Nous avons classé les particules pré et postfixés en trois classes :

- catégories des particules pré et postfixées qui s'emploient avec le nom (PN). Le tableau suivant donne un exemple de quelques particules de cette classe (voir l'annexe D pour la liste complète). Le code de la catégorie comprend trois champs, le premier champ désigne le type d'emploi de la particule (PN : cette particule s'emploie avec seulement un nom), le deuxième champ indique la position de la particule (P pour Proclitique, R pour Préfixe, S pour Suffixe et E pour Enclitique), le troisième champ désigne le code de la catégorie à proprement parler de la particule.

  Exemple : le code de la catégorie particule " بِ " est : PN_P_PRE

  Il signifie :

  - ✓ PN : cette particule s'emploie avec seulement un nom.
  - ✓ P : particule en position de proclitique.
  - ✓ PRE : c'est la catégorie préposition.

Particule	**Catégorie**	**Code catégorie**
بِ	Préposition	PN_P_PRE
ال	Article	PN_P_ART
ة	Suffixe du féminin singulier	PN_S_FS

ان	Suffixe du duel nominatif	PN_S_DN
ي	Nisba	PN_S_NISBA

- Catégories des particules pré et postfixées qui s'emploient avec un verbe (PV). Le tableau suivant donne quelques exemples de particules de cette classe (voir la liste complète dans l'annexe D). Le principe de la codification de ces catégories est le même que celui des catégories précédentes.

Particule	**Catégorie**	**Code catégorie**
س	Préfixe de l'inaccompli	PV_P_FUT
أ، ن، ي، ت	Préfixe inaccompli	PV_R_INA
تُ	Suffixe accompli 1er personne du singulier	PV_S_ACC1

- Catégories des particules pré et postfixées qui s'emploient avec le nom et le verbe (PNV). Le tableau suivant représente un exemple de quelques particules de cette classe (voir la liste complète dans l'annexe D).

Particule	**Catégorie**	**Code catégorie**
أ	Interrogatif	PNV_P_INT
و	Coordonnant	PNV_P_CORD
لَ	Corroborateur	PNV_P_CORB

2.2.2 Catégories des particules isolées

Nous avons recensé 22 catégories de particules isolées (voir la liste complète de ces catégories dans l'annexe E), elles se composent de trois classes :

- catégories des particules isolées qui s'emploient avec le nom (PN). Le tableau suivant présente un exemple de quelques particules de cette classe.

Particule	Catégorie	Code catégorie
من	Préposition	PN_PRE
إن	Ina et ses analogues	PN_INA

- Catégories des particules isolées qui s'emploient avec le verbe (PV). Le tableau suivant donne un exemple de quelques particules de cette classe.

Particule	Catégorie	Code catégorie
لم	Négation	PV_NEG
سوف	Future	PV_FUT
قد	Corroborateur	PV_CORB

- Catégories des particules isolées qui s'emploient avec le nom ou le verbe (PNV). Le tableau suivant présente quelques particules de cette classe.

Particule	Catégorie	Code catégorie
حتى	Coordonnant	PNV_CORD
هل	Interrogatif	PNV_INT

3. Les variables morpho-syntaxiques

Dix variables morpho-syntaxique caractérisent la langue arabe, On trouve celles qui sont propres aux formes nominales, celles qui sont propres aux formes verbales, et éventuellement, celles qui marquent indifféremment une forme nominale ou une forme verbale. Ces variables sont :

- ✓ CAS
- ✓ GENRE
- ✓ NOMBRE

- ✓ PERSONNE
- ✓ MODE
- ✓ ASPECT
- ✓ VOIX
- ✓ TRANSITIVITE
- ✓ FORME DERIVEE
- ✓ HUMAIN

a) Le CAS

On distingue : le nominatif, l'accusatif, le génitif et éventuellement le non marqué par le cas.

CAS = {Nominatif (N), Accusatif (A), Génitif (G), non marqué par le CAS[4] (NMC)}

La valeur NMC (non marqué par le cas) regroupe une liste finie de noms (voir l'annexe A). On trouve entre autres :

- les pronoms du nominatif ضمائر الرفع
- les pronoms de l'accusatif ضمائر النصب
- les démonstratifs أسماء الإشارة
- les relatifs الإسم الموصول
- les conditionnels إسم الشرط
- les interrogatifs إسم الإستفهام
- les noms de nombres (de 11 à 19 sauf 12) الأعداد المركبة
- quelques noms de lieu et de temps بعض الضروف
- noms verbaux إسم الفعل
- noms métonymiques كنايات
- nom de serment إسم قسم: تالله

[4] En arabe : مبني

Une liste (au nombre de 380) de tous ces noms se trouve dans l'annexe 'A' de cette thèse.

Dans un texte arabe, il y a lieu de constater que certains noms de cette catégorie sont agglutinés (affixés) à d'autres formes. Donc, ces noms possèdent les caractéristiques des particules, ce qui nous a conduit à les mettre dans une classe intermédiaire qui dérive (au sens paradigme objet) des deux classes (Nom et Particule). La reconnaissance des noms isolés ne pose aucun problème, par contre une opération de segmentation est nécessaire pour la reconnaissance des noms agglutinés.

La figure 23 illustre la classe des noms qui ne sont pas marqués par la variable « CAS », cette classe possède des propriétés de la classe Nom et en même temps des propriétés de la classe Particule. Autrement dit, la classe des noms (NMC) est obtenue par héritage multiple des deux sous classes supérieures (Nom et Particule).

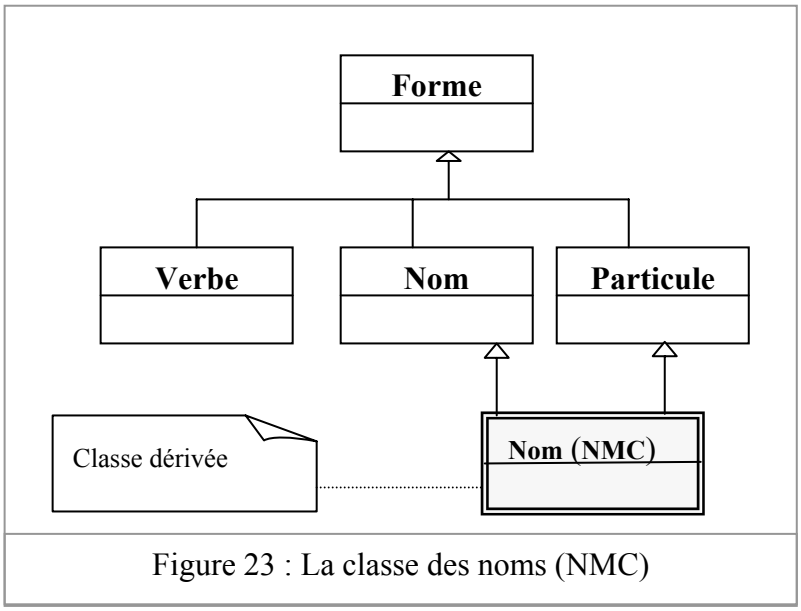


Figure 23 : La classe des noms (NMC)

Il faut noter toutefois que cette liste des noms (NMC) et la liste des particules non liées (au nombre de 74) sont regroupées dans un lexique spécifique appelé lexique des mots outils, il représente en quelque sorte un dictionnaire des constantes. Le rôle de ce lexique est décrit par la suite dans la section description de l'analyseur. Au niveau de notre modèle, une conséquence directe de ceci est que les deux classes, à savoir la classe des particules non liées et celle des noms (NMC) seront regroupées dans une classe unique nommée « Mot outil », ce qui va produire le changement suivant dans notre diagramme de classes (voir figure 24).

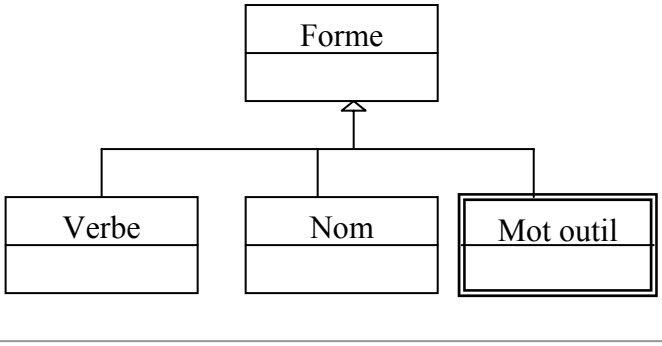


Figure 24 : La classe des mots outils

b) Le GENRE

On distingue : le masculin, le féminin et éventuellement le non marqué en genre.

GENRE = {Masculin (M), Féminin (F), non marqué (NMG)}

La valeur NMG (non marqué par le genre) concerne les noms qui peuvent avoir les deux genres indifféremment.

c) Le NOMBRE

On distingue : le singulier, le duel, le pluriel, le pluriel brisé et éventuellement le collectif.

NOMBRE = {Singulier (S), Duel (D), Pluriel (P), Pluriel Brisé (PB), Collectif (C)}

d) La PERSONNE

On distingue la première, la deuxième et la troisième personne.

PERSONNE = {Première personne (1P), Deuxième p. (2P), Troisième p.(3P)}

e) Le MODE

On distingue : l'indicatif, le subjonctif, l'apocopé, l'énergique I, l'énergique II et éventuellement un non marqué en mode.

MODE = {Indicatif (I), Subjonctif (S), Apocopé (A), Energique I (EI), Energique II (EII), non marqué (NMM) }

MODE = {I, S, A, EI, EII, NMM}

La valeur NMM (non marqué par le mode) concerne les verbes : à l'accompli, à l'impératif et seulement deux[5] cas pour l'inaccompli.

f) L'ASPECT

Il s'agit de : l'accompli, l'inaccompli et l'impératif.

ASPECT = {Accompli (A), Inaccompli (I), Impératif (M)}

Il faut noter que l'accompli n'a qu'une modalité.

g) La VOIX

Deux voix sont possibles : la voix active et la voix passive.

VOIX = {Active (A), Passive (P)}

Le passif arabe n'a pas d'impératif.

h) La TRANSITIVITE

On distingue sept cas :

- **I**ntransitive (I) : le verbe n'a pas besoin d'un complément,
- **T**ransitive **S**imple **D**irect (TSD) : le verbe utilise un seul complément direct,
- **T**ransitive **S**imple **I**ndirect (TSI) : le verbe utilise un seul complément indirect,
- **T**ransitive **D**ouble **D**irect **D**irect (TDDD) : le verbe utilise deux compléments directs,
- **T**ransitive **D**ouble **D**irect **I**ndirect (TDDI) : le verbe utilise deux compléments, le premier est direct, le deuxième est indirect,
- **T**ransitive **D**ouble **I**ndirect **D**irect (TDID) : le verbe utilise deux compléments, le premier est indirect, le deuxième est direct,
- **T**ransitive **T**riple (TT) : le verbe utilise trois compléments.

[5] Le premier cas : si le verbe est lié au suffixe « Noun niswa » « نون النسوة » pour la deuxième et troisième personne du féminin pluriel.

Le deuxième cas : si le verbe est lié au suffixe « Noun de corroboration » « نون التوكيد ».

TRANSITIVITE= {Intransitive (I), Transitive Simple Direct (TSD), Transitive Simple indirect (TSI), Transitive double (TDDD), Transitive double (TDDI), Transitive double (TDID), Transitive Triple (TT)}

Il y a lieu de noter qu'un ensemble de règles (ajout de la shadda, ajout d'une consonne au début de la forme, …) permet de faire passer un verbe d'une transitivité à une autre. Par contre, pour connaître la transitivité d'un verbe, et en dehors de toute règle, ces transitivités doivent être renseignées dans un dictionnaire.

i) Les FORMES DERIVÉES

Les grammairiens comptent quinze formes dérivées : une forme dérivée simple et quatorze autres formes dérivées. Il y a lieu de remarquer qu'un verbe n'a pas nécessairement les quatorze formes dérivées, et en l'absence de toute loi, ces formes doivent être renseignées dans un dictionnaire.

FORME DERIVÉE = {1, 2, 3, …15}

Pour un verbe trilitère arabe, on compte six formes de conjugaison et dix formes dérivées. Par ailleurs un verbe quadrilitère accepte une forme de conjugaison et trois formes dérivées. Une liste de toutes ces formes se trouve dans l'annexe 'B' de cette thèse.
A une racine arabe correspond un ou plusieurs verbes obéissant au vingt schèmes cités dans l'annexe 'B'. L'ensemble de ces verbes est appelé champ dérivationnel verbal.

j) HUMAIN

Trois possibilités caractérisent cette variable : humain, non humain et éventuellement un non marqué.

HUMAIN = {Humain (H), Non Humain (NH), non marqué (NMH)}

La valeur NMH (non marqué par la variable humain) concerne les noms ou les verbes qui peuvent prendre indifféremment les deux valeurs (H) ou (NH).

3.1 Affectation des variables morpho-syntaxiques pour les formes Nominales et verbales

Certaines variables syntaxiques marquent une forme nominale d'autres marquent une forme verbale.

Une forme nominale possède les variables :

- ✓ CAS
- ✓ GENRE
- ✓ NOMBRE
- ✓ HUMAIN

Par ailleurs une forme verbale possède les variables :

- ✓ GENRE
- ✓ NOMBRE
- ✓ PERSONNE
- ✓ MODE
- ✓ ASPECT
- ✓ VOIX
- ✓ TRANSITIVITE
- ✓ FORME DERIVEE
- ✓ HUMAIN

A la réalité une forme possède des variables morpho-syntaxiques (ou traits morpho-syntaxiques) nous avons associé une classe « Tms Forme », cette dernière contient donc l'ensemble des traits morpho-syntaxiques d'une forme. La figure 25 illustre cette classe.

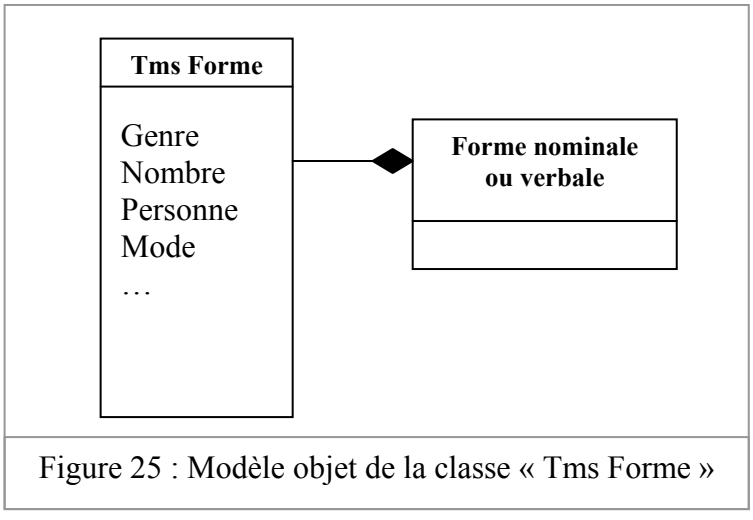


Figure 25 : Modèle objet de la classe « Tms Forme »

3.2 Les variables morpho-syntaxiques des particules pré et postfixées

Comme les formes Nominales et verbales, les particules pré et postfixées sont aussi marquées par des variables syntaxiques. Dans cette section nous allons, dans un premier temps, décrire ces variables avec leurs valeurs, qui en principe, sont les mêmes à quelques différences près. Toutes les variables sont censées avoir une valeur en plus « Indéterminée (X)» exprimant une indétermination de la variable. Dans un deuxième temps nous donnerons pour chaque classe de particule (proclitique, préfixe, suffixe, enclitique) les variables qu'elle peut avoir.

a) Le CAS

Pas de changement pour cette variable, elle garde les mêmes valeurs (que les formes nominales et verbales) plus la nouvelle valeur X.

CAS = {Nominatif (N), Accusatif (A), Génitif (G), non marqué (NMC), Indéterminée(X)}

b) Le GENRE

Pas de changement pour cette variable, elle garde les mêmes valeurs (que les formes nominales et verbales) plus la nouvelle valeur X.

GENRE = {Masculin (M), Féminin (F), non marqué (NMG), Indéterminée(X)}

c) Le NOMBRE

On distingue : le singulier (S), le duel (D), le pluriel (P), le duel ou pluriel (DP), et éventuellement une valeur indéterminée (X).

NOMBRE = {Singulier (S), Duel (D), Pluriel (P), Duel ou Pluriel (DP), Indéterminée (X)}

La valeur DP (duel ou pluriel) concerne les particules qui peuvent avoir les deux nombres (le duel ou le pluriel) indifféremment.

d) La PERSONNE

Pas de changement pour cette variable, elle garde les mêmes valeurs (que les formes nominales et verbales) plus la nouvelle valeur X.

PERSONNE = {Première personne (1P), Deuxième p. (2P), Troisième p. (3P), Indéterminée (X)}

e) Le MODE

Pas de changement pour cette variable, elle garde les mêmes valeurs (que les formes nominales et verbales) plus la nouvelle valeur X.

MODE = {Indicatif (I), Subjonctif (S), Apocopé (A), Energique I (EI), Energique II (EII), non marqué (NMM), Indéterminé (X) }

MODE = {I, S, A, EI, EII, NMM, X}

f) L'ASPECT

Pas de changement pour cette variable, elle garde les mêmes valeurs (que les formes nominales et verbales) plus la nouvelle valeur X.

ASPECT = {Accompli (A), Inaccompli (I), Impératif (M), Indéterminée (X)}

g) La VOIX

Pas de changement pour cette variable, elle garde les mêmes valeurs plus la nouvelle valeur X.

VOIX = {Active (A), Passive (P), Indéterminée (X) }

h) La TRANSITIVITE

Pour cette variable nous avons gardé seulement deux valeurs, à savoir, intransitive et transitive.

TRANSITIVITE= {Intransitive (I), Transitive (T)}

i) Les FORMES DERIVÉES

Cette variable est tout simplement annulée, les particules pré et postfixées ne sont pas marquées par cette variable.

j) HUMAIN

Cette variable est annulée, les particules pré et postfixées ne sont pas marquées par cette variable.

3.3 Affectation des variables pour les particules pré et postfixées

Le tableau 10 regroupe les différentes classes des particules pré et postfixées (ligne) et les différentes variables syntaxiques sur la colonne. Avec Tableau [i, j] = X, signifie que la particule de la ligne « i » est marquée par (possède) la variable de la colonne « j »

	Cas	**Genre**	**Nombre**	**Personne**	**Mode**	**Aspect**	**Voix**	**Transitivité**	**Forme dérivée**	**Humain**
Préfixe		X	X	X	X	X	X			
Suffixe Verbe		X	X	X	X	X	X			
Suffixe Nom	X	X	X							
Proclitique Verbe					X	X				
Proclitique Nom	X									
Enclitique Verbe		X	X	X				X		
Enclitique Nom		X	X	X						

Tableau 10 : Affectation des variables pour les particules pré et postfixés

Pour la modélisation de ces variables nous avons ainsi construit une classe « Tms Particule », cette dernière contient les attributs : genre, nombre… la figure 26 illustre cette classe.

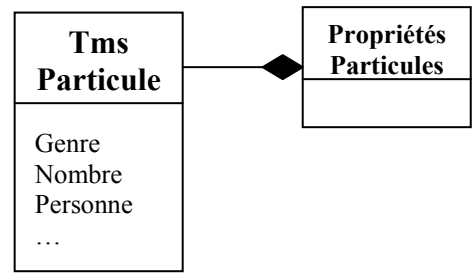


Figure 26 : Modèle objet de la classe « Tms Particule »

4. Spécification de l'analyseur

4.1 Les objets de base

Consonne = { ا، ء، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل، م، ن، ه، ة، و، ي}

Voyelle = {َ، ِ، ُ، ْ، ً، ٍ، ٌ، ّ، ا، و، ي}

Glide (voyelle longue) = { ا، و، ي }

Voyelle sans Glide (VSG) = *Voyelle \ Glide* = { َ، ِ، ُ، ْ، ً، ٍ، ٌ، ّ }

Alphabet = *Consonne* ∪ *Voyelle*

Chiffre = { 0, 1, 2,... 9}

AC = *Alphabet* ∪ *Chiffre*

Séparateurs : C'est l'ensemble de tous les autres caractères utilisables pour l'écriture de l'arabe.

Séparateur = {Ponctuation, Blanc, Opérateurs, Caractères spéciaux}

SB = *Séparateur*\ {Blanc}

Ponctuation = {Virgule, Point virgule, Point d'Exclamation, Point d'Interrogation, Point, Deux Points }

Opérateurs = {+, -, /, *, <, >, =, ≤, ≥}

Caractères spéciaux = {#, (,), {, }, …, -, @, \, [,], ‘}

Un mot non voyellé : c'est une suite contiguë finie de symboles pris dans l'ensemble des consonnes (noté *Consonne*).

Un mot voyellé : c'est une suite contiguë finie de symboles pris dans l'ensemble *Alphabet*.

Un nombre : est une suite contiguë de chiffres, auxquels peuvent se mêler des points et une virgule.

Un mot simple : peut être un mot non voyellé ou un mot voyellé.

Une forme : une forme peut être un mot simple, un mot composé ou éventuellement un *SB* (*Séparateur*\ {Blanc}).

Unité lexicale : elle permet de réunir les différentes formes issues d'un même mot simple, représentant la même catégorie grammaticale et un même concept.

Le représentant d'une unité lexicale (ou lemme) : peut être dans l'ordre en fonction de l'existence de la forme :

- la forme au masculin singulier, sinon la forme au masculin pluriel, sinon la forme au féminin singulier, sinon la forme au féminin pluriel.

- La forme conjuguée à l'accompli (troisième personne du singulier) dans le cas d'un verbe.
- La forme unique.

4.2 Les catégories grammaticales

Les catégories grammaticales permettent d'identifier les formes rencontrées dans un texte arabe. On distingue :

- VER Verbe
- NM Masdar,
- NC Nom commun
- NP Nom propres
- NTL Nom de temps et de lieu
- NA Adjectif assimilé
- NPA Participe actif
- NPP Participe passif
- NI Nom d'instrument

(NM, NC, NP, NTL, NA, NPA, NPP, NI) → Nom

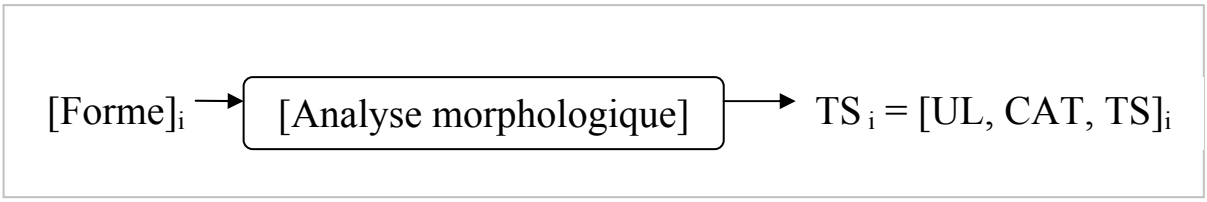
- NQ Qualificatif de supériorité
- PAR_i Particule pré et postfixées (201 catégories)
- MO_i Mots outils (22 catégories)
- NBN Nombre
- ABR Abréviation (pour classer les abréviations, exemple : ض.إ.= الضمان الإجتماعي),
- SEP Séparateur (pour classer les ponctuations orthographiques et éventuellement quelques caractères spéciaux, voir la liste dans la page précédente),
- INC Inconnue (pour classer tout ce qui n'appartient pas aux catégories précédentes)

La liste complète de toutes ces catégories se trouve dans les annexes de cette thèse.

4.3 Solution morphologique

A l'issue de l'analyse morphologique d'une forme, l'analyseur produit un triplet (unité lexicale, catégorie grammaticale, ensemble de traits syntaxiques), ce triplet noté TS

= (UL, CAT, TS) représente une solution morphologique hors-contexte calculée dans le modèle linguistique utilisé.



La solution TS_i dépend du modèle linguistique utilisé, autrement dit, elle est calculée à partir des données linguistiques manipulées par le programme d'analyse dans le modèle linguistique utilisé.

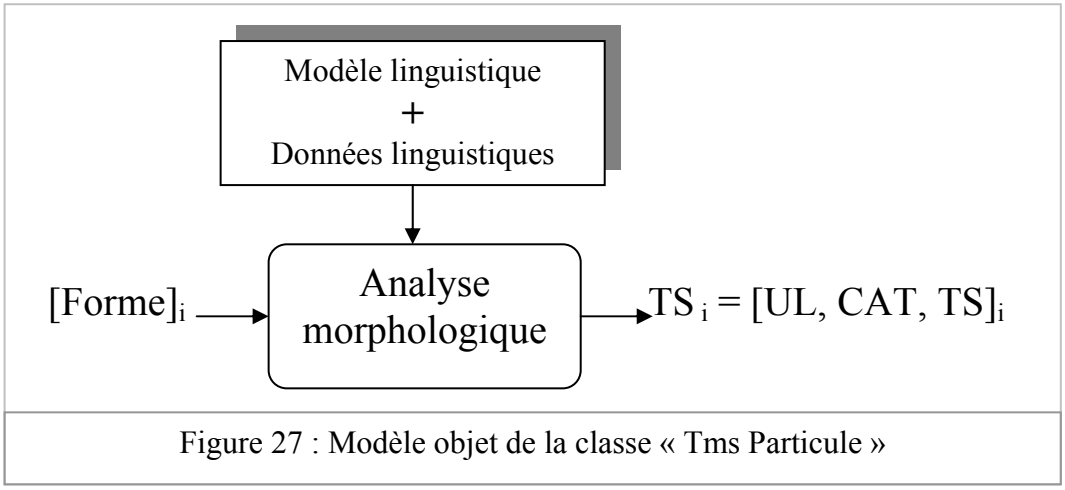


Figure 27 : Modèle objet de la classe « Tms Particule »

Trois cas se posent lors de l'analyse morphologique :

a) Le cas où il y a une solution morphologique unique pour une forme.

$$TS_i = TS_{i1}$$

[Forme] $_i$ *Possède une solution unique* $\Leftrightarrow$ $|TS_i| = 1$ où $|TS_i|$ désigne cardinal de TS_i.

b) Le cas où il n'y a pas de solution morphologique unique pour une forme : la solution morphologique est représentée par plusieurs triplet TS.

$$TS_i = TS_{i1}, TS_{i2}, \ldots TS_{in}$$

Dans ce cas on parle de forme ambiguë, autrement dit, une forme est ambiguë (appelée aussi forme homographe) si l'analyseur lui associe plusieurs triplets TS.

[Forme] $_i$ *est ambiguë* $\Leftrightarrow$ $|TS_i| > 1$ où $|TS_i|$ désigne cardinal de TS_i.

c) Le cas où l'analyseur manque d'informations pour traiter une forme : dans ce cas la solution morphologique est inexistante et on parle d'une forme inconnue par l'analyseur (ou non reconnue par l'analyseur).

$$[\text{Forme}]_i \ \underline{\textit{est inconnue}} \Leftrightarrow |TS_i| = 0 \quad \text{où } |TS_i| \text{ désigne cardinal de } TS_i.$$

4.4 Forme inconnue

Une forme peut être inconnue pour deux raisons :

- Soit la forme n'appartient pas à la langue (faute de frappe, faute d'orthographe).
- Soit la forme appartient à la langue mais l'analyseur ne peut pas lui attribuer une solution morphologique. Cette non-reconnaissance de la forme provient : soit d'une mauvaise segmentation de la forme, soit des données linguistiques incomplètes ou erronées. Le cas d'une mauvaise segmentation peut être résolu par la correction du programme de segmentation ou par l'amélioration des outils utilisés (le lexique par exemple), c'est le rôle de l'étape indispensable de la mise au point de l'analyseur, en génie logiciel on parle de l'étape de test du logiciel (contrôler le bon fonctionnement du logiciel).

5. L'analyseur morphologique

L'entrée de l'analyseur morphologique est un texte arabe non voyellé, non voyellé avec 'Shadda', partiellement voyellé ou éventuellement un texte complètement voyellé (pour une définition de ces différents types de textes arabes, voir le chapitre II pages 31 et 32 de cette thèse). Un premier traitement opéré par l'analyseur morphologique consiste à segmenter le texte d'entrée en formes. Le séparateur blanc étant la marque des frontières des formes, ce traitement ne devrait donc poser aucun problème. A la suite de l'identification de chaque forme, l'analyseur doit dans un deuxième temps dévoyeller chaque forme. La dévoyellation d'une forme est l'opération qui consiste à séparer les voyelles et les consonnes d'une forme de telle sorte à obtenir deux formes, une forme contenant les consonnes (schéma consonantique) de la forme de départ, l'autre contenant les voyelles (schéma vocalique). L'opération de dévoyellation permet le traitement, par l'analyseur, des différents types de textes en entrée. Il est évident que cette opération n'a de sens que si la forme en entrée est voyellée ou partiellement voyellée. La partie contenant les consonnes (schéma consonantique) va servir comme entrée pour les autres opérations qui seront appliquées par l'analyseur.

L'algorithme qui réalise la procédure de dévoyellation d'une forme se trouve dans l'annexe E.

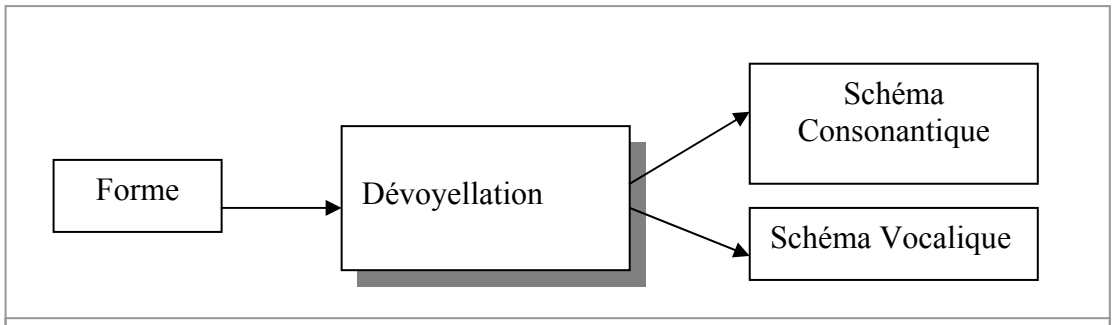


Figure 28 : Etape 1 : Dévoyellation de la forme

Après dévoyellation de la forme, l'analyseur morphologique effectue un accès au lexique des mots outils (le contenu de ce lexique est déjà décrit dans la partie modélisation dans les sections précédentes) pour savoir si la forme en question est un mot outil ou non (le lexique des mots outils regroupe toutes les particules et les noms qui ne sont pas marqués par la variable syntaxique "CAS", chaque entrée de ce lexique représente un mot outil, elle comprend entre autres, le schéma consonantique, le schéma vocalique et la catégorie du mot outil en question). La fréquence très élevée de ces mots outils dans un texte arabe, et leur nombre très réduit (nous avons recensé 380 mots outils), nous a conduit à les regrouper dans un lexique à part. Une conséquence à ceci est le gain en efficacité.

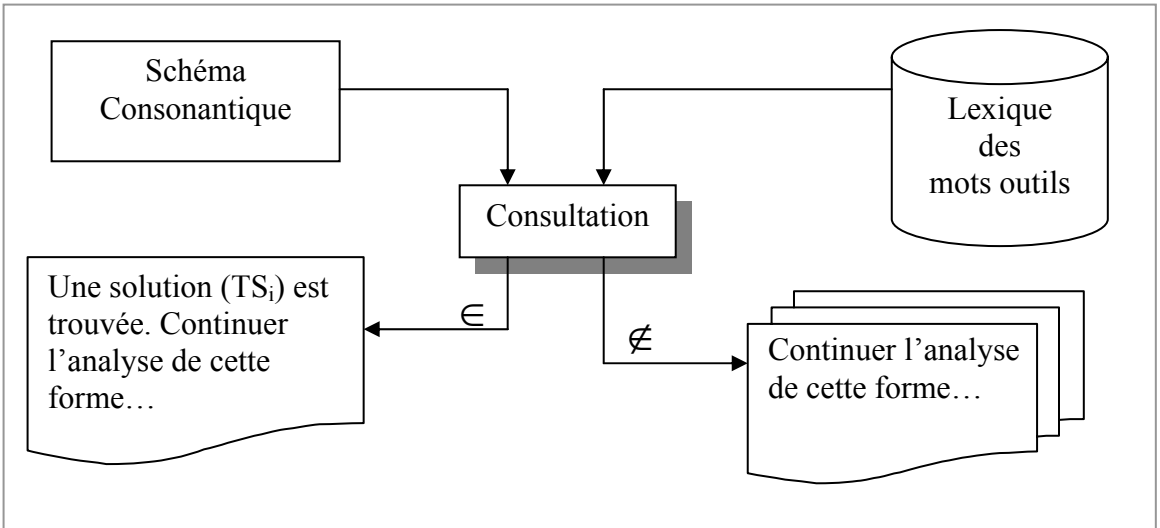


Figure 29 : Etape 2 : Consultation du lexique des mots outils

Si la recherche dans le lexique des mots outils n'aboutit pas, alors l'analyseur est devant quatre cas possibles pour cette forme :

- soit c'est une forme nominale,
- soit c'est une forme verbale,
- soit c'est une forme nominale et verbale,
- soit c'est une forme inconnue.

L'analyseur morphologique réalise un découpage (segmentation) du schéma consonantique de toutes les façons possibles pour extraire les différents segments qui composent le schéma consonantique en question. Ces segments vont permettre la détermination de la solution morphologique (le triplet TS_i) qui représente donc, les interprétations linguistiques hors-contexte, dans le modèle utilisé. Une solution morphologique TS_i est composée d'un représentant de l'unité lexicale, d'une catégorie grammaticale, et d'un ensemble de valeurs grammaticales. Pour réaliser sa tâche, l'analyseur morphologique utilise en plus du modèle linguistique, un ensemble de données consignées dans un dictionnaire de « bases », d'un ensemble de clitiques et d'affixes organisés en tables et éventuellement des matrices de compatibilité des clitiques, affixes et bases. A chaque base (dans le dictionnaire des bases) est associé un ensemble de propriétés et de traits linguistiques, et à chaque clitique ou affixe est associé un ensemble de propriétés et de traits linguistiques.

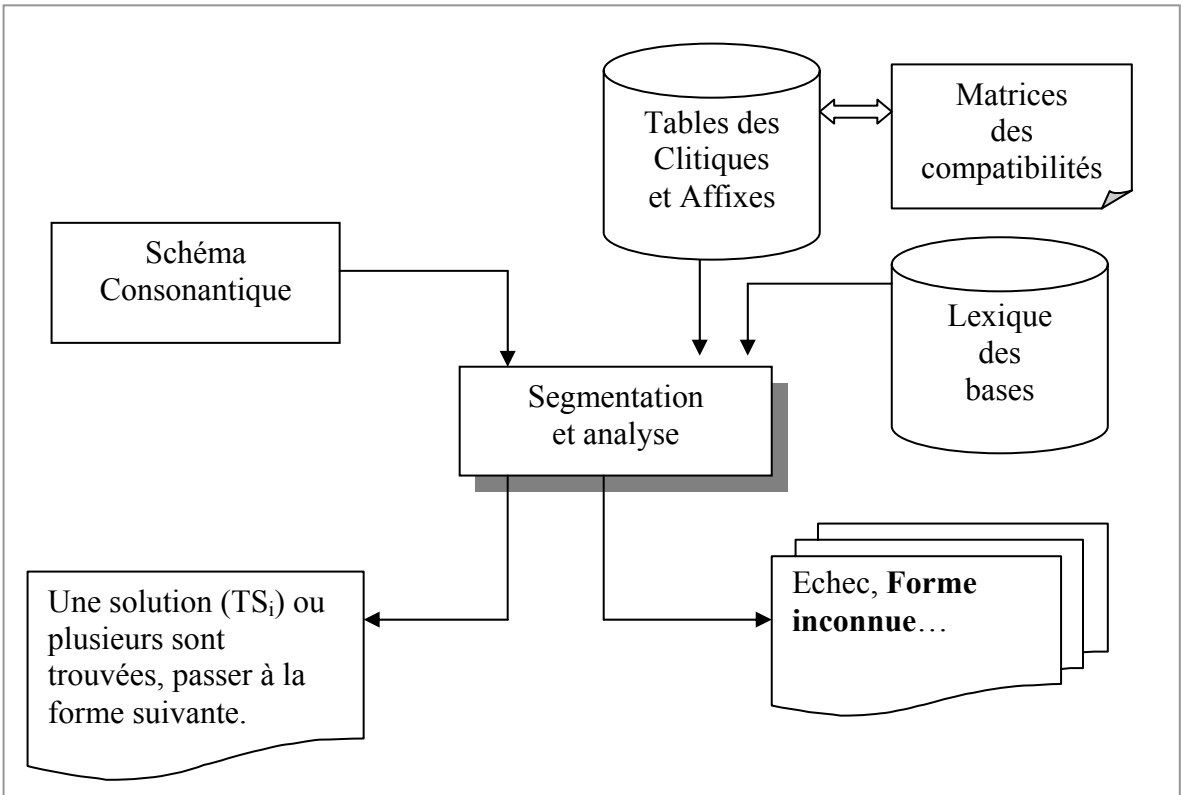


Figure 30 : Etape 3 : Segmentation et analyse de la forme

En se basant sur le modèle du mot graphique arabe décrit dans le chapitre II de cette thèse, nous allons proposer une modélisation en classe (dans le paradigme objet) pour les deux formes à savoir la forme nominale et la forme verbale. Le langage UML est utilisé pour l'élaboration des différents diagrammes.

5.1 Forme nominale (FN)

Une forme nominale est un mot dont le lexème est un nom. Elle se compose de quatre objets différents (proclitique, base nominale, suffixe, enclitique) :

- une particule ou plus parmi (ب، ل، ك، ف : fa, ka, li, bi), elle peut être absente (proclitique).
- Un article (ال : al), il marque la détermination. Il est omis quand le nom est indéterminé, déterminé par annexion ou un nom propre (proclitique).
- Un radical nominal (ou base nominale).
- Une terminaison nominale (flexion en genre et en nombre : suffixe).
- Un pronom possessif se réfère à un nom qui le précède s'il existe (enclitique).

La figure 31 montre le modèle objet d'une forme nominale.

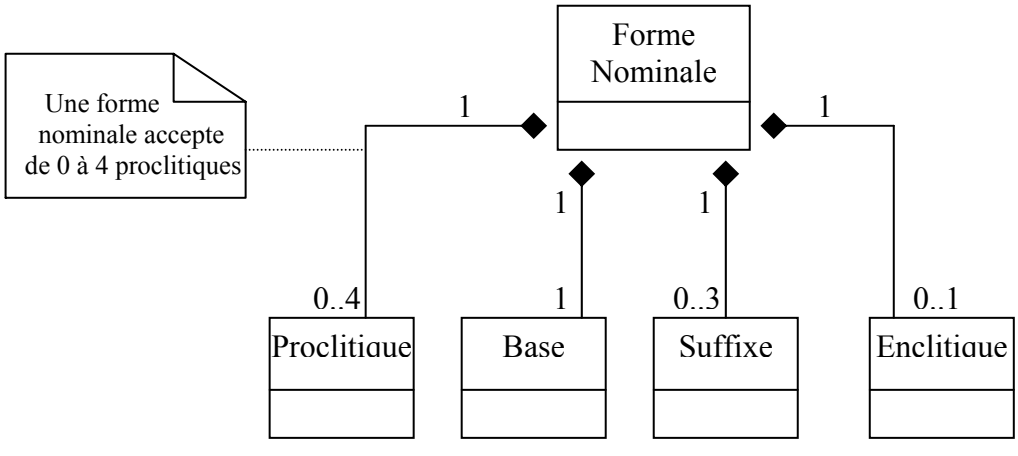


Figure 31: Modèle objet d'une forme nominale non voyellée

On peut décrire une forme nominale par la grammaire hors-contexte GN suivante :

$$GN = \{V_T, V_N, S, P\}$$

- ✓ V_T : représente le vocabulaire terminal fini,
- ✓ V_N : vocabulaire non terminal (ou auxiliaire) fini,
- ✓ S : l'axiome,
- ✓ P : représente un nombre fini de règles de production.

Avec:

- V_T = {liste des proclitiques simples, liste des bases, liste des suffixes simples, liste des enclitiques}
- V_N = {Forme nominale, Proclitiques, Proclitique, Base, Suffixes, Suffixe, Enclitique}
- S = {Forme nominale}
- P = règles de production :

1- <Forme nominale> → <Proclitiques> <Base> <Suffixes> <Enclitique>

2- <Proclitiques> → <Proclitiques> <Proclitique> | <Proclitique>

3- <Proclitique>→{liste finie des proclitiques simples[6] (classe fermée[7]) au nombre de 8} | ε

4- <Base> → {liste des bases (classe ouverte)}

5- <Suffixes> → <Suffixes> <Suffixe> | <Suffixe>

6- <Suffixe> →{liste finie des suffixes simples[8] (classe fermée) au nombre de 14} | ε

7- <Enclitique> → {liste finie des enclitiques (classe fermée) au nombre de 16} | ε

Le langage engendré par cette grammaire est un langage hors-contexte (ou algébrique), et le mécanisme de reconnaissance de ce langage est un automate d'états finis non déterministe, il est défini par le quintuplet (V_T, Q, I, T, F)

- ✓ V_T : représente l'alphabet terminal,

[6] Proclitiques qui s'emploient avec le nom et éventuellement ceux qui s'emploient avec le nom et le verbe indifféremment.

[7] Une classe fermée : est une classe dont l'énumération exhaustive des éléments est possible ; par opposition à une classe ouverte, dont le nombre d'éléments est forcément fini, mais pour laquelle une énumération exhaustive de ces éléments est impossible (ces classes sont continuellement enrichies : évolution de la langue, incorporation de mots d'origine étrangère…)

[8] Suffixes qui s'emploient avec le nom et éventuellement ceux qui s'emploient avec le nom et le verbe indifféremment.

- ✓ Q : est un ensemble fini dont les éléments représentent les états de l'automate,
- ✓ I : désigne l'état initial de l'automate,
- ✓ T : désigne l'état terminal de l'automate,
- ✓ F : désigne l'ensemble des transitions

Avec :

- ✓ V_T = {liste des proclitiques, liste des bases, liste des suffixes, liste des enclitiques}
- ✓ $Q = \{q_0, q_1, q_2, q_3, q_4\}$
- ✓ $I = q_0$
- ✓ $T = q_4$
- ✓ $F = \{(q_0, p_i, q_1), (q_1, p_i, q_1), (q_1, b_i, q_2), (q_2, s_i, q_3), (q_3, s_i, q_3), (q_3, e_i, q_4)\}$

$p_i \in$ { liste des proclitiques} $\cup \{\varepsilon\}$

$b_i \in$ { liste des bases }

$s_i \in$ { liste des suffixes} $\cup \{\varepsilon\}$

$e_i \in$ { liste des enclitiques}$\cup \{\varepsilon\}$

Cet automate peut être représenté par le graphe orienté étiqueté (appelé aussi graphe de transition) de la figure 32 :

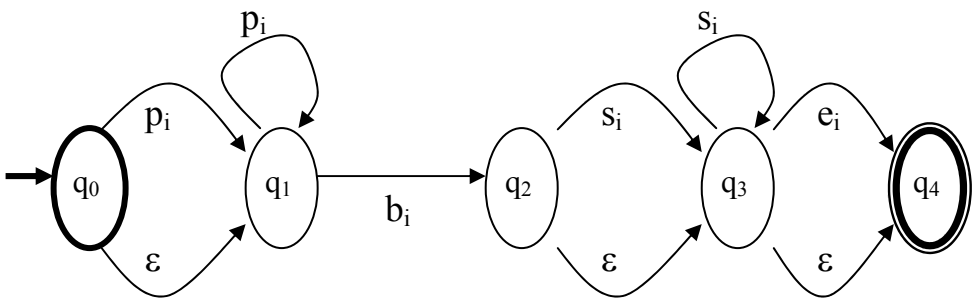


Figure 32 : Automate de reconnaissance d'une forme nominale

En examinant cette grammaire on s'aperçoit qu'elle génère du bruit (des formes non attestables par la langue). En effet, à cause de la multiplicité supérieure à 1 du côté proclitique, la règle de production numéro deux produit des proclitiques composés dans n'importe quel ordre, ce qui conduit à générer des proclitiques non valides.

Ce qui est dit pour la règle deux est aussi valable pour la règle numéro cinq.

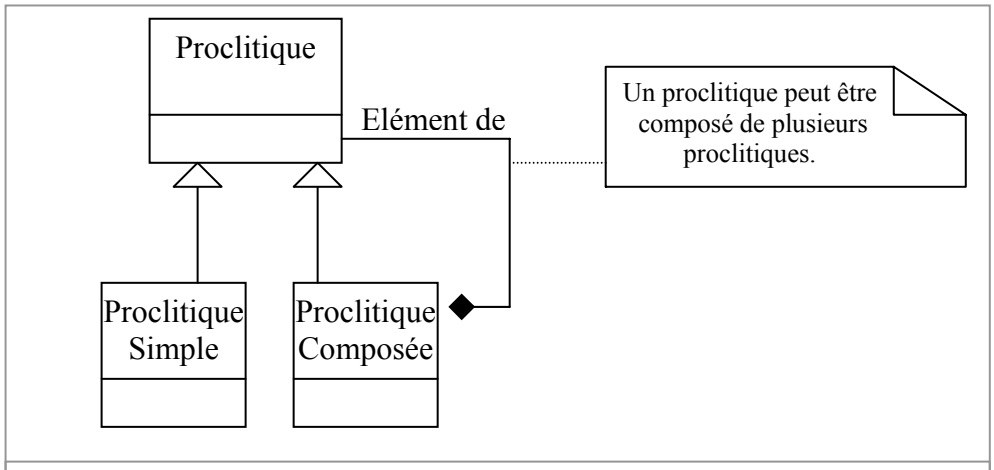


Figure 33 : Composition de proclitiques

Pour remédier au problème du bruit, autrement dit éliminer les règles de production deux et cinq, nous avons recensé tous les proclitiques composés (respectivement tous les suffixes composés) possibles et valides. Ces derniers vont avoir maintenant le statut d'un objet proclitique simple (respectivement un objet suffixe simple). Nous avons donc fusionné les deux objets (simple et composé) dans un seul objet, ce qui à pour conséquence la modification de la multiplicité du côté de l'objet proclitique (respectivement de l'objet suffixe). La nouvelle multiplicité devient donc [0..1]. Le modèle objet d'une forme nominale sera de la figure 34 :

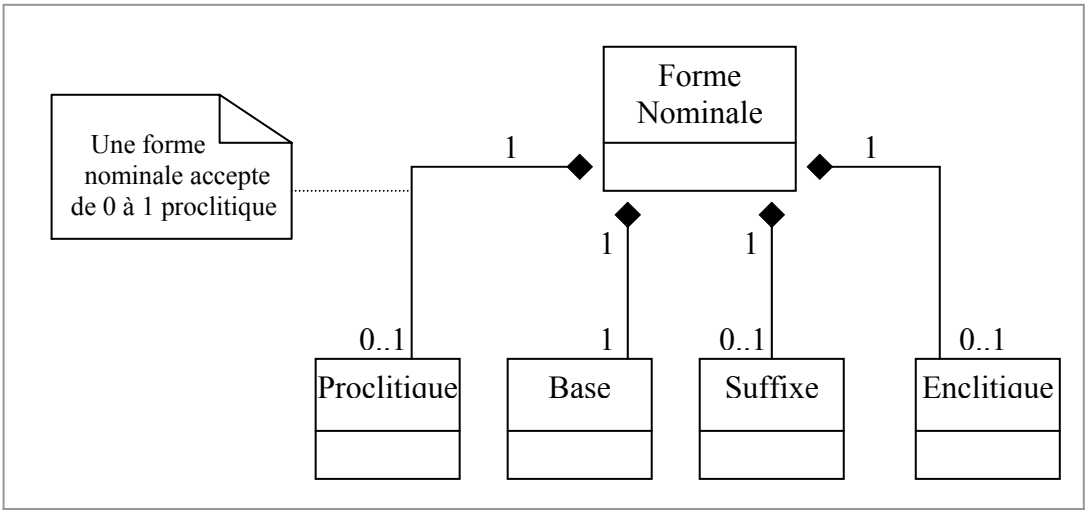


Figure 34 : Modèle objet (modifié) d'une forme nominale non voyellée

Par conséquent la nouvelle grammaire sera donc la suivante:

$$GN = \{V_T, V_N, S, P\ \}$$

Avec:

- V_T = {liste des proclitiques, liste des bases, liste des suffixes, liste des enclitiques}
- V_N = {Forme nominale, Proclitique, Base, Suffixe, Enclitique}
- S = {Forme nominale}
- P = règles de production :

1- <Forme nominale> → <Proclitique> <Base> <Suffixe> <Enclitique>

2- <Proclitique> → {liste finie (classe fermée) au nombre de 63} | ε

3- <Base> → {liste (classe ouverte)}

4- <Suffixe> → {liste finie (classe fermée) au nombre de 46} | ε

5- <Enclitique> → {liste finie (classe fermée) au nombre de 16} | ε

Liste des proclitiques = {proclitiques simples}∪ {proclitiques composés}

Liste des suffixes = {suffixes simples}∪ {suffixes composés}

L'automate de reconnaissance sera défini donc par le quintuplet (V_T, Q, I, T, F)

- ✓ V_T = {liste des proclitiques, liste des bases, liste des suffixes, liste des enclitiques}
- ✓ Q = $\{q_0, q_1, q_2, q_3, q_4\}$
- ✓ I = q_0
- ✓ T = q_4
- ✓ F = $\{(q_0, p_i, q_1), (q_1, b_i, q_2), (q_2, s_i, q_3), (q_3, e_i, q_4)\}$

$p_i \in$ { liste des proclitiques} $\cup$ {ε}

$b_i \in$ { liste des bases }

$s_i \in$ { liste des suffixes} $\cup$ {ε}

$e_i \in$ { liste des enclitiques}$\cup$ {ε}

Cet automate peut être représenté par le graphe de transition de la figure 35 :

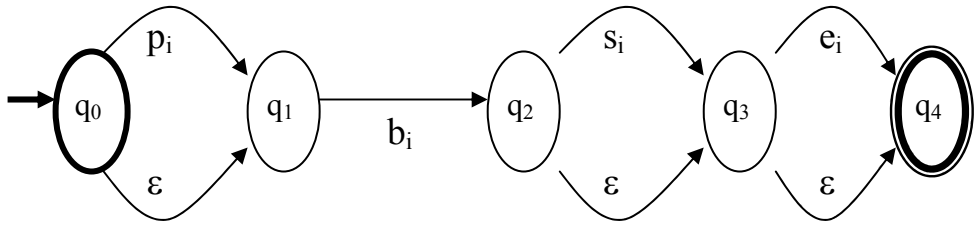


Figure 35 : L'automate (modifié) de reconnaissance d'une forme nominale

5.2 Forme verbale (FV)

Par opposition à une forme nominale, une forme verbale est un mot dont le lexème est un verbe. Elle se compose de cinq objets :

- une particule ou plus parmi (أ، و، ف، ل، س: ', wa, fi, la, sa), elle peut être absente (proclitique).
- Un préfixe parmi (أ، ن، ي، ت :', na, ya, ta), il marque les verbes à l'inaccompli. Il est omis quand le verbe est à l'accompli.
- Un radical verbal (ou base verbale).
- Une terminaison verbale (suffixe de conjugaison).
- Un ou deux pronoms objets (enclitique) elle peut être absente.

La figure 36 illustre le modèle objet d'une forme verbale.

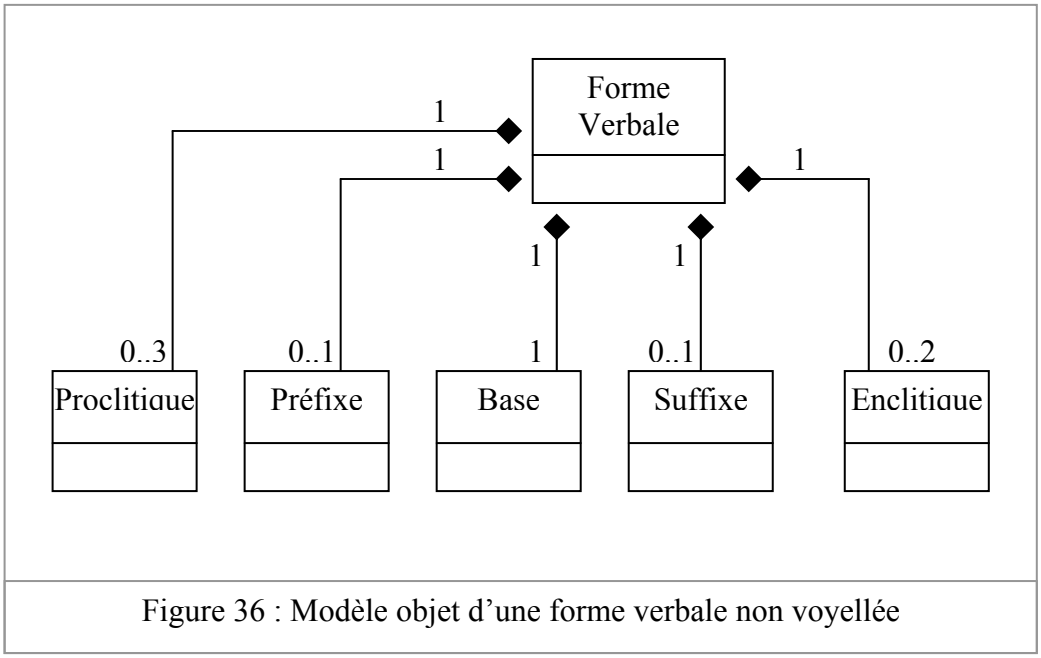


Figure 36 : Modèle objet d'une forme verbale non voyellée

Et éventuellement, une grammaire de génération d'une forme verbale peut se présenter de la façon suivante :

$$GV = \{V_T, V_N, S, P\ \}$$

Avec:

- V_T = {liste des proclitiques simples, liste des bases, liste des suffixes, liste des enclitiques simples, liste des préfixes}
- V_N = {Forme verbale, Proclitiques, Proclitique, Préfixe, Base, Suffixe, Enclitiques, Enclitique}
- S = {Forme verbale}
- P = règles de production :

1- <Forme verbale> → <Proclitiques><Préfixe><Base><Suffixe><Enclitiques>
2- <Proclitiques> → <Proclitiques> <Proclitique> | <Proclitique>
3- <Proclitique> → {liste finie[9] (classe fermée) au nombre de 7} | ε
4- <Préfixe> → {liste finie (classe fermée) au nombre de 10} | ε
5- <Base> → {liste finie (classe ouverte) }
6- <Suffixe> → {liste finie (classe fermée) au nombre de 87}
7- <Enclitiques> → <Enclitiques><Enclitique> | <Enclitique>
8- <Enclitique> → {liste finie[10] (classe fermée) au nombre de 23} | ε

Cette grammaire génère aussi du bruit à cause des deux règles de production numéro deux et sept. La démarche de la solution adoptée est similaire à la précédente (voir la section précédente forme nominale).

5.3 Forme unifiée (FU)

Dans un texte arabe, et avant toute analyse, on ne peut savoir à priori si la forme est nominale ou bien verbale, on est amené alors à définir un modèle général pour une

9 Proclitiques qui s'emploient avec le verbe et éventuellement ceux qui s'emploient avec le nom et le verbe indifféremment.

10 Enclitiques qui s'emploient avec le verbe et éventuellement ceux qui s'emploient avec le nom et le verbe indifféremment.

forme, ce dernier devra permettre d'analyser une forme quelconque. Il s'agit donc d'unifier les deux modèles, à savoir le modèle de la forme nominale et le modèle de la forme verbale dans un seul modèle qui prend en charge l'un et l'autre, c'est ce que nous avons nommé le modèle unifié de la forme. En réalité ce modèle n'est autre que le modèle de la forme verbale, puisqu'il permet de prendre en compte tous les objets des deux modèles précédents à une différence près qui consiste à regrouper les listes des proclitiques, des suffixes et des enclitiques correspondant aux deux formes en question. La figure 37 montre ce modèle.

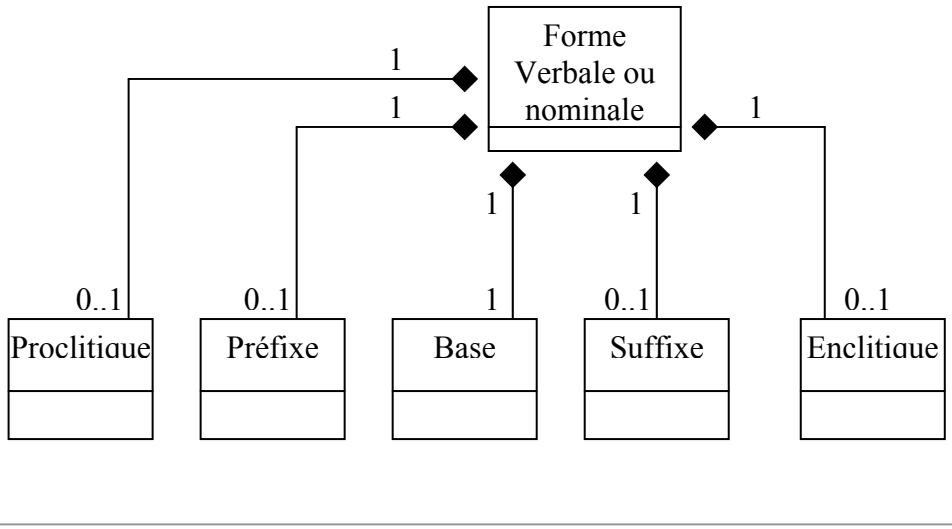


Figure 37 : Modèle objet de la forme unifiée

La grammaire de génération de la forme unifiée peut se présenter de la façon suivante :

$$GNV^{11} = \{V_T, V_N, S, P\ \}$$

Avec:

- $V_T = \{$liste des proclitiques, liste des bases, liste des suffixes, liste des enclitiques, liste des préfixes$\}$
- $V_N = \{$Forme unifiée, Proclitique, Préfixe, Base, Suffixe, Enclitique$\}$
- $S = \{$Forme unifiée$\}$
- P = règles de production :

1- <Forme unifiée> → <Proclitique><Préfixe><Base><Suffixe><Enclitique>

[11] En terme d'opérations entre grammaires, GNV n'est autre que l'union des deux grammaires GN et GV. $GNV = GN \cup GV$

2- <Proclitique> → {liste finie[12] (classe fermée) au nombre de 72} | ε

3- <Préfixe> → {liste finie (classe fermée) au nombre de 10} | ε

4- <Base> → {liste finie (classe ouverte)}

5- <Suffixe> → {liste finie[13] (classe fermée) au nombre de 133}

6- <Enclitique> → {liste finie[14] (classe fermée) au nombre de 35} | ε

L'automate de reconnaissance de cette grammaire sera défini donc par le quintuplet (V_T, Q, I, T, F)

- ✓ V_T = {liste des proclitiques, liste des préfixes, liste des bases, liste des suffixes, liste des enclitiques}
- ✓ Q = $\{q_0, q_1, q_2, q_3, q_4, q_5\}$
- ✓ I = q_0
- ✓ T = q_5
- ✓ F = $\{(q_0, p_i, q_1), (q_1, r_i, q_2), (q_2, b_i, q_3), (q_3, s_i, q_4), (q_4, e_i, q_5)\}$

$p_i \in$ { liste des proclitiques} $\cup \{\varepsilon\}$

$r_i \in$ { liste des préfixes} $\cup \{\varepsilon\}$

$b_i \in$ { liste des bases }

$s_i \in$ { liste des suffixes} $\cup \{\varepsilon\}$

$e_i \in$ { liste des enclitiques}$\cup \{\varepsilon\}$

Cet automate peut être représenté par le graphe de transition de la figure 38 :

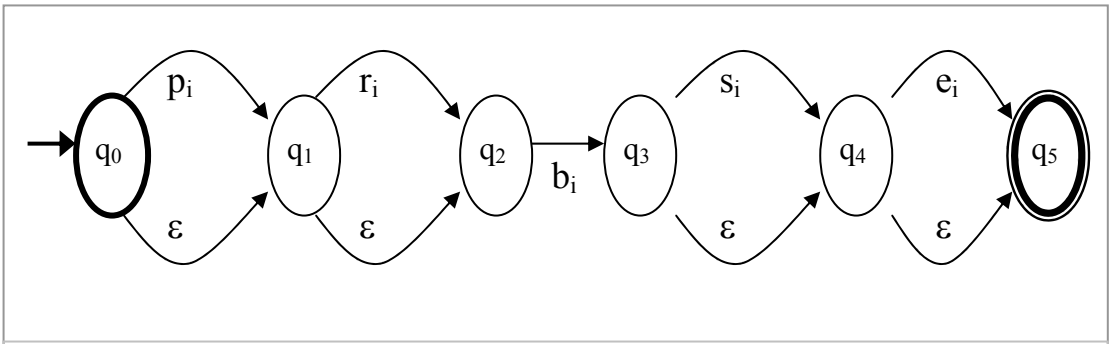


Figure 38 : L'automate (unifié) de reconnaissance d'une forme

[12] Tous les proclitiques (simples et composés).
[13] Tous les suffixes (simples et composés).
[14] Tous les enclitiques (simples et composés).

A priori on pouvait penser que cette dernière grammaire était capable de n'engendrer que des formes Nominales ou verbales attestées par la langue, mais il n'en est rien. En effet, la présence de certains proclitiques interdisent la présence de certains enclitiques, la présence de certains préfixes interdisent la présence de certains suffixes, certains proclitiques interdisent la présence de certains préfixes… Cette constatation nous a conduit à étudier les différents comportements (contraintes ou relations) entre les segments (proclitique, préfixe, suffixe et enclitique) qui composent une forme valide, cela va nous permettre, bien sûr, par la suite d'éliminer toutes les combinaisons invalides entre les segments (qui produisent une forme non attestable par la langue).

L'étude des relations et contraintes entre les objets de la « FU » nous a permis de distinguer les sept cas suivants :

– il existe une relation entre les proclitiques et les enclitiques, cette relation est souvent appelée « relation de compatibilité ». Un couple (proclitique, enclitique) compatible contribue à la formation d'une forme valide, par contre un couple (proclitique, enclitique) non compatible produit nécessairement une forme invalide. Trouver ces compatibilités revient donc à construire une matrice de dimension (72 x 35). Au niveau de notre modèle cela donne naissance à une classe-association « Compatibilité Proc-Enc » entre la classe « Proclitique » et la classe « Enclitique » ;

– il existe une relation entre les préfixes et les suffixes, cette relation est aussi de type compatibilité. Autrement dit, la présence d'un préfixe et un suffixe dans une forme exige que ces deux derniers soient compatibles. La matrice à construire est de dimension (10 x 133). De la même manière au niveau du modèle cette relation donne lieu à une classe-association « Compatibilité Pré-Suf » entre la classe « Préfixe » et la classe « Suffixe » ;

– il existe une autre relation de compatibilité entre les suffixes et les enclitiques. La matrice à construire est de dimension (133 x 35). Cette relation se traduit par une classe-association « Compatibilité Enc-Suf » entre la classe « Suffixe » et la classe « Enclitique » au niveau du modèle;

– il existe des proclitiques qui exigent la présence obligatoire d'un préfixe quelle que soit la forme (verbale ou nominale), par contre, il existe d'autres proclitiques qui exigent la présence obligatoire d'un préfixe seulement dans le cas où la forme est verbale. Cette contrainte donne lieu à un attribut « Liaison » au niveau de la classe « Proclitique », qui vaut « 1 » lorsque le proclitique exige la présence d'un préfixe, « 2 » lorsque le proclitique exige la présence d'un préfixe si la forme est verbale et bien entendu « 0 » dans le reste des cas ;

– il existe des suffixes qui exigent la présence obligatoire d'un enclitique, autrement dit, ces suffixes ne peuvent pas terminer une forme. Ils se trouvent toujours accompagnés d'un enclitique. Ce qui se traduit par l'ajout d'un attribut « Liaison » au niveau de la classe « Suffixe », qui vaut « 1 » lorsque le suffixe exige la présence d'un enclitique et « 0 » dans le cas contraire ;

– il existe des enclitiques qui exigent la présence obligatoire d'un suffixe quelle que soit la forme (verbale ou nominale), par contre il existe d'autres enclitiques qui exigent la présence obligatoire d'un suffixe seulement dans le cas où la forme est verbale. De la même façon que le cas précédent, cette contrainte se traduit par l'ajout d'un attribut « Liaison » au niveau de la classe « Enclitique », qui vaut « 1 » lorsque l'enclitique exige la présence d'un suffixe, « 2 » lorsque l'enclitique exige la présence d'un suffixe si la forme est verbale et « 0 » autrement ;

– il existe une relation de type « accepte » entre une base et les clitiques. Cette relation peut être représentée par un attribut au niveau de la classe « Base ».

Sur la base de ces différents cas et du modèle de la forme unifiée précédent, nous avons établi le nouveau modèle objet de la forme unifiée (voir figure 39).

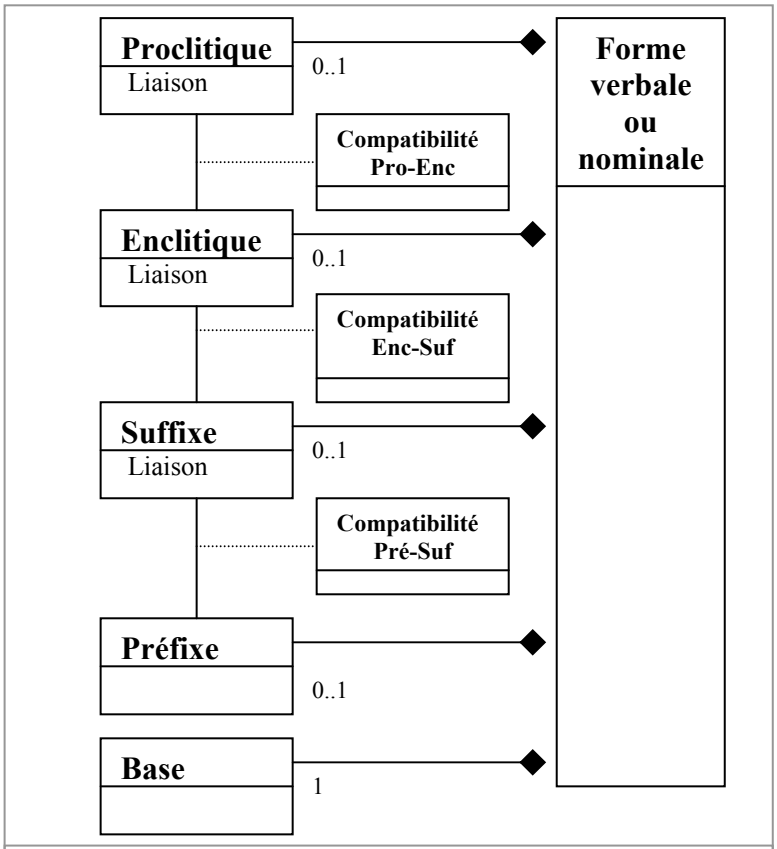


Figure 39 : Nouveau modèle objet de la forme unifiée

Nous avons répertorié tout d'abord, l'ensemble des clitiques, des affixes et des mots outils que nous avons consignés dans des tables. Ensuite nous avons essayé de recenser les bases, principalement à partir des dictionnaires [دار المشرق -2003] et [لسان العرب] avec lesquelles nous avons procédé à l'initialisation de notre dictionnaire des bases.

Le recensement des bases présente une tâche complexe et difficile, car en plus du nombre élevé de ces bases, cette tâche exige une très bonne connaissance de la langue arabe. A l'état actuel le travail de recensement est encore en cours.

5.4 Les clitiques

On distingue les proclitiques des enclitiques :

5.4.1 Les proclitiques

C'est une liste finie (voir tableau 11), elle comprend 72 éléments (deux étant homographes). Dans cette liste on a recensé toutes les combinaisons possibles entre les éléments suivants :

- Interrogatif : أ ,
- coordonnants : و، ف ,
- corroboratif : ل ,
- prépositions : ك، ل، ب ,
- article : ال,
- particule de l'impératif ou subordonnant : ل ,
- particule de futur : س .

أل	أك	أبِ	أل	أو	أفَ	أ
أولَ	أقَس	أقَكَ	أقَبِ	أقَل	أقَلَ	أس
أكَالْ	أوالْ	أقَالْ	أوس	أوك	أوبِ	أول
أوِيالْ	أفكَالْ	أقِيالْ	أقِلِالْ	أوِلِالْ	أِلِالْ	أِيالْ
فَ	كَالْ	ك	س	ـِيالْ	بِ	أوكَالْ
قَالْ	قَلْ	قَل	قَلَ	قَك	قَس	قَبِ
و	قَلِيالْ	قَلْ	قِلِالْ	قَلْبِ	فَكَالْ	قِيالْ
والْ	ولْ	ولَ	ول	وك	وس	وبِ
لَ	ولِيالْ	ولْ	وِلِالْ	ولْبِ	وكَالْ	وِيالْ
الْ	ـِلِالْ	ل	لِيالْ	لْلْ	لْك	لْبِ

Tableau 11 : Les proclitiques

5.4.2 Les enclitiques

C'est une liste finie, elle comprend 35 éléments. Le tableau 12 montre ces enclitiques.

كها	كِ	كِه	كه	ك	ـِني	ي
كِهما	كُما	نَا	ها	ه	ـِهِ	كِها
كُن	كِهم	كهم	كُم	هما	ـِهما	كهما
وه	هن	هِن	هم	هِم	كِهن	كهن
ا	وِني	وهن	وهم	وهما	ونَا	وها

Table 12 : Les enclitiques

5.5 Les affixes

On distingue deux types, les préfixes et les suffixes

5.5.1 Les préfixes

C'est une liste finie au nombre de dix, elle regroupe donc les éléments non voyellés suivants : أ، ن، ي، ت .

Ces préfixes ne concernent que les formes verbales. Ils marquent les verbes à l'inaccompli

5.5.2 Les suffixes

Les suffixes marquent soit une flexion en genre et en nombre, soit une terminaison verbale (suffixe de conjugaison). Les suffixes forment une liste finie au nombre de 133 éléments (67 suffixes étant homographes). Le tableau 13 montre des exemples de suffixes (voir la liste complète dans l'annexe D).

ة	اتْ	و	ون	ين	ين	ان	ا
تُن	تُم	تُما	نَا	تْ	تِ	تَ	تُ
ن	ي	وي	ي	ي	تَا	ن	وا
				ان	ن	نَان	ان

Tableau 13 : Les suffixes

5.6 Les bases

Le lexique des bases comprend :

- un lexique des bases verbales,
- un lexique des bases nominales.

5.6.1 Lexique des bases verbales

Ce lexique est organisé sous forme d'une liste d'enregistrements, chaque enregistrement correspond à une base verbale non voyellée auquel sont associées les informations suivantes :

- la racine,
- les voyellations possibles,
- pour chaque voyellation possible:
    - o les formes dérivées possibles,
    - o la transitivité,

- la variable humain

5.6.2 Lexique des bases nominales

D'une façon similaire au lexique précédent, le lexique des bases nominales est organisé sous forme d'une liste d'enregistrements dont chaque enregistrement correspond à une base nominale non voyellée auquel sont associées les informations suivantes :

- la racine,
- les voyellations possibles,
- pour chaque voyellation possible:
  - la catégorie grammaticale,
  - le genre,
  - le nombre,
  - la variable humain
  - un indicateur sur la flexion en genre,
  - et éventuellement, un indicateur sur la flexion en nombre

On constate que parmi les informations associées à une base verbale ou nominale certaines sont des traits morpho-syntaxiques, ces derniers vont donc nous servir à déterminer les traits morpho-syntaxiques de la forme a analyser. En d'autres termes, les traits morpho-syntaxiques d'une forme seront calculés à partir des traits morpho-syntaxiques de la base et des traits des particules. En terme de modèle on obtient la figure 40.

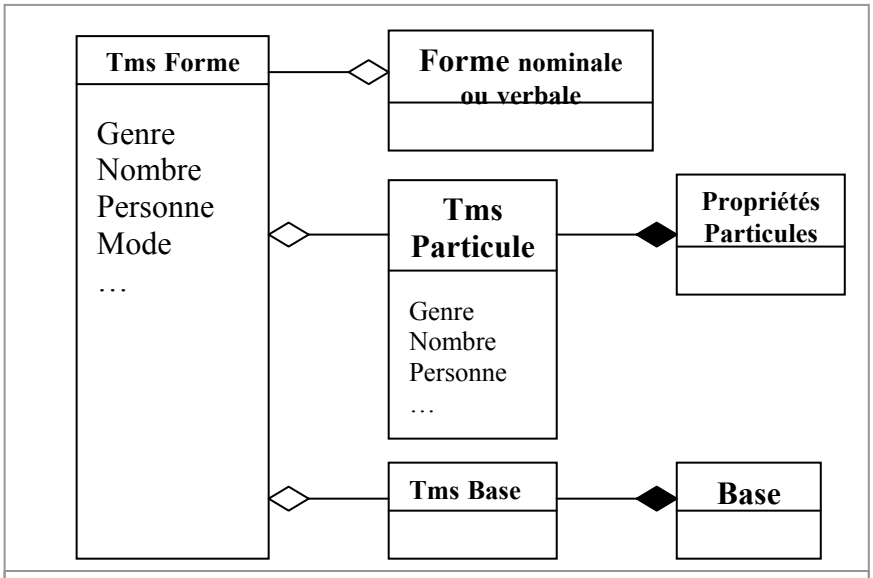


Figure 40: Modèle objet « Tms Base» de la base

5.7 Les mots outils

La liste des noms qui ne sont pas marqués par la variable syntaxique « CAS » (mots invariables au nombre de 167) et la liste des particules (au nombre de 74) sont regroupées dans un lexique à part, le lexique des mots outils (ou dictionnaire des constantes). Ce dernier comprend donc 241 mots.

Le lexique des mots outils est organisé sous une forme d'une liste d'enregistrements. Chaque enregistrement correspond à un mot outil non voyellé auquel sont associées les informations suivantes :

- les voyellations possibles,
- pour chaque voyellation possible :
    - une catégorie grammaticale.

5.8 Comptage des clitiques et affixes

Le tableau 14 donne les différents nombres de clitiques et d'affixes qui s'emploient avec une forme nominale, une forme verbale ou au deux en même temps.

		Forme nominale	**Forme verbale**	**Forme nominale ou forme verbale**	**Total**
Nombre de proclitiques	**Simples (1)**	4	3	4	11
	Composées (2)	42	6	13	61
	(1)+(2)	46	9	17	72
Nombre de Préfixes	*****	0	10	0	10
Nombre de suffixes	**Simples (1)**	14	87	0	101
	Composées (2)	32	0	0	32
	(1)+(2)	46	87	0	133
Nombre d'enclitiques	**Simples (1)**	1	8	15	24
	Composées (2)	0	10	0	10
	(1)+(2)	1	18	15	34

***** : Tous les préfixes sont simples.

Tableau 14 : Comptage des clitiques et affixes

5.9 Le modèle conceptuel d'une forme

L'union des différents modèles discutés dans les sections précédentes constitue donc notre modèle conceptuel pour le traitement automatique de l'arabe, voir la figure 41.

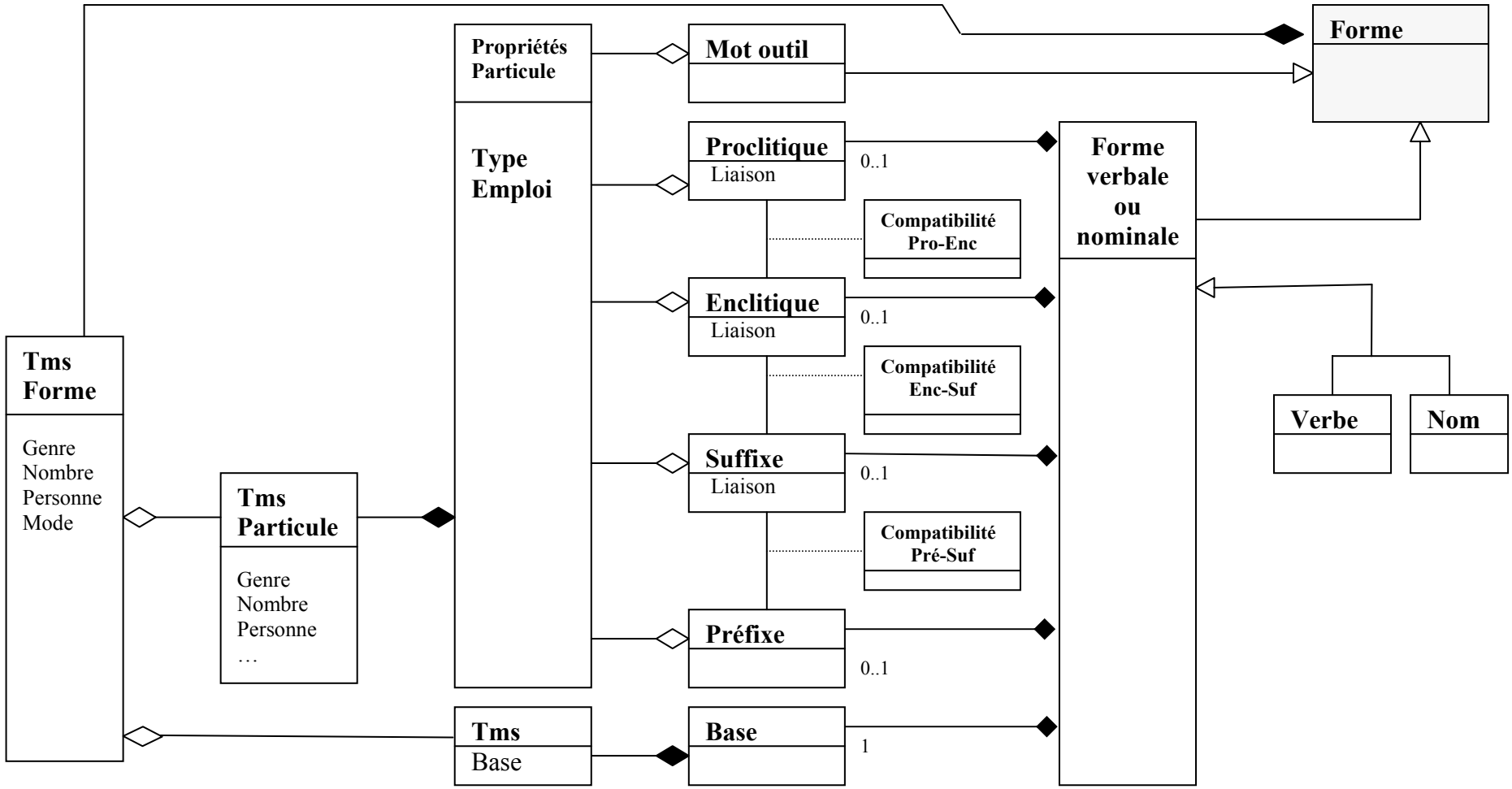


Figure 41 : Modèle conceptuel pour le traitement automatique de l'arabe

6. Principe de l'analyse morphologique

Le principe de l'analyse morphologique se résume dans quatre étapes :

1- Dévoyellation de la forme (un algorithme de dévoyellation d'une forme est proposé dans l'annexe F de cette thèse).

2- Consultation du lexique des mots outils.

3- Segmentation de la forme. Cette opération passe par l'accès aux différentes tables (clitiques et affixes) pour détecter la présence de proclitique, enclitique, préfixe et suffixe dans la forme. Le résultat étant un ensemble de cinq segments (proclitique, préfixe, radical, suffixe, enclitique).

4- Consultation des lexiques. Cette opération passe par l'accès au dictionnaire des formes simples pour vérifier l'existence du radical. S'il existe, l'analyseur lui associe l'ensemble de ses informations linguistiques (voyellations…). Il faut noter ici que l'analyseur applique sur le radical des règles de réécriture pour retrouver sa graphie initiale si ce dernier a été altéré suite à une agglutination à un clitique ou une affixe.

La figure 42 ci-dessous illustre le schéma général de notre analyseur.

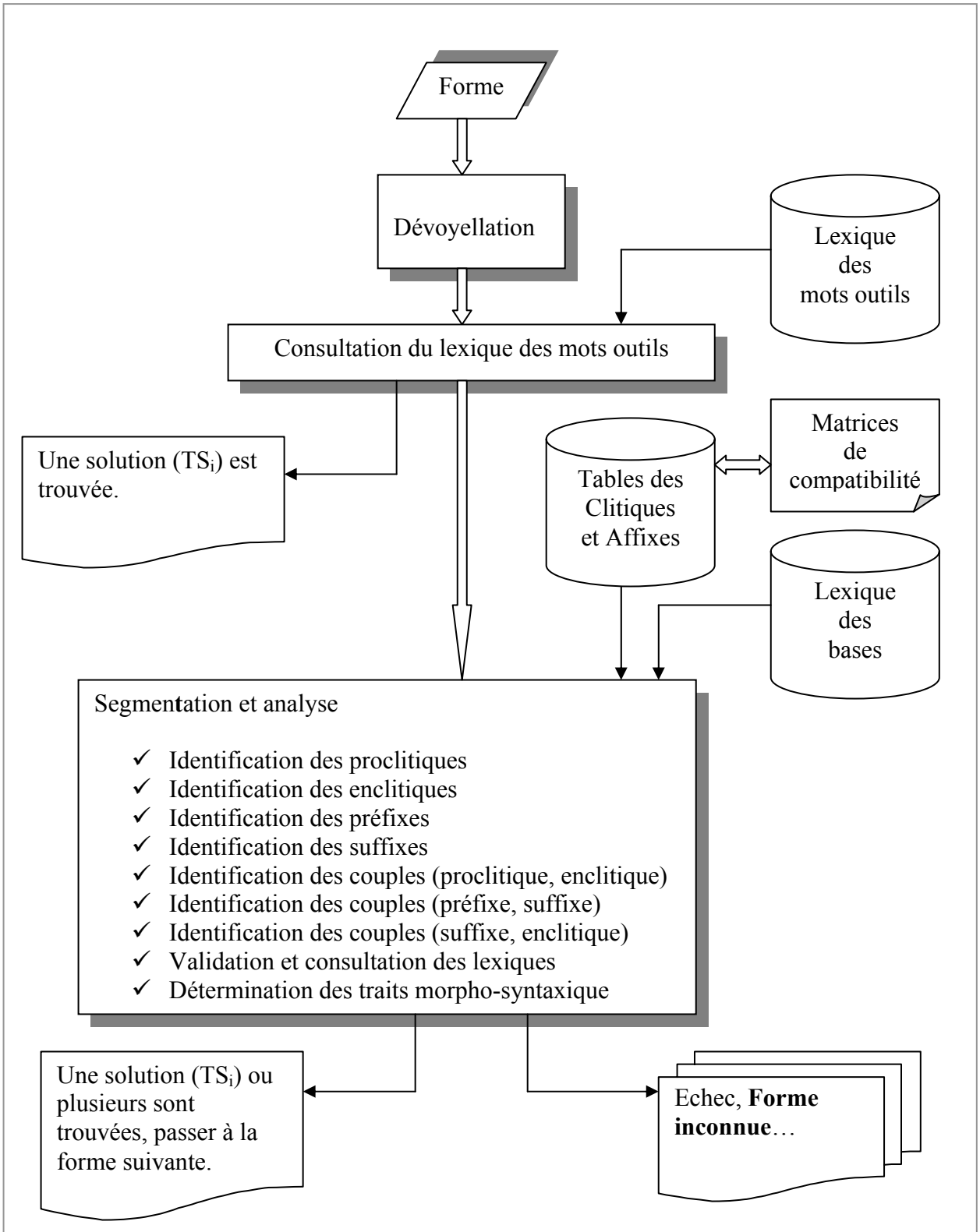


Figure 42: Architecture général de notre analyseur morphologique

6.1 Dévoyellation de la forme

(Voir page 108 de cette thèse)

6.2 Consultation du lexique des mots outils

(Voir page 109 de cette thèse)

6.3 Segmentation de la forme

A partir des tables de clitiques et d'affixes, l'analyseur localise tous les segments possibles appartenant à la forme. Autrement dit, dans cette étape l'analyseur essaye de détecter la présence d'éventuelle clitiques et/ou affixes dans la forme en question. Selon le modèle du mot graphique arabe, déjà décrit dans les sections précédentes de cette thèse {forme = enclitique + préfixe + radical ou base + suffixe + enclitique), plusieurs cas (16 cas) peuvent être envisagés :

1. Proclitique = ϕ et Enclitique = ϕ
2. Proclitique = ϕ et Enclitique ≠ ϕ
3. Proclitique ≠ ϕ et Enclitique = ϕ
4. Proclitique ≠ ϕ et Enclitique ≠ ϕ

   Et éventuellement, pour chaque cas (1, 2, 3 et 4) il faut envisager les cas suivants :

   a. Préfixe = ϕ et Suffixe = ϕ
   b. Préfixe = ϕ et Suffixe ≠ ϕ
   c. Préfixe ≠ ϕ et Suffixe = ϕ
   d. Préfixe ≠ ϕ et Suffixe ≠ ϕ

Il s'agit donc de construire les 16 cas possibles pour chaque forme, soit l'ensemble : {(1,a), (1,b), (1,c), (1,d), (2,a), (2,b), (2,c), (2,d), (3,a), (3,b), (3,c), (3,d), (4,a), (4,b), (4,c), (4,d)}.

On remarque par exemple que le cas (1, a) correspond à :

- pas de proclitique,
- pas d'enclitique,

- pas de préfixe,
- pas de suffixe.

Autrement dit, forme = radical.

Par ailleurs le cas (2, d) correspond à :

- pas de proclitique
- un ou plusieurs enclitiques
- un préfixe,
- un ou plusieurs suffixes

Donc : forme = préfixe + radical + suffixes + enclitiques.

L'algorithme de segmentation commence par l'identification des clitiques ensuite l'identification des affixes. Pour réaliser sa tâche de segmentation, l'analyseur manipule entre autres les structures de données suivantes :

- Table_Proclitiques_De_Forme [1..M], {*cette table contient tous les proclitiques possibles de la forme en entrée}*
- Table_Enclitiques_De_Forme [1..M], {*cette table contient tous les enclitiques possibles de la forme en entrée }*
- Table_Couples_Proclitique_Enclitique_De_Forme [1..M], { *cette table contient tous les couples (proclitique, enclitique) possibles de la forme en entrée. }*

Et éventuellement pour chaque table précédente :

- Table_Préfixes_De_Forme [1..M], {cette *table contient tous les préfixes possibles concernant un découpage possible en proclitique et enclitique de la forme en entrée.}*
- Table_Suffixes_De_Forme [1..M], {cette *table contient tous les suffixes possibles concernant un découpage possible en proclitique et enclitique de la forme en entrée.}*
- Table_Couples_Préfixe_Suffixe_De_Forme [1..M], {cette *table contient tous les couples (préfixe, suffixe) possibles concernant un découpage possible en proclitique et enclitique de la forme en entrée.}*

Une autre solution pour cette étape consiste d'abord à segmenter la forme de toutes les façons possibles, ensuite, valider chaque segmentation en utilisant les

différentes tables de clitiques et d'affixes. Cette dernière solution génère beaucoup de segmentations (par exemple, pour une forme de longueur 5 on a 21 segmentations possibles en trois segments), dont la plupart sont inutiles (par segmentation inutile on entend une segmentation qui ne contient pas un segment valide). Pour toutes ces raisons, il est plus logique d'adopter la première solution qui consiste à construire les 16 cas possibles selon le schéma décrit en haut.

Dans ce qui suit nous allons décrire la démarche de la première solution, celle que nous avons adoptée.

6.3.1 Identification des proclitiques

Il s'agit de détecter les différents proclitiques de la forme à analyser par la consultation de la table des proclitiques. Dans cette étape l'analyseur réalise un balayage de la table des proclitiques et compare chaque proclitique avec le segment début de la forme (qui a la même taille que le proclitique). S'il y a coïncidence l'analyseur retiendra ce proclitique.

L'algorithme suivant réalise la tâche d'identification des proclitiques. Cet algorithme utilise la table des proclitiques (déjà définie dans la partie précédente de ce chapitre), par ailleurs, il produit un tableau contenant des références (indice du proclitique dans la table des proclitiques) aux proclitiques identifiés à partir de la forme.

Algorithme Identification_Proclitiques (Forme)

Entrée : Forme {non *voyellé*}

Sortie : Table_Proclitiques_De_Forme : {*cette table contient tous les proclitiques possibles de la forme en entrée*. }

Utilise les fonctions

Taille (X : chaîne) : {cette *fonction retourne la taille de la chaîne X}*

Segment_Debut (Forme : chaîne, T : entier) *:* {cette *fonction retourne une chaîne de taille T à partir du début de la chaîne Forme* }

Constantes

Tables_Proclitiques *:* Tableau [1..72] de Proclitique {*Chaque élément de ce tableau contient entre autres : un schéma consonantique et un schéma vocalique.*

Proclitique = record of

Con : chaîne // Con : représente le schéma consonantique du proclitique

Voy : chaîne // Voy : représente le schéma vocalique du proclitique

…

Fin_Record }

Variables

I, J : entier

Table_Proclitiques_De_Forme : Tableau [1..N] d'entier { *entre 1 et 72* }

{*un élément de ce tableau contient l'indice d'un proclitique dans la table des proclitiques*}

Début

J← 1

Pour I **allant de** 1 **à** 72 **Faire**

Si Tables_Proclitiques[I].Con=Segment_Debut(Forme, aille(Proclitiques[I].Con))

Alors

Table_Proclitiques_De_Forme[j] ←I {*On a identifié le proclitique d'indice I*}

J←J+1 ;

Fin_Si

Fin_faire

Fin

6.3.2 Identification des enclitiques

Il s'agit de la même procédure que l'identification des proclitiques à une différence qui consiste à utiliser la table des enclitiques au lieu de la table des proclitiques. Donc, l'analyseur détecte les différents enclitiques de la forme à analyser par la consultation de la table des enclitiques. Pour cela, il réalise un balayage de la table des enclitiques et compare chaque enclitique avec le segment fin de la forme (qui possède la même taille que l'enclitique). S'il y a coïncidence l'analyseur retiendra cet enclitique.

L'algorithme qui réalise la tâche d'identification des enclitiques utilise la table des enclitiques (déjà définie dans la partie précédente de ce chapitre), par ailleurs, il produit un tableau contenant des références (indice de l'enclitique dans la table des enclitiques) aux enclitiques identifiés à partir de la forme. Cet algorithme est similaire à celui des proclitiques, la différence est que l'analyse de la forme est effectuée à partir de la fin de cette dernière.

6.3.3 Identification des préfixes

Il s'agit de la même procédure que l'identification des proclitiques à une différence qui consiste à utiliser la table des préfixes au lieu de la table des proclitiques.

L'algorithme est aussi similaire à celui de l'identification des proclitiques. On utilise la table des préfixes au lieu de la table des proclitiques.

6.3.4 Identification des suffixes

Il s'agit de la même procédure que l'identification des enclitiques avec la différence qui consiste à utiliser la table des suffixes au lieu de la table des enclitiques.

L'algorithme est aussi similaire à celui de l'identification des enclitiques. On utilise la table des suffixes au lieu de la table des enclitiques.

6.3.5 Identification des couples (proclitique, enclitique)

L'étape qui suit immédiatement l'identification des proclitiques et enclitiques, consiste à identifier les couples (proclitique, enclitique) possibles et ne retenir par la suite que ceux qui sont attestés par la langue. Pour valider les couples (proclitique, enclitique) on utilise une matrice de compatibilité entre proclitique et enclitique. Chaque élément [i, j] de cette matrice (de dimension 72 x 34) contient une indication :

- sur la compatibilité entre le proclitique (i) et l'enclitique (j).
- Sur l'éventuelle catégorie grammaticale (Nom, Verbe ou les deux) du radical présupposé (forme = proclitique (i) + radical + enclitique (j)). Cette indication permet d'orienter la recherche du radical dans les différents lexiques.

Une case de la matrice Compatibilité[i, j] peut valoir :

- 0 : si le proclitique (i) et l'enclitique (j) ne peuvent pas composer une forme attestée de la langue.
- 1 : Si le proclitique (i) et l'enclitique (j) peuvent composer une forme attestée de la langue, le radical de cette forme possède la catégorie Nom.
- 2 : Si le proclitique (i) et l'enclitique (j) peuvent composer une forme attestée de la langue, le radical de cette forme possède la catégorie Verbe.
- 3 : Si le proclitique (i) et l'enclitique (j) peuvent composer une forme attestée de la langue, le radical de cette forme peut avoir soit la catégorie Nom soit la catégorie Verbe indifféremment.

L'algorithme suivant réalise la tâche d'identification des couples (proclitique, enclitique) possibles. Cet algorithme utilise la table des proclitiques possibles et la table des enclitiques possibles (Table_Proclitiques_De_Forme, Table_Enclitiques_De_Forme), par ailleurs, il produit un tableau contenant des références aux couples (proclitique, enclitique) identifiés comme valides.

Algorithme Identification_Couples_Proclitiques_Enclitique

Entrée :

- Table_Proclitiques_De_Forme
- Table_Enclitiques_De_Forme

Sortie : Table_Couples_Proclitique_Enclitique_De_Forme

Utilise la fonction

Compatible (X, Y): { *cette fonction retourne 0, 1, 2 ou 3 ; voir signification en haut de la matrice de compatibilité entre proclitique et enclitique* }

Variables

I, J, K : entier

Table_Couples_Proclitique_Enclitique_De_Forme : Tableau [1..N] couple d'entier

{*un élément de ce tableau contient l'indice d'un proclitique dans la table des proclitiques, et l'indice d'un enclitique dans la table des enclitiques* }

Début

K← 1

Pour I **allant de** 1 **à** *nombre_proclitiques_possible* **Faire_1**
Pour J **allant de** 1 **à** *nombre_enclitiques_possible* **Faire_2**

Si Compatible (Table_Proclitiques_De_Forme[I],
Table_Enclitiques_De_Forme[J])<>0
Alors
Table_Couples_Proclitique_Enclitique_De_Forme [K] ←(I, J) {*On a identifié un couple (proclitique I, enclitique J)*}
K←K+1 ;
Fin_Si
Fin_faire_2
Fin_faire_1
Fin

6.3.6 Identification des couples (préfixe, suffixe)

D'une manière similaire à l'identification des couples (proclitique, enclitique), après identification des préfixes et suffixes de la forme, on essaye d'identifier les couples (préfixe, suffixe) possibles et ne retenir par la suite que ceux qui sont attestés par la langue. Pour valider les couples (préfixe, suffixe) on utilise une matrice de compatibilité entre préfixe et suffixe. Chaque élément [i, j] de cette matrice (de dimension 10 x 133) contient une indication sur la compatibilité entre le préfixe (i) et le suffixe (j). Il faut noter toutefois qu'on n'a pas besoin d'une indication sur la catégorie supposée du radical, car la présence d'un préfixe dans une forme suppose déjà que le radical a une catégorie verbe à l'inaccompli.

Une case de la matrice Compatibilité[i, j] peut valoir :

- 0 : si le préfixe (i) et le suffixe (j) ne peuvent pas composer une forme attestée de la langue.
- 1 : Si le préfixe (i) et le suffixe (j) peuvent composer une forme attestée de la langue, le radical de cette forme possède la catégorie Verbe à l'inaccompli.

L'algorithme qui réalise l'identification des couples (préfixe, suffixe) est similaire à celui de l'identification des couples (proclitique, enclitique). On utilise la matrice de compatibilité entre préfixes et suffixes au lieu de la matrice de compatibilité entre proclitiques et enclitiques.

6.3.7 Identification des couples (suffixe, enclitique)

Après identification des suffixes et enclitiques de la forme, on procède par identifier les couples (suffixe, enclitique) possibles et ne retenir par la suite que ceux qui sont attestés par la langue. Pour valider les couples (suffixe, enclitique) on utilise une matrice de compatibilité entre suffixe et enclitique.

Chaque élément [i, j] de cette matrice (de dimension 133 x 34) contient une indication sur la compatibilité entre le suffixe (i) et l'enclitique (j).

Une case de la matrice Compatibilité[i, j] peut valoir :

- 0 : si le suffixe (i) et l'enclitique (j) ne peuvent pas composer une forme attestée de la langue.
- 1 : Si le suffixe (i) et l'enclitique (j) peuvent composer une forme attestée de la langue.

L'algorithme qui réalise l'identification des couples (suffixe, enclitique) est similaire à celui de l'identification des couples (proclitique, enclitique). On utilise la matrice de compatibilité entre suffixes et enclitiques au lieu de la matrice de compatibilité entre proclitiques et enclitiques.

6.4 Validation et consultation des lexiques

Pour chaque solution ou segmentation valide retenue, l'analyseur accède aux différents lexiques (lexique des bases nominales, lexique des bases verbales) pour vérifier l'existence de la base. S'il existe, l'analyseur lui associe l'ensemble de ses informations linguistiques (catégorie, voyellations…). Avant d'accéder aux lexiques l'analyseur doit valider la segmentation pour s'assurer que le quadruplet (proclitique, préfixe, suffixe, enclitique) peut composer une forme attestable. En d'autres termes, dans cette étape l'analyseur effectue une série de contrôles pour éliminer les segmentations invalides d'une part et orienter les accès au lexique approprié selon le cas d'autre part (voir la fonction du moniteur dans les sections précédentes de ce chapitre).

La validation de la segmentation passe par le test des combinaisons possibles entre les attributs « Liaison »[15] et « Emploi » des segments comportant un proclitique ou un enclitique d'une part, et les valeurs de compatibilité du couple (proclitique enclitique) s'il existe d'autre part.

15 Voir la section « 5.3 Forme unifiée » de ce chapitre pour plus d'informations.

Rappelons que l'attribut « Liaison » peut avoir trois valeurs « 0 », « 1 » et « 2 », l'attribut « Emploi » peut avoir lui aussi trois valeurs « N », « V » et « NV ».

Sur la base de ces différents tests nous avons recensé vingt cas (I_1, I_2… I_{20}) de segmentations invalides, les deux tableaux 15 et 16 montrent ces cas.
Pour simplifier la lecture de ces deux tableaux, une segmentation représentée par le quadruplet (proclitique, enclitique, préfixe, suffixe) est notée (val, val, val, val) où val désigne la présence (val=1) ou l'absence (val=0) du segment en question. Par exemple la segmentation (1,0,1,1) désigne une segmentation comportant un proclitique, un préfixe et un suffixe. La partie commentaire du tableau 15 explicite le test effectué sur les attributs et les compatibilités.

N°	Cas	Désignation	Commentaire
1	I_1	Ce proclitique exige la présence d'un préfixe (voir l'autre tableau 2).	Proclitique.Liaison=1
2	I_2	Cet enclitique exige la présence d'un suffixe (voir l'autre tableau 2).	Enclitique.Liaison=1
3	I_3	Cet enclitique exige la présence d'un suffixe la forme est verbale (voir l'autre tableau 2).	Enclitique.Liaison=2
4	I_4	Ce proclitique exige la présence d'un préfixe la forme est verbale (voir l'autre tableau 2).	Proclitique.Liaison=2
5	I_5	Le suffixe exige la présence d'un enclitique (voir l'autre tableau 2).	Suffixe.Liaison=1
6	I_6	Le proclitique et le suffixe sont incompatibles ce proclitique s'emploi seulement avec un nom. (1,0,1,0)	Proclitique.Emploi=N
7	I_7	L'enclitique et le suffixe sont incompatibles cet enclitique s'emploi seulement avec le nom. (0,1,1,0)	Enclitique.Emploi=N Préfixe.Emploi=V
8	I_8	Le proclitique et l'enclitique ne s'emploient pas avec ce préfixe. (1,1,1,0)	Comp(Pro, Enc)=1 Préfixe.Emploi=V
9	I_9	Incompatibilité entre le proclitique {Verbe} et le suffixe {Nom}. (1,0,0,1)	Proclitique.Emploi=V Suffixe.Emploi=N
10	I_{10}	Incompatibilité entre le proclitique {Nom} et le suffixe {Verbe}. (1,0,0,1)	Proclitique.Emploi=N Suffixe.Emploi=V
11	I_{11}	Incompatibilité entre l'enclitique {Verbe} et le suffixe {Nom}. (0,1,0,1)	Enclitique.Emploi=V Suffixe.Emploi=N
12	I_{12}	Incompatibilité entre l'enclitique {Nom} et le suffixe {Verbe}. (0,1,0,1)	Enclitique.Emploi=N Suffixe.Emploi=V
13	I_{13}	Le proclitique et l'enclitique {Comp=1} sont incompatible avec ce suffixe {Verbe}. (1,1,0,1)	Comp(Pre, Enc)=1 Suffixe.Emploi=V
14	I_{14}	Le proclitique et l'enclitique {Comp=2} sont incompatible avec ce suffixe {Nom}. (1,1,0,1)	Comp(Pre, Enc)=2 Suffixe.Emploi=N
15	I_{15}	Le préfixe et le suffixe {Verbe} sont incompatible avec le proclitique {Nom}. (1,0,1,1)	Proclitique.Emploi=N Préfixe.Emploi=V Suffixe.Emploi=V
16	I_{16}	Le préfixe et le suffixe {Verbe} sont incompatible avec l'enclitique {Nom}. (0,1,1,1)	Enclitique.Emploi=N Préfixe.Emploi=V Suffixe.Emploi=V
17	I_{17}	Le proclitique et l'enclitique {Nom} sont incompatible avec le préfixe et le suffixe {Verbe} (1,1,1,1)	Comp(Pro, Enc)=1 Préfixe.Emploi=V Suffixe.Emploi=V
18	I_{18}	Le préfixe possède le trait {voix = Passive} qui n'autorise pas la présence d'un enclitique (voir l'autre tableau 2).	Préfixe.Tms.Voix = Passive
19	I_{19}	Le suffixe possède le trait {Aspect = Ina ou Imp} qui exige la présence d'un préfixe (voir l'autre tableau 2).	Suffixe.Tms.Aspect = 'I' ou 'M'
20	I_{20}	Le trait mode du suffixe et du proclitique d'un verbe à l'inaccompli sont incompatibles. (voir l'autre tableau 2).	Suffixe.Tms.Mode ≠ Proclitique.Tms.Mode

Tableau 15 : Les différents cas de segmentations invalides

N°	Proclitique	Enclitique	Préfixe	Suffixe	Pro. Liaison =1	Pro. Liaison =2	Enc. Liaison =1	Enc. Liaison =2	Suf. Liaison =1	Pré. Tms. Voix =Passive	Suf. Tms. Aspect = 'I' ou ='M'	Suf. Tms. Mode ≠ Pro. Tms. Mode
0	0	0	0	0								
1	0	0	0	1					I_5		I_{19}	
2	0	0	1	0								
3	0	0	1	1					I_5			
4	0	1	0	0			I_2	I_3				
5	0	1	0	1							I_{19}	
6	0	1	1	0			I_2	I_3		I_{18}		
7	0	1	1	1						I_{18}		
8	1	0	0	0	I_1	I_4						
9	1	0	0	1	I_1	I_4			I_5		I_{19}	
10	1	0	1	0								
11	1	0	1	1					I_5			I_{20}
12	1	1	0	0	I_1	I_4	I_2	I_3				
13	1	1	0	1	I_1	I_4					I_{19}	
14	1	1	1	0			I_2	I_3		I_{18}		
15	1	1	1	1						I_{18}		I_{20}

Proclitique = 0 → signifie pas de proclitique

Proclitique = 1 → signifie le proclitique existe

Tableau 16 : Les différents cas de segmentations invalides (I_1, I_2, I_3, I_4, I_5, I_{18}, I_{19}, I_{20})

Le programme de validation « Moniteur » complet se trouve dans l'annexe G.

6.5 Détermination des traits morpho-syntaxiques

Par détermination des traits morpho-syntaxique, on entend le calcul des traits de la forme analysée. Ces traits résultent directement à partir des traits morpho-syntaxiques de la base, des clitiques et des affixes. En effet après l'étape précédente (validation et consultation du lexique) l'analyseur dispose d'un quintuplet (Base, Proclitique, Enclitique, Préfixe, Suffixe) qui représente la forme en entrée et chaque élément de ce quintuplet dispose de ses propres traits. Une opération d'unification de ces différents traits est donc nécessaire pour déterminer les valeurs des traits qui seront associées à la forme. Deux cas sont à envisager : le cas où la base est nominale et le cas où la base est verbale.

a) Si la base est nominale alors ses traits sont : genre (G), nombre (N) et humain (H) donc : Tms[16] (base nominale) =[G, N, H].
   Par contre les traits des clitiques et affixes sont[17] :
   Tms (Proclitique) = [C]
   Tms (Enclitique) = [G, N, P]
   Tms (Préfixe) = [0]
   Tms (Suffixe) = [G, N, C]
   Les Tms de la forme seront donc :
   Tms(Forme) =[G, N, C, P, H]

b) Si la base est verbale alors :
   Tms(base verbale) =[T, H]
   Tms (Proclitique) = [A, M]
   Tms (Enclitique) = [G, N, P, T]
   Tms (Préfixe) = [G, N, P, A, M, V]
   Tms (Suffixe) = [G, N, P, A, M, V]
   Les Tms de la forme verbale seront donc :
   Tms(Forme verbale) =[G, N, P,A, P, M, T, H]

Nous constatons qu'il peut y avoir plusieurs valeurs pour une même variable, par exemple pour déterminer la valeur de la variable genre pour une forme verbale, nous

16 Tms = Traits morpho-syntaxiques
17 Voir la section : 3.3 Affectation des variables pour les particules pré et postfixés de ce chapitre

avons : une variable genre pour l'enclitique, le préfixe et le suffixe, alors laquelle de ces valeurs on doit prendre ?

Sachant qu'en principe les valeurs ne se contredisent pas, par exemple, nous ne pouvons pas trouver un préfixe de genre masculin et un suffixe de genre féminin, alors le seul cas qui peut surgir c'est d'avoir à choisir entre une valeur « Indéterminée » pour une variable et une autre valeur quelconque pour l'autre variable. Dans ce cas nous devons tout simplement ignorer la valeur « Indéterminée », par exemple entre un préfixe de genre « masculin » et un suffixe de genre « Indéterminée », nous retenons la valeur « masculin » pour le genre de la forme en question. Une solution simple à ce problème consiste à affecter une priorité à chaque valeur d'une variable, la détermination des valeurs de variables associées à la forme se fait par le choix des valeurs les plus prioritaires c'est celles qui ont le poids le plus fort.

7. Conclusion

Dans ce chapitre nous avons proposé un modèle en classe (classe dans le paradigme objet) pour le traitement automatique de l'arabe. Ce dernier peut être exploité par différentes applications dans le domaine du TALN arabe comme la traduction automatique, la correction orthographique, la recherche d'information, etc. L'utilisation du concept objet (très puissant et intuitif) nous a permis d'obtenir un modèle (pour le TALN arabe) clair et réutilisable. Une conséquence immédiate de cette modélisation est l'implémentation facile vue la disponibilité des langages de programmation supportant la notion d'objet et l'existence des outils pour la modélisation UML. Ce modèle sera considéré comme une plate forme commune et réutilisable par toutes les applications de TALN Arabe. Pour la validation de notre modèle, nous avons proposé de construire un analyseur morphologique pour l'arabe.

Dans le chapitre suivant, nous présentons la partie réalisation et expérimentation de notre analyseur.

Chapitre 5

Réalisation et expérimentation

Chapitre 5

Réalisation et expérimentation

1. Introduction

Pour valider notre modèle, nous l'avons intégré dans un framework nommé « MALA ». Ce dernier va nous servir comme une base pour le développement des applications TALN arabe. Dans cette optique un analyseur morphologique pour l'arabe ainsi qu'un ensemble d'outils sont construits.

Nous présentons dans les sections suivantes la partie réalisation et expérimentation de notre analyseur. En effet, à partir du modèle conceptuel de l'analyse d'une forme que nous avons développé dans le chapitre précédent (voir page 128), les classes ainsi que les relations d'association vont se traduire par des objets persistants (tables) dans le modèle physique. En raison de leur taille, certaines de ces tables sont chargées directement dans la mémoire au démarrage du programme d'analyse, ce qui permet d'optimiser le temps d'accès ; par ailleurs d'autres tables (comme les bases nominales, les bases verbales et les mots outils) seront consultées sur disque via un langage de manipulation de données, qui dans notre cas n'est autre que le langage SQL.

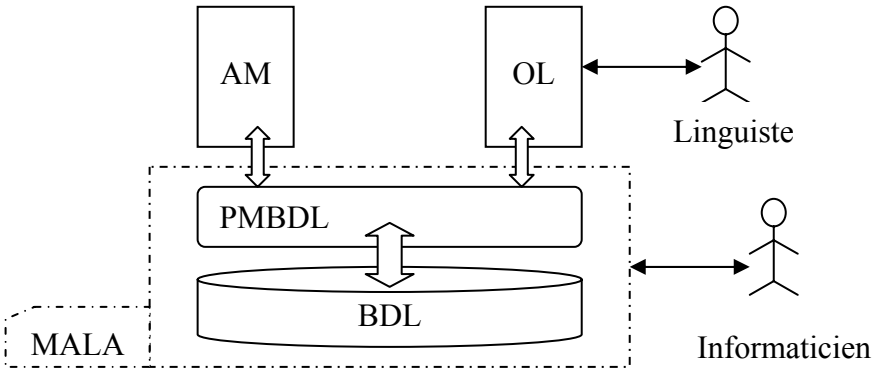
Dans la partie expérimentation et en l'absence constatée d'un consensus entre les chercheurs de ce domaine dans la définition des critères pour mesurer les performances d'un analyseur d'une part ; et par le manque (ou non disponibilité) des résultats obtenus par d'autres analyseurs d'autres part, on ne peut donc espérer faire des comparaisons avec d'autres analyseurs. Dans cette optique, la section suivante décrit les différents résultats de notre analyseur appliqué à différents textes. Ces textes étant recueillis à partir du web.

2. Réalisation

Notre analyseur morphologique est basé sur « MALA » ; un framework que nous avons construit et qui va nous servir comme une plate forme pour le développement des applications pour le TALN arabe, « MALA » repose sur deux composantes (couches) principales (voir figure 43) :

- une base de données linguistique (BDL) intégrant toutes les données linguistiques propres à la langue arabe. Conceptuellement la BDL est représentée par un modèle en classes (classe dans le paradigme objet). Par ailleurs son implémentation est réalisée en utilisant des tables dans le modèle relationnel, parmi ces tables on trouve entre autres la table des bases nominales, des bases verbales, des mots outils, des clitiques, des affixes, des compatibilités entre clitiques…
- Un ensemble de primitives ou de méthodes de base (PMBDL) pour la manipulation de la BDL.

Outre l'analyseur morphologique (AM) nous avons réalisé un outil (OL) destiné aux linguistes, il permet de faire la mise à jour de la BDL d'une manière très simple.



OL : **O**utils pour **L**inguiste
PMBDL : **P**rimitives de base pour la **M**anipulation de la **BDL**
BDL : **B**ase de **D**onnées **L**inguistique ;
MALA : framework pour TALN arabe
AM : **A**nalyseur **M**orphologique

Figure 43 : Architecture générale du framework MALA

Le développement de MALA présente de nombreux avantages comme par exemple :

- La séparation entre les données linguistiques et les programmes qui les manipulent,
- la réutilisation (plate forme commune pour toutes les applications de TALN arabe),
- la normalisation des développements (permettre de construire toutes les applications avec les mêmes normes, technologies,…),
- l'extension de la BDL et les PMBDL,
- et la facilité de la maintenance.

Nous avons réalisé cinq modules :

- un analyseur morphologique.

Cet analyseur est complété par un certain nombre d'utilitaires, indispensables pour la gestion des données que nécessite son fonctionnement. On trouve entre autres :

- pour le dictionnaire (des bases verbales, nominales et mots outils) :
  - Un utilitaire de mise à jour qui permet l'ajout, la correction et la consultation interactive de ce dictionnaire.
- Pour les matrices de compatibilité :
  - Un utilitaire de mise à jour qui permet l'ajout, la correction et la consultation interactive de ces matrices.
- Pour les clitiques et affixes :
  - Un utilitaire de mise à jour qui permet l'ajout, la correction et la consultation interactive des clitiques et affixes.
- Pour les propriétés particule :
  - Un utilitaire de mise à jour qui permet l'ajout, la correction et la consultation interactive de ces propriétés.

Ces programmes sont écrits en langage pascal (DELPHI Version 5). L'ensemble des modules représente environ 8000 lignes de code source, avec un programme exécutable d'une taille de 1447 Ko.

2.1 Les données

L'analyseur utilise un ensemble de données consignées dans différentes tables. On distingue deux types :

a) celles qui possèdent une taille réduite, ces dernières sont chargées dans la mémoire lors du premier démarrage du programme afin d'optimiser les temps d'analyse. On trouve :
   - la table des proclitiques (TProclitiques),
   - la table des enclitiques (TEnclitiques),

- la table des préfixes (TPréfixes),
- la table des suffixes (TSuffixes),
- la table de compatibilité proclitique-enclitique (TCompatible_PE),
- la table de compatibilité suffixes-enclitique (TCompatible_SE),
- la table de compatibilité préfixe-suffixe (TCompatible_PS),
- la table des propriétés particule (TPropriétés).

b) Celles qui possèdent une taille considérable et ne peuvent pas faire l'objet d'un chargement total dans la mémoire. Ces dernières font donc l'objet d'une gestion spéciale qui doit entre autres assurer l'optimisation des temps d'accès aux données. Dans notre cas, nous avons choisi (pour le moment dans cette phase de prototype) d'utiliser le noyau (moteur) de bases de données Borland (BDE) fourni avec DELPHI. De ce fait la manipulation des données est réalisée en utilisant des requêtes SQL. Dans cette catégorie de table on trouve :

- la table des bases nominales,
- la table des bases verbales,
- et la table des mots outils.

Dans ce qui suit nous allons décrire le contenu de ces différentes tables.

2.1.1 La table des proclitiques (TProclitiques)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre du proclitique
- Con : représente le schéma consonantique,
- Voy : représente le schéma vocalique,
- liaison : pouvant prendre comme valeur 0, 1 ou 2,
- Pro : un lien vers la table des propriétés particule.

2.1.2 La table des enclitiques (TEnclitiques)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre de l'enclitique
- Con : schéma consonantique,
- Voy : schéma vocalique,
- liaison : pouvant prendre comme valeur 0, 1 ou 2,
- Pro : un lien vers la table des propriétés particule.

2.1.3 La table des préfixes (TPréfixes)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre du préfixe
- Con : schéma consonantique,
- Voy : schéma vocalique,
- Pro : un lien vers la table des propriétés particule.

2.1.4 La table des suffixes (TSuffixes)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre du suffixe
- Con : schéma consonantique,
- Voy : schéma vocalique,
- liaison : pouvant prendre comme valeur 0 ou 1,
- Pro : un lien vers la table des propriétés particule.

2.1.5 Les tables de compatibilité (TCompatible_PE, TCompatible_SE, TCompatible_PS)

Chaque case de ces tables est un enregistrement contenant les champs:

- Num_Elem_1 : représente un numéro d'ordre de l'élément 1 selon la table,
- Num_Elem_2 : représente un numéro d'ordre de l'élément 2 selon la table,
- Comp : valeur de la compatibilité pouvant prendre 0, 1, 2 ou 3 selon la table.

2.1.6 La table des propriétés particule (TPropriétés)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre des propriétés,
- Désignation : Désignation de la propriété,
- Emploi : « N » pour Nom, « V » pour verbe ou « X » pour les deux,
- Type : « P » pour proclitique, « E » pour enclitique, « R » pour préfixe ou « S » pour suffixe,
- Tms : traits morphosyntaxiques, c'est un vecteur de huit cases. Son contenu est décrit dans les sections qui suivent.

2.1.7 La table des bases nominales (TBases_N)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre de la base nominale
- Con : schéma consonantique,
- Voy : schéma vocalique,
- Racine : la racine,
- Catégorie : Catégorie de la base,
- Genre : genre de la base,
- Nombre : le nombre de la base,
- Flex_Genre : indication si la base admet une flexion en genre,
- Flex_Nombre : indication si la base admet une flexion en nombre,
- Hum : humain/non humain.

2.1.8 La table des bases verbales (TBases_V)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre de la base verbale
- Con : schéma consonantique,
- Voy : schéma vocalique,
- Racine : la racine,
- Transitivité : Transitivité,
- Hum : Humain/ non humain,

2.1.9 La table des mots outils (TOutils)

Chaque case de cette table est un enregistrement contenant les champs:

- Num : représente un numéro d'ordre du mot outil,
- Con : schéma consonantique,
- Voy : schéma vocalique,
- Catégorie : la catégorie.

2.2 Le programme d'analyse

Contrairement à certains analyseurs morphologiques, l'analyse du texte en entrée s'applique directement sans prétraitement de ce dernier. Le texte du document est lu forme par forme. Pour chaque forme on détermine l'ensemble solution constitué de :

- la base,
- la racine,
- la segmentation en proclitique, préfixe, suffixe et enclitique,
- la catégorie principale (N si Nom, V si verbe, P si mot outil, NBN si nombre et SEP si séparateur),
- la catégorie secondaire (si la catégorie principale est Nom),
- l'ensemble des valeurs des variables morphosyntaxiques.

Le principe de l'analyse étant expliqué dans le chapitre précédent, nous allons, dans ce qui suit, donner un exemple d'analyse d'une forme et un autre exemple d'analyse d'une phrase.

2.2.1 Exemple d'analyse d'une forme

Pour pouvoir vérifier les résultats générés par l'analyseur lors des différentes étapes, et pour avoir une idée claire sur les différentes analyses possibles d'une forme (le détail de l'analyse d'une forme), nous avons prévu une procédure qui accepte en entrée une forme et produit en sortie un rapport complet d'analyse de cette forme. Pour faciliter la lecture de ce rapport, ce dernier est généré dans un fichier XML.

Le rapport se divise en deux parties, une partie quantitative, elle comprend les informations suivantes (voir la figure 44) :

a) le nombre de formes possibles,
b) le nombre de formes valides « Nom »
c) le nombre de formes valides « Verbe »
d) le nombre de formes valides « Nom » et « Verbe » en même temps,
e) le nombre de formes « mot outil »,
f) le nombre de formes inconnues « Nom » (il représente le nombre d'accès sans succès à la table des bases nominales),
g) le nombre de formes inconnues « Verbe » (il représente le nombre d'accès sans succès à la table des bases verbales),

h) le nombre de formes inconnues « Nom » et « Verbe » (il représente le nombre d'accès sans succès à la table des bases nominales et verbales simultanément, ce nombre étant différent des deux précédents, autrement dit, il ne s'agit en aucun cas de la somme de « g » et « h » précédents),
i) le nombre de segmentations invalides (ces segmentations sont détectées et rejetées par le moniteur avant la phase de consultation des bases nominales, verbales et mots outils,
j) et éventuellement le temps d'analyse sous la forme [Heure :Min :Sec :MSec].

La deuxième partie du rapport explicite chaque segmentation réalisée en donnant les informations suivantes :

a) la segmentation en proclitique, enclitique, préfixe et suffixe,
b) la base trouvée ainsi que les informations qui lui sont associées,
c) les traits morphosyntaxiques représentés par un vecteur de huit cases. Ces traits étant calculés à partir des affixes.

Exemple 1

L'exemple suivant montre le rapport d'analyse de la forme verbale ' فسيشكرونهم '. Le rapport étant trop long pour le reproduire en complet ici, une partie seulement de celui ci est donné, toutefois le rapport complet disponible sur le lien Internet :
http://www1.univ-tlemcen.dz/~ltala/index_fichiers/Page748.htm

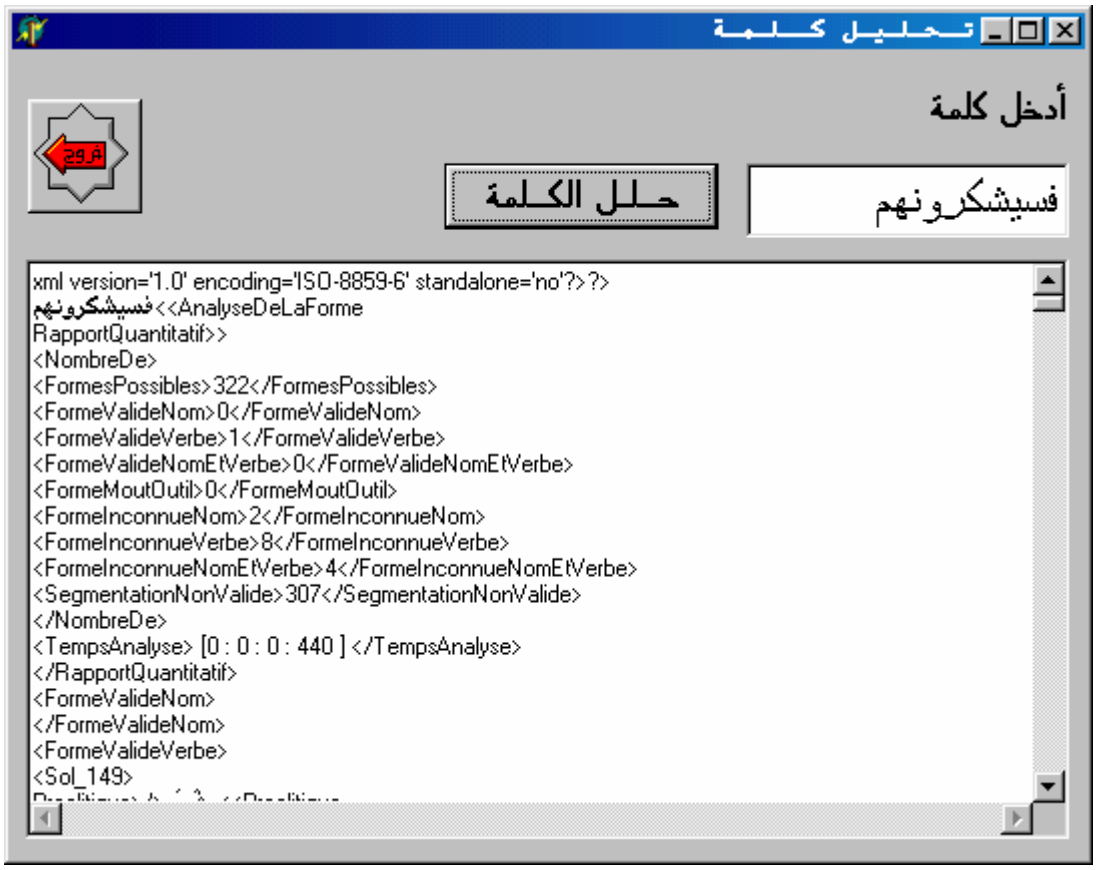


Figure 44 : Copie d'écran ; analyse de la forme verbale ' فسيشكرونهم '

Une partie de l'analyse de la forme « فسيشكرونهم » est donnée dans ce qui suit :

```
<?xml version='1.0' encoding='ISO-8859-6' standalone='no'?>
<AnalyseDeLaForme>فسيشكرونهم
<RapportQuantitatif>
<NombreDe>
<FormesPossibles>322</FormesPossibles>
<FormeValideNom>0</FormeValideNom>
<FormeValideVerbe>1</FormeValideVerbe>
<FormeValideNomEtVerbe>0</FormeValideNomEtVerbe>
<FormeMoutOutil>0</FormeMoutOutil>
<FormeInconnueNom>2</FormeInconnueNom>
<FormeInconnueVerbe>8</FormeInconnueVerbe>
<FormeInconnueNomEtVerbe>4</FormeInconnueNomEtVerbe>
<SegmentationNonValide>307</SegmentationNonValide>
</NombreDe>
<TempsAnalyse> [0 : 0 : 0 : 440 ] </TempsAnalyse>
</RapportQuantitatif>
<FormeValideNom>
</FormeValideNom>
<FormeValideVerbe>
<Sol_149>
<Proclitique>  فَسَ  </Proclitique>
<Des_Pro> حرف مركب  </Des_Pro>
<Prefixe>  يَ  </Prefixe>
```

```
<Des_Pre> حرف مضارعة </Des_Pre>
<Suffixe> ونَ </Suffixe>
<Des_Suf> "واو الجماعة " ضمير متصل بالفعل المضارع جمع غائب" </Des_Suf>
<Enclitique> هُمْ </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب" </Des_Enc>
<Base> [V/1](شكر,ccc,شكر,A,X) </Base>
<Tms> [MPC0IIA0] </Tms>
</Sol_149>
</FormeValideVerbe>
<FormeValideNomEtVerbe>
</FormeValideNomEtVerbe>
<FormeMotOutil>
</FormeMotOutil>
<FormeInconnueNom>
<Sol_37>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> هِمْ </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب" </Des_Enc>
<Base> [N/0] </Base>
<Tms> [00000000] </Tms>
</Sol_37>
<Sol_288>
<Proclitique> فَ </Proclitique>
<Des_Pro> حرف عطف </Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> هِمْ </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب" </Des_Enc>
<Base> [N/0] </Base>
<Tms> [00000000] </Tms>
</Sol_288>
</FormeInconnueNom>
<FormeInconnueVerbe>
<Sol_3>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> نَ </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل الماضى جمع غائبة" </Des_Suf>
<Enclitique> هُمْ </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب" </Des_Enc>
<Base> [V/0] </Base>
<Tms> [FPC0AZX0] </Tms>
</Sol_3>
<Sol_74>
<Proclitique> فَسَ </Proclitique>
<Des_Pro> حرف مركب </Des_Pro>
<Prefixe> يَ </Prefixe>
<Des_Pre> حرف مضارعة </Des_Pre>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [ZXC0IXA0] </Tms>
</Sol_74>
<Sol_75>
<Proclitique> فَسَ </Proclitique>
<Des_Pro> حرف مركب </Des_Pro>
<Prefixe> يُ </Prefixe>
<Des_Pre> حرف مضارعة </Des_Pre>
<Suffixe> -Pas de suffixe- </Suffixe>
```

```
<Enclitique>   -Pas d'enclitique-   </Enclitique>
<Base>   [V/0]   </Base>
<Tms>  [ZXC0IXP0]   </Tms>
</Sol_75>
<Sol_113>
<Proclitique>  فَسَ  </Proclitique>
<Des_Pro> حرف مركب  </Des_Pro>
<Prefixe>  يَ  </Prefixe>
<Des_Pre> حرف مضارعة  </Des_Pre>
<Suffixe>   -Pas de suffixe-   </Suffixe>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "ضمير متصل جمع غائب" هاء الغائب  </Des_Enc>
<Base>   [V/0]   </Base>
<Tms>  [ZXC0IXA0]   </Tms>
</Sol_113>
<Sol_150>
<Proclitique>  فَسَ  </Proclitique>
<Des_Pro> حرف مركب  </Des_Pro>
<Prefixe>  يَ  </Prefixe>
<Des_Pre> حرف مضارعة  </Des_Pre>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "ضمير متصل بالفعل المضارع جمع غائبة" نون النسوة  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "ضمير متصل جمع غائب" هاء الغائب  </Des_Enc>
<Base>   [V/0]   </Base>
<Tms>  [FPC0IIA0]   </Tms>
</Sol_150>
<Sol_153>
<Proclitique>  فَسَ  </Proclitique>
<Des_Pro> حرف مركب  </Des_Pro>
<Prefixe>  يَ  </Prefixe>
<Des_Pre> حرف مضارعة  </Des_Pre>
<Suffixe>  نَّ  </Suffixe>
<Des_Suf> "ضمير متصل بالفعل المضارع المؤكد الثقيل مفرد " نون التوكيد الثقيلة
غائب  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "ضمير متصل جمع غائب" هاء الغائب  </Des_Enc>
<Base>   [V/0]   </Base>
<Tms>  [MSC0IEA0]   </Tms>
</Sol_153>
<Sol_155>
<Proclitique>  فَسَ  </Proclitique>
<Des_Pro> حرف مركب  </Des_Pro>
<Prefixe>  يَ  </Prefixe>
<Des_Pre> حرف مضارعة  </Des_Pre>
<Suffixe>  نَّ  </Suffixe>
<Des_Suf> "ضمير متصل بالفعل المضارع المؤكد الثقيل جمع " نون التوكيد الثقيلة
الغائب  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "ضمير متصل جمع غائب" هاء الغائب  </Des_Enc>
<Base>   [V/0]   </Base>
<Tms>  [MPC0IEA0]   </Tms>
</Sol_155>
<Sol_166>
<Proclitique>  فَ  </Proclitique>
<Des_Pro> حرف عطف  </Des_Pro>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "ضمير متصل بالفعل الماضى جمع غائبة" نون النسوة  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "ضمير متصل جمع غائب" هاء الغائب  </Des_Enc>
```

```
<Base>   [V/0]   </Base>
<Tms>  [FPC0AZX0]   </Tms>
</Sol_166>
</FormeInconnueVerbe>
<FormeInconnueNomEtVerbe>
<Sol_1>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>   -Pas de suffixe-   </Suffixe>
<Enclitique>   -Pas d'enclitique-   </Enclitique>
<Base>   [N/0][V/0]   </Base>
<Tms>  [00000000]   </Tms>
</Sol_1>
<Sol_2>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>   -Pas de suffixe-   </Suffixe>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب  </Des_Enc>
<Base>   [N/0][V/0]   </Base>
<Tms>  [00000000]   </Tms>
</Sol_2>
<Sol_76>
<Proclitique>  فَ  </Proclitique>
<Des_Pro> حرف عطف  </Des_Pro>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>   -Pas de suffixe-   </Suffixe>
<Enclitique>   -Pas d'enclitique-   </Enclitique>
<Base>   [N/0][V/0]   </Base>
<Tms>  [00000000]   </Tms>
</Sol_76>
<Sol_165>
<Proclitique>  فَ  </Proclitique>
<Des_Pro> حرف عطف  </Des_Pro>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>   -Pas de suffixe-   </Suffixe>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب  </Des_Enc>
<Base>   [N/0][V/0]   </Base>
<Tms>  [00000000]   </Tms>
</Sol_165>
</FormeInconnueNomEtVerbe>
<SegmentationNonValide>
<Sol_4>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  ونَ  </Suffixe>
<Des_Suf> "واو الجماعة " ضمير متصل بالفعل المضارع جمع مخاطب  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب  </Des_Enc>
<Base>   0   </Base>
<Tms>  [MPB0IIX0]   </Tms>
</Sol_4>
<Sol_5>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع جمع مخاطبة  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب  </Des_Enc>
<Base>   0   </Base>
```

```
<Tms>  [FPB0IIX0]   </Tms>
</Sol_5>
<Sol_6>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  ونَ  </Suffixe>
<Des_Suf> "واو الجماعة " ضمير متصل بالفعل المضارع جمع غائب"  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [MPC0IIX0]   </Tms>
</Sol_6>
<Sol_7>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع جمع غائبة"  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [FPC0IIX0]   </Tms>
</Sol_7>
<Sol_8>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع المنصوب جمع مخاطبة"
</Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [FPB0ISX0]   </Tms>
</Sol_8>
<Sol_9>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع المنصوب جمع غائبة"
</Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [FPC0ISX0]   </Tms>
</Sol_9>
<Sol_10>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع المجزوم جمع مخاطبة"
</Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [FPB0IAX0]   </Tms>
</Sol_10>
<Sol_11>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَ  </Suffixe>
<Des_Suf> "نون النسوة" ضمير متصل بالفعل المضارع المجزوم جمع غائبة"  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
```

```
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [FPC0IAX0]   </Tms>
</Sol_11>
<Sol_12>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَّ    </Suffixe>
<Des_Suf> " نون التوكيد الثقيلة" ضمير متصل بالفعل المضارع المؤكد الثقيل مفرد
متكلم  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [ZSA0IEA0]   </Tms>
</Sol_12>
<Sol_13>
<Proclitique>   -Pas de Proclitique-   </Proclitique>
<Prefixe>   -Pas de prefixe-   </Prefixe>
<Suffixe>  نَّ    </Suffixe>
<Des_Suf> " نون التوكيد الثقيلة" ضمير متصل بالفعل المضارع المؤكد الثقيل مفرد
مخاطب  </Des_Suf>
<Enclitique>  هُمْ  </Enclitique>
<Des_Enc> "هاء الغائب" ضمير متصل جمع غائب"  </Des_Enc>
<Base>   0   </Base>
<Tms>  [MSB0IEA0]   </Tms>
</Sol_13>
```

Une partie du rapport d'analyse de la forme verbale ' فسيشكرونهم '.

Exemple 2

L'exemple suivant montre une partie du rapport d'analyse de la forme nominale ' والشاكرون '.

```
<?xml version='1.0' encoding='ISO-8859-6' standalone='no'?>
والشاكرون<AnalyseDeLaForme>
<RapportQuantitatif>
<NombreDe>
<FormesPossibles>113</FormesPossibles>
<FormeValideNom>1</FormeValideNom>
<FormeValideVerbe>0</FormeValideVerbe>
<FormeValideNomEtVerbe>0</FormeValideNomEtVerbe>
<FormeMoutOutil>0</FormeMoutOutil>
<FormeInconnueNom>3</FormeInconnueNom>
<FormeInconnueVerbe>12</FormeInconnueVerbe>
<FormeInconnueNomEtVerbe>2</FormeInconnueNomEtVerbe>
<SegmentationNonValide>95</SegmentationNonValide>
</NombreDe>
<TempsAnalyse> [0 : 0 : 0 : 390 ] </TempsAnalyse>
</RapportQuantitatif>
<FormeValideNom>
<Sol_100>
 </Proclitique>وَالْ<Proclitique>
 </Des_Pro>حرف مركب<Des_Pro>
```

<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>ونَ<Suffixe>
</Des_Suf>واو" جمع مذكر سالم مرفوع<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
,F,M,S,X,O,O) </Base>شكر,acic,شاكر<Base> [N/1](
<Tms> [MP0N0000] </Tms>
</Sol_100>
</FormeValideNom>
<FormeValideVerbe>
</FormeValideVerbe>
<FormeValideNomEtVerbe>
</FormeValideNomEtVerbe>
<FormeMotOutil>
</FormeMotOutil>
<FormeInconnueNom>
<Sol_22>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>ونَ<Suffixe>
</Des_Suf>واو" جمع مذكر سالم مرفوع<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [N/0] </Base>
<Tms> [MP0N0000] </Tms>
</Sol_22>
<Sol_58>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>ونَ<Suffixe>
</Des_Suf>واو" جمع مذكر سالم مرفوع<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [N/0] </Base>
<Tms> [MP0N0000] </Tms>
</Sol_58>
<Sol_79>
</Proclitique>وَالْ<Proclitique>
</Des_Pro>حرف مركب<Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [N/0] </Base>
<Tms> [00000000] </Tms>
</Sol_79>
</FormeInconnueNom>
<FormeInconnueVerbe>
<Sol_2>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بالفعل الماضى جمع غائبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [FPC0AZX0] </Tms>
</Sol_2>
<Sol_23>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نا الفاعلين" ضمير متصل بالفعل الماضى مثنى أو جمع<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>

<Base> [V/0] </Base>
<Tms> [XEA0AXA0] </Tms>
</Sol_23>
<Sol_37>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>ا<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [ZXB0MZA0] </Tms>
</Sol_37>
<Sol_38>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بالفعل الماضى جمع غائبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [FPC0AZX0] </Tms>
</Sol_38>
<Sol_59>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نا الفاعلين" ضمير متصل بالفعل الماضى مثنى أو جمع<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [XEA0AXA0] </Tms>
</Sol_59>
<Sol_72>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>ا<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بفعل الأمر جمع مخاطبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [FPB0MZA0] </Tms>
</Sol_72>
<Sol_73>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>ا<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نَّ<Suffixe>
</Des_Suf>نون التوكيد الثقيلة" ضمير متصل بفعل الأمر المؤكد الثقيل مفرد مخاطب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [MSB0MZA0] </Tms>
</Sol_73>
<Sol_74>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>ا<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>

</Suffixe>نْ<Suffixe>
</Des_Suf>نون التوكيد الخفيفة" ضمير متصل بفعل الأمر المؤكد الخفيف مفرد مخاطب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [MSB0MZA0] </Tms>
</Sol_74>
<Sol_75>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>اِ<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نَّ<Suffixe>
</Des_Suf>نون التوكيد الثقيلة" ضمير متصل بفعل الأمر المؤكد الثقيل مفرد مخاطبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [FSB0MZA0] </Tms>
</Sol_75>
<Sol_76>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>اِ<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نَّ<Suffixe>
</Des_Suf>نون التوكيد الثقيلة" ضمير متصل بفعل الأمر المؤكد الثقيل جمع مخاطب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [MPB0MZA0] </Tms>
</Sol_76>
<Sol_77>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>اِ<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نْ<Suffixe>
</Des_Suf>نون التوكيد الخفيفة" ضمير متصل بفعل الأمر المؤكد الخفيف مفرد مخاطبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [FSB0MZA0] </Tms>
</Sol_77>
<Sol_78>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
</Prefixe>اِ<Prefixe>
</Des_Pre>حرف أمر<Des_Pre>
</Suffixe>نْ<Suffixe>
</Des_Suf>نون التوكيد الخفيفة" ضمير متصل بفعل الأمر المؤكد الخفيف جمع مخاطب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [V/0] </Base>
<Tms> [MPB0MZA0] </Tms>
</Sol_78>
</FormeInconnueVerbe>
<FormeInconnueNomEtVerbe>
<Sol_1>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [N/0][V/0] </Base>
<Tms> [00000000] </Tms>
</Sol_1>

<Sol_36>
</Proclitique>وَ<Proclitique>
</Des_Pro>حرف عطف<Des_Pro>
<Prefixe> -Pas de prefixe- </Prefixe>
<Suffixe> -Pas de suffixe- </Suffixe>
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> [N/0][V/0] </Base>
<Tms> [00000000] </Tms>
</Sol_36>
</FormeInconnueNomEtVerbe>
<SegmentationNonValide>
<Sol_3>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>ونَ<Suffixe>
</Des_Suf>واو الجماعة " ضمير متصل بالفعل المضارع جمع مخاطب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> 0 </Base>
<Tms> [MPB0IIX0] </Tms>
</Sol_3>
<Sol_4>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بالفعل المضارع جمع مخاطبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> 0 </Base>
<Tms> [FPB0IIX0] </Tms>
</Sol_4>
<Sol_5>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>ونَ<Suffixe>
</Des_Suf>واو الجماعة " ضمير متصل بالفعل المضارع جمع غائب<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> 0 </Base>
<Tms> [MPC0IIX0] </Tms>
</Sol_5>
<Sol_6>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بالفعل المضارع جمع غائبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> 0 </Base>
<Tms> [FPC0IIX0] </Tms>
</Sol_6>
<Sol_7>
<Proclitique> -Pas de Proclitique- </Proclitique>
<Prefixe> -Pas de prefixe- </Prefixe>
</Suffixe>نَ<Suffixe>
</Des_Suf>نون النسوة" ضمير متصل بالفعل المضارع المنصوب جمع مخاطبة<Des_Suf> "
<Enclitique> -Pas d'enclitique- </Enclitique>
<Base> 0 </Base>
<Tms> [FPB0ISX0] </Tms>
</Sol_7>

Une partie du rapport d'analyse de la forme verbale والشاكرون.

2.2.2 Exemple d'analyse d'une phrase

Soit à analyser la phrase suivante :

فإذا جاء المـ ـ ـ ـطالب لهم لـ ـم يقاوموا مدافعتـ ـ ـه فيحتـ ـ ـاج صاحب الدولـ ـة حينئذ إلى الاستظهار بسواهم من أهل النجدة.

Avant d'exposer le résultat de l'analyse de cette phrase, nous allons décrire dans un premier temps les données de sortie de l'analyseur.

Chaque forme analysée est suivie par sa solution morphologique (ou ses solutions morphologiques si la forme est ambiguë). Une solution morphologique comprend :

- la base,
- la racine,
- la segmentation en proclitique, préfixe, suffixe et enclitique,
- la catégorie principale (N si Nom, V si verbe, P si mot outil, NBN si nombre et PONC si séparateur),
- la catégorie secondaire (si la catégorie principale est Nom), voir tableau numéro 18.
- l'ensemble des valeurs des variables morphosyntaxiques représenté par un vecteur de neuf cases organisé comme suit (voir figure 45) :

1	2	3	4	5	6	7	8	9
Genre	Nombre	Personne	Cas	Aspect	Mode	Voix	Transitivité	Humain

Figure 45 : Vecteur représentant les traits morphosyntaxiques de la solution morphologique

Le tableau ci-dessous (voir tableau 17) représente les valeurs que peut prendre chaque variable morphosyntaxique. Il y a lieu à noter que la valeur « 0 » pour une variable indique que la forme n'est pas concernée par cette variable. Par exemple une forme nominale ne peut pas avoir une valeur pour la variable « MODE ».

N°	Variable	Valeur	N°	Variable	Valeur
1	Genre	M : Masculin F : Féminin Z : Masculin ou Féminin X : Indéterminé	6	Mode	I : Indicatif S : Subjonctif A : Apocopé E : Energique I N : Energique II Z : Non Marqué X : Indéterminé
2	Nombre	S : Singulier D : Duel P : Pluriel C : Collectif T : Pluriel brisé X : Indéterminé	7	Voix	A : Active P : Passive X : Indéterminé
3	Personne	A : 1ère personne B : 2ème personne C : 3ème personne X : Indéterminée	8	Transitivité	I : Intransitive A : Tran. Simple Dire. B :Tran. Simple Indi. C :Tran. Double I D: Tran. Double II E : Tran. Double III F: Trans. Triple
4	Cas	N : Nominatif A : Accusatif G : Génitif Z : Non marqué X : Indéterminé	9	Humain	H : Humain N : Non humain X : Indéterminé
5	Aspect	A : Accompli I : Inaccompli M : Impératif X : Indéterminé			

Tableau 17 : Valeurs des variables morphosyntaxiques

La catégorie secondaire représente le deuxième niveau de catégorisation concernant la catégorie principale « NOM » ; elle peut prendre donc, les valeurs résumées dans le tableau 18.

Catégorie secondaire	**Désignation**
A	Masdar
B	Nom commun
C	Nom Propre
D	Nom de temps et de lieu
E	Adjectif assimilé
F	Participe actif
G	Participe passif
H	Nom d'instrument
I	Qualificatif de supériorité

Tableau 18 : Valeurs de la catégorie secondaire

A titre d'exemple, la figure 46 (une copie d'écran de notre analyseur) donne deux exemples de solution morphologique pour la forme verbale (فسيشكرونهم) et la forme nominale (والشاكرون).

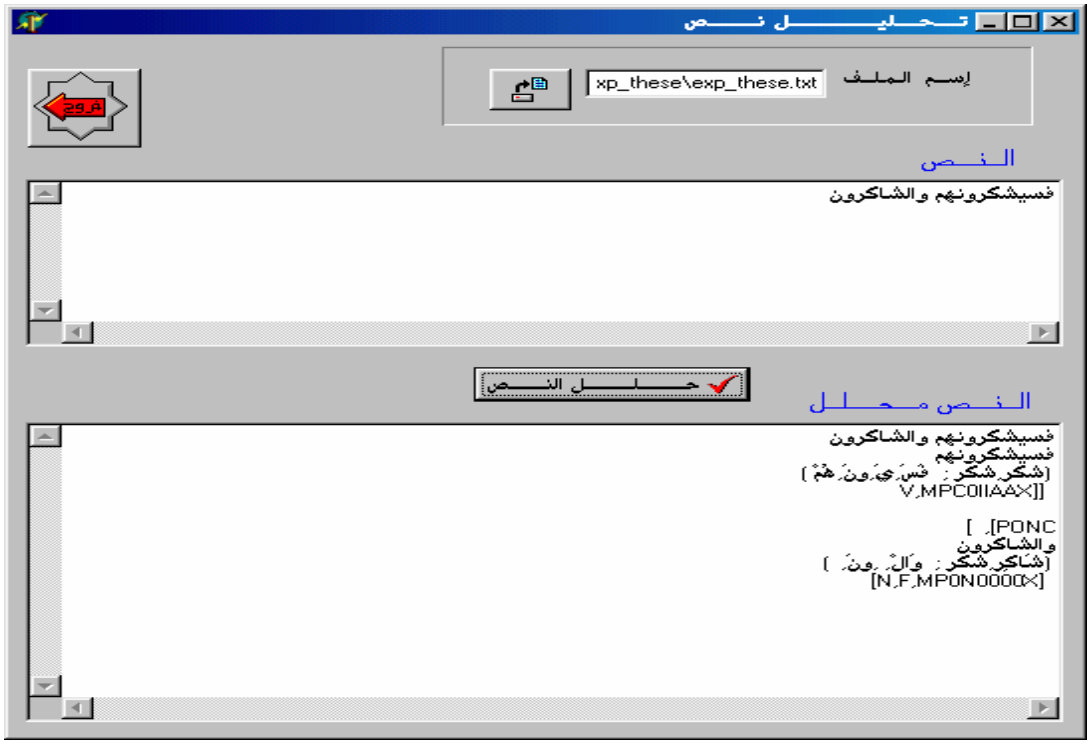


Figure 46 : Copie d'écran : Exemple d'analyse de deux formes

Les deux figures 47 et 48 illustrent respectivement les solutions morphologiques obtenues pour la forme verbale فسيشكرونهم et la forme nominale والشاكرون.

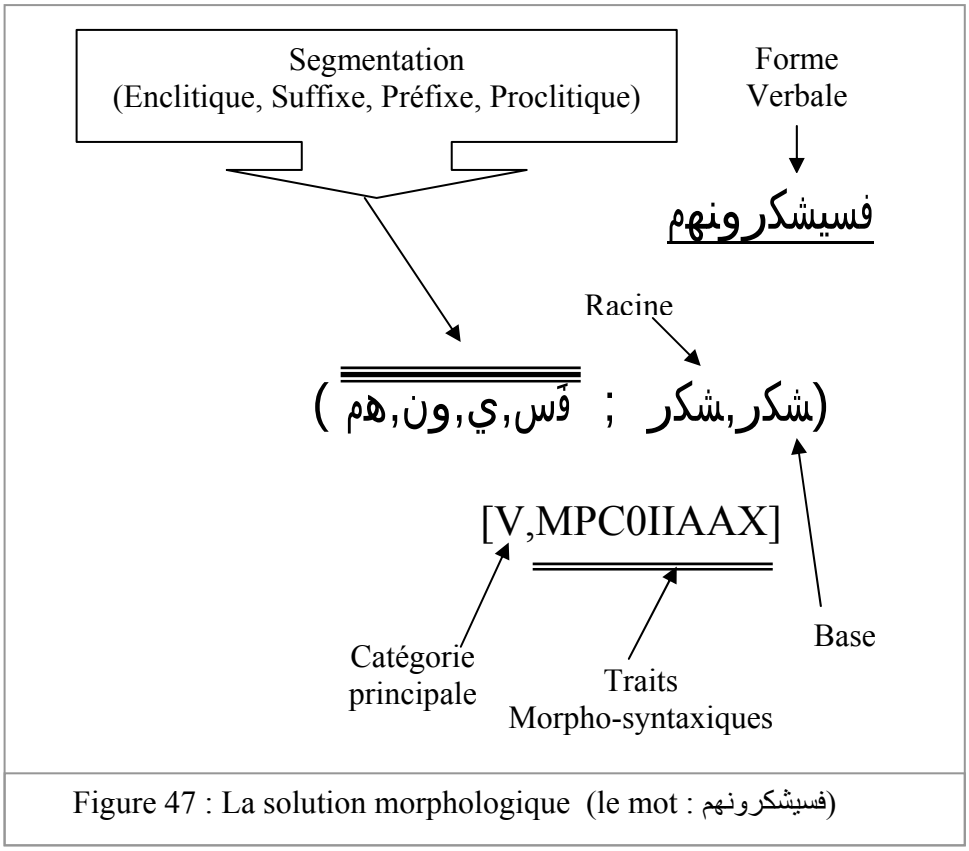


Figure 47 : La solution morphologique (le mot : فسيشكرونهم)

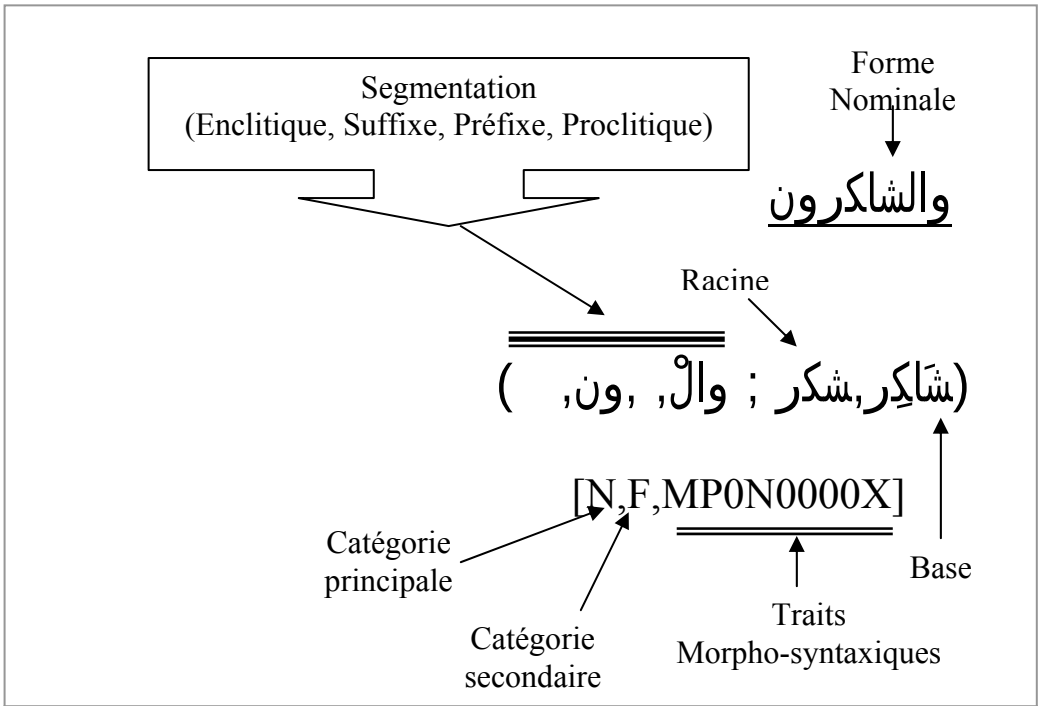


Figure 48 : Exemple d'une solution morphologique (le mot : والشاكرون)

Revenons maintenant à la phrase précédente, il s'agit d'une phrase reprise à partir du premier texte qui nous a servi dans la partie expérimentation (voir l'annexe G de cette thèse). Cette phrase présente les caractéristiques résumées dans le tableau 19

	Valeur	**%**
Nombre de mots	17	100%
Nombre de mots différents	17	100%
Nombre de formes nominales	7	41,18%
Nombre de formes verbales	3	17,64%
Nombre de mots outils	7	41,18%

Tableau 19 : Caractéristiques numériques de la phrase analysée

Le résultat de l'analyse de cette phrase est présenté dans ce qui suit.

فإذا جاء المطالب لهم لم يقاوموا مدافعتـه فيحتاج صاحب الدولـة حينئذ إلى الاستظهار بسواهم من أهل النجدة.

فإذا
(إِذَا ; فَ,)
[P,X]
(إِذًا ; فَ,)
[P,X]

[PONC,]
جاء
(جاء,جيأ ; , , ,)
[V,0000A00AX]

[PONC,]
المطالب
(مطَالِب,طلب ; الْ, , ,)
[N,B,MT000000H]
(مطَالِب,طلب ; الْ, , ,)
[N,F,MS000000X]
(مطَالَب,طلب ; الْ, , ,)
[N,G,MS000000X]

[PONC,]
لهم
(لَهم)
[P,N]
(هم,هم ; لَ, , ,)
[N,A,MS000000H]
(هِم,هم ; لَ, , ,)

[N,B,MS000000H]
(هم,هم ; ل, , ,)
[N,A,MS000000H]
(هِم,هم ; ل, , ,)
[N,B,MS000000H]

[,PONC]
لم
(لِم)
[P,X]
(لَم)
[P,V]
(لَم,لم ; , , ,)
[N,A,MC000000X]
(لَم,لم ; , , ,)
[V,0000A00AX]

[,PONC]
يقاوموا
(قاوم,قوم ; ,ي,وا,)
[V,MPC0ISAAX]
(قاوم,قوم ; ,ي,وا,)
[V,MPC0IAAAX]
(قاوم,قوم ; ,ي,وا,)
[V,MPC0ISPAX]
(قاوم,قوم ; ,ي,وا,)
[V,MPC0IAPAX]

[,PONC]
مدافعتـه
(مدافَعة,دفع ; , , ,ه)
[N,A,FS000000X]
(مدافِع,دفع ; , ,ت,ه)
[N,F,FS0X0000X]
(مدافَعة,دفع ; , , ,ِه)
[N,A,FS000000X]

[,PONC]
فيحتاج
(حتَاج,حوج ; فَ,ي, ,)
[V,ZXC0IXAAX]
(حتَاج,حوج ; فَ,ي, ,)
[V,ZXC0IXPAX]
(حتَاج,حوج ; فَ,ي, ,)
[V,ZXC0IXAAX]
(حتَاج,حوج ; فَ,ي, ,)
[V,ZXC0IXPAX]

[,PONC]
صاحب
(صاحِب,صحب ; , ,)
[N,F,MS000000H]
(صاحب,صحب ; , ,)
[V,0000A00AH]

[,PONC]
الدولـة
(دولة,دال ; ال, , ,)
[N,D,FS000000N]

[,PONC]
حينئذ
(حِينَئِذٍ)
[P,X]
[,PONC]
إلى
(إِلَى)
[P,N]

[,PONC]
الاستظهار
(اِستِظْهار,ظهر ; ال, , ,)
[N,A,MS000000X]

[,PONC]
بسواهم
(سِوى ; بِ,هِم)
[P,N]

[,PONC]
من
(من)
[P,V]
(من)
[P,V]
(من)
[P,X]
(مِن)
[P,N]
(من,من ; , ,)
[N,B,MS000000N]
(من,منن ; , ,)
[V,0000A00AX]

[,PONC]

أهل
(أهل,أهل ; , , ,)
[N,B,MC000000H]
(أهل,أهل ; , , ,)
[V,0000A00AH]
(أه ل,أهل ; , , ,)
[V,0000A00AH]
(أهلّ,هلل ; , , ,)
[V,0000A00AH]
(هلْ ; أ,)
[P,X]

[,PONC]
النجدة
(نَجدة,نجد ; الْ, , ,)
[N,B,FS000000X]
[. ,PONC]
[,PONC]

Résultat de l'analyse de la phrase

Figure 49 : Copie d'écran de l'analyseur : Analyse d'une phrase

Pour mesurer les performances de notre analyseur nous avons emprunté les mesures de bruit et de silence déjà utilisées dans le domaine de la recherche de l'information pour mesurer par exemple les performances des moteurs de recherches. Dans notre cas, on enregistre un bruit[1] par exemple pour la forme : (أهل) (voir forme n°16 dans la phrase de l'exemple), l'analyseur propose cinq solutions possibles (voir Tableau 20).

Solution 1 (**solution correcte**) (forme nominale : les proches, digne)	(أَهل,أهل ; , , ,) [N,B,MC000000H]
Solution 2 (forme verbale : marier)	(أَهل,أهل ; , , ,) [V,0000A00AH]
Solution 3 (forme verbale : qualifier)	(أَهل,أهل ; , , ,) [V,0000A00AH]
Solution 4 (forme verbale : apparaître)	(أَهلّ,هلل ; , , ,) [V,0000A00AH]
Solution 5 (**cas de bruit**) (mot outil : est-ce que est-ce que)	(هل ; أ,) [P,X]

Tableau 20: La solution morphologique de la forme: (أهل)

La cinquième solution présente un bruit, son origine étant une mauvaise segmentation dû au modèle linguistique utilisé et non à l'algorithme de segmentation. Autrement dit, le modèle linguistique ne prend pas en compte les cas d'agglutination invalides des particules aux mots outils. Le nombre de ces cas étant limité, la solution la plus simple à ce problème consiste à recenser tous ces cas et les consignés dans un dictionnaire. Dans ce cas l'analyseur doit consulter ce dictionnaire à chaque fois lorsqu'il rencontre une segmentation incluant un mot outil et une particule.

Les causes du silence[2] sont de deux types : soit les dictionnaires sont incomplets, soit les règles sont contraignantes. Pour résoudre le problème du silence on doit : mettre à jour les dictionnaires et revoir les règles. Dans notre cas nous avons procédé à la mise à jour de nos dictionnaires avec toutes les bases et mots outils qui existent dans les textes d'évaluation au départ pour éliminer la première cause du silence (Nous avons

[1] Pour la définition de "bruit" voir le chapitre trois, page 54.
[2] Pour la définition du "silence" voir le chapitre trois, page 54.

volontairement recensé manuellement tous les bases et mots outils contenant dans les textes d'évaluation et nous les avons consignés dans les différents dictionnaires utilisés par l'analyseur. Ce qui nous conduit dès le départ à éliminer l'hypothèse de l'incomplétude de nos dictionnaires comme source de silence).

En plus du bruit et de silence, nous avons calculé le taux d'ambiguïté qui correspond tout simplement au taux de formes qui possèdent plusieurs solutions morphologiques, c'est le cas de la forme précédente (cinq solutions morphologique sont proposées par l'analyseur).

L'analyse automatique montre les différents taux illustrés dans le tableau 21:

	Taux
Bruit	5.88%
Silence	0%
Temps analyse (Sec : Msec)	1 :600
Ambiguïté	58.82%

Tableau 21: Résultat de l'analyse automatique de la phrase.

Le temps (CPU) d'analyse est obtenu sur un micro-ordinateur P4 de 256 Mo de mémoire RAM.

Dans le tableau 21 on remarque qu'il y a plusieurs formes ambiguës (58.82%), c'est le cas des formes (1, 3, 4, 5, 6, 7, 8, 9, 15, 16).

On distingue quatre types d'ambiguïtés :

- formes homographes : formes qui différent par la catégorie grammaticale, c'est le cas des formes (4, 5, 9, 16),
- formes qui diffèrent par certaines valeurs de variables grammaticales, c'est le cas des formes (6, 8),
- formes qui diffèrent par la racine, c'est le cas des formes (15, 16),
- formes qui diffèrent par la segmentation, c'est le cas de la forme (7).

Les formes non reconnues par l'analyseur sont signalées par l'étiquette « Forme inconnue ».

3. Expérimentation

Pour valider la méthode que nous avons proposée, nous présentons dans ce qui suit les résultats obtenus par l'analyseur sur plusieurs textes.

Les textes ainsi que des parties de textes analysés se trouvent dans l'annexe G.

Le tableau 22 montre les différentes caractéristiques des textes de l'expérimentation.

	Texte I	**Texte II**	**Texte III**
Taille du texte en nombre de mots graphiques arabe (formes)	654	362	565
Nombre de formes différentes	396	239	370
Nombre de formes nominales	323	202	289
Nombre de formes verbales	96	62	79
Nombre de mots outils	229	98	276

Tableau 22 : Caractéristiques numériques des textes utilisés

Après analyse automatique nous avons observé les résultats illustrés dans le tableau 23 :

	Taux Texte 1	Taux Texte 2	Taux Texte 3
Bruit	2.14%	1.9%	1.59%
Silence	0%	0%	0%
Ambiguïté	51.22%	41.16%	44.60%

Tableau 23 : Résultat de l'analyse automatique des textes utilisés

Les résultats obtenus dans le tableau 23 suggèrent que l'analyseur propose pour chaque forme analysée au moins la solution morphologique correcte. Autrement dit, l'analyseur ne produit pas de silence (0%) en admettant bien sûr que nos dictionnaires sont théoriquement exhaustifs. Ce constat nous conduit à dire que notre algorithme d'analyse est correct. Par ailleurs le taux de bruit enregistré a une moyenne de 1.88% (2.14%, 1.9% et 1.59) suggère qu'il y a des segmentations invalides. L'origine de ce problème n'est pas l'algorithme de segmentation lui-même, mais c'est plutôt les règles de validations des segments du mot graphique qui sont mis en cause, ou tout simplement c'est le modèle linguistique utilisé qui comporte des incohérences et par conséquent il ne peut prétendre à la capture de toute la réalité linguistique propre à la structure du MGA, d'où l'inconvénient majeur de ce modèle. Finalement on remarque qu'il y a plusieurs formes ambiguës en moyenne 45.66% (51.22%, 41.16% et 44.60%). L'origine n'est pas

l'analyseur lui-même, mais c'est plutôt la langue arabe (à cause du manque des voyelles et du problème de l'agglutination). La seule façon de régler le problème de l'ambiguïté dans ce cas, c'est de développer des outils pour la désambiguïsation. Deux approches sont possibles, la première approche consiste à utiliser un modèle linguistique à base de règles contextuelles, alors que la seconde approche utilise un modèle statistique qui donne la séquence la plus probable des catégories trouvées lors de l'analyse. [LALL-1990]
Par exemple, dans le cadre de sa thèse, [KALL-1987] a choisi comme modèle statistique celui des chaînes de Markov pour la désambiguïsation du français.

4. Discussion

Comparer notre analyseur à d'autres est une tâche très difficile et reste toujours subjective du moment où il n'existe pas de standard en terme de critères pour pouvoir faire cette confrontation. Chaque analyseur possède sa propre sortie et cible bien une application spécifique. Cependant on pourra avancer ces quelques remarques.
Notre analyseur diffère des autres analyseurs qui existent pour la langue arabe par le fait qu'il :

- ne cible pas une application spécifique,
- ne réalise pas de prétraitement du texte en entrée,
- peut être utilisé pour les textes voyellé ou non,
- repose sur un modèle sûr et cohérent.

En plus de ces avantages, la réalisation de cet analyseur est simplifiée (seulement un algorithme de segmentation des formes et un autre pour la validation des segments) par le fait qu'il repose sur un framework qui implémente le modèle du mot graphique arabe. L'idée de la modélisation du MGA et par conséquent la séparation entre tâche du linguiste et développeur informaticien, n'a jamais était abordée dans les analyseurs existants. Elle ouvre une nouvelle perspective pour le développement d'une nouvelle génération d'applications pour le TALN arabe.
La taille de nos lexiques actuels étant limitée (comparée à d'autres lexiques), on ne peut donc prétendre à une large couverture. Une opération de mise à jour de nos lexiques est nécessaire.

5. Conclusion

Après avoir justifié notre choix pour l'organisation du lexique afin de construire notre modèle de départ pour la modélisation des objets linguistiques propre à la langue arabe, nous avons proposé d'élaborer une plate forme pour le développement des applications TALN arabe. Une conséquence remarquable est la séparation entre tache de linguiste et développeur. Pour la validation un analyseur morphologique basé sur cette plateforme est ainsi construit. Cet analyseur peut être exploité par des applications de TALN comme la traduction automatique, la correction orthographique, la recherche d'information, etc.

En faisant abstraction du problème de la taille limitée de nos lexiques actuels, une première expérimentation nous a montré que l'analyseur donne de bonnes performances Par ailleurs il nous semble qu'un outil pour la désambiguïsation est nécessaire pour le compléter.

Conclusion

Conclusion générale

L'objectif principal visé par ce travail était de concevoir et de réaliser un système pour la reconnaissance des unités linguistiques signifiantes qui peut être réutilisé dans les applications de TALN arabe. Nous avons proposé dans le cadre de cette thèse :

- une organisation adéquate pour le lexique, cette dernière nous a permet, d'une part, d'éviter de gérer un lexique trop volumineux et difficile à mettre à jour et d'autre part, de ne garder que les éléments essentiels nous permettant de générer les « mots » dont on a besoin.

- Une modélisation en terme de classe (classe dans le paradigme objet) du mot graphique arabe. La modélisation objet est un choix dicté par la nature même de l'arabe, car, à l'inverse de la structure simple du mot graphique d'une langue comme le français ou l'anglais, le mot graphique arabe possède une structure complexe et demande par conséquent une modélisation plus riche pour une prise en charge facile par un système de TALN. Le modèle que nous avons construit peut être exploité par différentes applications dans le domaine du TALN arabe comme la traduction automatique, la correction orthographique, la recherche d'information, etc. Par ailleurs l'utilisation du concept de classe (très puissant et intuitif) nous a permis d'obtenir un modèle (pour le TALN arabe) clair et réutilisable. Une conséquence immédiate de cette modélisation est l'implémentation facile vue la disponibilité des langages de programmation supportant la notion de classe et l'existence des outils pour la modélisation. Pour la description de nos modèles, nous avons opté pour l'utilisation du langage UML pour plusieurs raisons, la plus importante étant sa normalisation par l'OMG (www.omg.org) (les spécifications sont accessibles gratuitement), c'est un langage universel pouvant servir de support pour tout langage orienté objet. L''idée de la modélisation du mot graphique arabe et par conséquent la séparation entre tâche du linguiste et développeur informaticien, n'a jamais était abordée dans les analyseurs existants. Elle ouvre une nouvelle perspective pour le développement d'une nouvelle génération d'applications pour le TALN arabe.

- Un framework nommé « MALA ». Ce dernier étant fondé sur notre modélisation, il va nous servir comme une base pour le développement des applications TALN arabe. « MALA » repose sur deux composantes principales, une base de données linguistique intégrant toutes les données linguistiques propres à la langue arabe, et un ensemble de primitives ou de méthodes pour la manipulation de cette base de données linguistique. De nombreux avantages sont ainsi obtenus, comme par exemples, la séparation entre les données linguistiques et les programmes qui les manipulent, la réutilisation, la normalisation des développements…
- Un analyseur morphologique pour la validation, cet analyseur est capable d'analyser les textes arabes voyellés ou non. Nous estimons que les résultats obtenus par cet analyseur sont encourageants. Notre analyseur diffère des autres analyseurs qui existent pour la langue arabe par le fait qu'il repose sur un modèle sûr et cohérent et ne cible pas une application spécifique.
- Un ensemble d'outils indispensables pour la gestion des données linguistiques.

Dans l'état actuel de notre système et pour faire une évaluation détaillé, il est important de définir un domaine d'application et de faire une étude approfondie basée sur corpus. Toutefois, une évaluation réelle et détaillée des performances sur un corpus de textes tout-venant n'est pas une tâche évidente, elle nécessite une phase de préparation très importante.

Sur le plan des améliorations du travail présenté, nous préconisons les améliorations suivantes qui concernent essentiellement la phase de désambiguïsation dont la performance influe directement sur les taux de résolution des ambiguïtés. Donc, un travail important reste à faire :

- La prochaine étape consistera à compléter le système par une phase de désambiguïsation. Deux approches sont possibles, la première approche consiste à utiliser un modèle linguistique à base de règles contextuelles, alors que la seconde approche utilise un modèle statistique qui donne la séquence la plus probable des catégories trouvées lors de l'analyse. Le modèle statique le plus utilisé, notamment pour les autres langues comme le français, est celui des chaînes de Markov.
- Il faut ajouter au système d'analyse un module supplémentaire pour la reconnaissance des locutions. Cette reconnaissance présente un intérêt dans la réduction des ambiguïtés. Par locution on entend une suite de mots séparés

mais qui forment une unité de sens. Autrement dit : c'est une « *U*nité *fonctionnelle du langage, composée de plusieurs mots graphiques, appartenant à la langue et devant être apprise en tant que forme globale non divisible* » [http://fr.wiktionary.org].

Les locutions sont donc des expressions figées :

- ils forment une seule unité d'un point de vue sens et fonction grammaticale,
- leur sens est non décomposable, autrement dit la signification globale n'est pas obtenue à partir du sens de chaque composant.

Exemple

- كرة القدم (Football) :
- من حين لآخر (de temps en temps)
- بمجرد ما (dès que)

L'étape de la reconnaissance des locutions est donc intéressante car elle permet de lever un certain nombre d'ambiguïtés portant sur les mots qui forment l'expression.

Pour ceci, nous devons disposer d'un dictionnaire de locutions arabe ainsi que d'un algorithme qui permet de détecter ce type d'expressions dans le texte analysé.

- Le lexique à l'aide duquel nous avons mené nos expériences constitue un lexique initial et limité, il contient environ deux mille bases extraites d'une manière manuelle à partir de nos textes d'expérimentation. Donc, un travail important de mise à jour reste à faire pour que nous puissions prétendre à une large couverture de la langue.

Résumé

A l'heure actuelle, le Traitement Automatique des Langues Naturelles (TALN) et plus précisément la langue arabe, fait l'objet de nombreux travaux en ce qui concerne d'une part, la modélisation linguistique propre à la langue, d'autres part, la conception et la réalisation de logiciels pour divers domaines d'applications. Notre recherche participe à ces développements.

Plus spécifiquement, nous montrerons les limites des systèmes existants et proposons une approche basée objet pour la modélisation des connaissances linguistiques. Pour l'exploitation de ce modèle, nous l'avons intégré dans une plate forme, nommée « MALA », pour le développement des applications TALN arabe. Une conséquence remarquable est la séparation entre tache de linguiste et développeur. Pour la validation un analyseur morphologique basé sur cette plateforme est ainsi construit. En faisant abstraction du problème de la taille limitée de nos lexiques actuels, une première expérimentation nous a montré que notre analyseur donne de bonnes performances. Par ailleurs il nous semble qu'un outil pour la désambiguïsation est nécessaire pour le compléter.

Conjointement, un important travail de collecte de données linguistiques ainsi que l'élaboration d'un lexique arabe ont été réalisés.

MOTS CLES

TALN Arabe, Modélisation objet, Analyseur morphologique, Mot graphique arabe.

Abstract

Many different aspects of Natural Language Processing (NLP) and more precisely the Arabic language are presently studied: linguistic modelling, the design and the realization of software for various applications. Our research belongs to that stream of research.

We show the limits of actual systems and propose an object based approach for the language knowledge modelling. For the exploitation of this model, we integrated it in a framework, named "MALA", that allows the development of Arabic NLP applications. A remarkable consequence is the separation between the task of linguist and developer. For the validation, a morphological analyzer which can be based on the framework is built. By disregarding the problem of limited size of our current lexicon, a first experimentation showed us that the analyzer gives good performances. Furthermore it seems to us that a tool for the disambiguation is necessary to supplement it.

In addition, an important work of linguistic data collection as well as the development of an Arabic lexicon have been done.

KEY WORDS

Arabic NLP, Object Modelling, Morphological analyzer, Arabic graphical word.

ملخص

إن المعالجة الآلية للغات الطبيعية و بالخصوص اللغة العربية تشغل الكثير من البحوث في الوقت الحالي. هذه الأخيرة يمكن تلخيصها في جانبين، أما الأول فيتمثل في إيجاد نمادج لحوسبة اللغة، أما الثاني فهو يتعلق بتطوير برمجيات في مختلف الميادين التطبيقية للغة. إن موضوع بحثنا هذا يساهم في إثراء هاذين الجانبين.
في هذه الأطروحة سوف نبين نقائص الأنظمة الحالية و من ثم نقترح نموذج جديد لحوسبة البيانات اللغوية بالإعتماد على المقاربة بالكائنات. و للتمكين من إستغلال هذا النموذج قمنا بدمجه ضمن قاعدة للمعاجة الآلية للغة العربية وسمناها "مالا" من أبرز خصائصها التفرقة بين ماهو عمل الحاسوبي المطور للبرامج و ماهو عمل اللغوي. لقد قمنا بتطوير محال صرفي للنصوص العربية للتمكن من إثبات جدوى و مدى صلاحية النموذج المقترح. إن التجارب التي قمنا بها على المحال الصرفي سمحت لنا بالحصول على نسب مشجعة من التعرف الآلي وهذا بغض النظر عن الحجم الضئيل للقواميس المستعملة في هذه المرحلة. وبعد هذا كله تبين لنا ضرورة تطوير برمجيات لرفع اللبس أثناء تحليل الوحدات اللغوية الدالة. بالموازات مع هذا العمل قمنا بعملية ضخمة لتجميع المادة اللغوية المتمثلة في البيانات و القواعد اللغوية.

الكلمات المفتاحية

المعاجة الآلية للغة العربية، النمذجة عن طريق المقاربة بالكائنات، المحال الصرفي، الوحدة اللغوية الخطية.

Université Aboubekr Belkaid –Tlemcen-
Faculté des sciences de l'ingénieur
Département d'informatique

Résumé thèse de doctorat en informatique

Titre : Reconnaissance des unités linguistiques signifiantes
Présentée par : Mohammed El Amine ABDERRAHIM
Encadreur : Pr Fethi Breksi Reguig
Soutenue le : 08 juillet 2008

A l'heure actuelle, le Traitement Automatique des Langues Naturelles (TALN) et plus précisément la langue arabe, fait l'objet de nombreux travaux en ce qui concerne d'une part, la modélisation linguistique propre à la langue, d'autres part, la conception et la réalisation de logiciels pour divers domaines d'applications. Notre recherche participe à ces développements.

Plus spécifiquement, nous montrerons les limites des systèmes existants et proposons une approche basée objet pour la modélisation des connaissances linguistiques. Pour l'exploitation de ce modèle, nous l'avons intégré dans une plate forme, nommée « MALA », pour le développement des applications TALN arabe. Une conséquence remarquable est la séparation entre tache de linguiste et développeur. Pour la validation un analyseur morphologique basé sur cette plateforme est ainsi construit. En faisant abstraction du problème de la taille limitée de nos lexiques actuels, une première expérimentation nous a montré que notre analyseur donne de bonnes performances. Par ailleurs il nous semble qu'un outil pour la désambiguïsation est nécessaire pour le compléter.

Conjointement, un important travail de collecte de données linguistiques ainsi que l'élaboration d'un lexique arabe ont été réalisés.

MOTS CLES - KEY WORDS - الكلمات المفتاحية

TALN Arabe, Modélisation objet, Analyseur morphologique, Mot graphique arabe.
Arabic NLP, Object Modelling, Morphological analyzer, Arabic graphical word.
المعاجة الآلية للغة العربية، النمذجة عن طريق المقاربة بالكائنات، المحال الصرفي، الوحدة اللغوية الخطية.

Résumé

A l'heure actuelle, le Traitement Automatique des Langues Naturelles (TALN) et plus précisément la langue arabe, fait l'objet de nombreux travaux en ce qui concerne d'une part, la modélisation linguistique propre à la langue, d'autres part, la conception et la réalisation de logiciels pour divers domaines d'applications. Notre recherche participe à ces développements.

Plus spécifiquement, nous montrerons les limites des systèmes existants et proposons une approche basée objet pour la modélisation des connaissances linguistiques. Pour l'exploitation de ce modèle, nous l'avons intégré dans une plate forme, nommée « MALA », pour le développement des applications TALN arabe. Une conséquence remarquable est la séparation entre tache de linguiste et développeur. Pour la validation un analyseur morphologique basé sur cette plateforme est ainsi construit. En faisant abstraction du problème de la taille limitée de nos lexiques actuels, une première expérimentation nous a montré que notre analyseur donne de bonnes performances. Par ailleurs il nous semble qu'un outil pour la désambiguïsation est nécessaire pour le compléter.

Conjointement, un important travail de collecte de données linguistiques ainsi que l'élaboration d'un lexique arabe ont été réalisés.

MOTS CLES : TALN Arabe, Modélisation objet, Analyseur morphologique, Mot graphique arabe.

Abstract

Many different aspects of Natural Language Processing (NLP) and more precisely the Arabic language are presently studied: linguistic modelling, the design and the realization of software for various applications. Our research belongs to that stream of research.

We show the limits of actual systems and propose an object based approach for the language knowledge modelling. For the exploitation of this model, we integrated it in a framework, named "MALA", that allows the development of Arabic NLP applications. A remarkable consequence is the separation between the task of linguist and developer. For the validation, a morphological analyzer which can be based on the framework is built. By disregarding the problem of limited size of our current lexicon, a first experimentation showed us that the analyzer gives good performances. Furthermore it seems to us that a tool for the disambiguation is necessary to supplement it.

In addition, an important work of linguistic data collection as well as the development of an Arabic lexicon have been done.

KEY WORDS: Arabic NLP, Object Modelling, Morphological analyzer, Arabic graphical word.

ملخص

إن المعالجة الآلية للغات الطبيعية و بالخصوص اللغة العربية تشغل الكثير من البحوث في الوقت الحالي. هذه الأخيرة يمكن تلخيصها في جانبين، أما الأول فيتمثل في إيجاد نمادج لحوسبة اللغة، أما الثاني فهو يتعلق بتطوير برمجيات في مختلف الميادين التطبيقية للغة. إن موضوع بحثنا هذا يساهم في إثراء هاذين الجانبين.

في هذه الأطروحة سوف نبين نقائص الأنظمة الحالية و من ثم نقترح نموذج جديد لحوسبة البيانات اللغوية بالإعتماد على المقاربة بالكائنات. و للتمكين من إستغلال هذا النموذج قمنا بدمجه ضمن قاعدة للمعاجة الآلية للغة العربية وسمناها "مالا" من أبرز خصائصها التفرقة بين ماهو عمل الحاسوبي المطور للبرامج و ماهو عمل اللغوي. لقد قمنا بتطوير محال صرفي للنصوص العربية للتمكن من إثبات جدوى و مدى صلاحية النموذج المقترح. إن التجارب التي قمنا بها على المحال الصرفي سمحت لنا بالحصول على نسب مشجعة من التعرف الآلي وهذا بغض النظر عن الحجم الضئيل للقواميس المستعملة في هذه المرحلة. وبعد هذا كله تبين لنا ضرورة تطوير برمجيات لرفع اللبس أثناء تحليل الوحدات اللغوية الدالة. بالموازات مع هذا العمل قمنا بعملية ضخمة لتجميع المادة اللغوية المتمثلة في البيانات و القواعد اللغوية.

الكلمات المفتاحية : المعاجة الآلية للغة العربية، النمذجة عن طريق المقاربة بالكائنات، المحال الصرفي، الوحدة اللغوية الخطية.

Annexe A

La variable NMC (non marqué par le cas) rassemble une liste finie de noms. On trouve :

- les pronoms du nominatif ضمائر الرفع

أنا	أنت	أنت	هو	هي
نحن	أنتما	أنتما	هما	هما
نحن	أنتم	أنتن	هم	هن

- les pronoms de l'accusatif ضمائر النصب

إياي	إيانا	إياك	إياك
إياكما	إياكم	إياكن	
إياه	إياها	إياهم	إياهن

- les démonstratifs أسماء الإشارة

هذا	هذه	ذاك	ذلك	تلك	هؤلاء
أولئك	هنا	ههنا	هناك	هنالك	

- les relatifs الإسم الموصول

الذى	التى	الذين	اللاتى
اللائى	من	ما	

- les conditionnels إسم الشرط

من	ما	مهما	متى	أيان	أين
أينما	أنى	حيثما	كيفما	أى	

- les interrogatifs إسم الإستفهام

من	ما	متى	أين

كم	كيف	أى	

- les noms de nombres (de 11 à 19 sauf 12) الأعداد المركبة

أحد عشر	ثلاثة عشر	أربعة عشر	خمسة عشر
ستة عشر	سبعة عشر	ثمانية عشر	تسعة عشر

- quelques noms de lieu et de temps بعض الضروف

حيث	أمس	الآن	إذ

- noms verbaux إسم الفعل

هيهات	شتّان	سرعان	آه	أف	آمين
عليك	حذار	صه	إيه	حى	

Annexe B

Pour un verbe trilitère arabe, on compte six formes de conjugaison et dix formes dérivées. Par ailleurs un verbe quadrilitère accepte une forme de conjugaison et trois formes dérivées.

1. Schèmes des formes du verbe trilitère arabe (au nombre de 6)

- فَـعَـلَ – يَـفْـعَـلُ
- فَـعَـلَ – يَـفْـعَـلُ
- فَـعَـلَ – يَـفْـعَـلُ
- فَـعَـلَ – يَـفْـعَـلُ
- فَـعَـلَ – يَـفْـعَـلُ
- فَـعَـلَ – يَـفْـعَـلُ

2. Schèmes des formes dérivées du verbe trilitère (au nombre de 10)

- فعل - يفعل
- فاعل – يفاعل
- أفعل – يفعل
- تفعل – يتفعل
- تفاعل – يتفاعل
- إنفعل – ينفعل
- إفتعل – يفتعل
- إفعل – يفعل
- إستفعل – يستفعل
- إفعوعل – يفعوعل

3. Schème des formes du verbe quadrilitère arabe

- فعلل – يفعلل

4. Schèmes des formes dérivées du verbe quadrilitère (au nombre de 3)

- تفعلل – يتفعلل
- إفعنلل – يفعنلل
- أفعلل – يفعلل

A une racine arabe correspond un ou plusieurs verbes obéissant au vingt schèmes cités ci-dessus. L'ensemble de ces verbes est appelé champ dérivationnel verbal.

Annexe C

Les particules (Catégorie P)

La catégorie P se compose de trois sous catégories : PN, PV, et éventuellement PNV.

SCATP = {PN, PV, PNV}

a) La sous catégorie PN

حروف الجر

من	إلى	عن	فى	الباء	الكاف
اللام	واو القسم	تاء القسم	حتّى	رب	مذ
واو رب	عدا	خلا	حاشا		

إن و أخواتها

إن	أن	كأن	لكن
ليت	لعلّ	لا	

حروف النداء

يا	أيا	هيا	أى	الهمزة	

حرف الإستثناء: " إلا "

واو المعية

لام الإبتداء

b) La sous catégorie PV

حروف النصب

أن	لن	كى	إذن
لام الجحود	لام التعليل	فاء السببية	حتى

حروف الجزم

لم	لما	لام الأمر	لا الناهية	إن

قد

ما، لا

السين و سوف

c) La sous catégorie PNV

حروف العطف

الواو	الفاء	ثم	أو	أم
لكن	لا	بل	حتى	

حرفا الإستفهام: الهمزة و هل

واو الحال

لام القسم

Annexe D

Catégories des particules pré et postefixées

Nous avons recensé 201 catégories de particules pré et postfixées.

LES SUFFIXES

Particule	Catégorie	Code catégorie
تُ	Accompli 1er personne du singulier	PV_S_ACC1
تَ	Accompli 2ème personne masculin singulier	PV_S_ACC2
تِ	Accompli 2ème personne féminin singulier	PV_S_ACC3
تْ	Accompli 3ème personne féminin singulier	PV_S_ACC5
نَا	Accompli 1er personne duel & pluriel	PV_S_ACC6
تُما	Accompli 2ème personne duel	PV_S_ACC7
ا	Accompli 3ème personne masculin duel	PV_S_ACC8
تَا	Accompli 3ème personne féminin duel	PV_S_ACC9
تُم	Accompli 2ème personne masculin pluriel	PV_S_ACC10
تُن	Accompli 2ème personne féminin pluriel	PV_S_ACC11
وا	Accompli 3ème personne masculin pluriel	PV_S_ACC12
ن	Accompli 3ème personne féminin pluriel	PV_S_ACC13
ين	Inaccompli indicatif 2ème personne féminin singulier	PV_S_INA_IND3
ان	Inaccompli indicatif : 2ème per. duel, 3ème per. masculin duel, 3ème per. féminin duel	PV_S_INA_IND_7_8_9
ون	Inaccompli indicatif :2ème pre. masculine pluriel, 3ème per. masculin pluriel	PV_S_INA_IND_10_12
ي	Inaccompli subjonctif, apocopé :2ème per. féminin duel	PV_S_INA_SUB_APO_3
ا	Inaccompli subjonctif apocopé : 2ème per. duel, 3ème per. masculin duel, 3ème per. féminin duel	PV_S_INA_SUB_APO_7_8_9
وا	Inaccompli subjonctif apocopé :2ème pre. masculine pluriel, 3ème per. masculin pluriel	PV_S_INA_SUB_APO_10_12
ن	Inaccompli indicatif subjonctif, apocopé :2ème per. féminin pluriel, 3ème per. féminin pluriel	PV_S_INA_IND_SUB_APO_11_13

ـِني	Pronom personnel 1er personne singulier	PV_S_PRO1
ن	Inaccompli Energique I :1, 2, 3, 4, 5, 6,10, 12	PV_S_INA_EI_1_2_3_4_5_6_10_12
ان	Inaccompli Energique I :7, 8, 9	PV_S_INA_EI_7_8_9
نَان	Inaccompli Energique I :11, 13	PV_S_INA_EI_11_13
ن	Inaccompli Energique II :1, 2, 3, 4, 5, 6, 10, 12	PV_S_INA_EII_1_2_3_4_5_6_10_12
ان	Inaccompli Energique II :9	PV_S_INA_EII_9
ي	Impératif :3	PV_S_IMP_3
ا	Impératif :7	PV_S_IMP_7
وا	Impératif :10	PV_S_IMP_10
ن	Impératif :11	PV_S_IMP_11
ن	Impératif Energique I:2, 3, 10	PV_S_IMP_EI_2_3_10
ان	Impératif Energique I:7	PV_S_IMP_EI_7
نَان	Impératif Energique I:11	PV_S_IMP_EI_11
ن	Impératif Energique II:2, 3, 10	PV_S_IMP_EII_2_3_10
و	Accompli 3ème personne masculin pluriel	PV_S_ACC12-1
ن	Accompli 1er personne duel & pluriel	PV_S_ACC6[1]
ا	Inaccompli indicatif : 2ème per. duel, 3ème per. masculin duel, 3ème per. féminin duel	PV_S_INA_IND_7[2]_8[2]_9[2]
و	Inaccompli subjonctif apocopé :2ème pre. masculine pluriel, 3ème per. masculin pluriel	PV_S_INA_SUB_APO_10[1]_12[1]
و	Accompli 3ème personne masculin pluriel	PV_S_ACC12[1]
ا	Inaccompli Energique I :7, 8, 9	PV_S_INA_EI_7[3]_8[3]_9[3]
نَا	Inaccompli Energique I :11, 13	PV_S_INA_EI_11[3]_13[3]
و	Impératif :10	PV_S_IMP_10[1]
ا	Impératif Energique I:7	PV_S_IMP_EI_7[3]
نَا	Impératif Energique I:11	PV_S_IMP_EI_11[3]

N[1] : enclitique حذف "ا" الألف لاتصالها ب
N[2] : enclitique حذف " ن " النون " لاتصالها ب " ني "
N[3] : enclitique حذف " نِّ " النون " لاتصالها ب " ني "

ة,ت	Féminin Singulier	PN_S_FS
ان	Duel Nominatif	PN_S_DN
ين	Duel Accusatif Génitif	PN_S_DAG
ون	Pluriel Masculin Sain Nominatif	PN_S_PMSN
ين	Pluriel Masculin Sain Accusatif Génitif	PN_S_PMSAG
اتْ	Pluriel Féminin Sain	PN_S_PFS
ا	Duel Nominatif Modaf	PN_S_DNM
ي	Duel Accusatif Génitif Modaf	PN_S_DAGM
و	Pluriel Masculin Sain Nominatif Modaf	PN_S_PMSNM
ي	Pluriel Masculin Sain Accusatif Génitif Modaf	PN_S_PMSAGM
تَان	Duel Feminin Nominatif	PN_S_DFN
تَين	Duel Feminin Accusatif Génitif	PN_S_DFAG
تَا	Duel Feminin Nominatif modaf	PN_S_DFNM
تَي	Duel Feminin Accusatif Génitif modaf	PN_S_DFAGM
ا	Tanwin accusatif	TANWIN
ي، ويَّ	Nisba masculin	PN_S_NISBAM
ية،ـويّة	Nisba féminin	PN_S_NISBAF
يتان،ـويَّتَان	Nisba Duel Feminin Nominatif	PN_S_NISBA1
يتَين،ـويَّتَيْن	Nisba Duel Feminin Accusatif Génitif	PN_S_NISBA2
يتَا،ـويَّتَا	Nisba Duel Feminin Nominatif modaf	PN_S_NISBA3
يتَي،ـويَّتَيْ	Nisba Duel Feminin Accusatif Génitif modaf	PN_S_NISBA4
ياتْ،ـويَّتَا	Nisba Pluriel Féminin Sain	PN_S_NISBA5
يان،ـويَّان	Nisba Duel Masculin Nominatif	PN_S_NISBA6
يين،ـويَّيْن	Nisba Duel Masculin Accusatif Génitif	PN_S_NISBA7
يين،ـويِّيْن	Nisba Pluriel Masculin Sain Accusatif Génitif	PN_S_NISBA8
يون،ـويُّونَ	Nisba Pluriel Masculin Sain Nominatif	PN_S_NISBA9
يا،ـويًّا	Nisba Duel Masculin Nominatif Modaf	PN_S_NISBA10
يي،ـويَّيْ	Nisba Duel Masculin Accusatif Génitif Modaf	PN_S_NISBA11
يو،ـويُّو	Nisba Pluriel Masculin Sain Nominatif Modaf	PN_S_NISBA12
يي،ـويِّيْ	Nisba Pluriel Masculin Sain Accusatif Génitif Modaf	PN_S_NISBA13

LES PREFIXES

Particule	Catégorie	Code catégorie
أ	Préfixe inaccompli P1	PV_P_INA
ن	Préfixe inaccompli P6	PV_P_INA
ي	Préfixe inaccompli P4, P8, P12, P13	PV_P_INA
ت	Préfixe inaccompli P2, P3, P5,P7, P9, P10, P11	PV_P_INA
أ	Préfixe inaccompli P1	PV_P_INA
ن	Préfixe inaccompli P6	PV_P_INA
ي	Préfixe inaccompli P4, P8, P12, P13	PV_P_INA
ت	Préfixe inaccompli P2, P3, P5,P7, P9, P10, P11	PV_P_INA
أ,إ	Préfixe impératif	PV_P_IMP

LES ENCLITIQUES

Particule	Catégorie	Code catégorie
ي	Pronom personnel 1^{èr} personne masculin singulier	PN_E_PRO1
ك	Pronom personnel 2ème personne masculin singulier	PNV_E_PRO2X
كِ	Pronom personnel 2ème personne féminin singulier	PNV_E_PRO3X
ه	Pronom personnel 3ème personne masculin singulier	PNV_E_PRO4X1
ها	Pronom personnel 3ème personne féminin singulier	PNV_E_PRO4X2
نَا	Prohnohm personneu 1er per. duel, pluriel	PNV_E_PRO5X
كُما	Pronom personneu 2ème personne féminin masculin duel	PNV_E_PRO6X
هما	Pronom personneu 3ème personne féminin masculin duel	PNV_E_PRO7X
كُم	Pronom personneu 2ème personne masculin pluriel	PNV_E_PRO8X1,PRO9X1
كُن	Pronom personneu 2ème personne féminin pluriel	PNV_E_PRO8X2,PRO9X2
هم	Pronom personneu 3ème personne masculin pluriel	PNV_E_PRO10X
هن	Pronom personneu 3ème personne féminin pluriel	PNV_E_PRO11X

ـهِ	Pronom personneu 3ème personne masculin singulier	PNV_E_PRO12X1
ـهِما	Pronom personneu 3ème personne masculin féminin duel	PNV_E_PRO12X2
ـهِم	Pronom personnel 3ème personne masculin pluriel	PNV_E_PRO13X1
ـهِن	Pronom personnel 3ème personne féminin pluriel	PNV_E_PRO13X2
ـنِي	Pronom personnel composé P1	PV_E_EC1
ـكَه	Pronom personnel composé P2	PV_E_EC2
ـكِه	Pronom personnel composé P3	PV_E_EC3
ـكَها	Pronom personnel composé	PV_E_EC4
ـكِها	Pronom personnel composé	PV_E_EC5
ـكَهما	Pronom personnel composé	PV_E_EC6
ـكِهما	Pronom personnel composé	PV_E_EC7
ـكَهم	Pronom personnel composé	PV_E_EC8
ـكِهم	Pronom personnel composé	PV_E_EC9
ـكَهن	Pronom personnel composé	PV_E_EC10
ـكِهن	Pronom personnel composé	PV_E_EC11
وه	Pronom personnel composé	PV_E_EC12
وها	Pronom personnel composé	PV_E_EC13
ونَا	Pronom personnel composé	PV_E_EC14
وهما	Pronom personnel composé	PV_E_EC15
وهم	Pronom personnel composé	PV_E_EC16
وهن	Pronom personnel composé	PV_E_EC17
ونِي	Pronom personnel composé	PV_E_EC18
ا	Pronom personnel composé	PV_E_EC19

LES PROCLITIQUES

Particule	Catégorie	Code catégorie
أ	Interrogatif	PNV_P_INT
و,فَ	Coordonnant	PNV_P_CORD
لَ	Corroborateur	PNV_P_CORB
س	Future	PV_P_FUT
فَ	Subordonnant	PV_P_SUB

ل	Impératif ou subordonnant	PV_P_IMPSUB
بِ، ك، ل	Préposition	PN_P_PRE
الْ	Article	PN_P_ART
أبِ	Proclitique composé	PN_P_PC1
أكَ	Proclitique composé	PN_P_PC2
أوَبِ	Proclitique composé	PN_P_PC3
أوَكَ	Proclitique composé	PN_P_PC4
أوبِ	Proclitique composé	PN_P_PC5
أوكَ	Proclitique composé	PN_P_PC6
أقَالْ	Proclitique composé	PN_P_PC7
أوالْ	Proclitique composé	PN_P_PC8
أكَالْ	Proclitique composé	PN_P_PC9
أبِالْ	Proclitique composé	PN_P_P10
أَلِالْ	Proclitique composé	PN_P_PC11
أوِلالْ	Proclitique composé	PN_P_PC12
أقِلِالْ	Proclitique composé	PN_P_PC13
أقِبالْ	Proclitique composé	PN_P_PC14
أقَكَالْ	Proclitique composé	PN_P_PC15
أوِبالْ	Proclitique composé	PN_P_PC16
أوكَالْ	Proclitique composé	PN_P_PC17
بِالْ	Proclitique composé	PN_P_PC18
كَالْ	Proclitique composé	PN_P_PC19
وَبِ	Proclitique composé	PN_P_PC20
وَكَ	Proclitique composé	PN_P_PC21
قَالْ	Proclitique composé	PN_P_PC22
قِبالْ	Proclitique composé	PN_P_PC23
فَكَالْ	Proclitique composé	PN_P_PC24
فَلْبِ	Proclitique composé	PN_P_PC25
فَلْ	Proclitique composé	PN_P_PC26
فَلِ	Proclitique composé	PN_P_PC27
فَلْبالْ	Proclitique composé	PN_P_PC28
وبِ	Proclitique composé	PN_P_PC29
وكَ	Proclitique composé	PN_P_PC30
والْ	Proclitique composé	PN_P_PC31
وِبالْ	Proclitique composé	PN_P_PC32

وكَالْ	Proclitique composé	PN_P_PC33
وٱبِ	Proclitique composé	PN_P_PC34
وٱلْ	Proclitique composé	PN_P_PC35
وِلْ	Proclitique composé	PN_P_PC36
وٱلِيَالْ	Proclitique composé	PN_P_PC37
ٱبِ	Proclitique composé	PN_P_PC38
ٱلْ	Proclitique composé	PN_P_PC39
ٱلْ	Proclitique composé	PN_P_PC40
ٱلِيَالْ	Proclitique composé	PN_P_PC41
ٱلِلْ	Proclitique composé	PN_P_PC42
أَاس	Proclitique composé	PN_P_PC43
أَاوَس	Proclitique composé	PN_P_PC44
أَاوس	Proclitique composé	PN_P_PC45
وَس	Proclitique composé	PN_P_PC46
وِس	Proclitique composé	PN_P_PC47
أَافَ	Proclitique composé	PN_P_PC48
أَاو	Proclitique composé	PN_P_PC49
أَالَ	Proclitique composé	PN_P_PC50
أَال	Proclitique composé	PN_P_PC51
أَاوَلَ	Proclitique composé	PN_P_PC52
أَاوَل	Proclitique composé	PN_P_PC53
أَاولَ	Proclitique composé	PN_P_PC54
أَاول	Proclitique composé	PN_P_PC55
وَلَ	Proclitique composé	PN_P_PC56
وَل	Proclitique composé	PN_P_PC57
وَلْ	Proclitique composé	PN_P_PC58
ولَ	Proclitique composé	PN_P_PC59
ول	Proclitique composé	PN_P_PC60
ولْ	Proclitique composé	PN_P_PC61

Annexe E

Catégories des particules isolées

Remarque :

Emploi = N ➔ Cette particule s'emploie avec seulement un nom.
Emploi = V ➔ Cette particule s'emploie avec seulement un verbe.
Emploi = NV ➔ Cette particule s'emploie avec le nom et le verbe.

N° code catégorie	Catégorie particule	Emploi	Code
1	Pronom personnel ضمير رفع منفصل	NV	PRO1
2	Pronom personnel ضمير نصب أو جر منفصل	V	PRO2
3	Pronom démonstratif اسم إشارة	N	DEM
4	Pronom relatif اسم موصول	V	REL
5	Nom de nombres اسم العدد	N	NOM
6	Nom conditionnel اسم الشرط	V	CON
7	Nom interrogatif اسم استفهام	NV	INT
8	Métonymique كناية	NV	KIN
9	Nom verbaux أسماء الأفعال	NV	VER
10	Cordonnant حروف العطف	NV	CORD
11	Préposition حروف الجر	N	PRE

12	Inaa et ses analogues إن و أخواتها	N	INAA
13	Particule d'appel حروف النداء	N	APP
14	Particule d'exclusion حرف استثناء	N	EXC
15	Particule accusatif subordonnant حروف النصب	V	SUB
16	Particule composé حرف مركب	V	PC
17	Particule génitif de négation حروف الجزم	V	NEG
18	Corroborateur حروف التأكيد		CORB
19	Particule de futur حرف مستقبل	V	FUT
20	Nom de confirmation affirmatif حروف الإثبات		AF
21	ظروف		TH
22	قسم دعاء		SE

Annexe F

Algorithme Dévoyellation (Forme)

Entrée : Forme { *non voyellé, non voyellé avec 'Shadda', partiellement voyellé ou complètement voyellé* }
Sortie : Schéma_Consonantique, Schéma_Vocalique

Utilise les fonctions
Taille (X :chaîne) : { *cette fonction retourne la taille de la chaîne X* }
X+Y : { *cette fonction retourne une chaîne qui contient la concaténation de deux chaînes X et Y* }

Constantes
Voyelle_sans_Glide :{ *Ensemble des voyelles de l'arabe sans les glides.* }

Variables
I : entier
Schéma_Consonantique, Schéma_Vocalique : chaîne

Début
Schéma_Consonantique ← ''
Schéma_Vocalique ← ''
Pour I **allant de** 1 **à** Taille (Forme) **Faire**
Si Forme[I] ∈ Voyelle_sans_Glide **Alors**
Schéma_Vocalique ← Schéma_Vocalique + Forme[I]
Sinon Schéma_Consonantique ← Schéma_Consonantique + Forme[I]
Fin_Si
Fin_Faire
Fin

Annexe G

Texte 1

الفصل الرابع عشر في أن الدولة لها أعمار طبيعية كما للأشخاص

اعلم أن العمر الطبيعي للأشخاص على ما زعم الأطباء والمنجمون مائة وعشرون سنةً وهي سنو القمر الكبرى عند المنجمين ويختلف العمر في كل جيل بحسب القرانات فيزيد عن هذا وينقص منه فتكون أعمار بعض أهل القرانات مائة تامة وبعضهم خمسين أو ثمانين أو سبعين على ما تقتضيه أدلة القرانات عند الناظرين فيها وأعمار هذه الملة ما بين الستين إلى السبعين كما في الحديث ولا يزيد على العمر الطبيعي الذي هو مائة وعشرون إلا في الصور النادرة وعلى الأوضاع الغريبة من الفلك كما وقع في شأن نوح عليه السلام وقليل من قوم عاد وثمود.

وأما أعمار الدول أيضاً وإن كانت تختلف بحسب القرانات إلا أن الدولة في الغالب لا تعدو أعمار ثلاثة أجيال والجيل هو عمر شخص واحد من العمر الوسط فيكون أربعين الذي هو انتهاء النمو والنشوء إلى غايته قال تعالى: حتى إذا بلغ أشده وبلغ أربعين سنة ولهذا قلنا أن عمر الشخص الواحد هو عمر الجيل ويؤيده ما ذكرناه في حكمة التيه الذي وقع في بنى إسرائيل وأن المقصود بالأربعين فيه فناء الجيل الأحياء ونشأة جيل آخر لم يعهدوا الذل ولا عرفوه فدل على اعتبار الأربعين في عمر الجيل الذي هو عمر الشخص الواحد و إنما قلنا أن عمر الدولة لا يعدو في الغالب ثلاثة أجيال لأن الجيل الأول لم يزالوا على خلق البداوة وخشونتها وتوحشها من شظف العيش والبسالة والافتراس والاشتراك في ا لـد فلا تزال بذلك سورة العصبية محفوظةً فيهم فحدهم مرهف وجانبهم مرهوب والناس لهم مغلوبون والجيل الثاني تحول حالهم بالملك والترفه من البداوة إلى الحضارة ومن الشظف إلى الترف والخصب ومن الاشتراك في ا لـد إلى انفراد الواحد به وكسل الباقين عن السعي فيه ومن عز الاستطالة إلى ذل الاستكانة فتنكسر سورة العصبية بعض الشيء وتؤنس منهم المهانة والخضوع ويبقى لهم الكثير من ذلك بما أدركوا الجيل الأول وباشروا أحوالهم وشاهدوا اعتزازهم وسعيهم إلى ا لـد ومراميهم في المدافعة والحماية فلا يسعهم ترك ذلك بالكلية وإن ذهب منه ما ذهب و يكونون على رجاء من مراجعة الأحوال التي كانت للجيل الأول أو على ظن من وجودها فيهم وأما الجيل الثالث فينسون عهد البداوة والخشونة كأن لم تكن ويفقدون حلاوة العز والعصبية بما هم فيه من ملكة القهر ويبلغ فيهم الترف غايته بما تبنقوه من النعيم وغضارة العيش فيصيرون عيالاً على الدولة ومن جملة النساء والولدان المحتاجين للمدافعة عنهم وتسقط العصبية بالجملة وينسون الحماية والمدافعة والمطالبة ويلبسون على الناس في الشارة

والزي وركوب الخيل وحسن الثقافة يموهون با وهم في الأكثر أجبن من النسوان على ظهورها فإذا جاء المطالب لهم لم يقاوموا مدافعته فيحتاج صاحب الدولة حينئذ إلى الاستظهار بسواهم من أهل النجدة ويستكثر بالموالى ويصطنع من يغني عن الدولة بعض الغناء حتى يتأذن الله بانقراضها فتذهب الدولة بما حملت فهذه كما تراه ثلاثة أجيال فيها يكون هرم الدولة وتخلفها ولهذا كان انقراض الحسب في الجيل الرابع كما مر في أن ا لـد والحسب إنما هو أربعة آباء وقد أتيناك فيه ببرهان طبيعي كاف ظافر مبني على ما مهدناه قبل من المقدمات فتأمله فلن تعدو وجه الحق إن كنت من أهل الإنصاف وهذه الأجيال الثلاثة عمرها مائة وعشرون سنة على ما مر ولا تعدو الدول في الغالب هذا العمر بتقريب قبله أو بعده إلا أن عرض لها عارض آخر من فقدان المطالب فيكون الهرم حاصلاً مستولياً والطالب لم يحضرها ولو قد جاء الطالب لما وجد مدافعاً فإذا جاء أجلهم لا يستأخرون ساعةً ولا يستقدمون فهذا العمر للدولة بمثابة عمر الشخص من التزايد إلى سن الوقوف ثم إلى سن الرجوع ولهذا يجري على ألسنة الناس في المشهور أن عمر الدولة مائة سنة وهذا معناه فأعتبره واتخذ منه قانوناً يصحح لك عدد الآباء في عمود النسب الذي تريده من قبل معرفة السنين الماضية إذا كنت قد استربت في عددهم وكانت السنون الماضية منذ أولهم محصلة لديك فعد لكل مائة من السنين ثلاثة من الآباء فإن نفدت على هذا القياس مع نفود عددهم فهو صحح وإن نقصت عنه بجيل فقد غلط عددهم بزيادة واحد في عمود النسب وأن زادت بمثله فقد سقط واحد وكذلك تأخذ عدد السنين من عددهم إذا كان محصلاً لديك فتأمله تجده في الغالب صحيحاً والله يقدر الليل والنهار.

Texte 2

الحياء

صاحبه حيُ الضمير , مرهف المشاعر , محب للتواضع , عاشق للهدوء والسكينة 0

الحياء مزية إسلامية لكل مؤمن خالص إيمانه لله , وشعور يخالج النفس فيترجم على ملامح الوجه 0 والحياء خُلُق الإسلام كما قال صلى الله عليه وسلم " لكل دين خلقا وخلق الإسلام الحياء " رواه مالك والحياء أنواع وأولها وأجلها هو الحياء مع الله سبحانه وتعالى عن ابن مسعود رضي الله عنه قال : قال الرسول صلى الله عليه وسلم " استحيوا من الله حق الحياء 0قلنا إنا نستحي من الله يا رسول الله ـ ـ والحمد لله قال : ليس ذلك0 الاستحياء من الله حق الحياء أن تحفظ الرأس وما عوى , والبطن وما حوى , وتذكر الموت والبلى , ومن أراد الآخرة ترك زينة الدنيا ,وآثر الآخرة على الأولى فمن فعل ذلك فقد استحيا من الله حق الحياء " رواه الترمذي 0

هب البعث لم تأتنا رسله *** وحاطمة النار لم تضرم

أليس من الواجب المستحق *** حياء العباد من المنعم

ثم يأتي الحياء مع رسول الله صلى الله عليه وسلم باتباع أوامره واجتناب نواهيه ومحبته والصلاة عليه دائماً , وقراءة سيرته وآثاره , وإفهام الناس وتنويرهم بعلو شأنه وسمو قدره وجلال سنته صلى الله عليه وسلم 0 وهناك الحياء في الكلام بتـ يه اللسان عن كل ما يعيبه ويضجه , والحياء في المعاملات مع الناس بإظهار خُلُق المسلم الحقيقي , والابتعاد عن النفاق والبدع وا املات الرخيصة التي لا تعدو إلا سفا سف أمور لا يلتفت إليها 0

قال علي كرم الله وجهه : من كسى بالحياء ثوبه لم يرى الناس عيبه 0

وقيل : إن الحياء دليل الخير كلّه 0

وقال بعض العلماء : حقيقة الحياء " خُلُق يبعث على ترك القبيح , ويمنع من التقصير في حق ذي الحق 0

والحياء ابتعاد عن الشبهات , وترفّع على الدنيئات , والتشبث بالمرؤات , والخوف على المكارم , والحرص على المحامد 0

ومن لزم الحياء فهو مؤمن , ومن فارقه فقد وقع أسير الذنوب والمعاصي والله المستعان 0

يقول ابن القيم رحمه الله : ومن عقوبا ا - أي الذنوب والمعاصي - ذهاب الحياء الذي هو مادة الحياة للقلب , وهو أصل كل خير , وذهابه ذهاب كل خير بأجمعه 0

وفي الصحيح أنه قال صلى الله عليه وسلم " الحياء خير كله " وقال " أن مما أدرك الناس من كلام النبوة الأولى :"إذا لم تستحي فاصنع ما شئت " أخرجه البخاري وأحمد

وعن أبي هريرة رضي الله عنه أن رسول الله صلى الله عليه وسلم :" الإيمان بضع وسبعون شعبه أو بضع وستون شعبه , فأفضلها قوا لا إله إلا الله , وأدناها إماطة الأذى عن الطريق , والحياء شعبة من الإيمان" متفق عليه0

Texte 3

الأزمنة في اللغة

أما بعد: فإن الأزمنةَ ذاتُ أهمّيةٍ بالغةٍ من حيث علاقة الفعل با، وإنْ لم تكن من المسائل الأساسية للّغةِ . فإن علماءَ العربيةِ وإنْ كانوا قد خاضوا مثار البحثِ والجدل في دقائقِ علومِ اللغةِ، إلاّ أنهم لم يكترثوا لهذا الأمرِ أو لم ينـتبهوا له بكل ما يستحقهُ من اهتمامٍ. ولعلّ التقليد قد حجبهم عن اكتشافهِ، وقد يكون ذلك ناشئاً من أسبابٍ أُخرى.

هذا، ولا نقولُ أنّ علماءَ العربيةِ لم يفْطُنوا إلى حقيقةِ الزمان من حيث علاقة الفعل به، إذ لا يجوز أن يكون قد التبس عليهم الفرقُ بين معاني الفعل الخالي عن القيودِ الزمانية، وبين الفعل المقيدِ با مثل (طَلَب) و(سبق أنْ طَلَب) و (يكُونُ قَد طَلَب) و(لَولاهُ لَما طَلَب). فإنّ الفرق بين هذه التراكيبِ المختلفةِ واضحةٌ بينةٌ. ولابد أن يكونوا قد أدركوا هذا الفرق، إلاّ أنهم لم يدخلوا في تفاصيلها ولم يجعلوا لكل من الصيغ الزمنية باباً خاصاً كما قد تم ذلك في كثيرٍ من اللّغات.

ويغلبُ أنْ يكونَ هذا ناشِئاً من عدمِ علمِهم بِلُغاتٍ أجنبيةٍ ، إذْ أن كثيراً من المفاهيمِ والحاجاتِ والْمُقْتضياتِ لا يتبادرُ إلى الذهنِ تلقائياً وبسهولةٍ إلا بعد ظهورِ سببٍ يذكرُهُ أو ضرورةٍ تستدعيه، فهذه المنجزاتُ الحضاريةُ والتقنيةُ الضخمةُ التي نلمِسُها في مجالاتٍ مختلفةٍ من حياتنا، تُبرهِنُ على هذه الحقيقةِ، إذْ ليس من القليل ما قد كشفه العلماءُ والخبراءُ وما عثَر عليها الباحثون وأبدعها أهلُ الذّوق والفن من صناعاتٍ وعلومٍ وآلاتٍ وأجهزةٍ بقرينةٍ بسيطةٍ في بداية الأمرِ، أو بالقياس بين شيئين ِاستحسوا المكنوناتِ في أحدِهما بفضل ما في الآخرِ من أمثالِها.

هذا، لابد وأنْ نؤكد بأن للفعل علاقةٌ عُضويةٌ بمفهومِ الزمان، وما من فعلٍ إلا ولِحُدوُثِهِ وقت؛ والوقتُ ِقسطٌ من الزمان.

أما الزمانُ فإنه مفهوم معقد لم يتمكن العلماءُ من الوصول إلى حقيقته بعد. وهو ناشئ من دوران الكرة الأرضية حول محورها وعلى مدارٍ مُعينٍ مرتبطةً فيهما بالشمس، يعني أن الأرض تجري في ذات الوقت حول الشمس على مدارٍ مُعينٍ، إضافةً إلى جريا ا حولَ محورِها فيتمخضُ عن الأول المواسمُ الأربعةُ، وعن الثاني اللّيلُ والنهارُ المتعاقبان.

والوحدة القياسية للزمان هي الساعة ولا يسع المقام لحصر ما يدخل في هذا الباب من تفاصيلَ جانبيةٍ.

أما الزمانُ بالنسبة للفعل، فإنه جدير بالإهتمام؛ ذلك أن للفعل مراتبُ زمنيةٌ مختلفةٌ. ولهذا فإن علاقة الفعل بالزمان أشمل بكثير من القدر الذي حصرها علماء العربية في صِيغ ثلاثٍ. وربما لم يكن عَرضُهُمُ الوقوف على مفهوم الزمان مباشرةً، وإنما أرادوا أن يتوصلوا إلى تعريفٍ للفعل يميزُهُ عن الإسمِ والحرفِ، فاقتصروا في هذه المحاولة على تقسيمه إلى الماضي والحال والإستقبال فحسب. وقد سمى بعضُهُم الصيغَ بالأزمنة الثلاثة. لذا، فإنّ المسئلةَ غيرُ واضحةٍ؛ حيث أن طالب اللّغةِ العربيةِ قد يكون متردداً حول هذه القضية، فيتسائل عما إذا كانت هذه التسميةُ تعبيراً عن الصيغ الفعليةِ أم المراتبِ الزمنيةِ؟

نعم إنّ الفعلَ بوقوعه حقيقة، لا ينفك عن الزمان على الإطلاق، وهذا لا يدخل في نطاق دراستنا. إذ أنّ الفعلَ حالي عند وقوعه البتة، ويصبح ماضياً بعد وقوعه حقيقة. وهو مستقبل ما لم يقع. فهذه لا تتعدى عن تصورات منطقية بحتة.

وإنما نحن بصدد الصيغةِ الزمنيةِ الموضوعةِ للفعل. فالصيغة الزمنـيةُ لها دلالات مقرونة بأحد الأزمنة المطلقة؛ وهي الماضي والحال والإستقبال؛ كبناء الماضي على الفتح، واستهلال المضارع بأحد حروف المضارعة، ودخول السين وسوف على المستقبل. فإذا كانت هذه القرائن تدل على المراتب الزمنية للفعل، فإنّ علاقة الفعل بالزمان أكثر شمولا من هذا التقسيم. إذ لا يخفى أن للصِيغةِ الواحدةِ من الفعل دلالات متعلقةٌ بأزمنةٍ مختلفةٍ على حسب ما يصطحب الفعل من كلمات أو تركيب. مثلاً فإن (إنتهى)، وهو فعلٌ ماضٍ، لكن قد يفيد في كل مثالٍ من الأمثلة الآتيةِ الخمسةِ معنى لمراتب الزمان الغابر، يختلف ذلك المعنى في كل واحدٍ منها عن الدلالاتِ الزمنيةِ في بقيةِ الأمثلةِ بسببِ الأدواتِ التي لُحِقَت بـا وهي:

1ـ انتهى

2ـ قد انتهى

3ـ إذ انتهى

4ـ كان قد انتهى

5ـ لولاه لما انتهى

إنّ المرتبةَ الزمنيةَ التي يدل عليها فعلُ (انتهى) في كل مثالٍ من الأمثلةِ المذكورةِ، تختلفُ عن بقيهِ المراتبِ الزمنيةِ في الأمثلة الأخرى ، مع أن فعل (انتهى) قد ورد في كل من هذهِ الأمثلةِ على السواء. وأغربُ من هذا فإن فعلَ (انتهى) ـ على سبيل المثال ـ في جملة : "إذا انتهى الأجل، إنتهى الوجل" يدلّ على مستقبلٍ مُعلقٍ مع أنه فعلٌ ماضٍ.

Partie du texte « 2 » analysée

الحياء

صاحبه حي الضمير , مرهف المشاعر , محب للتواضع , عاشقٌ للهدوء والسكينة 0 الحياء مزية إسلامية لكل مؤمن خالص إيمانه لله , وشعور يخالج النفس فيترجم على ملامح الوجه .

الحياء	[N,E,MS000000X]
(حياء,حيي ; الْ , , ,)	[PONC,]
[N,A,MS000000H]	المشاعر
صاحبه	(مشَاعِر,شعر ; الْ , , ,)
(صاحِب,صحب ; , , ,ه)	[N,B,MT000000H]
[N,F,MS000000H]	[PONC,]
(صاحب,صحب ; , , ,ه)	[PONC, ,]
[V,0000A00AH]	[PONC,]
(صاحِب,صحب ; , , ,ِه)	محب
[N,F,MS000000H]	(محِب ,حب , , , ;)
[PONC,]	[N,F,MS000000H]
حي	[PONC,]
(حي ,حيي , , , ;)	للتواضع
[N,B,MS000000X]	(تَواضع,وضع ; لْ , , ,)
[PONC,]	[N,B,MS000000H]
الضمير	(تَواضع,وضع ; لِلْ , , ,)
(ضمِير,ضمر ; الْ, , ,)	[N,B,MS000000H]
[N,B,MS000000H]	[PONC,]
[PONC,]	[PONC, ,]
[PONC, ,]	[PONC,]
[PONC,]	عاشقٌ
مرهف	(عاشِقْ,عشق , , , ;)
(مرهف,رهف , , , ;)	[N,F,MS000000H]

[PONC,]
للهدوء
(هدوء,هدأ ; لِلْ , , ,)
[N,A,MS000000X]
(هدوء,هدأ ; لِلْ , , ,)
[N,A,MS000000X]

[PONC,]
والسكينة
(سكِينَةْ,سكن ; والْ , , ,)
[N,A,FS000000X]

[PONC,]
[CHIF, 0]
[PONC,]

[PONC,]

الحياء
(حياء,حيي ; الْ , , ,)
[N,A,MS000000H]

[PONC,]
مزية
(مزي ةْ,مزي , , , ;)
[N,B,FS000000H]

[PONC,]
إسلامية
(إسْلام,سلم ; , , ي ةُ ,)
[N,A,FS0X0000H]

[PONC,]
لكل
(كلّ,كل ; لَ , , ,)
[N,B,MS000000X]
(كُل ; لَ ,)
[P,X]
(كلّ,كل ; ل , , ,)
[N,B,MS000000X]
(كُل ; ل ,)
[P,X]

[PONC,]
مؤمن
(مؤْمِن,أمن , , , ;)
[N,B,MS000000H]
(مؤَم ن,أمن , , , ;)
[N,B,MS000000X]
(مؤَم ن,أمن , , , ;)
[N,E,MS000000X]

[PONC,]
خالص
(خَالِص,خاص , , , ;)
[N,F,MS000000X]
(خالص,خاص , , , ;)
[V,0000A000]

[PONC,]
إيمانه
(إيمان,أمن ; , , , ه)
[N,A,MS000000H]
(إيمان,أمن ; , , , هِ)
[N,A,MS000000H]

[PONC,]

لله

(لِل ه,أله , , , ;)
[N,A,MS000000H]
(لَه ; لَ ,)
[P,N]
(لُه ; ل ,)
[P,N]

[PONC,]
[PONC, ,]
[PONC,]

وشعور

(شُعور,شعر ; و(, , ,
[N,A,MS000000H]

[PONC,]

يخالج

(خالج,خلج ; ,ي , ,)
[V,ZXC0IXA0]
(خالج,خلج ; ,ي , ,)
[V,ZXC0IXP0]

[PONC,]

النفس

(نَفْس,نفس ; الْ , , ,)
[N,B,FS000000X]
(نَفَس,نفس ; الْ , , ,)
[N,B,MS000000X]

[PONC,]

فيترجم

(ترجم,ترجم ; فَ,ي , ,)
[V,ZXC0IXA0]
(ترجم,ترجم ; فَ,ي , ,)
[V,ZXC0IXP0]
(ترجم,ترجم ; فَ,ي , ,)
[V,ZXC0IXA0]
(ترجم,ترجم ; فَ,ي , ,)
[V,ZXC0IXP0]

[PONC,]

على

(علٰى)
[P,N]
(علٰى,على , , , ;)
[N,B,MS000000X]
(عِلى,على , , , ;)
[N,C,MS000000H]
(علّى,علو , , , ;)
[V,0000A00AX]

[PONC,]

ملامح

(ملْامِح,لمح , , , ;)
[N,B,MT000000X]

[PONC,]

الوجه

(وجه,وجه ; الْ , , ,)
[N,B,MS000000X]

[PONC.]

Partie du texte « 3 » analysée

أما بعد: فإن الأزمنةَ ذاتُ أهميةٍ بالغةٍ من حيث علاقة الفعل بها، وإن لم تكن من المسائل الأساسية للّغةِ . فإن علماء العربيةِ وإن كانوا قد خاضوا مثار البحثِ والجدل في دقائق علومِ اللغةِ،

أما
(أما)
[P,X]
(أم,أمم ; , ,ا ,)
[V,MDC0AZX0]
(ٌأم,أم ; , ,ا ,)
[N,B,MD0N0000H]
(ٌأم,أم ; , ,ا ,)
[N,B,FS0A0000H]
[PONC,]
بعد
(بعد)
[P,X]
(بعد,بعد , , , ;)
[N,A,MS000000N]
(بِعد,بعد , , , ;)
[V,0000A00AX]
(بعد,بعد , , , ;)
[V,0000A00AX]
(عد,عدد ; ٍ , , , ,)
[N,A,MS000000X]

[PONC, :]
[PONC,]
فإن
(إن ; فَ ,)
[P,N]
(إن ; فَ ,)
[P,V]
[PONC,]
الأزمنةَ
(أزمِنَة,زمن ; الْ , , , ,)
[N,B,MT000000X]
[PONC,]
ذاتُ
(ذَات,ذات , , , ;)
[N,B,MS000000X]
[PONC,]
أهميةٍ
(أهمِي ة,همم , , , ;)
[N,B,FS000000X]
(هِم,هم ; أ, ,ي ةُ ,)
[N,B,FS0X0000H]
[PONC,]
بالغةٍ
(بالِغَة,بلغ , , , ;)
[N,F,FS000000X]
[PONC,]
من
(من)
[P,V]
(من)

[P,V]
(من)
[P,X]
(مِن)
[P,N]
(من,من , , , ;)
[N,B,MS000000N]
(من,منن , , , ;)
[V,0000A00AX]
[PONC,]
حيث
(حيثُ)
[P,X]
[PONC,]
علاقة
(علَاقَة,علاق , , , ;)
[N,B,FS000000X]
[PONC,]
الفعل
(فِعل,فعل ; الْ , , ,)
[N,B,MS000000X]
(فِعل,فعل ; الْ , , ,)
[N,A,MS000000H]
(فُع ل,فعل ; الْ , , ,)
[N,I,MT000000X]
[PONC,]
بها
(بِها)
[P,N]

],[PONC,
[PONC,]
وإن
(إن ; و ,)
[P,N]
(إن ; و ,)
[P,V]
[PONC,]
لم
(لِم)
[P,X]
(لَم)
[P,V]
(لَم,لم , , , ;)
[N,A,MC000000X]
(لَم,لم , , , ;)
[V,0000A00AX]
[PONC,]
تكن
(كُن,كون ; ,تَ , ,)
[V,XXX0IXAAX]
(كن,كون ; ,تَ , ,)
[V,XXX0IXAAX]
(كن,كنن ; ,تَ , ,)
[V,XXX0IXAAH]
(كُن,كون ; ,تُ , ,)
[V,ZXX0IXPAX]
(كن,كون ; ,تُ , ,)
[V,ZXX0IXPAX]
(كن,كنن ; ,تُ , ,)

[V,ZXX0IXPAH]
[PONC,]
من
(من)
[P,V]
(من)
[P,V]
(من)
[P,X]
(مِن)
[P,N]
(من,من , , , ;)
[N,B,MS000000N]
(من,منن , , , ;)
[V,0000A00AX]
[PONC,]
المسائل
)مسائِل,سئل ; الْ , , , (
[N,B,FT000000X]
[PONC,]
الأساسية
(أساسِية,أسس ; الْ , , ,)
[N,B,FS000000X]
[PONC,]
للّغةِ
(لُغَة,لغو ; لَ , , ,)
[N,A,FS000000H]
(لُغَة,لغو ; ل , , ,)
[N,A,FS000000H]
[PONC,]
[PONC, .]

[PONC,]
فإن
(إن ; فَ ,)
[P,N]
(إن ; فَ ,)
[P,V]
[PONC,]
علماء
(علْماء,علم , , , ;)
[N,B,MP000000H]
[PONC,]
العربيةِ
(عربِي ة,عرب ; الْ , , ,)
[N,B,FS000000H]
[PONC,]
وإن
(إن ; و ,)
[P,N]
(إن ; و ,)
[P,V]
[PONC,]
كانوا
(كان,كون ; , ,وا ,)
[V,MPC0AZXAX]

[PONC,]
قد
(قَد)
[P,X]
(قَد,قد , , , ;)

[N,B,MS000000X]
(قُد,قد , , , ;)
[N,B,MS000000N]
(قِد,قد , , , ;)
[N,B,MS000000N]
(قد,قدد , , , ;)
[V,0000A00AX]

[PONC,]
خاضوا
(,خاض ; , ,وا ,)
[V,MPC0AZX0]

[PONC,]
مثار
(مثّار,ثور , , , ;)
[N,B,MS000000X]

[PONC,]
البحثِ
(بحث,بحث ; الْ , , ,)
[N,A,MS000000X]
[PONC,]
والجدل
(جدل,جدل ; والْ , , ,)
[N,B,MS000000X]
(جدِل,جدل ; والْ , , ,)
[N,E,MS000000H]
(جد ل,جدل ; والْ , , ,)
[N,E,MT000000H]

[PONC,]
في
(فِي)
[P,N]

[PONC,]
دقائق
(,دقق , , , ;)
[N,B,FT000000X]

[PONC,]
علومِ
(علُوم,علم , , , ;)
[N,B,MT000000H]

[PONC,]
اللغةِ
(لُغَة,لغو ; الْ , , ,)
[N,A,FS000000H]

],[PONC,
[PONC,]

Bibliographie

[ABDE-2008] Med. El A. Abderrahim; Un analyseur morphologique pour l'arabe voyellé ou non ; SIIE'2008 : 1ère Conférence Internationale "Systèmes d'Information et Intelligence Economique" SIIE 2008 Hammamet – Tunisie, 14-16 Février 2008, Proceedings tome II, IHE éditions, pp 324-339.

[ABDE-2007] Med. El A. Abderrahim & al. Un modèle objet pour le traitement automatique de l'arabe voyellé ou non. JeTIC'2007. Bechar 21/22 avril, 2007.

[ABEI-1993] Anne ABEILLE ; Les nouvelles syntaxes, Grammaire d'unification et analyse du français ; Armand Colin, 1993.

[AFNO-1993] Norme AFNOR NF ISO 233-2 Translittération des caractères arabes en caractères latins ; Z46-002 AFNOR ; Association Française de Normalisation Décembre 1993.

[AGNE-1997] E. AGNEL, M. Bertier ; Modélisation à objet et implantation sous O2 d'un spécimen du dictionnaire CRISTAL ; aide à la recherche n° 94.K.6427, université Stendhal, ICM, novembre 1997.

[ALJL-2002] Mohammed ALJLAYL & al ; On Arabic Search :Improving the Retrieval effectiveness via a Light Stemming Approach ; CIKM'02n November 4-9, 2002, M clean, Virginia, USA.

[AL-S-1998] R.Al-Shalabi & al; *A computational morphology system for Arabic*. http://acl.ldc.upenn.edu/W/W98/W98-1009.pdf

[ALOU-2003] Chafik ALOULOU ; Analyse syntaxique de l'arabe : Le système MASPAR ; RECITAL 2003, Batz-sur-Mer, 11-14 juin 2003.

[AMMA-1999] Sam AMMAR & al ; Les verbes arabes ; Collection Bescherelle, édition HATIER, 1999.

[ANTO-1991] Ambassador El-Dahdah ANTOINE; A dictionary of Arabic verb conjugation; El-Dahdah encyclopedia of Arabic grammar; Librairie du Liban, 1991.

[AUDI-2003] Laurent AUDIBERT ; Outils d'exploration de corpus et désambiguïsation lexicale automatique ; Thèse université d'Aix Marseille I, université de Provence, décembre 2003.

[ATTI-2000] M. A. Attia ; A large-scale computational processor of the Arabic morphology, and applications. These faculty of engineering Cairo University, Egypt, 2000.

[ATTI-2005] M. A. Attia ; Developing a Robust Arabic Morphological Transducer Using Finite State Technology. 8th Annual CLUK Research Colloquium, Manchester.

[ATTI-2006] M. A. Attia ;An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. The Challenge of Arabic for NLP/MT Conference, October 2006. The British Computer Society. London.

[ATTI-2007] M. A. Attia ;Arabic Tokenization System. ACL-Workshop on Computational Approaches to Semitic Languages. Prague.

[BALI-1993] Laurence BALICCO; Génération de répliques en français dans une interface homme-machine en langue naturelle; Thèse, université Pierre Mendès-France Grenoble,1993.

[BEN-1998] Chiraz BEN OTHMANE ZRIBI; De la synthèse lexicographique à la détection et la correction des graphies fautives arabes; Thèse, université de Paris-Sud, décembre 1998.

[BENH-1993] Boualem BENHAMOUDA ; Les clés de la langue arabe ; office des publications universitaires Alger, 1993.

[BERR-1990] Alain BERRENDONNER ; Grammaire pour un analyseur Aspects morphologiques ; université des sciences sociales de Grenoble ; les cahiers du CRISS N° 15, novembre 1990.

[BEES-1997] K. R. BEESLEY, Arabic morphological analysis on the internet.
http://www.cis.upenn.edu/~cis639/docs/mltt-97-03.ps

[BEES-2001] Kenneth R. BEESLEY, *Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001*. *Proceedings of the Arabic Language Processing:* Status and Prospect-39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France.
http://www.elsnet.org/arabic2001/beesley.pdf

[BLAC-2004] R. BLACHERE ; Grammaire de l'arabe classique ; Editions Maisonneuve-LAROSE, 2004.

[BLAH-2005] Michael BLAHA & al. ; « Modélisation et conception orientées objet avec UML 2 » 2ème édition PEARSON Education France, 2005.

[BOOC-2000] Grady BOOCH & al. ; Le guide de l'utilisateur UML, Edition Eyrolles 2000.

[BOUI-1998] Pierrette BOUILLON ; Traitement automatiques des langues naturelles ; Edition Duculot, 1998.

[BRUS-1989] Jean BRUSSET et al. ; Création d'une base de données lexicale de l'arabe écrit utilisable par un système morpho-syntaxique ; Linguistique arabe et informatique / Actes du IV colloque international de linguistique, Tunis, 9-12 Novembre 1989 ; Université de Tunis, Centre d'études et de recherches économiques et sociales, p.9-27.

[CANT-1960] Jean CANTINEAU ; Etudes de linguistique arabe ; mémorial Jean

CANTINEAU ; librairie C. KLINCKSIECK, Paris, 1960.

[CAVA-2000] Violetta CAVALLI-SFORZA & al, Arabic morphology generation using a concatenative strategy, NAACL-2000.

[CHAP-1983] Sylviane CHAPPUY ; Formalisation de la description des niveaux d'interprétation des langues naturelles. Etude menée en vue de l'analyse et de la génération au moyen de transducteurs ; Thèse, institut national polytechnique de Grenoble, juillet 1983.

[CHAR-2005] Benoît CHARROUX & al.; UML2, Collection Synthex, PEARSON Education France 2005.

[CLAV-1995] Viviane CLAVIER ; Modélisation de la suffixation en vue du traitement automatique du français. Application à la recherche d'informations ; Thèse, université Stendhal Grenoble III, octobre 1995.

[COHE-1970] David COHEN ; Essai d'une analyse automatique de l'arabe ; Etudes de linguistique sémitique et arabe, mouton 1970 Paris, p.49-78.

[DEBI-2001] Fathi DEBILI et al. ; Sur l'ambiguïté grammaticale de l'arabe et sa résolution automatique ; Linguistique arabe et sémitique, tome 2; centre d'étude des langues et littératures du monde arabe, ENS Editions, 2001, p.79-113.

[DEBR-2005] Laurent DEBRAUWER & al.; UML2 initiation, exemples et exercices corrigés; Editions ENI, Janvier 2005.

[DELS-2001] Philippe DELSARTE & al. ; « Logique pour le traitement de la langue naturelle, application à la langue française » Edition HERMES Science, 2001.

[DEND-2003] J. DENDIEN et al. ; Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence ; in Traitement automatique des langues vol. 44-n°2/2003 ; Germes, Lavoisier, 2003.

[DICH-1990] Joseph DICHY, « L'écriture dans la représentation de la langue :la lettre et le mot en arabe » Thèse université lumière LYON 2, 1990.

[DICH-2001] Joseph DICHY; On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases.
http://www.elsnet.org/arabic2001/dichy.pdf

[DOUZ-2004] Fouad Soufiane Douzidia ; Résumé automatique de texte arabe ; Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de M.Sc en informatique ; Université de Montréal, Faculté des arts et des sciences, Septembre, 2004

[FEUI-1988] Jack FEUILLET « Introduction à l'analyse morphologique » Presses universitaire de France, 1988.

[FOUA-1973] Niima FOUAD ; Grammaire de langue Arabe ; Le Caire, 1973.

[FOWL-2004] Martin FOWLER; UML 2.0; CampusPress, 2004.

[FREE-2001] Andrew FREEMAN ; Brill's POS tagger and a Morphology parser for Arabic. http://www.elsnet.org/arabic2001/freeman.pdf

[GAL-1986] A. GAL et al. Compréhension automatique du langage naturel le cas de l'interrogation simple en français, Masson, 1986.

[GRAI-2003] Jonathan GRAINGER et al. ; Approche expérimentale de la reconnaissance du mot écrit en arabe ; Fait de langue N° 122, 2003, p.77-86.

[GRAN-1975] Ernest GRANDJEAN ; Conception et réalisation d'un dictionnaire pour un analyseur interactif de langues naturelles ; Mémoire d'ingénieur C.N.A.M. Conservatoire national des arts et métiers centre régional associé de Grenoble, février 1975.

[HABA-2004] Nizar HABASH, Large scale lexeme based arabic morphological generation ; JEP-TALN 2004, Traitement automatique de l'arabe, Fès, 20 avril 2004.

[HASS-1987] Mohamed Omar HASSOUN ; Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application ; Thèse université Claude Bernard – Lyon I ; Juillet 1987.

[JACC-1997] André JACCARINI ; Grammaires modulaires de l'arabe modélisations, mise en œuvre informatique et stratégies ; Thèse de doctorat, université Paris IV, Février 1997.

[JACC-2003] André JACCARINI & al ; Un logiciel pour la mise au point de grammaire pour le filtrage d'information en arabe (cas de l'information citationnelle) TALN 2003, Batz-sur-Mer, 11-14 juin 2003.

[JAYE-1985] J-H. JAYEZ, Compréhension automatique du langage naturel le cas du groupe nominal en français, Masson, 1985

[KALL-1987] Ghassan KALLAS ; Résolution des solutions multiples en analyse morphologique automatique des langues naturelles utilisation des modèles de Markov ; Thèse, université des sciences sociales de Grenoble, juin 1987.

[KHOJ-2001] Shereen KHOJA & al ; A tagset for the morphosyntactic tagging of Arabic. http://archimedes.fas.harvard.edu/mdh/arabic/CL2001.pdf

[KOUL-1994] Djamel KOULOUGHLI ; Grammaire de l'arabe d'aujourd'hui ; Pocket – Langues pour tous, 1994.

[LAI-2004] Michel LAI ; Penser objet avec UML et JAVA ; Edition DUNOD 3ème édition, 2004.

[LALL-1990] Geneviève LALLICH-BOIDIN et al. ; Analyse du français achèvement et implantation de l'analyseur morpho-syntaxique ; université des sciences sociales de Grenoble ; les cahiers du CRISS N° 16, novembre 1990.

[LECO-2005] Sébastien LECOMTE; XML par la pratique bases indispensables, concepts et cas pratiques; Editions ENI, Septembre 2005.

[LEE-2003] Young-Suk LEE & al ; Language Model Based Arabic Word Segmentation; Proceedings of the 4 st annual meeting of the association for computationanl linguistics, july 2003, pp. 399-406.

[MANK-1989] Chafia MANKI et al. ; Un système de compréhension automatique de langue arabe ; Linguistique arabe et informatique / Actes du IV colloque international de linguistique, Tunis, 9-12 Novembre 1989 ; Université de Tunis, Centre d'études et de recherches économiques et sociales, p.161-195.

[MERL-1982] Alain MERLE ; Un analyseur pré-syntaxique pour le levée des ambiguïtés dans des documents écrits en langue naturelle : Application à l'indexation automatique ; Thèse, institut national polytechnique de Grenoble, septembre 1982.

[MOKH-1995] Hassan MOKHLIS ;Théorie du tasrif et traitement du lexique chez les grammairiens arabes anciens ; Thèse université des sciences humaines de Strasbourg 1995.

[MONE-1989] Walid MONEIMNE ; TAO vers l'arabe ; Spécification d'une génération standard de l'arabe ; réalisation d'un prototype anglais-arabe à partir d'un analyseur existant ; Thèse, université Joseph Fourier Grenoble I, juin 1989.

[MULL-1997] Pierre-Alain MULLER ; Modélisation objet avec UML ; Edition Eyrolles 1998.

[NEYR-1996] Michel NEYRENEUF & al ; Grammaire active de l'arabe littéral ; Librairie Générale Française, Méthode 90, 1996.

[NOGI-1991] Jean-François NOGIER ; Génération automatique de langage et graphes conceptuels ; HERMES, 1991.

[OUER-2002] Riadh OUERSIGHNI ; La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : Utilisation pour la détection et le diagnostique des fautes d'accord. Thèse université Lumière-Lyon2, 2002.

[PALM-1990] Patrick PALMER ; Etude d'un analyseur de surface de la langue naturelle. Application à l'indexation automatique de textes ; Thèse, université Joseph Fourier, Grenoble I, 1990.

[PITR-1985] Jacques PITRAT, Textes ordinateurs et compréhension, Eyrolles, 1985.

[POIB-2003] Thierry POIBEAU ; Extraction automatique d'information du texte brut au web sémantique ; publication Germes LAVOISIER, 2003.

[PONT-1995] Claude PONTON ; Un système de génération morphologique linguistiquement justifié permettant une large couverture du français écrit ; Equipe CRISTAL-GRESEC Université Stendhal ; TALN'95

[QUIN-1997] Julien QUINT ; Morphologie à deux niveaux des noms du français ; reconnaissance et indexation de termes par réseaux à états finis ; DEA Informatique Systèmes et Communication Laboratoire d'accueil : Rank Xerox Research Centre, équipe MLTT 19 Juin 1997.

[RAFE-1993] Ahmed A. RAFEA; Lexical Analysis of Inflected Arabic Words using Exhaustive Search of an Augmented Transition Network; Software-Practice and Experience, vol. 23(6), 567-588 june 1993.

[RASS-1994] Gilles S.RASSET ; SUBLI : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions ; Thèse JOSEPH FOURIER, GRENOBLE 1, 1994

[REIG-1983] Daniel REIG ; La conjugaison arabe ; éditions MAISONNEUVE & LAROSE, 1983.

[ROLE-2005] François ROLE ; Modélisation et manipulation de documents XML ; Germes, Lavoisier, 2005.

[ROQU-2004] Pascal ROQUES & al. ; « UML 2 en action de l'analyse des besoins à la conception J2EE » EYROLLES, 2004.

[SABA-1988] Gérard SABAH ; L'intelligence artificielle et le langage ; volume 1, Représentation des connaissances ; HERMES, 1988.

[SABA-1989] Gérard SABAH ; L'intelligence artificielle et le langage ; volume 2, Processus de compréhension ; HERMES, 1989.

[SARO-1989] Abdelghani SAROH ; Base de données lexicales dans un système d'analyse morpho-syntaxique de l'arabe –SYAMSA- ; Thèse, université Pail Sabatier de Toulouse, octobre 1989.

[SILB-1993] Max SILBERZTEIN ; Dictionnaires électroniques et analyse automatique de textes, le système INTEX ; collection informatique linguistique, MASSON, 1993.

[SOUI-1989] Dalila SOUILEM et al. ; Un système d'enseignement assisté par ordinateur de la grammaire arabe « S.E.A.G.A. » ; Linguistique arabe et informatique / Actes du IV colloque international de linguistique, Tunis, 9-12 Novembre 1989 ; Université de Tunis, Centre d'études et de recherches économiques et sociales, p.209-228.

[STEF-1993] Marie-Hélène STEFANINI ; TALISMAN :une architecture multi-agents pour l'analyse du français écrit ; Thèse informatique ; université Pierre Mendès-France, Grenoble, 1993.

[TAHI-2004] Youssef TAHIR *& al.* Modélisation à objets d'une base de données morphologiques pour la langue arabe ; JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, 20 avril 2004

[TITT-2003] Ed TITTEL ; XML ; EdiScience, SCHAUM'S, 2003.

[TUER-2004] Laurence TUERLINEKX, La lemmatisation de l'arabe non classique ;

JADT 2004 :7ème journées internationales d'analyse statiques des données textuelles.

[VERO-2001] Jean VERONIS, Informatique et Linguistique 1 ; unité d'enseignement INF Z18, université de Provence, centre informatique pour les lettres et sciences humaines, 2001.

[YAVA-1988] Gholam Réza YAVARI-SARTAKHTI; Générateur – Conjugueur de la morphologie verbale arabe ; Thèse, université de la Sorbonne nouvelle Paris III, 1988.

[ZAAF-2002] Riadh ZAAFRANI ; Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère, thèse, université LYON II, 18 janvier 2002.

[ZAAF-2004] Riadh ZAAFRANI ; Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004, Traitement automatique de l'arabe, Fès, 20 avril 2004.

[ZEMI-2004] Z. ZEMIRLI & al ; TAGGAR : Un analyseur morphosyntaxique destiné à la synthèse vocale de texte arabes voyellés ; JEP-TALN 2004, Traitement automatique de l'arabe, Fès, 20 avril 2004.

[ZOUA-1989] Lotfi ZOUARI ; Construction automatique d'un dictionnaire orienté vers l'analyse morpho-syntaxique de l'arabe, écrit voyellé ou non voyellé ; Thèse, université de Paris XI Orsay, avril 1989.

المراجع باللّغـــــــة العـــــربيـــــــــــــــة

[دار المشرق -2003] المنجد في اللغة والأعلام، دار المشرق – بيـــــــروت لبنان، الطبعة الأربعون 2003.

[الدحداح-1991] السفير أنطوان الدحداح، معجم تصريف الأفعال العربية، موسوعة الدحداح في علم العربية، مكتبة لبنان، 1991.

[الدحداح-1991] السفير أنطوان الدحداح، معجم قواعد العربية العالمية عربي-فرنسي، موسوعة الدحداح في علم العربية، مكتبة لبنان، 1991.

[فؤاد-2003] فؤاد نعمة، ملخص قواعد اللغة العربية، المكتب العلمى للتأليف والترجمة، مصر، الطبعة الثانية والعشرون.