Contribution au traitement automatique des langues : cas de l'arabe

Mounir Zrigui

THESE D'HABILITATION À DIRIGER DES RECHERCHES

Université Stendhal-Grenoble 3

Spécialité : Informatique Linguistique (Industries de la langue)

Présentée par

MOUNIR ZRIGUI

Pour l'obtention du diplôme
D'habilitation à diriger des recherches
Sur le thème :

Contribution au traitement automatique des Langues : cas de l'Arabe

Soutenue le 28 Avril 2008 devant le jury composé de :

Mme Rita MAZEN : Professeur, Université Stendhal-Grenoble 3, France
M. Mathieu GUIDERE : Professeur, Université de Genève, Suisse
M. Ahmed JERRAYA : Directeur de Recherche CNRS, CEA-LETI, Minatec, Grenoble, France
M. Mohamed JEMNI : Professeur, Université de Tunis, Tunisie
M. Christian ABRY : Professeur, Université Stendhal-Grenoble 3, France
Mme Gabrièle SAUCIUER : Professeur Emérite, INPG, Grenoble, France

Laboratoire LIDILEM(Grenoble)

REMERCIEMENTS

Prologue

Ce document résume mes activités de recherche et d'enseignement auxquelles je me suis consacré depuis plus de 20 années, d'abord au centre TOBIA à l'Université Paul Sabatier de Toulouse ou j'ai préparé ma thèse de doctorat en informatique, puis en tant qu' assistant contractuel à l'Université de Bretagne Occidentale en France, puis à l'Institut Régional des Sciences Informatiques et Télécommunications à Tunis(IRSIT) en tant que maître de Recherche, chef du projet synthèse de la parole , et enfin en tant qu'enseignant chercheur à la Faculté des Sciences de Monastir (FSM) en Tunisie.

Mes travaux de recherche ont commencé avec mes travaux de DEA et thèse au centre TOBIA à l'Université Paul Sabatier de Toulouse, laboratoire LSI, sur la conception et la transcription d'un nouveau système Braille abrégé arabe basé sur la phonétique. Ma recherche à l'IRSIT constituait la suite de ces travaux de thèse puisque j'étais chef du projet de synthèse de la parole arabe. Depuis mon affectation comme enseignant chercheur à la FSM, en septembre 1993, je m'intéresse à plusieurs aspects du traitement automatique des langages naturels (TALN) dont notamment les aspects phonétique, morphologique, syntaxique et sémantique.

En parallèle, j'ai également exercé une activité d'enseignement. Entre 1987 et 1989, j'ai été affecté à l'Université de Bretagne occidentale en tant qu'assistant contractuel en Informatique; Depuis 1990, j'ai enseigné l'informatique et l'informatique linguistique en Tunisie dans différents écoles (ENIG, ENSI), institut (ISG de Tunis) et facultés (FST, FSM) et en Master Recherche à l'ISG de Tunis (DEA Management information system), et à l'ENSI (cours de traitement automatique des langages naturels)

Ce mémoire présente dans sa première partie, une synthèse de mes travaux de recherche, menés en collaboration avec plusieurs doctorants que j'ai co-encadrés (Tahar Saidane, Anis Zouaghi, Kheirallah Khouja et Ahmed Haddad en Tunisie et Mohsen Maraoui, Mourad MARS et Mohamed BELGACEM au LIDILEM, et avant eux, lassaad Ghanmi, Arafat Ghrab, Mourad Hamdoun et Brahim Missaoui au CSI (INPG)à Grenoble en France) avec 21 Mastères de recherche à l'ISG, l'ENSI et l'ENIS en Tunisie, et en France.

La deuxième partie du mémoire décrit mes activités et mes responsabilités administratives et collectives au sein de l'IRSIT, la FSM et l'Université du centre.

La troisième partie est un recueil des principales publications scientifiques de ces dernières années et elle est présentée dans un document annexe.

Résumé

Ce document retrace mes activités de recherche depuis ma thèse soutenue en juin 1987 à l'Université Paul Sabatier à Toulouse, laboratoire LSI.
Certains des travaux présentés sont achevés, d'autres sont en cours ou encore dans un stade exploratoire.

De 1984 à 1992, je me suis intéressé à la phonétique puis à la synthèse de la parole. Depuis 1992, mes travaux ont été étendus à l'ensemble du domaine de traitement automatique des langages naturels et plus particulièrement les aspects phonétique, morphologique, syntaxique, sémantique et pragmatique. A partir de l'année 2005, tous mes travaux de recherche et d'encadrement sont dans le cadre d'un projet intitulé "**Oréllodule**": un système temps réel de synthèse, traduction et reconnaissance de l'arabe dans un contexte finalisé. Mes recherches sont donc focalisées sur les trois fonctions principales qui composent le projet: Reconnaissance, Traduction et Synthèse de la parole. Pour ces trois fonctions ou axes, des méthodes et des outils ont été définis et développés.

Des travaux toujours en cours portent sur l'amélioration, le raffinement des outils et méthodes développés.

Mes travaux futurs s'orientent vers deux axes: le choix de l'architecture finale et la définition des modèles et techniques pour le système final qui doit être un système embarqué.
Ce travail s'appuie sur une expérience acquise depuis plusieurs années sur la normalisation, et la manipulation de ressources linguistiques comme les corpus, les dictionnaires électroniques et les techniques du Taln au sein des équipes de recherche tel que RIADI-Monastir, UTIC-Tunis, LIDILEM-Grenoble, TIMA-Grenoble, DESIGN&REUSE-Grenoble, CLIPS-Grenoble et autres laboratoires et industriels arabes et français.

SOMMAIRE

TABLE DES FIGURES

I. INDUSTRIE DES LANGUES

I.1. Contexte et motivations

L'information joue aujourd'hui un rôle de plus en plus important dans toutes les activités du monde contemporain. En effet, de nombreux facteurs, agissant de manière synergique, ont contribué au cours des trente dernières années, à accroître son influence dans tous les domaines. Simultanément, l'internationalisation des marchés et le développement des Nouvelles Technologies de l'Information et de la Communication (NTIC) ont favorisé le multilinguisme et l'augmentation du volume d'information.

Dans ce contexte, les aspects d'acquisition, de gestion, d'analyse, d'exploitation, etc. des informations, multilingues ou non, pour la plupart sous formes textuelles ou orales, sont au centre des grands débats du monde de la recherche et de l'économie. La société de l'information donne par ailleurs au support de l'information, la langue elle-même, une importance nouvelle sur le plan de son traitement informatique.

Le domaine des technologies de la langue est ainsi devenu un des domaines-clés qui propose des voies de recherche susceptibles d'apporter des solutions à ces problèmes et répond ainsi aux besoins actuels de ce que les experts dénomment la « société de l'information ». C'est dans ce cadre que nous situons notre travail qui consiste à réaliser un système temps réel de synthèse, traduction et reconnaissance de la parole arabe intitulé OREILLODULE:

Informellement parlant, l'Oréllodule est une prothèse qu'on met à l'oreille et qui permet de traduire la parole dans le langage de celui qui la porte.

Formellement parlant c'est un système temps réel de reconnaissance, de traduction et de synthèse de la parole. Ce système devrait reconnaître, traduire et synthétiser une voix la plus naturelle possible avec un minimum d'erreurs d'une langue vers une autre.

Cette prothèse aurait l'aspect d'une oreillette qui se portera sans dérangement et qui aura pour tâche de tenir une conversation naturelle entre deux ou plusieurs individus communicants en langues différentes, tout en respectant des critères économiques et énergétiques évidents.

Un tel système est composé d'une couche logicielle utilisant plusieurs techniques gourmandes en mémoire et en puissance de calcul. On citera en exemple les fonctions d'accès aux bases de données, les traitements lexicaux, syntaxiques, sémantiques, pragmatiques et phonétiques**.**

Le schéma suivant illustre l'architecture générale de l'Oréllodule :

.

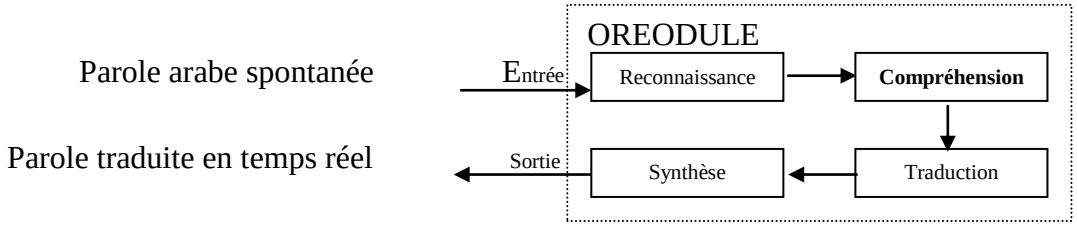


Figure 1: Architecture générale du système ORELLODULE

Comme le montre la figure 1, le système ORELLODULE est composé de quatre principaux éléments:

- Reconnaissance de la parole: ce module a pour but de transcrire le message vocal énoncé par le guide touristique en un message orthographique.

- Compréhension de la parole: ce module permet de comprendre et de désambiguïser le sens du message.
- Traduction de la parole: ce module a pour objectif de traduire le message énoncé en langue arabe à une autre langue étrangère: allemand, anglais, français, etc.
- Synthèse de la parole: ce module a pour objectif de prononcer le message donné en entrée, dans une autre langue autre que l'arabe.

I.2. Les technologies de la langue

L'expression « technologies de la langue » désigne l'ensemble des produits et services qui ont pour objet le Traitement Automatique des Langues Naturelles (TALN).

Le TALN s'effectue sur des données linguistiques (textes écrits, corpus oraux, lexiques, phrases, etc.) et met en œuvre des outils et des techniques de traitement qui sont de trois ordres : linguistiques (description et explicitation des connaissances de la langue relevant des niveaux de la phonétique/phonologique, de la morphologie, de la syntaxe, de la sémantique), formels (expression et représentation de ces connaissances dans un formalisme susceptible d'être implémenté par la machine) et informatiques (élaboration de techniques et de stratégies informatiques de traitement effectif).

I.3. Les enjeux des technologies da la langue

Les enjeux des technologies de la langue sont étroitement liés à l'avènement de la société de l'information pour laquelle les « autoroutes de l'information » jouent un rôle majeur. Ces enjeux sont aussi bien culturels, économiques et géopolitiques.

-Du point de vue culturel : les langues qui ne seront pas informatisées, c'est-à-dire pour lesquelles des outils performants de traitement automatique ne seront pas disponibles, risquent à terme d'être exclues des médias modernes de production et de diffusion de l'information professionnelle et non professionnelle.

-Du point de vue économique : le marché des technologies de la langue connaît une forte croissance due en particulier à des secteurs comme la gestion de l'information et les technologies vocales. Ils peuvent par ailleurs conduire à des gains de productivité dans de nombreux secteurs.

I.4. Etat de l'art

Cette partie traite des méthodes et techniques mises en œuvre dans le domaine des technologies de la langue. Il s'agit de recenser les techniques et méthodes de modélisation et de traitement de l'écrit et de l'oral tel que la reconnaissance et la synthèse de la parole, la traduction automatique, de faire le point sur les méthodes d'évaluation des technologies, de voir quelles sont les méthodes les plus appropriées.

On peut résumer les problèmes posés par le traitement des langages naturels en trois catégories :

- L'absence de définition explicite : il n'y a aucune façon en langage naturel qui permet de représenter le langage de façon complète et exacte (car celui ci est implicite)
- L'influence du contexte : la compréhension d'un mot dans une phrase est rarement abordée sans la prise en compte du contexte
- L'ambiguïté : A un mot ou une phrase peuvent être associées plusieurs interprétations possibles.

Ces problèmes peuvent être considérés comme étant des caractéristiques du langage naturel, et pour bien les aborder il faut trouver le formalisme adéquat qui permet de les représenter correctement pour pouvoir les résoudre: donc le problème devient celui de la représentation de ces connaissances linguistiques en vue de les traiter automatiquement: ce terme apparaît de plus en plus souvent dans le vocabulaire informatique. En effet, connaître est une opération active, qui implique certes des capacités de mémorisation, mais surtout une faculté d'inférence.

La vraie connaissance suppose l'utilisation à bon escient des informations dont on dispose, et le problème consiste justement à trouver des structures informatiques permettant le stockage et l'utilisation, par la machine de ces informations. Partant du principe "représenter veut dire résoudre", nous pouvons imaginer la problématique de la représentation des connaissances.

Toute fois le formalisme de représentation est lié à la nature de la connaissance manipulée, sans oublier que celle-ci dépend entièrement du point de vue du concepteur du système d'informations.

Plusieurs méthodes de formalisation et de représentation des connaissances sont utilisées pour modéliser les connaissances linguistiques, dont notamment les mathématiques et particulièrement les modèles de Markov et les logiques classiques et non classiques, les modèles psychologiques et plus particulièrement la sémantique psychologique, les prototypes et la dépendance conceptuelle et ses dérivées, et enfin les modèles informatiques comme les bases de données, les réseaux sémantiques, les langages orientés objet, les règles de production.

Pour le traitement de la parole, nous avons optés pour le Modèle de Markov Caché (Hidden Markov Model) qui est une méthode statistique puissante pour caractériser les échantillons de données observés d'un processus à temps discret. Elle apporte non seulement un moyen efficace de construction de modèles paramétriques, mais elle incorpore aussi le principe de programmation dynamique pour unifier la segmentation et la classification de séquence de données variant dans le temps.

Dans la modélisation d'un processus par un HMM, les échantillons peuvent être caractérisés par un processus paramétrique aléatoire dont les paramètres peuvent être estimés dans un cadre de travail bien défini. La théorie de base des HMM a été publiée dans une série de papiers par L. Baum.

Les HMMs sont devenus la méthode la plus couramment utilisée pour la modélisation des signaux de parole dans les applications suivantes : reconnaissance automatique de la parole, suivi de la fréquence fondamentale et des formants, synthèse vocale, traduction automatique, étiquetage syntaxique, compréhension du langage oral, traduction automatique... Dans une chaîne de Markov, chaque état correspond à un événement à observation déterministe (la sortie de ses sources pour un état donné n'est pas aléatoire). Une extension naturelle à la

chaîne de Markov introduit un processus non déterministe qui génère des symboles de sortie pour chaque état. L'observation est donc une fonction probabiliste de l'état. Le nouveau modèle est appelé HMM, pouvant être vu comme deux processus stochastiques imbriqués dont l'un (la séquence d'états) est non observable directement. Ce processus sous-jacent est donc associé de façon probabiliste à un autre processus produisant la séquence de trames, qui elle, est observable. Ci-dessous, nous présentons les trois problèmes de base à résoudre pour l'application de cette méthode :

- Le problème d'évaluation : Quelle est la probabilité d'un modèle générant une séquence d'observation ? Ce problème est résolu par l'application de l'algorithme FORWARD.
- Le problème de décodage : Quelle est la séquence d'états la plus probable pour un modèle et une séquence d'observation donnés ? On utilise l'algorithme VITERBI pour effectuer cette tâche.
- Le problème d'apprentissage : Comment peut-on ajuster les paramètres du modèle pour maximiser la vraisemblance (probabilité jointe) de génération d'une séquence d'observation ? Les algorithmes de BAUM-WELCH et de VITERBI permettent d'effectuer l'apprentissage. Dans les applications de la parole, on utilise fréquemment les HMM continus, où l'observation n'appartient pas à un ensemble discret mais à une distribution (le plus souvent normale). Ainsi, une topologie gauche droite pour un HMM continu permet de modéliser les états successifs d'un phonème pour un signal de parole. Plus généralement, l'objectif à atteindre est la détermination à partir de vecteurs acoustiques de la séquence phonétique prononcée. Nos travaux ont permis de tester cette méthode et de conclure qu'elle est la plus appropriée au traitement automatique de l'arabe.

Certains travaux ont utilisés les réseaux de neurones qui constituent depuis une vingtaine d'année une technique utilisée dans les systèmes de reconnaissance automatique de la Parole. Ils sont basés sur une modélisation grossière du neurone biologique (neurone formel). Tout comme le neurone biologique, le neurone formel calcule sont activation en fonction des signaux qu'il reçoit d'autres neurones, pondérés par des « poids synaptiques » et d'une fonction d'activation plus ou moins complexe.

L'ensemble de ces neurones est organisé selon des architectures plus ou moins complexes matérialisés par les connexions entre ces neurones. Selon cette architecture, ainsi que le type de la fonction d'activation, les réseaux de neurones peuvent résoudre un certain nombre de problèmes tels que des problèmes de classification, de mémorisation et de résolution de contraintes. Une particularité des réseaux de neurones est qu'ils sont dotés d'algorithmes d'apprentissages qui leur permettent d'apprendre les formes, les classes à reconnaître et à classer les problèmes à résoudre. Ces algorithmes sont soit supervisés lorsque l'on connaît déjà les classes associées aux exemples du corpus d'apprentissage, soient non supervisés. Le but recherché est de faire en sorte que les réseaux de neurones répondent correctement à des stimuli jamais rencontrés. Etant donné le large spectre des possibilités des réseaux de neurones, ils peuvent être employés à de nombreux niveaux dans un système de traitement automatique de la parole. De nombreuses études ont été menées pour les utiliser pour le traitement de signal (filtrage, annulation d'échos, séparation de sources), la modélisation acoustique mais aussi pour des tâches de plus hauts niveaux telles que la modélisation linguistique.

I.5. Traitement automatique de l'arabe

La langue arabe, par ses propriétés phonologique, morphologique, syntaxique et sémantique est considérée comme une langue difficile à traiter automatiquement [Alj02] et [Lar02].

Depuis plusieurs décennies déjà, des recherches se sont poursuivies dans le cadre du traitement automatique de la langue arabe. L'un des premiers théoriciens de ce domaine,

David Cohen propose un essai d'analyse automatique dès 1961 [Coh61, 70]. Partant d'une analyse morphologique minimaliste, basée sur le principe que toute forme linguistique arabe se traduit en schème et racine, les recherches vont se développer depuis les premiers travaux sur le lexique et la morphologie jusqu'à la mise au point d'analyseurs automatiques, de systèmes d'indexation, de correcteurs, etc. De nombreux projets sont en cours et il existe des bases de données disponibles proposant des corpus divers sous forme électronique, traités automatiquement. Un traitement automatique de l'arabe est donc non seulement possible, mais réalisé et en cours de perfectionnement (notamment dans le domaine de la traduction automatique).

Les premiers travaux de recherche concernaient le lexique et la morphologie.

Au cours de la dernière décennie et avec la diffusion de la langue arabe sur le web, on a remarqué qu'il y a plus d'intérêt à l'étude de cette langue. On voit de plus en plus de recherches et d'outils nécessaires au développement de son traitement automatique, nous citons les références suivantes et il y'en a beaucoup d'autres: [Alo02] [Ali04] [Bac03] [Bou93] [Deb01] [Dic02] [Gau01] [Jac97].

A partir de ces dernières années, des conférences et des workshops ont été même organisés spécialement pour le traitement automatique de la langue arabe tels que JEP-TALN' 04 (traitement automatique de la langue arabe à fés au Maroc), COLING'04 (computational approaches to arabic script-based language à Gen7ve suisse) Arabic language ressourcesand tools Conference au caire Egypte, et d'autres à travers le monde arabe: le dernier est le CITALA07 au Maroc en juin 2007.

Mais malheureusement et contrairement aux langues latines, le problème de l'analyse sémantique reste encore très peu abordé au niveau de la recherche scientifique. La majorité des efforts déployés par les chercheurs arabes et autres sont concentrés sur la réalisation des analyseurs morphologique et syntaxique tels que [Alo03] [Bee01] [Oue01] [Dit01] [Fed89) etc. Nous pouvons citer quelques travaux sur la sémantique dans [Had01] [Had05] [Mef01] [Joh97] [Man96] [Des88]

I.6. Contributions

Mes recherches autour du TALN visent à concevoir et réaliser un système temps réel de synthèse, traduction et reconnaissance de l'arabe. Dans un premier temps, la traduction va se faire entre l'arabe et le français. Pour cela, je me suis intéressé aux différents niveaux de traitement des langues naturelles, à savoir phonétique, morphologique, syntaxique, sémantique et pragmatique.

Plusieurs outils logiciels ont été conçu et réalisés pour monter le prototype final:à savoir la prothèse intitulé "Oreillodule".

Nous avons réalisé un dictionnaire linguistique **étiqueté** d'analyse, de dérivation et de conjugaison de la langue arabe. Ce dictionnaire contient plus que 264 milles mots et il est muni de 7 types d'étiquettes qui sont : **اسم الفاعل، اسم المفعول، اسم التفضيل، اسم المكان، اسم المرّة، اسم آلة، اسم الهيئة.**

En plus, nous disposons d'un système de synthèse de la parole arabe fonctionnelle ainsi que d'un prototype d'un système de traduction de l'oral.

I.7. Plan du mémoire pour la partie recherche

La suite de ce mémoire comporte cinq chapitres. Le chapitre II introduit la conception et la réalisation d'un système de synthèse de la parole arabe et détaille son fonctionnement ainsi que les outils associés développés dans notre équipe de recherche à Monastir. Le chapitre III

rappelle et décrit les principes utilisés pour générer automatiquement une suite de dictionnaires linguistiques et thématiques de l'arabe en se basant sur les conditions de structure morphématique (CSM). Le chapitre IV résume les travaux sur l'analyse sémantique de l'arabe et décrit le système de traduction de l'oral du français vers l'arabe développé dans notre équipe de recherche. Le dernier chapitre conclut ces travaux, trace un bilan et donne quelques perspectives et orientations

II. GENERATION ET ETIQUETAGE AUTOMATIQUE D'UN SYSTEME DE DICTIONNAIRES LINGUISTIQUES ET THEMATIQUES DE L'ARABE

Les travaux sur la génération automatique de dictionnaires linguistiques et thématiques de l'arabe et sur la lemmatisation ont débuté avec mon travail d'après thèse au sein de l'institut régional des sciences de l'informatique à Tunis et donc depuis 1990. J'ai commencé par explorer la possibilité de générer automatiquement des dictionnaires de l'arabe en utilisant les conditions de Structures Morphématiques (CSM): caractéristique qui fait l'originalité de ce travail.

Plusieurs projets de fin d'études, trois masters et deux thèses en cours sont sur ce projet dont le projet de Salim JOUILI et Ammar MAHDHAOUI et les deux thèses de Mohsen MARAOUI au LIDELEM à Grenoble et Ahmed HADDAD à l'ENSI de Tunis

II.1. Problématique

Le dictionnaire est un élément capital pour une bonne performance de tout système de traitement automatique du langage naturel, que ce soit en termes de couverture ou de précision. De même le jeu de catégories grammaticales utilisé dans l'étiquetage a une forte influence sur la qualité du système (Cha95). Il est clair qu'un système réduit de catégories conduira souvent à de meilleurs taux de réussite qu'un système plus détaillé (Mer95).

Plusieurs linguistes ont poussé l'idée de génération automatique de lexique à partir des caractéristiques morphologiques et dérivationnelles de la langue arabe et les conditions de structures morphématiques (CSM), comme Cantinau et Greenberg (HAB76). Cette idée est à la base de ce travail où nous présenterons un système de génération et d'étiquetage automatique de dictionnaires arabes, en se basant sur les CSM, les matrices lexicales (ML), ainsi que leurs structures et leurs modes d'accès (ZAA04) (SIL93) et les schèmes de dérivation de la langue arabe. Le travail a été focalisé au départ sur le dictionnaire des racines admissibles et attestées. Nous avons appliqués des procédures autant que possible automatisées, pour engendrer un maximum d'entrées et d'informations et éliminer les bruits : l'intervention manuelle des spécialistes de la langue s'est avérée nécessaire dans certains cas bien déterminés et limités.

II.1.1. Génération automatique du dictionnaire

II.1.1.1. Les conditions de structures morphématiques (CSM)

Les phonèmes de l'arabe sont liés à des restrictions combinatoires et des restrictions séquentielles très strictes qui sont énoncées sous la forme de CSM. Ces conditions sont des règles qui régissent la génération des mots dans la langue arabe : un mot qui enfreint une condition ne peut pas appartenir à l'arabe (HAB76).

Tout phonème de l'arabe est une colonne de x spécifications correspondant à ces x traits, les (x-quatorze) spécifications qui ne sont pas représentées découlent automatiquement des quatorze présentes en vertu de conditions propres à l'arabe classique

Le tableau suivant illustre un exemple:

Graphème	ب	ت	ط
Phonème	b	t	
sonnant	-	-	-
syll	-	-	-
cons	+	+	+
cont	-	-	-
nas	-	-	-
ant	+	+	+
cor	-	+	+
voisé	+	-	-
strident	-	-	-
rhizo	-	-	+
haut	-	-	-
bas	-	-	+
arriére	-	-	+
rond	-	-	-

sonore	مجهور
syllabique	مقطعيّة
consonantique	صوامتي
continu	مستمر
nasal	أنفي
voisé	مجهور
antérieur	أمامي
arriére	خلفي
bas	خافت
haut	صائت

Extrait du tableau T.10

Les linguistes dénombrent cinq CSM qui régissent la formation des mots arabes. Ces conditions sont classées en deux types: les restrictions combinatoires et les restrictions séquentielles.

Avant de détailler ces restrictions, commençant tout d'abord par donner une formulation mathématique du problème :

Soit x l'ensemble des traits possibles définis par la théorie linguistique.

Soit C l'ensemble des 28 consonnes de la langue arabe. Soit C1C2C3 une racine trilitère, avec C1, C2 et C3 $\in$ C. Soit MP[j] [k] la matrice phonologique (avec $1 \leq j \leq 14$ et $1 \leq k \leq 28$) cette matrice représente l'ensemble des traits des consonnes de l'arabe. Soit V l'ensemble des 6 voyelles de la langue arabe. Soit C1V1C2V2C3V3 une racine trilitère voyellée, avec V1, V2 et V3 $\in$ V. Soit MPv[j][k] la matrice phonologique des voyelles(avec $1 \leq j \leq 14$: l'ensemble des traits des voyelles de l'arabe et $1 \leq k \leq 6$).

II.1.1.1.1 Restrictions combinatoires

Ces restrictions régissent les spécifications des traits correspondant aux phonèmes de la langue arabes. Dans ce cas trois règles sont à énoncer :

1) CSM1 : tous les phonèmes sont [-aspirés]

. La condition CSM1 distingue l'arabe classique de nombreuses langues naturelles qui opposent phonèmes aspirés et non aspirés. C'est l'existence de telles restrictions valables pour tous les phonèmes de l'arabe classique, qui a permis de ne faire figurer que quatorze traits (HABAILI, 1976), parmi x traits possibles définis par la théorie linguistique.

Si ci $\in$ C et ci $\subset$ $C_1C_2C_3$ (avec $1 \leq i \leq 28$) alors MP[aspiré][i] = [0] . (**1**)

2) CSM2 : tous les phonèmes vocaliques sont [-nasal]

La condition CSM2 exclut les voyelles nasales de l'inventaire des phonèmes de l'arabe classique.

Si vi $\in$ V et vi $\subset$ $C_1V_1C_2V_2C_3V_3$ (avec $1 \leq i \leq 6$) alors MPv[nasale][i] = [0] . (**2**)

3) CSM3 : tous les phonèmes qui sont [+consonantiques] sont aussi [-syllabiques]
La condition CSM3 exclut les consonnes [+syllabiques]. Cette règle est formulée de la manière suivante:

Si MP[consonante][i] = [-] alors MP[syllabique][k] = [0] . **(3)**

Outre les restrictions combinatoires entre les valeurs des traits appartenant à un même segment, il existe aussi des restrictions séquentielles.

II.1.1.1.2 Restrictions séquentielles

Ce sont des restrictions qui lient les spécifications de traits appartenant à des segments successifs de la matrice de l'arabe classique, ces restrictions reflètent le fait que n'importe quelle séquence de phonèmes de l'arabe n'est pas un morphème-racine ou un allomorphe possible (variante combinatoire d'un phonème). Par exemple مدّ et كجب sont des séquences de consonnes permises par la structure de la langue, mais pas خخذ et تتت.

Des deux séquences possibles que nous nous sommes données, seule مدّ est effectivement attestée. C'est la racine du verbe مدّ (tendre). Il existe une infinité de séquences de phonèmes qui n'enfreignent aucune des restrictions combinatoires qu'impose la structure de l'arabe classique, mais comme le lexique ne contient jamais que quelques milliers d'allomorphes ou de morphème-racine, la plupart de ces séquences possibles n'y figurant pas , comme كجب par exemple. Le fait qu'il n'existe aucun morphème-racine dont la représentation phonologique soit كجب n'est la séquence d'aucune contrainte structurelle, il s'agit seulement d'une lacune accidentelle : Il s'agit d'une combinaison admissible par la structure de la langue, mais qui est absente du lexique .

En revanche, des séquences telles que "خخذ" ou "ذبذ" ne sont pas des morphèmes-racines possibles en arabe classique. La première enfreint la restriction qui est exprimée par la condition CSM4 et la seconde celle qui est exprimée par CSM5 :

CSM4 : La condition CSM4 exclut de l'ensemble des morphèmes-racines possibles en arabe classique toute séquence de phonèmes formés de deux segments identiques, en première et en deuxième consonne radicale.

Soit $C_1C_2C_3$ une racine trilitère, avec C_1, C_2 et $C_3 \in C$. soit A[j] le vecteur des consonnes de l'arabe classique, avec $(1 \leq j \leq 28)$

$$\forall\ c_i,d_j \in C \text{ et } c_i,d_j \subset \{C_1C_2C_3\} \text{ (avec } 1 \leq i,j \leq n) \text{ et } (c_i = C_1\,,\, d_j = C_2\,) \text{ alors } A[\,i\,] \neq A[\,j\,].$$

CSM5 : La condition CSM5 interdit des consonnes identiques qui sont [+continu, +voisé] en première et troisième consonnes radicales.

Soit $C_1C_2C_3$ une racine trilitère quelconque, avec C_1, C_2 et $C_3 \in C$. soit MP[j] [k] la matrice phonologique (avec $1 \leq j \leq 14$: ensemble de traits des consonnes de l'arabe et
$1 \leq k \leq 28$: ensemble des consonnes de l'arabe, n=28)

$\forall$ $c_i, d_l \in C$ (avec $1 \leq i,l \leq n$) alors
Si [(MP[j][i] = [+ continu], MP[k][i] = [+ voisé]) et (MP[j][l] = [+ continu], MP[k][l] = [+ voisé]) et ($1 \leq j,k \leq 14$)] alors $c_i, d_l \not\subset C_1C_2C_3$.

Ainsi, en liant entre elles certaines spécifications de traits dans les matrices phonologiques, les conditions de structures morphématiques permettent de prédire certaines spécifications à partir d'autres. Prenons par exemple la représentation phonologique du morphème-racine du verbe منع , qui est une certaine matrice de x lignes et de trois colonnes. Pour la distinguer des autres matrices x*3 possibles en arabe classique, il n'est pas besoin de donner la liste exhaustive de x*3 spécifications qu'elle contient, ni même des 14*3 = 42 spécifications de la matrice phonologique du tableau T.11 , il suffit de donner les 11 spécifications contenues dans la table en matrice T.11-b , qui a été obtenue en effaçant de T.11-a toutes les spécifications qui peuvent être prédites à partir d'autres à l'aide des conditions de structures morphématiques.

	ع	ن	م
Son	-	+	+
Syll	-	-	-
Cons	+	+	+
Cont	+	-	-
Nas	-	+	+
Ant	-	+	+
Cor	-	+	-
Voix	+	+	+
Stri	-	-	-
Rhizo	-	-	-
Haut	-	-	-
Bas	+	-	-
Arr	+	-	-
Rond	-	-	-

Tableau T.11-a

ع	ن	م
-		
+		
	+	+
-	+	-
+		
-		
+		
+		

Tableau T11

Tableau T.11. *Matrice phonologique*

En surcroît les conditions de structure morphématique allégent considérablement l'acquisition de nouveaux mots soit en ce qui concerne la langue maternelle ou la langue seconde. Celles-ci permettent aux sujets d'éviter la mémorisation de toutes les spécifications phonologiques

qui sont redondantes, pour ne retenir que le nombre de traits restreint de la représentation lexicale, qui permet d'en déduire les autres spécifications redondantes. Il faut donc distinguer entre les règles phonologiques et les conditions de structures morphématiques. Elles n'ont pas le même pouvoir. Les règles phonologiques ont le pouvoir de changer les signes, d'invertir les colonnes, d'ajouter ou de retrancher des colonnes entières. Par contre, les conditions de structures morphématiques ne peuvent que prédire les spécifications redondantes pour permettre le passage d'une matrice lexicale à une matrice phonologique [HAB 76] : le passage de la matrice lexicale d'un morphème à sa prononciation effective, s'effectue selon le schéma suivant :

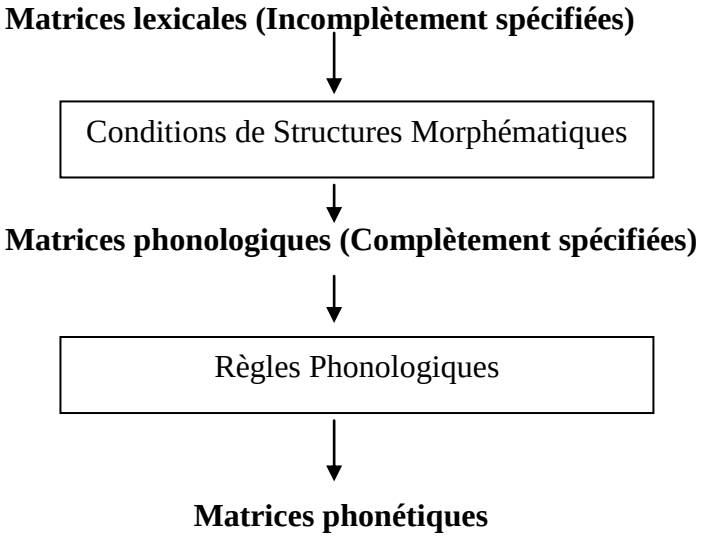


Figure 2 les matrices

Plusieurs linguistes ont poussé l'idée de génération de lexique à partir des conditions de structures morphématiques, comme Cantinau et Greenberg [HAB 76].

II.1.1.2. Les Matrices lexicales :

II.1.1.2.1. Matrices Lexicales Trilitères (MLT)

Ce sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine trilitère, ces matrices sont extraites de la référence "تاج العروس ", avec quelques transformations afin de l'utiliser dans ce travail (HADDAD, 2004). Aux 28 consonnes de la langue arabe correspondent 28 MLT. Les 28 matrices sont issues d'une statistique élaborée par Amr Helmi MOUSSA sur le dictionnaire تاج العروس, en transformant les racines trilitères du dictionnaire en des matrices décrivant les racines attestées par ce dictionnaire. Ce sont des matrices binaires Mi, avec $1 \leq i \leq n$ (n = 28 : nombre des consonnes).
Mi [j][k] exprime les raines $C_iC_jC_k$ (avec i,j et k Є [1..28])(exemple كتب KTB), tel que :
Mi[][] indique la lettre qui est en première position dans la racine $C_iC_jC_k$ (ك) K
Mi[j][] indique la lettre qui est en deuxième position dans la racine $C_iC_jC_k$ (ت) T
Mi[][k] indique la lettre qui est en troisième position dans la racine $C_iC_jC_k$ (ب) B

Nous distinguons les cas suivants :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée par le dictionnaire تاج العروس (exemple كتب)
- Sinon (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " تاج العروس ".(exemple طضذ).

Nous pouvons schématiser comme suit cette représentation:

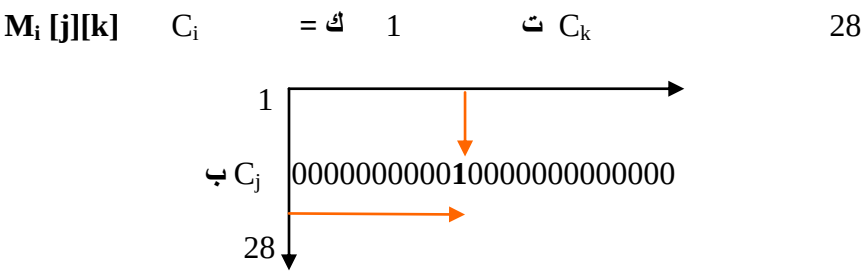


Figure 3 : Représentation de la matrice lexicale

II.1.1.2.2. Matrices Lexicales Quadrilitères (MLQ)

Les matrices lexicales quadrilitères sont des matrices bidimensionnelles qui représentent la position des consonnes dans une racine quadrilitère. En s'inspirant de " تاج العروس ", et du "الشامل في تصريف الأفعال العربيّة ", nous avons pu établir 28 matrices comme suit :

Soit Mi une matrice, avec $1 \leq i \leq 28$. Soit Q une représentation d'une racine quadrilitère quelconque attestée par la langue arabe, soit C1C2C3 une représentation d'une racine trilitère attestée et qui a donnée la racine quadrilitère Q, avec C_1, C_2 et $C_3 \in C$. Mi [j][k] exprime les racines $C_iC_jC_k$ (avec i, j et k $\in$ [1..28])

Ces matrices bidimensionnelles sont formulées de la manière suivante :

- Si Mi [j][k]= 1 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "فاعل", comme "كاتب".
- Si Mi [j][k]= 2 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعّل", comme "بعّد".
- Si Mi [j][k]= 3 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de $C_iC_jC_k$ par dérivation avec le schème "أفعل" , comme " أبعد".
- Si Mi [j][k]= 4 alors la racine $C_iC_jC_k$ est une racine attestée et Q est la racine quadrilitère générée de CiCjCk par dérivation avec le schème "فعلل" , comme "زلزل".
- Si Mi [j][k]= x, avec x $\in$ [أ ب ج د...ه و ي] , alors Q = $C_iC_jC_k$ x , comme "حوقل".
- Sinon (Mi [j][k]= 0) alors la racine $C_iC_jC_k$ n'est pas attestée par le dictionnaire " " تاج العروس.

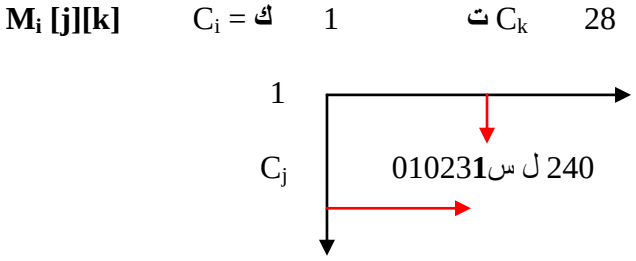


Figure 4 : Représentation de la matrice lexicale quadrilitère

II.1.2. Etiquetage automatique du dictionnaire

L'étiquetage consiste à affecter à chaque lexème toutes les informations morphologique, syntaxique et statistique (classe grammaticale, code de conjugaison pour le verbe, code de déclinaison pour le nom, indice d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, fréquence d'apparition,…). Nous présentons dans ce qui suit quelques observations concernant la morphologie de l'arabe, en rappelant toutefois qu'il ne s'agit pas de discuter le bien-fondé linguistique des divers concepts que nous allons présenter, mais nous voulons simplement apprécier dans quelle mesure chaque concept nous aide à atteindre notre objectif, à savoir l'étiquetage automatique du dictionnaire arabe.

II.1.2.1. Caractère dérivationnel

Il semble généralement admis qu'en arabe il existe un mécanisme morphologique qui permet de décrire une large partie des formes canoniques du lexique. Ce mécanisme est basé sur la notion de schème. Un schème est généralement défini comme un modèle qui décrit un groupe de mots partageant certaines propriétés linguistiques (phonologiques, morphologiques, syntaxiques et sémantiques). Sous l'angle morphologique et selon notre centre d'intérêt, le schème est simplement une sorte de fonction dans laquelle viennent se couler les racines pour former des mots. La combinaison d'un petit nombre de schèmes (14 schèmes nominaux et 10 schèmes verbaux) avec l'ensemble des racines attestées suffirait donc pour décrire la majorité des mots arabes. Cette observation nous conduit à la possibilité de générer de façon automatique la presque totalité du lexique arabe théorique en partant de l'ensemble des racines et de l'ensemble des schèmes.

Désormais, le problème n'est pas aussi simple. L'application systématique du principe de dérivation décrit ci-dessus nous permet de générer artificiellement des mots ne répondant pas aux critères d'appartenance à la langue. La raison en est qu'en réalité, certains schèmes ne peuvent pas aller avec certaines racines. Si la majorité des mots de la langue peuvent toujours être ramenés à une racine et à un schème, toutes les racines et tous les schèmes ne peuvent être croisés pour former des mots de la langue arabe.

Toutefois, pour que le lexique généré soit exempt d'erreurs, nous devons accompagner cette procédure automatique d'une autre procédure manuelle. Selon les démarches possibles suivantes :

Soit mener à posteriori une opération de correction, qui vise à éliminer les mots générés à partir de croisements inopportuns entre racines et schèmes.

Soit effectuer en amont, un travail préparatoire permettant de définir les appariements possibles entre racines et schèmes; et procéder en aval, à l'élimination des mots éventuellement générés de façon abusive. Plus cet appariement initial sera fin, moins il y aura de mots incorrects générés.

Nous avons choisi la seconde démarche qui nous semble plus exhaustive. Les appariements possibles sont assurés par les matrices de dérivation verbales et nominale, ces matrices binaires définissent pour chaque racine les dérivées possibles (verbales pour la matrice de dérivation verbale et nominale pour la matrice de dérivation nominale) ces matrices sont formulées de la manière suivante :

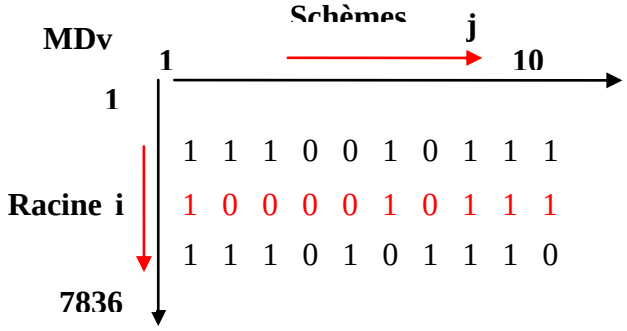


Figure 5 : Matrice de dérivation Verbale

Si MDv[i][j] = 1 alors le schème chj coincide avec vi et le dérivé Dvi,j existe

Si MDv[i][j] = 0 alors le schème chj ne coïncide pas avec vi et le dérivé Dvi,j n'existe pas

Ces matrices seront générées une seule fois par un linguiste, mais leur utilisation est indispensable dans la génération automatique des dérivés verbaux et nominaux.

II.1.2.2. Caractère flexionnel

"L'arabe est une langue à flexions. Elle emploie, pour la conjugaison du verbe et pour la déclinaison du nom, des indices d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, qui sont en général des suffixes." [CHI 98]

Les principes de suffixation et de préfixation des mots permettant la conjugaison des verbes et la déclinaison des noms sont connus et, ils donnent l'impression d'être facilement automatisables. Nous serions donc tentés d'automatiser l'opération de génération de manière complète les formes fléchies.

II.1.2.3. Fréquence d'apparition

Bien que le dictionnaire ne comporte pas véritablement de probabilités, il indique pour chaque mot sa fréquence d'apparition , cette fréquence est déduite à partir d'un corpus étiqueté, le corpus sur lequel nous avons travaillé est constitué d'un texte, représentant un volume global de X mots.

La complexité du dictionnaire est de la forme :

$$\text{Complexité} = \frac{\sum_{i=1}^{V} A(m_i)}{V}$$

Où V désigne la taille du vocabulaire, $A(m_i)$ dénote l'ambiguïté grammaticale du mot m_i , c'est-à-dire le nombre du classe grammaticale différentes qui peuvent être affectés au mot m_i.

II.2. Description du système réalisé

Le système développé est articulé autour des trois fonctions suivantes :

- La génération automatique des dictionnaires (Figure A.28)

Le schéma suivant illustre le diagramme des données relatif à la génération automatique des différents dictionnaires :

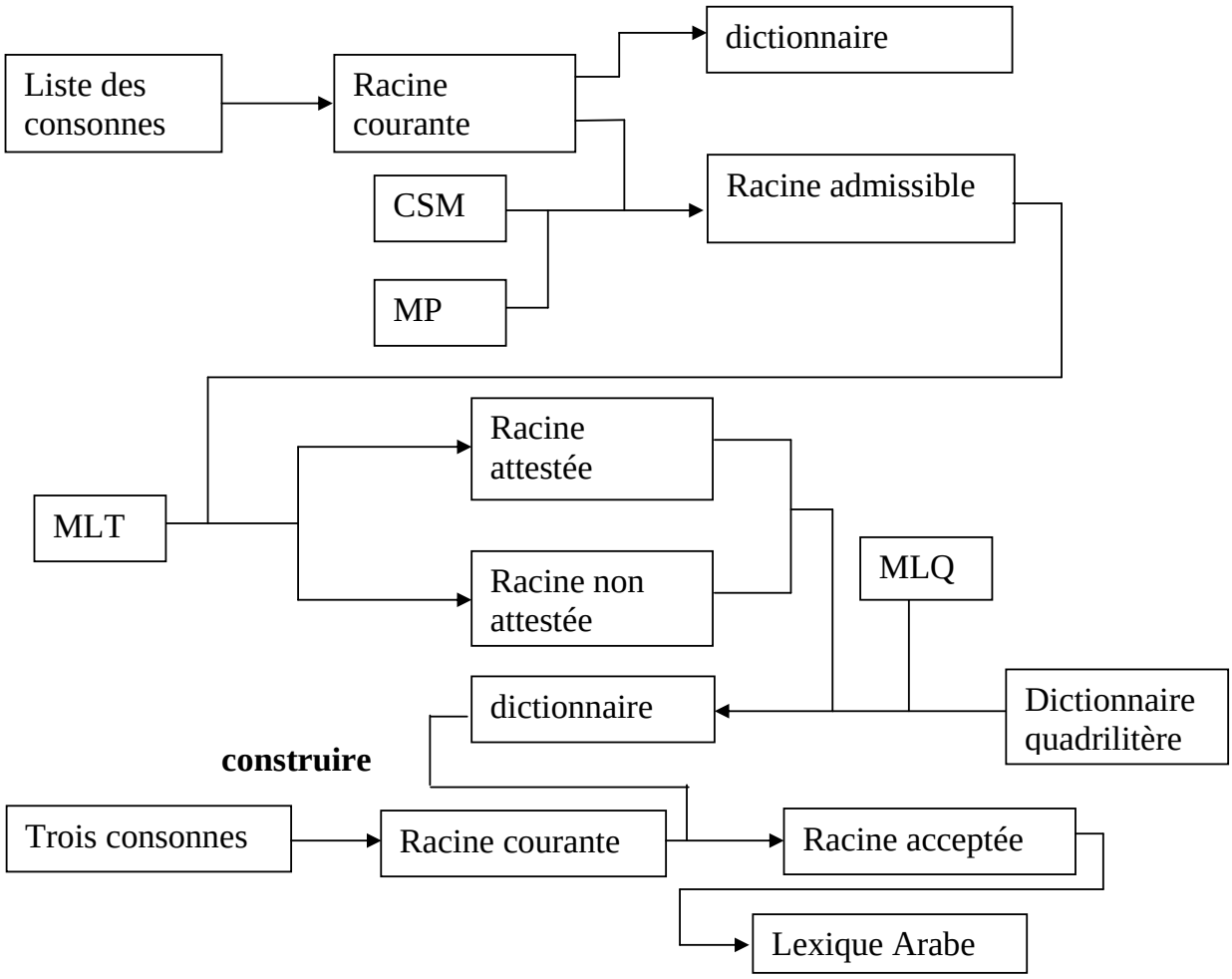


Figure 6 Diagramme des données relatif à la génération des différents dictionnaires du lexique arabe

- La consultation de ces dictionnaires dans le but, de la recherche d'une racine trilitère donnée, ou l'affichage d'une liste des racines d'un dictionnaire spécifié.
- L'affichage de toutes les racines attestées (utilisées dans la langue arabe), construites à partir de trois consonnes données : cette méthode s'appelle " اللإشتقاق الأكبر" .

II.2.1 Les entrées et les informations associées :

La forme générale associée à chaque entrée du dictionnaire est composée des huit zones suivantes (fig.A.29) :

- Zone 1 : la forme canonique de l'entrée (entrée sous forme normée c'est à dire la racine trilitère).
- Zones 2,3 et 4 : l'ensemble des dérivées de la forme canonique (dérivés verbaux), ces dérivés respectent des modèles (schèmes) précis et sont générés selon le besoin par un algorithme approprié, avec le code de dérivation (tableau T.14) et du code de conjugaison (tableau T.15) associé à chaque dérivé verbal.
- Zone 5, 6,7 et 8 l'ensemble des dérivées de la forme canonique (dérivés nominaux), ces dérivés respectent des modèles (schèmes) précis et sont générés selon le besoin par un algorithme approprié, en plus du code morpho-syntaxique(tableau T.16), code de dérivation(tableau T.17) et une information flexionnelle(tableau T.18) associée à chaque dérivé nominal.

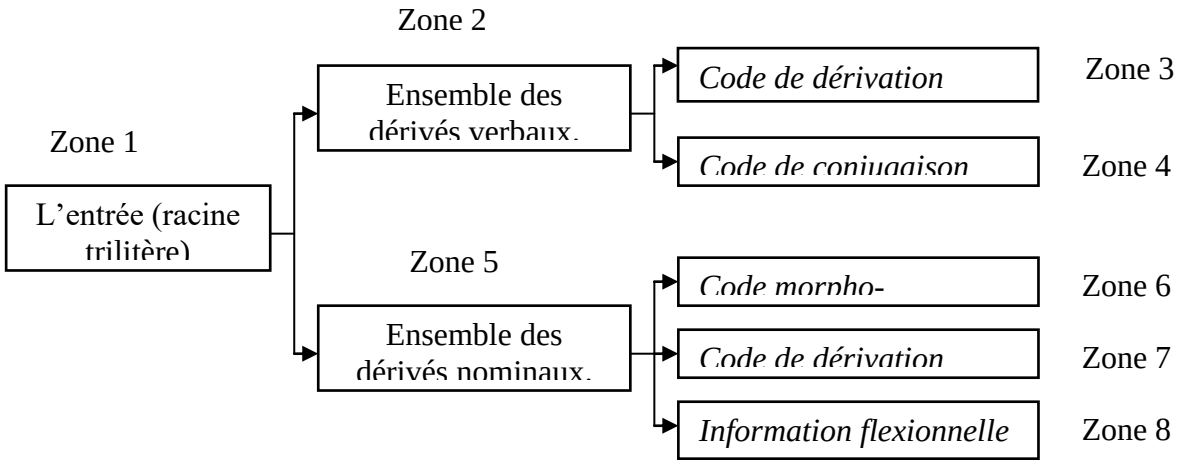


Figure 7 Association entrée / information

CODE DE DERIVATION	ف1	ف2	ف3	ف4	ف5	ف6	ف8	ف9	ف10
schème	لعف	لعاف	لعفأ	لّعفا	لعفنا	لعتفا	لعافتـ	لعفتـ	لعفتسا

Tableau T.14

CODE DE CONJUGAISON	ت1	ت2	ت3					ت127	ت128
schème	يفعل	يفعل	يفعّل					يفعلل	يتفعلل

Tableau T.15

CODE MORPHOSYNTAXIQUE	ص1	ص2	ص3	ص4	ص5	ص6
schème	اسم الفاعل	اسم	اسم الزمان	جمع	الصّفة	المصدر

		المفعول	والمكان	الزمان والمكان		

Tableau T.16

CODE DE DERIVATION	1م	2م	3م	4م	5م	6م	8م	9م	10م
schème	لاعف	لوعف	نلاعف	لاعفم	ةلاعف	لوعفم	ةلوعف	ةلاعفم	لعف

Tableau T.17

code	1ذ	2ذ	3ذ	4 ذ	5 ذ	6ذ
Information flexionnelle	مذكّر	مؤنّث	مفرد	1جمع	2جمع	3جمع

Tableau T.18

II.2.2 Structure interne des données du dictionnaire capital :

Pour réaliser notre dictionnaire capital, nous avons adopté la structure de listes chaînées. Cette structure permet un accès simple et une recherche facile des informations.

La figure suivante décrit cette représentation interne ainsi que les différents niveaux des données de notre dictionnaire :

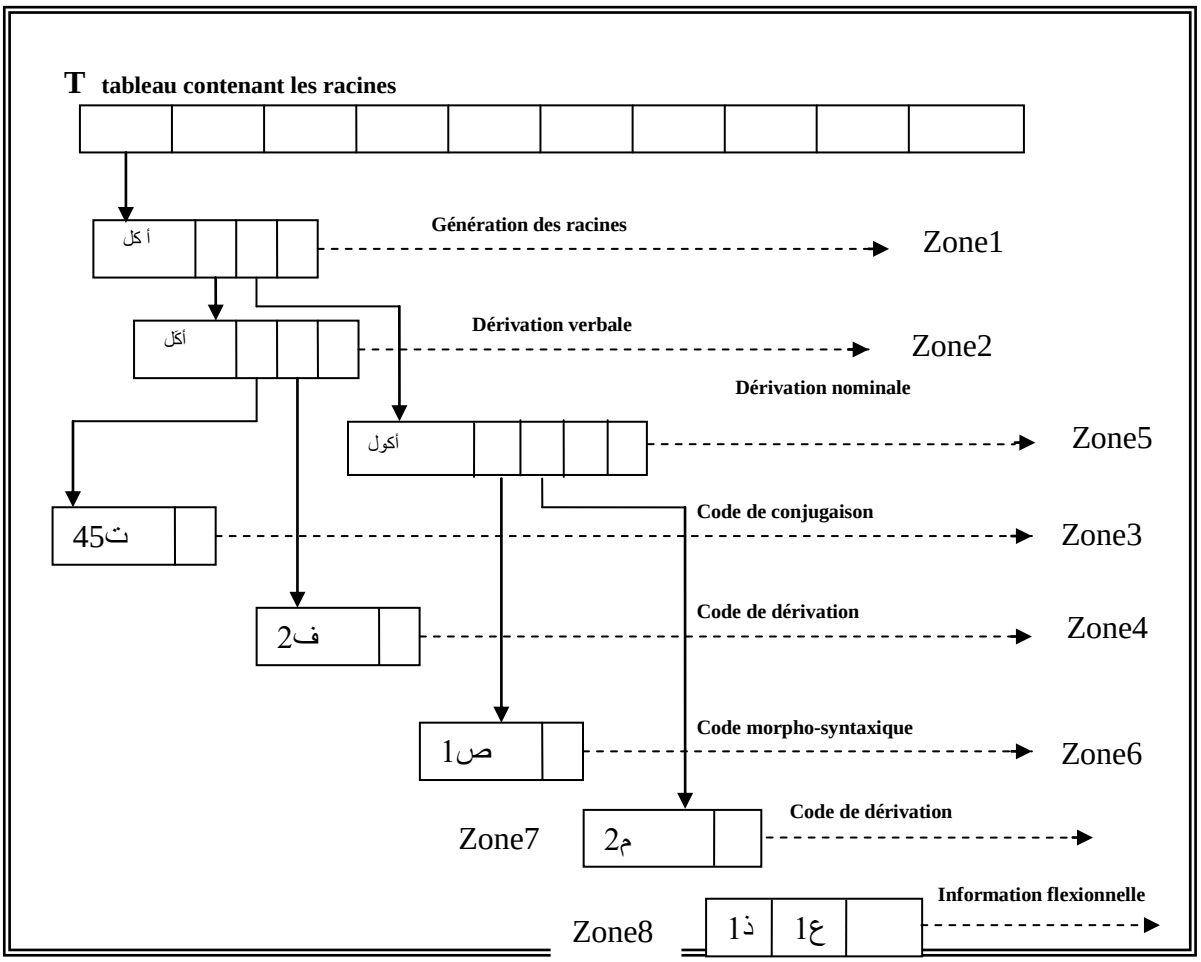


Figure 8 : Structure interne des données

Dans ce schéma, T est un tableau contenant des racines trilitères ; il est de type Entrée[][n] où n est la taille de la fonction de hachage Fhc qui associe à chaque racine de clé c un code égale à Fhc (c) , donc les racines ne sont pas enregistrées dans le tableau T par ordre de génération mais par leurs codes déduits de Fhc, et en cas de collision (deux racines ayant la même image par la fonction de hachage Fhc), on associe à l'entrée Fhc (c) un tableau contenant toutes les racines ayant le même code.

II.2.3 Le HASH-CODING

Le hash-coding se préocupe de localiser directement un élément à partir de l'information dont on dispose. Ce terme qui provient du mot anglais « TO HASH » (hacher, découper) désigne la méthode dont le principe est d'associer à une racine, une adresse et une seule où elle sera stockée en mémoire.

II.2.4 Formalisme du problème :

On détermine l'adresse réelle associée à une racine r_n avec $1 \leq n \leq N$ (nombre des racines) à partir d'un argument I_n. Cet argument étant une valeur calculée directement à partir de la racine.

Soit $I = \{ I_1,\ldots,I_n\}$ l'ensemble de ces arguments qui constituent des indices d'aiguillage pour la recherche de la racine.

Soit $A = A_1,\ldots,A_n$ l'ensemble des adresses effectives associées à tout élément de I.

On est ramené à déterminer une fonction fhc : I → A telle que

$$Fhc(I_i) = A_i : 1 \leq i \leq N ; 1 \leq A_i \leq M ; I_{inf} \leq Ii \leq I_{sup}$$

- N étant le nombre de racines à placer et M la taille de la mémoire où elles seront stockés ($N \leq M$)
- I_{inf} et I_{sup} dépendent de la méthode de calcul de l'indicatif.

Notation :

On notera cont(X) : le contenu de l'emplacement mémoire relatif au champ de la racine uniquement. Avec les notations et les restrictions adoptées, on doit donc s'efforcer de trouver une fonction fhc telle que :

$$Cont(Fhc(IND(Ri))) = Ri \quad 1 \leq i \leq N.$$

Le schéma suivant illustre le Hash-Coding :

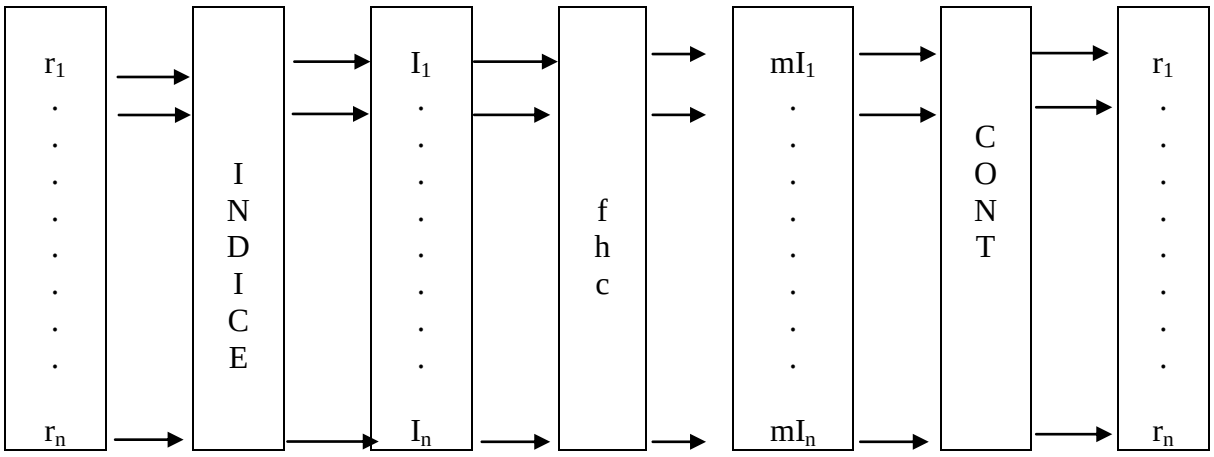


Figure 9 traitement lors du hash-coding

On voit à travers cette figure que les différents traitements à effectuer lors de la recherche d'une racine r_j $1 \leq j \leq n$ sont les suivantes :

- Calcul de l'indicatif correspondant
- Recherche de l'emplacement mémoire correspondant

- Accès au champ racine

Lors du traitement de la deuxième étape, fhc essaiera de placer n éléments en n adresses à choisir parmi m. Certains de ces I_i donneront lieu par une première transformation à une même adresse X. on dira que ces I_i sont synonymes ou encore qu'il y a eu collision en X. la préoccupation du hash-coding sera de trouver :

a) Une fonction notée fhc qui, pour une taille M du fichier, donnera lieu à un petit nombre de collisions et donc étalera bien les valeurs entrées.

b) Résoudre au mieux le problème des collisions en cherchant un prochain espace de libre p pas plus loin.

Le système développé est composé de deux fonctions qui sont :
1. la génération des dictionnaires,
2. la consultation des dictionnaires.

Notre système permet de générer 5 types de dictionnaires:

- Le premier dictionnaire est théorique (21952 racines = (28)^3). Il contient toutes les racines trilitères théoriquement possibles de l'arabe standard.
- Le deuxième dictionnaire (20415 racines) : c'est le dictionnaire des racines trilitères admissibles. C'est-à-dire les racines qui n'enfreignent aucune des (CSM).
- Le troisième dictionnaire (7836): c'est le dictionnaire des racines trilitères attestées ; c'est-à-dire utilisées dans la langue arabe et qui sont tirées des tableaux de répartitions construits à partir du grand dictionnaire arabe(الصحاح لإبن الجوهري).
- Le Quatrième dictionnaire (13023 racines): c'est le dictionnaire des racines admissibles par la langue arabe mais non attestées. Ces racines peuvent être utilisées pour enrichir la langue arabe par d'autres mots nouveaux.
- Le cinquième dictionnaire (4000 racines) : c'est le dictionnaire des racines quadrilitères attestées ; qui sont tirées des matrices lexicales quadrilitéres.

Certaines racines trilitères attestées n'obéissent pas à une ou plusieurs CSM : nous avons créé un sixième dictionnaire (203 racines) qui regroupe ces racines, avec pour chacune, l'affichage de la CSM qui n'est pas vérifiée. Exemple : la racine (ببب) est attestée mais ne vérifie pas la condition CSM4.

Pour le stockage des dictionnaires nous avons étudié plusieurs méthodes et essentiellement les principaux standards utilisés, à savoir : Traitement de texte (Word), ASCII, RTF, Bitmap(TIFF, GIF,...) , SGML, HTML, XML et PDF).
Chacune de ces méthodes a des avantages et des inconvénients. Le tableau suivant décrit cela :

Format	Principale application	Principal avantage	Principal inconvénient
Traitement de texten et PAO (Word, WordPerfect, FrameMaker,...)	Éditique	Richesse des mises en formes	Format exclusif à son application

ASCII	Echange	Compatible avec toutes les applications	Aucune mise en forme
RTF	Echangé/Conversion	Formatage de base conservé	Restreint à certaines applications et plate-formes
Bitmap(TIFF, GIF,...)	Numérisation	Informatisation du document papier	Fichiers volumineux
SGML	Edition structurée	Norme officielle	Complexité
HTML	Diffusion Web	Très répandu	Très changeant
XML	Diffusion Web	Simplification de SGML	Nouveauté
PDF	Distribution	Ne nécessite aucun balisage ou travail d'édition	Sous licence (propriété d'Adobe)

Tableau T.19 *Formats numériques de documents textuels : utilisation, avantage et inconvénient principaux*

Nous avons choisis XML pour la représentation des dictionnaires fléchies et le dictionnaire capital. Ce choix a été retenu pour les raisons suivantes:

- La possibilité de donner du " sens " au contenu d'un document
- l'utilisation du jeu de caractère Unicode qui a la prétention de coder l'ensemble des caractères utilisés dans toutes les langues de la planète
- XML est issu des travaux effectués sur SGML (Structured Generalized Markup Language) le format de fichier défini dans la norme ISO 8879 de 1986. Le but recherché par XML est de permettre un balisage sémantique d'un document.
- La possibilité de définir les balises necessaires.

II.3 Discussion

Les travaux sur la génération automatique des dictionnaires nous a permis d'identifier les limites des outils actuels et d'imaginer des voies de recherche possibles pour nos futures expérimentations.

Pour la consultation de dictionnaires, il serait très intéressant de pouvoir accéder au méta information sur les ressources afin de distinguer leur qualité et leur couverture. Les utilisateurs aimeraient aussi pouvoir consulter plusieurs dictionnaires avec la même interface même si ces dictionnaires ont des formats hétérogènes. Ils pourraient de ce fait comparer plus facilement les articles des différents dictionnaires. Il nous semble aussi nécessaire de proposer des outils d'aide à la consultation, en amont ou en aval en proposant des correcteurs orthographiques et des lemmatiseurs (pour la recherche) ou des conjugueurs (pour l'utilisation). Enfin, il est indispensable que l'utilisateur puisse personnaliser le résultat

de ses requêtes au niveau de la structure (informations à cacher, etc.) et de la présentation (style, couleurs, polices, etc.) afin de sélectionner uniquement les informations dont il a besoin dans une grande quantité d'information.

Pour la construction de dictionnaires, il est possible de distinguer deux démarches : la rédaction d'articles entiers et les contributions localisées sur des parties d'articles. Pour la rédaction, il faut proposer des outils d'aide à la rédaction et aussi un mécanisme d'aller/retour entre les rédacteurs et la base pour pouvoir réviser le travail accompli. Pour les contributions, il faut disposer d'outils simples fonctionnant directement en ligne et permettant de partager les contributions entre plusieurs utilisateurs. Ensuite, il faut mettre au point une procédure de validation /correction /intégration des données.
Pour la structure interne des dictionnaires, les standards choisis en vue de garantir la portabilité et la compatibilité avec un maximum d'outils sont les standards UNICODE et XML.

Signalons en fin que la génération automatique du dictionnaire des racines trilitères et quadrilitères en utilisant les CSM et les ML fait l'originalité de ce travail. Ce dictionnaire sera à la base de toute analyse morphosyntaxique de l'arabe, il regroupe les racines du grand dictionnaire (معجم الصحاح), auquel on peut ajouter d'autres dictionnaires.

Le dictionnaire capital, résultat de ce système, contient 36876216 verbes et mots de l'arabe : ce dictionnaire est généré automatiquement à la demande de l'utilisateur donc ne pose pas de problèmes d'encombrement en mémoire.

II. CONTRIBUTION A LA SYNTHESE AUTOMATIQUE DE LA PAROLE ARABE

Les travaux en synthèse de la parole concernent surtout la synthèse à partir du texte. Les efforts se sont portés sur plusieurs points. Les traitements linguistiques du synthétiseur ont été particulièrement développés. Cela a fait l'objet de plusieurs travaux de projets de fin d'études, de mémoire de DEA et plus particulièrement la thèse de Tahar SAIDANE, soutenue en octobre 2006 à l'université de la Manouba Tunis, (Ecole Nationale des Sciences de l'Informatique de Tunis).

II.1. Problématique

Ces dernières années, les techniques de traitement de la parole ont connu plusieurs grandes révolutions. La première, est celle qui touche le plus grand nombre d'utilisateurs: la téléphonie mobile. Un nombre important d'utilisateurs transporte, souvent sans le savoir, un ordinateur de poche spécialisé dans l'analyse-synthèse LPC.

Les algorithmes de codage sont par ailleurs également utilisés dans les boites vocales: nos paroles y sont stockées sous forme de suites de vecteurs de paramètres LPC. Le marché du codage de la parole est donc à présent largement ouvert, la reconnaissance et la synthèse vont suivre.

La seconde révolution est celle des grandes bases de données de parole et de textes. Depuis 1995, sous l'égide de LDC (Language Data Consortium) aux Etats-Unis et de l'ERLA (European Language Resource Agency) en Europe, de nombreux laboratoires de recherche mettent en commun leurs ressources. Il en résulte une abondance de données propices à l'établissement de modèles numériques de la parole.

Une troisième révolution, est celle des outils d'ingénierie pure (modèles de Markov cachés, réseaux de neurones artificiels, synthèses par sélection d'unités dans une grande base de données), qui tendent à supplanter de plus en plus de l'expertise humaine (reconnaissance analytique, synthèse par règles), laquelle intervient plutôt au second plan, en permettant d'affiner les résultats.

Les technologies de la synthèse de la parole modernes impliquent des méthodes et des algorithmes sophistiqués. Une des méthodes actuelles appliquées est les modèles de Markov cachés (HMM: hidden Markov models). Les HMM ont été appliqués à la reconnaissance de la parole vers la fin des années 70 [Cae99].

Les réseaux de neurones ont eux aussi été sollicités lors des dix dernières années et les résultats obtenus sont encourageants. Cependant cette méthode n'a pas été explorée suffisamment. Comparativement aux HMM, cette méthode n'est pas gourmande en espace mémoire mais présente une difficulté importante au niveau du choix des paramètres du réseau et de son architecture (Caw96). Plusieurs travaux s'y intéressent[Ben98], [Jau98], [Caw96].

Nous avons classé les méthodes de synthèse en trois groupes:

La synthèse articulatoire qui essaie de modéliser directement le système de la production de la parole humain.

La synthèse par règles qui modélisent les transitions entre les phonèmes, pour commander un synthétiseur à formants.

La synthèse par concaténation qui utilise des segments d'enregistrement, de parole naturelle, de longueurs différentes.

Les méthodes à formants et par concaténation sont les plus communément utilisées dans les systèmes de synthèse actuels [cal89], [Ben98]. La méthode à formants était pendant

longtemps dominante, mais aujourd'hui la méthode par concaténation devient de plus en plus populaire. La méthode articulatoire est plus compliquée surtout pour la synthèse de haute qualité, mais peut devenir la méthode du futur (à cause de la complexité des calculs mis en jeu et d'une nécessité de performance matérielle accrue).

La synthèse par concaténation consiste à synthétiser le signal par concaténation d'unités acoustiques, c'est-à-dire de segments de parole préenregistrés. Cette technique repose sur l'utilisation de segments de signaux extraits de la parole naturelle, ce qui permet de synthétiser des voix dont le timbre s'approche de celui d'un locuteur humain [Lem02]. Pour pouvoir effectuer de la synthèse, il va falloir disposer d'éléments minimaux acoustiques permettant de constituer n'importe quelle phrase faute de devoir mémoriser toutes les phrases d'une même langue.

Un des aspects les plus importants de la synthèse par concaténation est de trouver la longueur optimale de l'unité à utiliser. Avec les unités longues on réussi à avoir un haut niveau de naturel, moins de points de concaténation et un bon contrôle de la coarticulation, mais le nombre d'unités exigées et la quantité de mémoire nécessaire sont plus grands. Avec les unités courtes, moins de mémoire est exigée, mais le rassemblement et l'étiquetage des échantillons deviennent plus difficiles et plus complexes. Le choix des unités joue donc un rôle primordial pour ce type de synthèse.

Une voie de recherche actuelle est la synthèse par unités de taille variable. L'idée est de se fixer une taille globale du dictionnaire d'unités et d'inférer, à partir d'un corpus de texte représentatif des phrases à synthétiser, l'ensemble optimal des unités. Ces unités peuvent être au choix, des segments de phrases, des mots ou des fragments de mots, des syllabes, des diphones ou même des sons isolés [Zri91].

Pour notre part, nous avons choisi d'établir notre propre stratégie de sélection d'unités acoustiques. En effet, nous avons abouti à ce choix suite à une étude approfondie de la langue arabe et en s'inspirant des méthodes récentes de synthèse par unités de taille variable. Il s'agit d'utiliser trois types d'unité : le phonème, le diphone et le triphone. La combinaison des trois unités est bien sûr régie par un algorithme d'optimisation qui à chaque situation présente la combinaison idéale. Ce choix nous a permis non seulement d'obtenir une meilleure qualité de naturel mais aussi de limiter le nombre d'unités à stocker pour un vocabulaire illimité.

II.2. Le système hybride de synthèse de la parole

Un système de synthèse est divisé en deux grandes parties : une partie linguistique (transcription) et une partie acoustique (concaténation). Le schéma suivant présente les constituants d'un tel système :

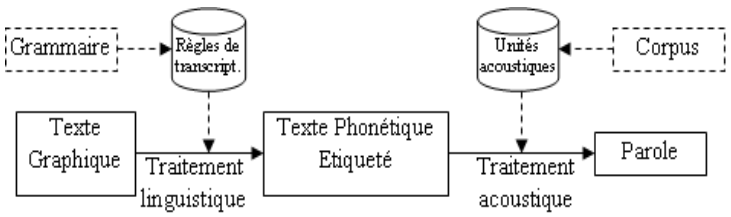


Figure 10 Schéma de principe de notre système de synthèse de la parole.

La partie linguistique ou symbolique du système de synthèse de la parole permet, à partir d'un texte écrit (graphique), de générer un texte phonétique étiqueté (allophones). Ce passage comprend [Gue98] le prétraitement du texte, la correction phonétique et morphologique, le traitement phonologique (conversion graphème phonème, détermination des allophones et l'accentuation en fonction des structures syllabiques des mots) :

A l'issue du traitement linguistique appelé encore transcription orthographique phonétique, nous aboutissons à un texte phonétique étiqueté.

Le module de transcription comporte plusieurs phases présentées dans le schéma suivant :

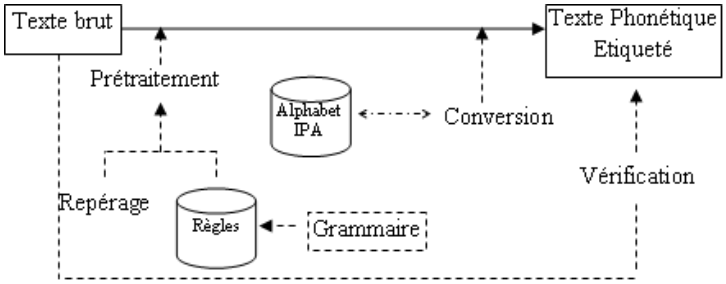


Figure 11 Schéma de principe du système de transcription.

Une des premières recherches à effectuer avant toute autre démarche consiste à formaliser au mieux les problèmes posés par la langue arabe. Nous pouvons alors mettre en évidence les règles de transcription les plus générales, les exceptions, etc. En ce qui suit le principe d'utilisation des règles de transcription dans notre système de synthèse de la parole :

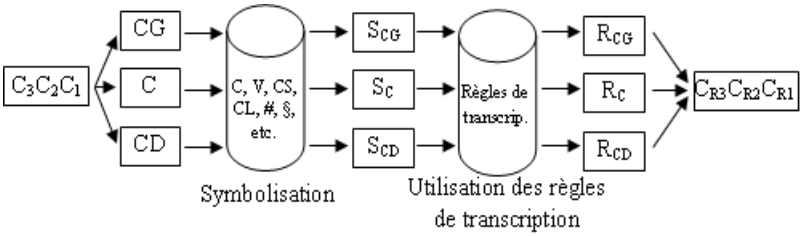


Figure 12 Les différentes phases de fonctionnement du système de transcription.

Notre analyse linguistique nous a permis d'établir pour notre système un ensemble de 133 règles. Il est à noter que l'ordre d'application de ces règles est très important et influe énormément sur le résultat final [Sai02]; comme règles nous pouvons donner les deux suivantes:

1. [uu]={CS}+{ُ}+{و} [uu]={CL}+{ُ}+{و}

Lorsque leو est précédé par la voyelleُ et qu'il est suivi par une consonne, on obtient le phonème de la voyelle longue [uu]. Exemple : حُوتٌ , دُونَ.

2. [uu]= §+{ُ}+{و} [uu]= $+{ُ}+{و}

Lorsque leو est précédé par la voyelleُ et qu'il est en fin de mot, on obtient le phonème de la voyelle longue [uu]. Exemple : كَتَبُوا , لَبَسُوا.

II.3. La syllabation

Les méthodes de synthèse directe se limitent à concaténer un certain nombre de segments à partir d'éléments temporels stockés pour chaque phonème [Zri90]. Le résultat de cette concaténation mène à la reconstitution phonème par phonème, ensuite à l'association entre ces phonèmes pour former un mot. Toutefois, cette technique ne permet pas de prendre en compte les phénomènes de coarticulation traduits par l'influence d'un phonème sur un autre voisin. L'intelligibilité de la parole se trouve ainsi limitée [Zri90]. Le principe de notre système de syllabation se présente comme suit :

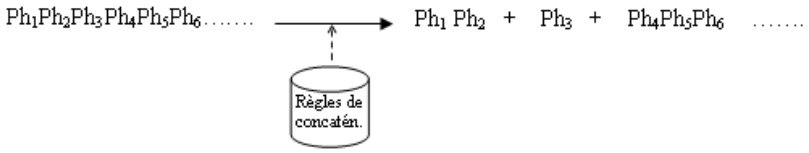
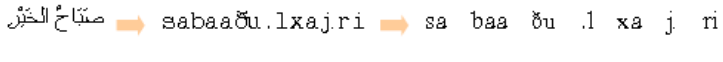


Figure 13 Schéma de principe du système de syllabation.

Nous avons alors conclu qu'il n'est pas viable de faire de la synthèse par simple concaténation de phonèmes car précisément, ce sont les transitions entre les phonèmes qui transportent l'information pertinente. L'option retenue est alors une application de règles de syllabation établies par nos soins et propres à notre système [8].

II.3.1. Choix des unités acoustiques

Dans notre système de synthèse par concaténation, les unités acoustiques sont de trois types : les triphones, les diphones et les phonèmes. Ceci nous a permis d'apporter plus de souplesse et surtout une meilleure qualité à notre module acoustique. Par ailleurs nous avons pu limiter considérablement le nombre d'unités acoustiques. On a établi un ensemble de règles de concaténation à partir desquelles les différentes occurrences de trois phonèmes pouvaient se transformer en : un triphone, un diphone suivi d'un phonème, un phonème suivi d'un diphone, ou éventuellement trois phonèmes. L'entrée du module de sélection est une séquence de phonèmes, l'algorithme converge alors vers une suite optimale d'unités acoustiques à concaténer. La sélection dynamique des unités se traduit alors par la recherche de la séquence optimale de représentants, visant à minimiser les discontinuités au point de concaténation [9], [8]. Le schéma suivant présente un exemple de syllabation pour l'expression «صَبَاحُ الخَيْرِ » (Bonjour) :



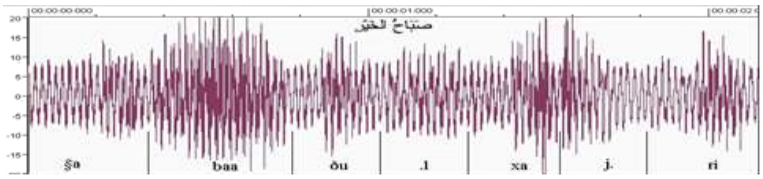


Figure 13 Exemple de syllabation

II.3.2. Les règles de syllabation

La problématique de la sélection des unités a été formalisée en établissant six règles de syllabation [Sai04] qui sont :

1. [CVV] = {V} + {V} +{C} : lorsqu'une consonne est suivie de deux voyelles les trois graphèmes constituent une unité acoustique de notre système.

2. [CV]={C}+{V}+{C} : lorsqu'une consonne est suivie d'une voyelle puis d'une consonne les deux premiers graphèmes constituent une unité acoustique de notre système.

3. [CC]={C} +{C} +{C} : lorsque nous avons une succession de trois consonnes les deux premiers graphèmes constituent une unité acoustique de notre système.

4. [C]= {V} +{C} +{C} : lorsque nous avons deux consonnes suivies par une voyelle seul le premier graphème constitue une unité acoustique de notre système.

5. [VV]= {V} + {V} : lorsque nous avons une succession de deux voyelles, les deux constituent une unité acoustique de notre système.

6. [V]= {V} : lorsque nous avons une voyelle isolée elle constitue une unité acoustique de notre système.

L'utilisation de ces différentes règles de syllabation se présente sous la forme suivante :

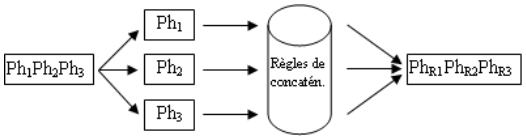


Figure 14 Utilisation des règles de syllabation

Il est à noter que l'ordre d'application de ces règles ainsi établies est très important pour une bonne syllabation et donc une meilleure concaténation sonore. C'est à partir de ces résultats que nous avons recueilli les échantillons sonores susceptibles de nous aider à la constitution de la base d'enregistrement nécessaire à notre synthèse vocale [9], [11].

II.3.3. Elaboration du dictionnaires des Unités Acoustiques (UA)

Les six règles de syllabation élaborées vont imposer les types d'unités acoustiques à utiliser pour la synthèse de la parole.

Théoriquement, le dictionnaire ainsi établi contient 196 unités acoustiques suffisantes pour la réalisation des différentes occurrences possibles: 28 phonèmes de type C, 84 diphones de type CV et 84 triphones de type CVV.

Néanmoins, la pratique et l'étude de la langue arabe ont permis de dégager d'autres unités dues principalement aux contraintes de la langue [Sai04].

Pour constituer un dictionnaire d'unités acoustiques il faut disposer de toutes les combinaisons réalisables. Le module de concaténation a besoin de la totalité des unités acoustiques sous la forme d'enregistrements sonores. Ces enregistrements constituent le dictionnaire de notre système.

C'est à partir du corpus de mots et d'expressions choisies que commence l'élaboration du dictionnaire d'unités acoustiques. Les étapes de réalisation peuvent se résumer en ce qui suit :

- La saisie du corpus de mots et d'expressions qui permettra d'obtenir, sous de bonnes conditions, la totalité des unités acoustiques nécessaires.

- L'enregistrement sonore des expressions.

- La segmentation des enregistrements sonores obtenus en phonèmes, diphones et triphones.

- Le test du dictionnaire obtenu.

Le schéma suivant explicite les étapes de constitution de notre dictionnaire :

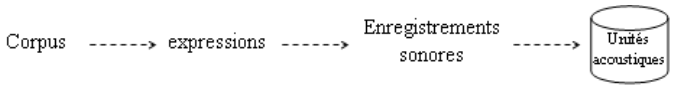


Figure 15 Etapes de constitution du dictionnaire

La qualité du résultat final de la synthèse dépend directement de la qualité des enregistrements effectués lors de l'élaboration du dictionnaire d'unités acoustiques ; quelques précautions ont alors été prévues [Lem00], [Mou96] :

- L'utilisation d'un seul locuteur par dictionnaire et la limitation des séances d'enregistrement : pour l'homogénéité du timbre.

- La prononciation sur un ton monocorde et par petites périodes afin d'éviter l'effet de liste au cours de l'enregistrement.

- La minimisation du risque de perte d'information lors de la phase de numérisation (choix de la bonne fréquence d'échantillonnage).

Le dictionnaire d'unités acoustiques ainsi établi a une taille de 9 MØ (en moyenne un phonème prend 20 kØ, un diphone 40 kØ et un triphone 60 kØ).

Pour l'extraction de la totalité des polyphones nous avons utilisé les enregistrements de près de 137 phrases et expressions (d'une taille de 1 MØ) utilisant le vocabulaire arabe usuel. Nous avons par la suite relevé la fréquence d'utilisation des différents polyphones dans ces expressions afin de se donner le maximum de possibilités pour une bonne extraction des unités. Le choix d'un corpus composé de mots artificiels est soutenu par les raisons suivantes:

• Il n'est pas aisé de trouver des mots réels qui contiennent toutes les unités acoustiques escomptées.

• Les mots artificiels ont l'avantage de pouvoir être créés automatiquement à l'aide d'un logiciel. Les logatomes, sont construits automatiquement, à l'aide de règles.

En ce qui nous concerne, des phrases contenant toutes les unités acoustiques de l'arabe standard, nécessaires à notre système, ont été enregistrées par deux locuteurs différents (une femme et un homme), tous deux maîtrisant la langue et l'accent local. Ce choix a été considéré afin de pouvoir évaluer le type de locuteur qui se prête le mieux à ce genre de système.

II.3.4. Les opérations d'enregistrement

Pour notre système nous avons utilisé des fichiers WAV en format PCM échantillonné à 44,1 kHz en mode 16 bits et en stéréo soit à 172 k bits/s. Nous avons utilisé, lors de nos enregistrements, un matériel standard pour pouvoir juger de la dépendance matériel – qualité, mais aussi dans l'optique d'un système peu contraignant visant un maximum d'utilisateurs.

II.3.5. La segmentation et le Lissage

L'approche utilisée traditionnellement par la synthèse par concaténation de diphones est la segmentation manuelle, en plaçant des limites dans les régions de relative stabilité dans le milieu des phonèmes. La tâche de l'opérateur est, à partir d'un continuum de parole, d'isoler les sons en segments phonétiques et de leur associer un symbole [11], [15]. En ce qui suit un exemple de cheminement type pour l'obtention d'une unité acoustique donnée :

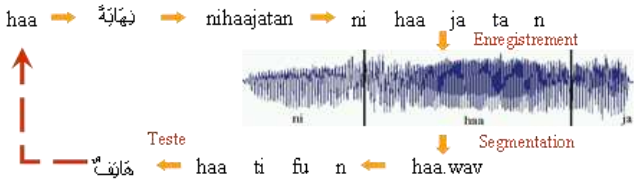


Figure 16 Un exemple de traitement pour l'obtention du triphone « haa » de l'identification au test en passant par l'enregistrement et la segmentation

La difficulté de cette tâche provient de la multitude de réalisations possibles des sons et de leur caractère transitoire, entraînant une grande difficulté à établir une classification des sons d'une langue. Le plus grand problème rencontré lors de la segmentation manuelle était l'énorme effort nécessaire, et le grand délai (de l'ordre de quelques mois) pour la segmentation d'une nouvelle base d'unités acoustiques. Un autre problème est la détermination exacte des limites qui peuvent se produire dans des régions de changement rapide de la parole. Assurer l'égalité au niveau des frontières lors de la concaténation devient alors problématique.

La méthode de segmentation que nous avons utilisée dans notre travail est une méthode de segmentation manuelle. Au cours de cette étape, l'identification des différentes unités s'est faite à travers l'utilisation de plusieurs outils parmi lesquels :

• La forme temporelle de l'onde acoustique correspondant à l'enregistrement.

- Le spectrogramme de l'enregistrement.
- L'audition, qui reste le critère de choix majeur pour la segmentation.

La concaténation de deux unités successives se fait dans un fichier résultat qui se verra ajouter une unité acoustique à chaque nouvelle étape (passage d'une syllabe à une autre), pour contenir à la fin du traitement la phrase à synthétiser [Gha04].

Pour obtenir une qualité de son comparable à celle d'un disque compact nous avons voulu utiliser le format Wav avec un échantillonnage à 44100 Hz et une résolution de 16 bits.

La simple concaténation d'unités de parole extraites de contextes différents ne produit en général pas une parole de bonne qualité. Deux types de solution se présentent actuellement: choisir les segments dont le contexte est le plus proche de la chaîne phonétique à synthétiser ou effectuer un traitement temporel ou spectral sur le résultat obtenu.

Pour notre système nous avons voulu commencer par un traitement temporel pour mesurer l'effet d'un post traitement sur la qualité de la parole obtenue. Un traitement spectral du type TD-PSOLA (Time-Domain Pitch-Synchronous Overlap and Add), MBROLA (Multiband Resynthesis OverLap-Add), WI (Waveform interpolation) ou autres.

Après l'analyse des différentes unités acoustiques de l'arabe, il s'avère que celles-ci présentent une atténuation aux niveaux de leurs extrémités. L'idée retenue consiste alors à procéder, lors de la concaténation, à une accentuation aux niveaux d'un certain nombre de valeurs d'extrémités avant le collage en bout à bout. Ce traitement touchera évidement la fin de la première unité et le début de la suivante.

Un signal numérique de la parole étant [Cal89] :

$$s(t) = \sum_1^N s_n \, \delta(t - nT) \qquad \textbf{(4)}$$

s(t) : signal numérisé de la parole (échantillonné).

$s_n = s(nT)$: la valeur du signal à l'instant nT.

$\delta(t)$: impulsion de Dirac.

La concaténation de deux unités sera :

$$s(t) = s_1(t) + s_2(t) = \sum_1^N s_{1n} \, \delta(t - nT) + \sum_1^M s_{2n} \, \delta(t - nT) \qquad \textbf{(5)}$$

L'idée consiste alors à isoler X valeurs du premier signal et Y valeurs du second :

$$s(t) = \sum_1^{N-X} s_{1n} \, \delta(t - nT) + \sum_{N-X+1}^{N} s_{1n} \, \delta(t - nT) + \sum_1^{Y} s_{2n} \, \delta(t - nT) + \sum_{Y+1}^{M-Y} s_{2n} \, \delta(t - nT) \qquad \textbf{(6)}$$

Ces valeurs subiront alors une atténuation proportionnelle définie par :

$$s_i^{atténué} = s_i \frac{K - i}{K} \quad i = 1..K \qquad \textbf{(7)}$$

Le résultat se présentera sous la forme :

$$s(t) = \sum_{1}^{N-X} s_{1n}\,\delta(t-nT) + \sum_{N-X+1}^{N} s_{1n}\,\frac{N-n}{N}\delta(t-nT) + \sum_{1}^{Y} s_{2n}\,\frac{Y-n}{Y}\delta(t-nT) + \sum_{Y+1}^{M-Y} s_{2n}\,\delta(t-nT) \quad \textbf{(8)}$$

La fonction d'atténuation ainsi définie a été appliquée pour un nombre de points représentants 10 % de la durée du signal de l'unité acoustique. Les sonagrammes suivants montrent les résultats obtenus:

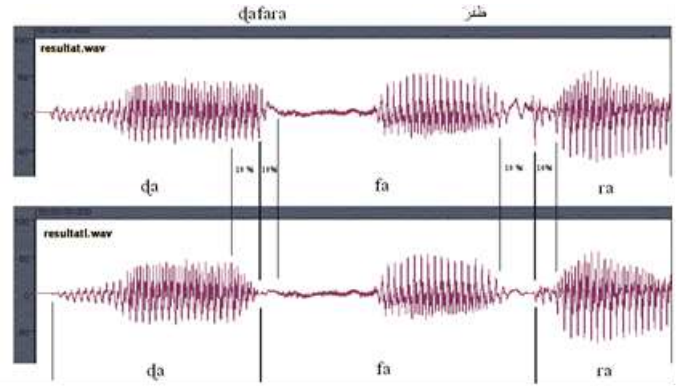


Figure 17 Effet du lissage temporel sur la forme d'onde au niveau des points de discontinuités.

Les courbes précédentes montrent l'effet de ce lissage temporel sur un exemple de synthèse du mot « ظَفَرَ » (il a gagné). En effet, la première courbe montre une concaténation bout à bout sans aucune intervention et nous constatons une discontinuité flagrante aux niveaux des points de jointures. La courbe du bas introduit, quant à elle, le résultat d'une concaténation lissée et la fluidité aux niveaux des points de concaténation.

Le résultat obtenu en introduisant une fonction de lissage a sensiblement amélioré la qualité de la voix synthétisée. Néanmoins, nous constatons un chevauchement entre les unités, ce qui influe sur le résultat. Pour éviter un tel problème nous avons introduit un temps de silence de 10 millièmes de seconde. Le résultat de cette opération est présenté en ce qui suit :

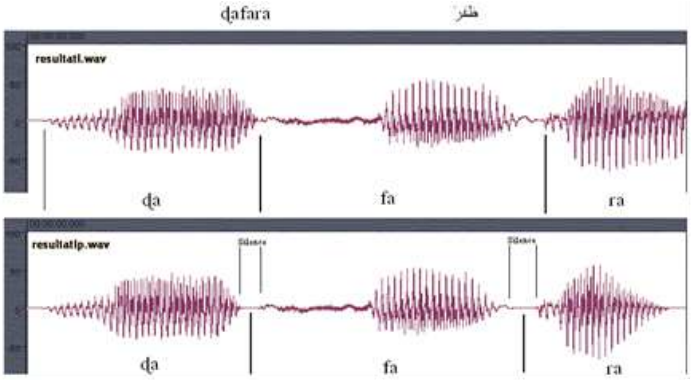


Figure 18 Introduction d'une pause au niveau des points de discontinuités.

La première courbe présente le résultat d'une concaténation avec lissage sans temps de pause entre les unités. L'effet de l'insertion d'une pause de 10 ms entre les unités avec

l'utilisation d'une atténuation est présenté sur la figure du bas. Cette intervention, nous a permis d'obtenir une meilleure intelligibilité

III.4 Discussion et évaluation:

Lors de l'élaboration de notre dictionnaire d'unités acoustiques, nous avons pu constater plusieurs difficultés et obstacles de nature à ralentir le travail et surtout à influer énormément sur la qualité de la voix synthétisée en aval. La majorité de ces contraintes survient lors de l'étape cruciale et hargneuse de segmentation [Zou93]. Parmi ces problèmes nous avons voulu citer les points suivants :

- L'unité acoustique à extraire doit être au milieu d'un mot, afin d'éviter les variations incontrôlées d'intonation du début et de la fin du mot.

- L'unité acoustique ne doit pas précéder un phonème pour éliminer toute ambiguïté lors de son identification au cours de la segmentation.

- Un diphone ne doit pas être précédé par un triphone, la prononciation de ce qui suit devient naturellement plus rapide.

- Des lettres comme (ف ه ح خ ض ظ ذ) sont prononcées au moyen d'une forte expiration, la qualité du microphone peut influer sur le résultat.

- Des lettres comme (غ) et (ر) posent des problèmes de naturel lors des essais de synthèse à cause de leur nature de prononciation.

- Des lettres comme (ف و م أ) se prononcent de manières différentes dans des positions différentes : عمارة_مسافة ; سفه_سفارة ; واضحة -وافق. Les deux prononciations sont à prendre en considération séparément (une unité pour chaque cas).

- La voix féminine est plus nette que celle du locuteur masculin, ce qui influe sur la qualité de la parole produite.

- La qualité de synthèse ne dépend pas que de la nature de la voix d'origine mais principalement de la qualité de la segmentation.

- Un prétraitement des enregistrements est souvent nécessaire : il faut notamment une normalisation du volume des unités.

- Le matériel peut affecter énormément la qualité de la parole synthétisée.

Afin d'évaluer notre système de synthèse de la parole, nous avons établi une procédure de test basée sur l'écoute et l'identification de phrases synthétisées. Pour ce faire nous avons utilisé un corpus de référence [Bou93]. Ce corpus est un ensemble de vingt listes de dix phrases arabes phonétiquement équilibrées chacune. De ce corpus nous avons extrait 20 phrases, soit 53 mots, 211 unités acoustiques dont 73 différentes ce qui constitue 37.2 % de la totalité des unités acoustiques qu'utilise notre système. Nous les avons fait écouter à 8 personnes (4 femmes et 4 hommes) ce qui a permis une évaluation statistique réaliste du résultat. Chaque phrase est écoutée trois fois, à chaque passage le sujet doit orthographier ce qu'il entend. En ce qui suit le résumé de ces résultas [Sai04] :

Les candidats ont été conviés à orthographier tout ce qu'ils arrivaient à comprendre spontanément et sans efforts. Le calcul des valeurs mentionnées a été fait sur le pourcentage d'identification des mots dans une phrase donnée. Pour une meilleure utilisation de ces résultats, nous les avons présentés sous la forme de trois courbes correspondant aux trois phases du test:

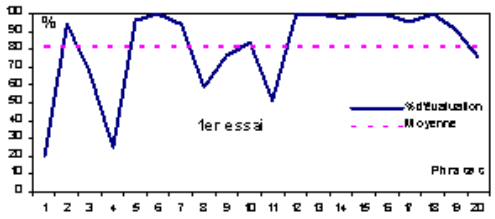


Figure 19 Les résultats de la 1ère phase de test

Lors de la première phase nous avons noté une moyenne d'identification de plus de 81 %. Nous constatons une meilleure compréhension vers les dernières phrases, ce qui laisse penser qu'une phase d'adaptation a été nécessaire. Le résultat de la 2ème phase est représenté dans le graphique suivant :

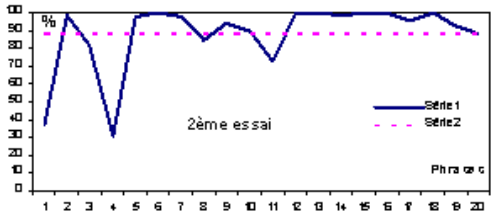


Figure 20 : Les résultats de la 2ème phase de test

La deuxième phase présente une moyenne d'identification de plus de 88 %. Néanmoins, un problème de compréhension est persistant au niveau de la 11ème phrase سَيُؤْذِيهِمْ زَمَانُنَا (notre époque leurs fera du mal). Nous avons expliqué ce fait par un sens peu commun de la phrase ainsi que par une mauvaise qualité de synthèse du ذ. La courbe suivante présente les résultats de la 3ème phase :

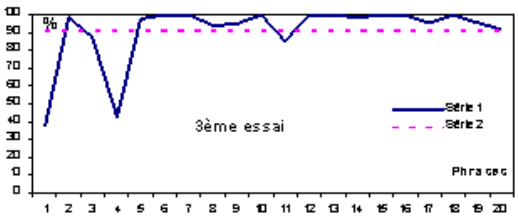


Figure 21 Les résultats de la 3ème phase de test

Nous avons alors pu conclure à un pourcentage d'identification de plus de 80 % dès la première écoute, ce taux passe à plus de 91% pour la troisième phase. Par ailleurs nous avons remarqué qu'une phase d'adaptation de 2 à 3 phrases a été nécessaire pour avoir une stabilisation des taux de reconnaissance. De ces relevés nous avons aussi constaté que :

• Les mots non courants sont difficilement identifiables (exp : لَذَعْتَهُ),

• Quelques consonnes sont plus difficiles que d'autres pour l'identification comme le : ذ

IV. ANALYSE SEMANTIQUE DE L'ORAL: (CAS DE L'ARABE)

Les travaux sur la traduction automatique de l'oral ont fait apparaître le besoin d'entamer nos recherches sur deux axes: la compréhension de et la reconnaissance de la parole. Ces deux axes de recherche ont commencé avec les travaux de DEA des étudiants Walid JAOUADI et Rami AYADI à l'ENSI(Tunis) puis en thèse sur l'analyse sémantique de l'arabe (cas de l'oral) avec Anis ZOUAGHI à l'ENSI (date prévue de soutenance fin 2007) et Mohamed BELGACEM en thèse au LIDILEM à Grenoble en Co-encadrement avec Mr. Georges ANTONIADIS

Contrairement aux systèmes de compréhension guidés par une analyse syntaxique détaillée (par exemple [Jain et al. 92], [Spriet et al. 97], [Lopez et al. 98] et [Goulian et al. 01]), ou aux systèmes guidés essentiellement par la sémantique (par exemple [Pieraccini et al. 95], [Minker 99], [Hacioglu et al. 01] et [Bousquet 02a]), nous proposons un système de compréhension guidé en même temps par une analyse syntaxique et sémantique. Nous adoptons ainsi, le même point de vue que [Seneff 92], [Antoine 94] et [Pepelnjak 95], où la compréhension est le fruit d'une coopération entre la syntaxe et la sémantique.

Cette approche a l'avantage d'être en même temps robuste aux difficultés de l'oral spontané: caractéristique des approches sémantiques et performante face à des énoncés ayant une structure syntaxique complexe: caractéristique des approches syntaxiques. Dans le paragraphe suivant nous exposons les sources de difficultés de la compréhension de la parole.

III.1. Problématique de la compréhension de la parole :

La compréhension automatique de la langue naturelle écrite ou parlée est une tâche très difficile. Les principales difficultés auxquelles est confronté un système de compréhension sont dues essentiellement:

- à la caractéristique des langues naturelles: l'utilisation de prédicats ambigus, de mots polysémiques...
- à la particularité de l'oral spontané: possibilités d'avoir des hésitations, des autocorrections, des phrases n'ayant aucune structure syntaxique juste…
- aux performances du module de reconnaissance ou à la présence de mots inconnus du système.

IV.1.1 Les difficultés d'ordre linguistique

Les difficultés d'ordre linguistiques sont communes à la compréhension de l'écrit et de l'oral. Ces difficultés sont généralement dues à l'utilisation des:

- références : qui sont des mots qui font référence à d'autres mots ou groupes de mots présents dans le discours, tels que:

  Les anaphores: هي - هو - etc.
  Les démonstratifs: هنا - هذه - etc.
  Les déictiques de localisations spatiales: هنا - هنالك - etc.
  Les déictiques de localisations temporelles: غدا - البارحة - etc.

Ces références permettent à l'énonciateur de simplifier et minimiser la quantité d'informations pour construire son énoncé. Dans de tels énoncés, il est très important de tenir compte des données contextuelles, pour lever les ambiguïtés.
Par exemple dans l'énoncé (1) suivant:

(1) غدا سوف أسافر إلى فرنسا (Demain, je voyage à la France.)

Dans l'énoncé (1), le déictique temporel غدا (Demain) ne peut pas être interprété correctement, si on ne connaît pas le moment de l'énonciation.
[Sabah et al. 89], [Gaiffe et al. 92 et 00] et [Salmon-Alt 01] traitent ce genre de problèmes (traitement des énoncés référentiels) dans le dialogue homme-machine.

- mots polysémiques: qui sont des mots qui ont plusieurs sens possibles. Par exemple, la lexie طار (voler) dans l'énoncé métaphorique (3), ne doit pas être interprétée de la même manière que dans l'énoncé (2). En effet dans l'énoncé (2) طار est l'action de voler dans l'air, alors que dans l'énoncé (3), ce même mot permet d'exprimer le degré (sentiment fort) de joie éprouvé par la personne (voir [Ferrari 97] en ce qui concerne le traitement automatique des métaphores, qui cherche à les localiser à partir de marqueurs linguistiques dans des documents écrits).

(2) طار العصفور (l'oiseau a volé).
(3) طار من شدة الفرحة (il a sauté de joie).

Le problème de la polysémie engendre bien la différence entre les langues naturelles et les langages formels, où un prédicat logique ne peut avoir qu'une seule signification. On ne peut donc pas calculer le sens d'un énoncé en langue naturelle, comme ce qu'on fait avec un programme à l'aide d'un compilateur. Si on cherche à déterminer le sens d'un énoncé, en composant l'ensemble des significations possibles pour chaque mot le long de l'arbre résultant de l'analyse syntaxique, il y a un risque d'une explosion combinatoire. Ces mêmes problèmes de combinatoire sont rencontrés avec les modèles classiques, basés essentiellement sur la logique [Morris 39] et [Van Dijk 77].

- prédicats linguistiques vagues et floues: qui sont des mots imprécis et vagues, tel que : مسن - كبير - صغير - يمكن.. Ce caractère vague et flou de la langue a emmené à la naissance de la théorie de la croyance et des possibilités et aux logiques modales.

En plus dans les communications finalisées, les messages sont souvent implicites. Dans ce cas seuls les inférences permettent d'interpréter et de comprendre l'implicite de l'énoncé. [Sabah 89] expose les différentes méthodes utilisées en intelligence artificielle pour faire des inférences. Les scripts [Schank Abelson 77] sont souvent utilisés pour la reconstitution de l'implicite.

Une autre difficulté s'ajoute à la compréhension automatique de la langue naturelle d'une façon générale et de la langue arabe en particulier, est la complexité structurelle de certains énoncés.

IV.1.2 Les difficultés de l'oral spontané

Une des sources de difficulté en compréhension de la parole est la particularité de l'oral spontané par rapport à l'écrit. [Biber 86] essaye de classifier les différents modes d'expression suivant leur degré d'explicité et d'interactivité. Cette classification est représentée par la figure deux suivante:

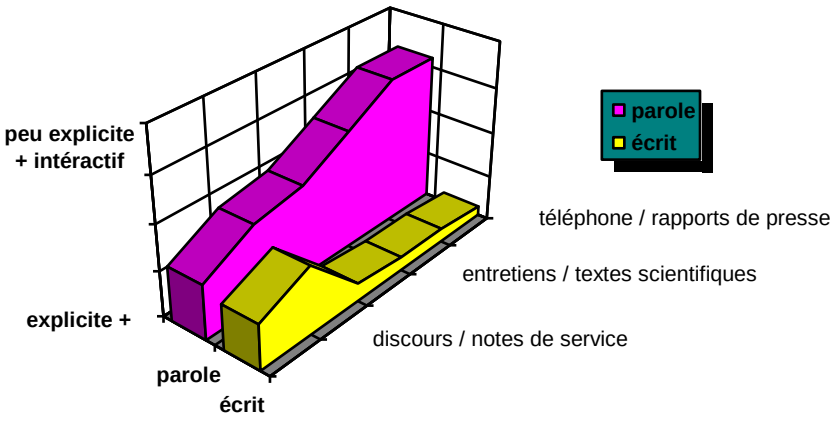


Figure 22 Classification des différents modes d'expression suivant leur degré d'explicité et d'interactivité.

En effet les énoncés transcrits par le module de reconnaissance, possèdent souvent une:

- une structure syntaxique qui ne respecte pas les règles de la grammaire, tel que les fautes d'accords.
- en plus, ils peuvent contenir des hésitations, des autocorrections, des répétitions de mots. On notera une grande utilisation de termes d'appui du discours, tels que إذا (*alors)* ou نعم (oui). Suite à une étude de corpus de discours, nous avons remarqué que l'utilisation de ces marqueurs de structuration est très variable suivant les locuteurs. Certains individus y font recours très fréquemment à ces marqueurs, alors que d'autres ne les utilisent presque pas.

Par exemple, à la suite d'une question ou d'une suggestion la plupart des locuteurs utilisent le terme "hein", comme dans l'énoncé (4) suivant:

(4) hein لما لا تذهب (Pourquoi tu ne vas pas, hein?)

[Kurdi 00] par exemple fait appel à des prétraitements pour corriger automatiquement ce genre d'énoncés.

IV.1.3 Les performances du module de reconnaissance

Les modules de la reconnaissance de la parole spontanée ne sont pas encore parfaitement puissants. Ils font encore beaucoup d'erreurs. Ces erreurs sont dues souvent à:

- un environnement bruité.
- des accents particuliers: les locuteurs n'ont pas tous le même accent.
- une couverture lexicale insuffisante: emploi de mots inconnus du lexique du module de reconnaissance et par conséquence du module de compréhension.

Le problème de traitement des mots mal reconnu est exposé dans [Bousquet 02b].

IV.1.4 Conclusion

La compréhension de la parole confronte ainsi de nombreuses et diverses difficultés. Certaines difficultés sont communes à l'écrit et à l'oral, et qui sont des problèmes d'ordre linguistique. Et d'autres sont dues aux propriétés de l'oral spontané et aux erreurs dues au module de reconnaissance. La figure trois résume les sources de difficultés que confronte la compréhension de la parole.

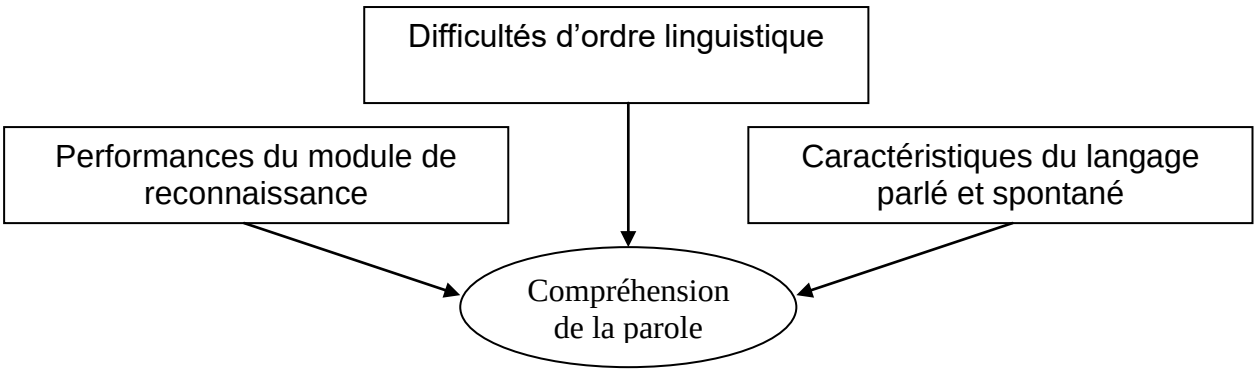


Figure 23 Les sources de difficultés de la compréhension de la parole.

Les systèmes de communication vocale touchent plusieurs domaines d'applications. Dans le cas des communications homme/machine, un grand nombre d'applications concerne actuellement les serveurs vocaux et les bornes d'information interactives, permettant par exemple de fournir les horaires d'avions [Bou94] ou de trains [Ben02]. Les systèmes de communication homme/homme concernent les systèmes de traduction automatique permettant à deux interlocuteurs de langues différentes de communiquer.

Par exemple le système JANUS [wai96] permet de traduire de l'anglais, de l'allemand ou de l'espagnol vers l'allemand, l'anglais, le chinois, le coréen, l'espagnol ou le japonais.

Quant à la langue arabe, il n'existe que des systèmes de traduction de l'arabe écrit mais pas de l'oral, tels que par exemple le système AJEEB qui permet de traduire l'arabe vers l'anglai, et le moteur de traduction de SAKHR qui permet de naviguer sur Internet en utilisant la langue arabe sans avoir besoin d'un pré requis préalable en langue anglaise. Le système TORJMANE [Lab94] développé à l'institut de recherche des sciences informatique

Et des télécommunications (IRSIT) en Tunisie, permet aussi une traduction assistée par ordinateur d'une phrase de l'anglais vers l'arabe, ainsi que de proposer un ensemble de traductions possibles. Le système ALMISBAR est aussi un traducteur de l'anglais vers l'arabe.

Les systèmes de communication vocale sont aussi utilisés dans d'autres types d'applications tels que pour commander une chaîne HIFI [Fer98] ou pour interagir avec un robot [Cou92]. Une liste plus exhaustive des différents projets et applications de recherche concernant les technologies de reconnaissance et de compréhension de la parole est détaillée dans [Min99] .

Nous signalons toutefois qu'il n'existe pas encore à nos jours des conversations en langage naturel libre. La plupart des systèmes de communication vocale actuels sont conçus pour des des taches relativement limitées. Ce sont les connaissances du domaine et de la tache qui permettent de résoudre les difficultés rencontrées lors de la réalisation d'un dialogue.

IV.2 Etat de l'art des systèmes de compréhension :

IV.2.1 Introduction: les approches utilisées

On distingue généralement en compréhension automatique de la parole [Luzzati 89] et [De Mori 94] deux approches :

- la 1ère approche: utilisation pour le traitement de l'oral des mêmes méthodes d'analyses développées en traitement automatique des langues naturelles. Les systèmes utilisant cette approche commencent alors par construire un arbre syntagmatique à la suite d'une analyse syntaxique. A partir de cet arbre est extrait dans un deuxième temps le sens des énoncés à partir d'une autre analyse. Nous pouvons comme titre d'exemples, le système *Parsec* développé [Jain et al. 92] ou le modèle [Spriet et al. 97], etc.

- la 2ème approche: extraction directement du sens sans ou (avec peu) de considération syntaxique, afin de construire une structure sémantique de représentation du sens de l'énoncé ou du message, tels que le système *Phoenix* de [Ward 91], le système *Philips* de renseignement sur les horaires de train de [Aust 95], le système *CACAO* de [Bousquet et al. 99], ainsi que les systèmes de [Minker 99], [Hacioglu et al. 01], etc.

La question qui se pose alors, quelle est l'approche la plus adéquate pour la compréhension automatique de l'oral spontané. En tenant compte de la caractéristique fondamentale de l'oral spontané par rapport à l'écrit, qui généralement ne respecte la grammaire, nous avons opté pour la deuxième approche qualifiée de sémantique globale par De Mori et sélective par Luzzati.

On peut aussi classer les systèmes de compréhension de la parole, selon si l'approche adoptée est stochastique ou non [Bousquet 02a].

Dans ce qui suit, nous allons présenter quelques approches sémantiques utilisées pour l'interprétation, afin d'introduire notre modèle qui est plutôt basé sur une approche sémantique
componentielle et différentielle. [De Mori 99] donne un état de l'art des formalismes de représentation des connaissances linguistiques d'une façon générale.

IV.2.2 Les divers formalismes et sémantiques de représentation du sens

La définition de qu'est ce que le sens ?, ou que voulons dire par le mot sens ?, est une chose qui n'est pas facile du tout. Cette difficulté explique l'existence d'un grand nombre d'approches et de paradigmes essayant de formaliser cette notion. La sémantique s'intéresse à l'étude du sens.

Sabah par exemple distingue les sémantiques suivantes :

- la sémantique véri-conditionnelle : permet de préciser les conditions de vérité de l'expression traitée. Elle permet de donner une description formelle des situations dans lesquelles l'expression peut être considérée comme vraie.

- la sémantique intensionnelle : donne une description du sens d'une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants.

- la sémantique extensionnelle : donne une description d'une expression comme l'ensemble des objets ou des situations du monde de référence que cette expression peut désigner.

- la sémantique componentielle : ne considère pas forcément les mots comme des entités primitives et cherche à les décomposer en éléments de sens plus primitifs.

- la sémantique procédurale : donne une description du sens d'une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné.

- la sémantique argumentative : cherche à dépasser la description d'actes de langage isolés pour étudier les enchaînements d'actes dans le discours et les connecteurs qui marquent ces enchaînements.

La compréhension automatique des énoncés est constituée en faite de deux étapes :

- La première étape: correspond à la phase d'apprentissage, c'est-à-dire la construction de la structure de représentation de sens notée SRS, sur laquelle se base après le processus de construction de sens noté PCS.
- La deuxième étape: est l'interprétation (PCS) proprement dite des énoncés reconnus par le module ″Reconnaissance″, et la résolution des ambiguïtés d'ordre linguistique tel que la présence de mots polysémiques, des anaphores, des déictiques et des problèmes dues aux erreurs de reconnaissance de la parole.

C'est au niveau de la première étape du processus de compréhension qu'intervient implicitement l'analyse structurelle. La sémantique lexicale décrit le sens littéral des mots, elle permet la structuration du lexique sémantique de l'application. Elle est alors la première étape fondamentale du processus de compréhension. Il existe plusieurs paradigmes proposant une description de cette connaissance (sémantique lexicale), dont on peut citer :

- La sémantique du prototype.
- Les réseaux sémantiques.
- La sémantique componentielle différentielle.
- La sémantique componentielle référentielle.

Dans ce qui suit, nous nous focaliserons plutôt sur la description des réseaux sémantiques et des sémantiques componentielles, dont notre modèle est inspiré de cette dernière approche.

IV.2.2.1. Les réseaux sémantiques

Les réseaux sémantiques ou graphes étiquetés entre concepts tirent leur origine de modélisation des expériences psychologiques de Collins et Quillian sur la mémoire associative [Quillian 68]. Ces réseaux permettent de mettre en évidence les liens sémantiques suivants:

- le lien taxonomique traduit par la relation *Est-Un.*
- les lien traduit par la relation *Partie*-D*e.*

- des relations de causalité C*ause-De*

- ainsi que des relations d'antinomie, de synonymie, etc.

Selon ce modèle, calculer le sens d'un mot revient à déterminer les relations sémantiques que possède ce mot avec les autres mots. Ainsi la structure sémantique d'un énoncé est, elle aussi, représentée sous la forme d'un réseau sémantique. Cette structure est obtenue en faite, à la suite du parcours du réseau initial.

Malheureusement, ces réseaux représentent plusieurs faiblesses, malgré le grand succès qu'ils ont connu en Intelligence Artificielle tel que graphes conceptuels de [Sowa 84].

Le principal reproche adressé aux réseaux sémantiques est qu'ils ne font pas distinction entre concepts et signifiés lexicaux [Rastier 87].

Dans la pratique, cette approche entraîne des problèmes de polysémie. Chaque mot peut correspondre à un plusieurs concepts, et le choix du bon nœud sémantique sera difficile. D'ailleurs c'est le même problème qu'affronte la logique des prédicats.

IV.2.2.2. Sémantique componentielle référentielle

Katz et Fodor [Katz 63] sont parmi les premiers à avoir développer ce paradigme. Ils reprennent le modèle aristotélicien, suivant lequel le sens d'un lexème est présenté par un ensemble d'unités élémentaires de sens appelées de diverses façons selon les disciplines:

- traits sémantiques.

- primitives.

- atomes de sens.

- marqueurs.

- sèmes, etc.

Katz et Fodor proposent pour décrire les deux significations différentes du mot polysémique "balle", les sèmes suivants:

(1) balle = /objet physique/ + /forme sphérique/
(2) balle = /objet physique/ + /projectile solide lancé par un appareil militaire/

L'idée de base est donc de définir a priori un ensemble de traits et d'indiquer dans le dictionnaire, pour chaque sens d'un mot, le sous-ensemble des traits présents. Ce type d'analyse fut systématisé principalement par [Greimas 66], [Katz 72], [Jackendoff 75], [Miller et al. 76], et [LeNy 79].

Le problème majeur de cette approche est la définition de ces traits primitifs. En plus, [Pitrat 85] indique qu'en s'intéressant à des textes de taille courte, le nombre de ces primitives est presque égal au nombre des mots de la langue. A l'opposé, la sémantique componentielle différentielle propose un critère opératoire pour la définition des traits sémantiques.

IV.2.2.3 Sémantique componentielle différentielle

La sémantique componentielle différentielle utilise aussi la notion de traits sémantiques pour la description de la signification des mots. Mais selon ce paradigme la finalité des sèmes est de permettre à distinguer entre l'ensemble des éléments du lexique.

[Cavazza 91, p 68] représente le lexique sémantique par un arbre taxonomique structuré par une hiérarchie de sèmes. La figure quatre présente un petit échantillon de cet arbre taxonomique.

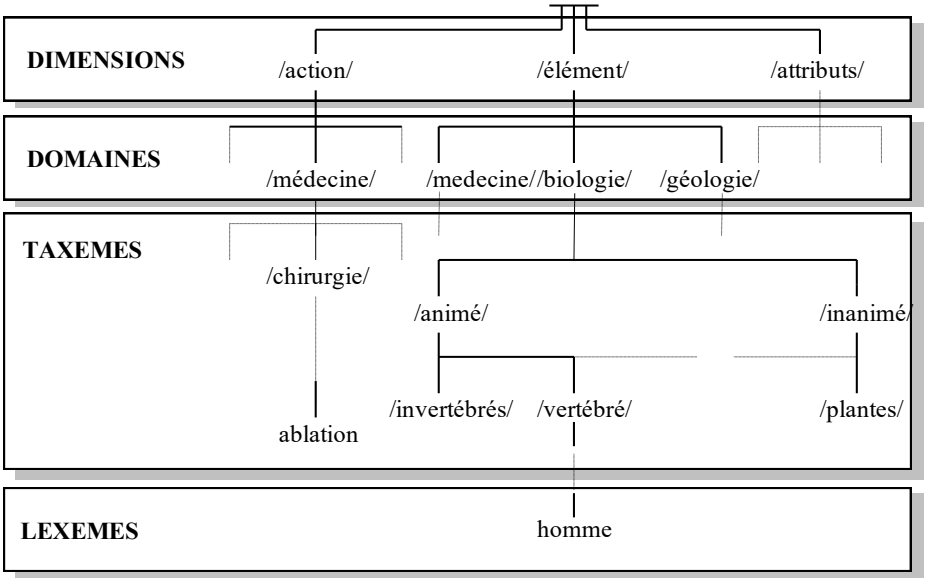


Figure 24 Structuration du lexique selon une approche componentielle différentielle.

On remarque qu'il existe selon cet arbre quatre niveaux différents:

- Le premier niveau : représente les dimensions et ça correspond aux sèmes macrogénériques.
- Le deuxième niveau : représente les domaines et ils permettent une partition du lexique sémantique en sous lexiques de spécialité. Des sèmes mésogénériques sont utilisés pour la description de la signification dans ce niveau.
- Le troisième niveau: représente les taxèmes qui permettent de regrouper les lexèmes dans des classes partageant un même champ de signification et se trouvant en opposition sémantique directe.
- Et enfin le quatrième niveau: représente les lexèmes qu'on cherche à déterminer leur signification. Nous avons alors des lexèmes bien structués.

Cette classification est très répandue, on la retrouve par exemple dans [Schank 72], [Andry 92] ainsi que dans le système *MICRO* de [Antoine 92]. Nous, nous avons adopté ce paradigme pour la première phase du processus de compréhension, c'est-à-dire la phase de structuration du lexique sémantique.

IV.3 Présentation des principaux systèmes de compréhension

Après une étude de l'état de l'art des systèmes de compréhension, nous avons remarqué qu'il existe de nombreux systèmes de compréhension de la parole. Ces systèmes utilisent des

formalismes de natures différentes pour la détermination de sens des énoncés. Parmi les formalismes utilisés, nous trouvons:

- les grammaires hors-contexte: tels que [Gavignet 92], [Seneff 92b], [Aust 95], [Mayfield 95b], [Kompe 97], [Nöth 99], [Wang 99], [Carpenter 00], [Yan 01], etc.
- les grammaires de cas: tels que [Haton 91], [Bennacef 94], [Minker 99], etc.
- les modèles de Markov cachés: tels que [Pieraccini 95], [Minker 99], [Macherey 01], etc.
- les réseaux de neurones: tels que [Jain 92], [Stévenin 95], [Jamousi 01], etc.
- les modèles de langage N-grammes: tels que [Pepelnjak 95], [Knight 01], etc.
- le l-calcul: tels que [Villaneau 01], etc
- les logiques: tels que [Sadek 95], etc.
- les grammaires d'unifications sous ses diverses formes: tels que [Eckert 94], [Brondsted 98], [Lopez 98], [Kurdi 01], etc.

Cependant, les approches stochastiques sont les plus utilisées pour des systèmes guidés par la sémantique. Selon une analyse faite par [Bousquet 02], la représentation des connaissances est essentiellement réalisée à l'aide de grammaires hors contexte probabilisées, de modèles de Markov cachés, plus rarement avec des réseaux neuronaux, en ce qui concerne les approches stochastiques. Pour les approches non stochastiques de divers formalismes sont utilisés, mais le plus courant est la grammaire hors contexte.

En ce qui concerne le traitement des mots inutiles dans l'énoncé, qui sont dus comme par exemple aux hésitations et aux répétitions de différentes approches sont utilisées. Certaines utilisent la notion de classes de mots ou de concepts de rebut pour récupérer ces mots inutiles tels que [Aust 95], [Fereiros 98], [Boros 99] et [Hacioglu 01]. D'autres consistent à dépasser les mots inconnus à l'aide de règles spécifiques dans des grammaires hors contexte tels que [Mayfield 95] ou [Wang 99].

En tenant compte que notre modèle est guidé plutôt par la sémantique, nous décrivons dans ce qui suit, quelques systèmes existant et utilisant cette théorie.

IV.3.1 Le système TINA

Le système *TINA* [Seneff 92], utilise en faite une approche hybride. Selon cette approche, le système est analysé syntaxiquement, si cette analyse est échouée, alors le système procède à une analyse sémantique.
Pour l'analyse syntaxique Seneff utilise une grammaire hors contexte transformée de façon automatique en un automate portant des probabilités sur les arcs. Ces arcs "probabilisés" permettent de donner priorité aux constructions les plus courantes.
Si la première analyse (syntaxique) ne permet pas d'extraire le sens de l'énoncé, une deuxième analyse est utilisée. Dans ce cas, on ne tient compte que des segments porteurs de sens. Si plusieurs solutions sont possibles, c'est celle qui a consommé le plus de mots qui est retenue.

IV.3 .2 Le système PHOENIX

Le système *PHOENIX* utilise une approche conceptuelle, et utilise à la fois le module de reconnaissance de la parole et celui de compréhension pour la compréhension automatique de la parole.

Dans ce système, la représentation sémantique est décrite sous la forme d'un schéma (frame). Et les connaissances linguistiques sont représentées par une grammaire hors contexte partielle non stochastique et guidée essentiellement par la sémantique. Chaque paire attribut / valeur du Frame est représentée par un automate à états finis. Ces automates permettent d'indiquer toutes les manières de dire une séquence de mot dont le sens correspond à cette paire. Un automate peut en appeler un autre.
Les mots ne correspondant à aucune paire ne sont pas interprétés, ceci permet de résoudre les ambiguïtés dues à l'oral spontané. Le principe d'interprétation consiste à détecter les segments correspondant aux paires attribut / valeur et à en déduire le schéma. Si plusieurs solutions sont possibles, on retient alors celle qui a le meilleur score, ce score étant calculé en fonction du nombre de mots de l'énoncé qui ont été pris en compte pour l'interprétation.

IV.3.3 Le système CHRONUS

Le système *CHRONUS* (Conceptual Hidden Representation of Natural Unconstrained Speech) de [Pieraccini 95] est basé sur une approche stochastique et conceptuelle. Il utilise une grammaire sémantique de concepts, pour être robuste face aux problèmes de l'oral spontané. Un concept est une unité élémentaire de sens. Dans ce système, on suppose que tout énoncé peut être représenté par une suite de concepts. Le processus de compréhension se décompose en deux modules correspondant au découpage classique de compréhension littérale et compréhension contextuelle. Le processus de compréhension littérale, appelé analyse locale, est réalisé en trois étapes (voir figure 5). L'étape de compréhension contextuelle consiste en une analyse globale de l'énoncé, et est réalisée avec le module appelé interpréteur. Ce module permet de résoudre les ambiguïtés à l'aide de règles écrites manuellement. Par exemple, seule la dernière ville est considérée, dans le cas où il existe plusieurs villes de départ possibles.

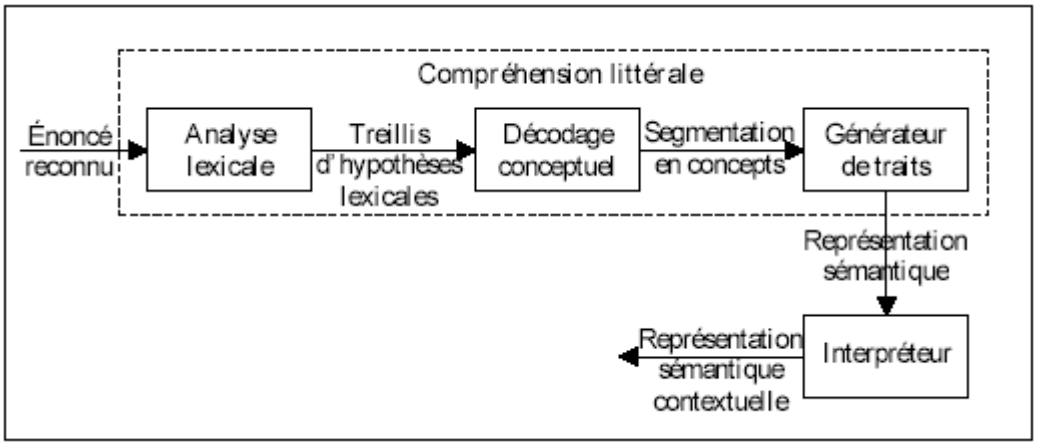


Figure 25 Architecture du système CHRONUS.

IV.3.4 Le système de Philips

Le système *Philips* [Aust et al. 95] est guidé plutôt par la sémantique. Il est basé sur une grammaire hors contexte. Le sens de l'énoncé dans ce système de compréhension est exprimé en fonction de séquences de mots types appelés concepts.
Le processus de compréhension prend en entrée le graphe de mots proposé par le module de reconnaissance de la parole. La compréhension se déroule en cinq étapes (voir Figure 5):

- 1ère étape: prétraitement du graphe de mots consistant à construire un chemin dans le graphe permettant de sauter les mots inutiles.

- 2ème étape: ressemble au décodage conceptuel du système CHRONUS, elle cherche à construire le graphe de concepts à partir du graphe de mots. La grammaire utilisée est un ensemble de grammaires de concepts dont les règles sont probabilisées et attribuées. À chaque concept est associé le sens correspondant sous la forme d'une paire attribut / valeur.
- 3ème étape: complète la construction du graphe de concepts avec des arcs dits filler correspondant aux mots inutiles. Cette étape permet de construire un chemin du premier au dernier nœud du graphe.
- 4ème étape: permet de déterminer le meilleur chemin, en déterminant le chemin le plus probable graphe de concept.
- 5ème étape: permet de désambiguïser le sens de l'énoncé lorsqu'il y a par exemple des autocorrection de la part de l'usager ou des répétitions. Cette étape est réalisée à l'aide de règles de combinaison.

La figure six, suivante permet d'illustrer le résultat, suite aux cinq étapes du processus de compréhension utilisées dans le système *Philips* [Aust et al. 95].

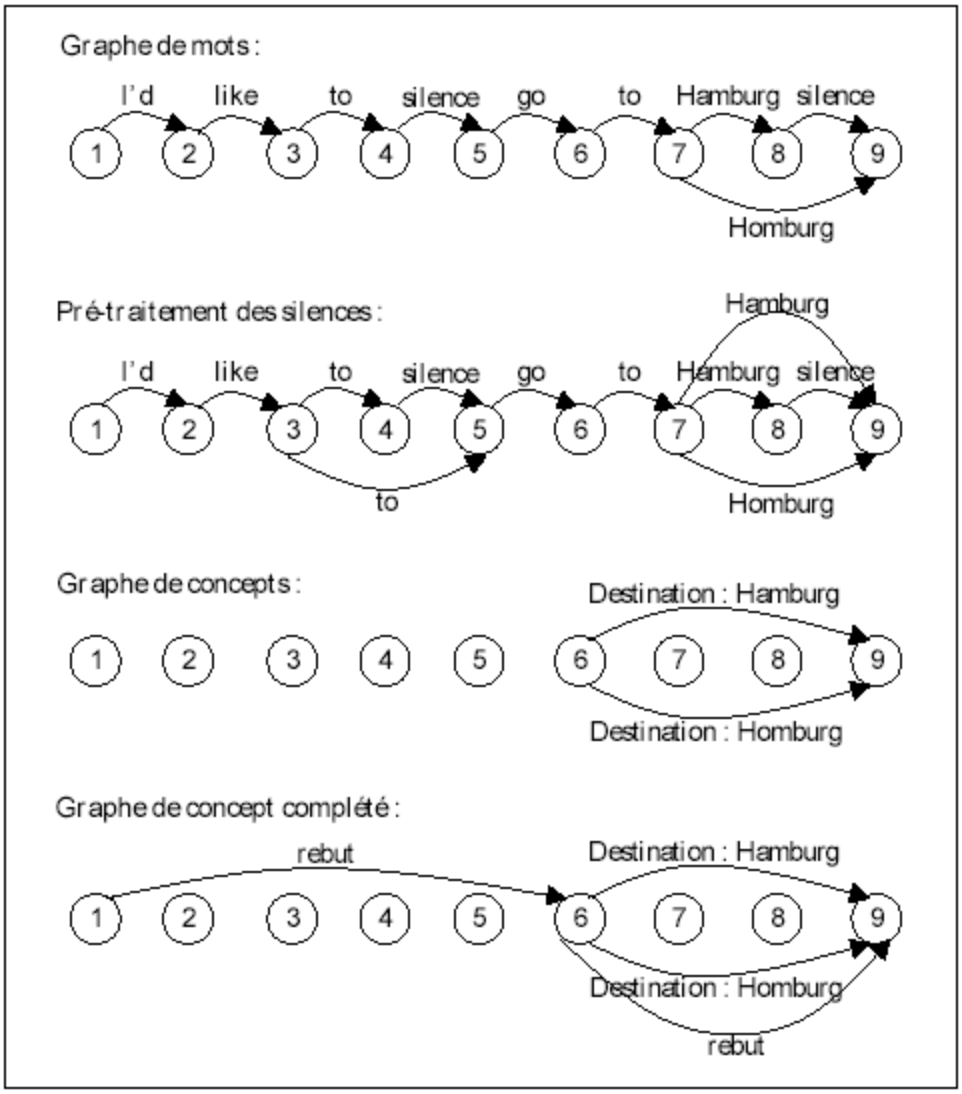


Figure 26 Résultat du processus de compréhension du système Philips.

IV.3.5 Le système CACAO

Le système de compréhension *CACAO* [Bousquet 02a] est guidé par la sémantique et il est basé sur une modélisation stochastique et conceptuelle. Le principe d'interprétation d'un énoncé selon ce modèle revient à le découper en segments conceptuels et en classes de mots. Généralement un segment conceptuel correspond à un concept, où un concept est défini comme une représentation mentale générale et abstraite d'un objet, et est indépendant de la langue. Et chaque segment conceptuel regroupe les séquences de classes de mots exprimant une même unité de sens.

On distingue trois types de segments conceptuels :

- Les segments conceptuels illocutoires: ils correspondent aux actes de langage illocutoires lorsque ceux-ci sont explicitement prononcés par le locuteur. Par exemple dans: *Je voudrais partir pour Toulouse*, le groupe de mots *je voudrais* de cette énoncé est un segment conceptuel illocutoire de demande.
- Les segments conceptuels référentiels: ils font référence aux informations en relation avec le domaine de l'application. Par exemple dans le cadre d'une application sur les horaires de trains, ils expriment les notions de villes de départ et de destination, de date, d'horaire, etc.
- Les segments conceptuels de rebut: ils regroupent les mots et les groupes de mots considérés comme inutiles pour la compréhension de l'énoncé, tels que les hésitations, etc.

Deux formalismes de représentation sémantique sont utilisés, afin de pouvoir choisir la plus appropriée par rapport à la complexité de l'application. Ces deux formalismes sont les structures de traits et les ensembles de paires attribut / valeur. La figure sept, donne un exemple de la représentation sous la forme d'une structure de traits, et la figure huit donne une description sémantique sous forme de paires attribut / valeur.

L'énoncé à interpréter est le suivant:
Je veux un train partant de Toulouse lundi à 10h et arrivant à Paris à 18h.

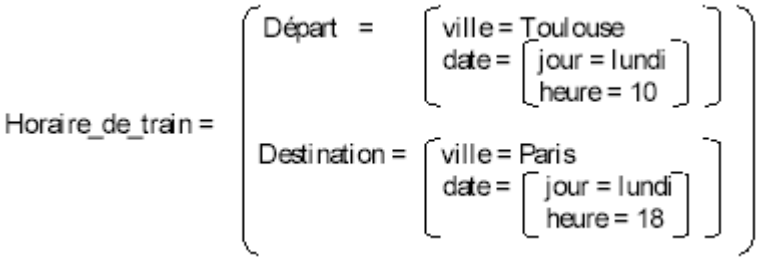


Figure 7: Représentation sémantique sous forme d'une structure de traits du Système *CACAO*.

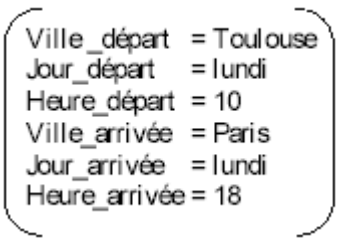


Figure 8: Représentation sémantique sous forme de paires attribut / valeur.

Conclusion :

On distingue généralement en compréhension automatique de la parole [Luz89] et [Mor94] deux types d'approche:

* les approches basées sur une analyse syntaxique profonde.
* les approches sélectives.

On peut aussi classer les systèmes de compréhension de la parole, selon que l'approche adoptée est stochastique ou non [Bou02].

Les approches basées sur une analyse syntaxique profonde sont caractérisées par l'utilisation pour le traitement de l'oral des mêmes méthodes d'analyses développées en traitement automatique des langues naturelles. Les systèmes utilisant ce type d'approche commencent par construire un arbre syntagmatique à la suite d'une analyse syntaxique. A partir de cet arbre est extrait dans un deuxième temps le sens des énoncés par une analyse adaptée. Nous donnons à titre d'exemples, le système PARSEC développé par [Jai92] ou le modèle de Spriet [Spr97

Les approches sélectives consistent à profiter du caractère finalisé du dialogue pour développer des approches orientées par la tache. Ces approches se basent sur l'extraction directe du sens sans ou (avec peu) de considération syntaxique, afin de construire la structure de représentation sémantique de l'énoncé ou du message. Nous citons comme exemples, le système PHOENIX [War91], le système PHILIPS de renseignement sur les horaires de train [Aus95], le système CACAO [Bou99], ainsi que les systèmes [Min99] [Hac01].

IV.4 Approche adoptée

La question qui se pose alors: quelle est l'approche la plus adéquate pour la compréhension automatique de l'oral spontané? Sachant que l'oral spontané, par rapport à l'écrit, ne respecte pas la grammaire, et puisque nous nous intéressons aux communications finalisées, nous avons opté pour le deuxième type d'approches. Par ailleurs, certains utilisent des approches hybrides combinant une analyse syntaxique profonde avec une analyse sélective tel que par exemple le système TINA de [Sen92].

Le système de compréhension automatique de la parole arabe spontanée a pour objectif principal de contribuer à l'interprétation littérale des énoncés oraux en langue arabe reconnus par le module de reconnaissance de la parole ou de leurs constituants en contexte. On se limitera à l'analyse sémantique, les phénomènes liés à la pragmatique telles que la métaphore, l'ellipse ou les aspects de conversation ne seront pas traités. Ils feront l'objet de nos prochains travaux

La figure 16, ci-dessous présente par exemple le résultat de l'interprétation sémantique de l'énoncé:

نهاركم سعيد هل يمكن حجز ثلاث تذاكر درجة اولى في القطار الذي ينطلق من صفاقس على الساعة الحادية عشر صباحا و يصل الى تونس على الساعة السادسة و نصف مساءا عبر سوسة يوم الجمعة واحد ماي.

Figure 27 Un exemple d'interprétation sémantique

Pour réaliser ce système, nous avons opté pour les choix suivants:

- une représentation componentielle
- une grammaire probabiliste
- une analyse sélective
- une méthode anthropocentrée ;

IV.4.1 Représentation componentielle du sens

Chaque mot significatif pour l'application est représenté à partir d'un ensemble de traits sémantiques noté TSe = {domaine, classe sémantique, trait micro sémantique} et un ensemble de traits syntaxiques noté TSy = {genre, nombre, nature}. Les traits de l'ensemble TSe indiquent respectivement le domaine de l'application, la classe sémantique à laquelle appartient le mot à interpréter, et le dernier trait c'est un trait micro sémantique qui permet de différencier le sens des mots appartenant à une même classe sémantique. L'avantage de cette représentation par rapport aux représentations logiques et aux réseaux sémantiques, est qu'elle correspond le mieux au résultat d'interprétation obtenu par notre système, qui permet de représenter le sens d'un énoncé sous la forme d'un ensemble de paires attributs / valeurs. Ci-dessous un exemple de la description du sens du mot الذاهب ″allant″ en utilisant cette représentation (c-à-d les ensembles de traits TSe et TSy) comme suit:

الذاهب ″ethaheb″ → Tse = {(transport) نقل ″naql″, (mouvement) حركة ″haraka″, (destination) وجهة ″wijha″} + Tsy = {(masculin) مذكر, (singulier) مفرد, (nom) اسم}.

IV.4.2 Analyse sélective

C'est une analyse qui ne se base pas sur une analyse syntaxique détaillée: pour le décodage sémantique des énoncés, nous nous sommes basés plutôt sur une analyse sémantique et nous avons considéré seulement les éléments significatifs pour l'application. Les mots vides sont éliminés lors de la phase du prétraitement de l'énoncé. Cette analyse est plus tolérante face au phénomène d'agrammaticalité qui caractérise la parole spontanée. En plus, elle ne nécessite pas des connaissances linguistiques très approfondies. Néanmoins, son problème est au niveau de son adéquation à des applications ouvertes (non finalisées), et lorsqu'il s'agit de traiter des énoncés à structure syntaxique très complexe. C'est pour cela que dans le cas des applications non finalisées, certains systèmes combinent une analyse

syntaxique profonde avec une analyse sélective tel que le système TINA de [Sen92]. D'autres systèmes utilisent les stratégies d'analyses du TAL robuste [Ant03]. Ces systèmes sont plus performants dans des applications plus ouvertes.

IV.4.3 Méthode anthropocentrée

Elle se base sur une analyse de corpus: Pour la construction de la structure de représentation du sens SRS [Zou04a], nous avons développé une méthode basée sur une analyse de corpus pour l'extraction des mots significatifs, des mots de référence et des classes sémantiques de l'application, et sur une coopération homme/machine pour l'interprétation des mots. Les mots vides sont éliminés en utilisant un filtre lexical. Selon cette méthode le rôle de l'utilisateur est de définir et d'attribuer l'ensemble des TSe et de Tsy aux mots. Et le rôle de la machine est de satisfaire les contraintes d'intégrités afin d'aboutir à une SRS non ambiguë et cohérente. Notre système se base en tout sur une dizaine de contraintes. Cette méthode nous a permis de faciliter la tâche d'interprétation des mots, ainsi que de la tâche de construction de la SRS et de la maintenance de sa cohérence. Un exemple de contrainte à vérifier est que: Deux mots différents ne peuvent pas être décrits par un même ensemble de Tse sauf dans le cas où ils sont considérés comme synonymiques ou possédant un même rôle sémantique.

La figure 17 suivante résume le principe et les différentes étapes de traitements qui aboutissent à la construction d'une SRS (c à d l'ensemble des paires mot / Tse + Tsy) cohérente et non ambiguë d'une application donnée.

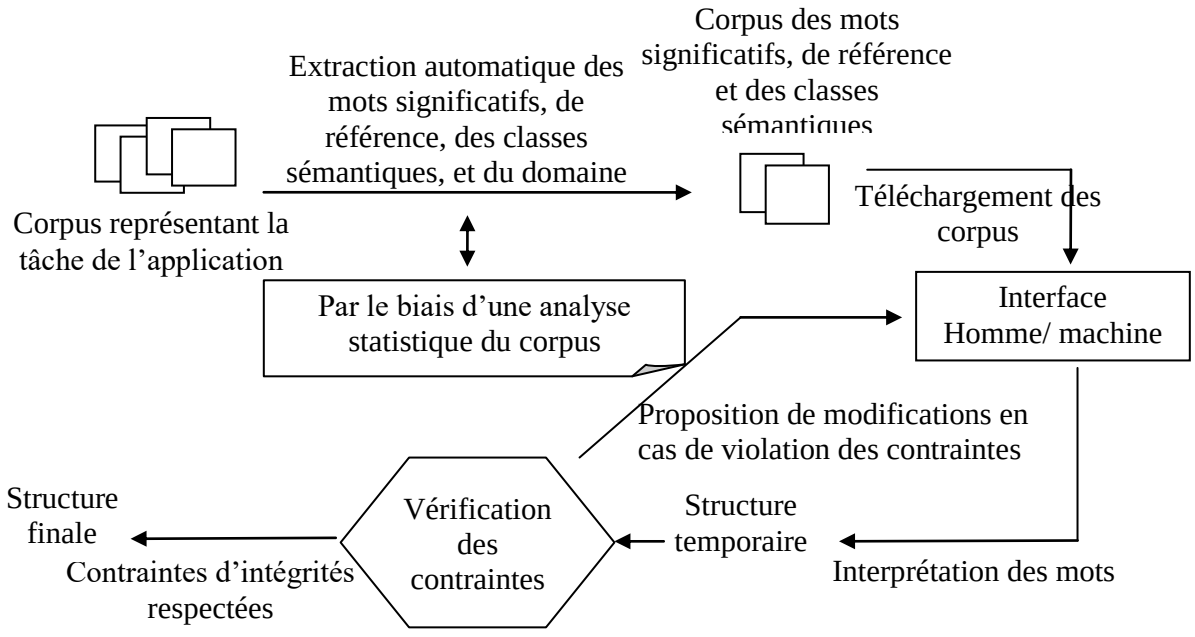


Figure 28 Les étapes de construction d'une SRS d'une application.

L'intervention du sujet humain dans le processus de construction de la structure de représentation du sens, permet alors de réduire le degré de complexité de la réalisation de cette tâche. En plus cette approche est validée par les sciences cognitives, puisque le résultat d'interprétation dépend du sujet humain en interaction avec la machine. Cette approche s'inscrit dans le courant des travaux réalisés en informatique par [Tan97] et [Beu98, 02].

Nous signalons que pour l'extraction des mots de référence (mots indiquant le type et la nature illocutoire d'un énoncé), nous avons utilisé la méthode Tf×Idf (Term frequency×Inverse document frequency) dont la formule est la suivante :

$p_{ij} = [tf(m_i, D_j).\log(n/df(m_i))]/[tf(m_i, D_j)+0{,}5+(1.5.n.l(D_j)/\sum_{Dk} l(Dk)).\log(n+1)]$

Où, n et l(Dk) désignent respectivement le nombre des types de requêtes considérées et la longueur de l'ensemble de tous les requêtes du corpus représentant le domaine Dk de l'application finalisée considérée. Le terme tf (mi, Dj) désigne le nombre d'occurrences de mi dans Dj. Df (mi) correspond au nombre des types de requêtes où apparaît mi.

Ainsi les mots de référence associés aux requêtes de type Dj = {mi / p_{ij} > seuil donné}, où p_{ij} correspond au poids du mot mi dans les requêtes de type Dj.

La figure 18 suivante illustre la manière de partitionnement des requêtes d'une application traitant des demandes de renseignements ferroviaires, et la manière de calcul du degré de pertinence p_{ij} de chaque terme mi pour chaque type de requêtes.

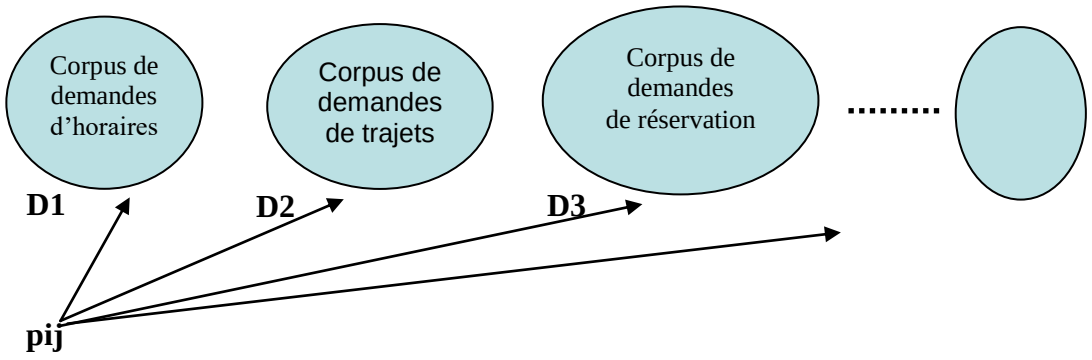


Figure 29 Partitionnement des requêtes de l'application selon leur type, et calcul des pij pour l'extraction des mots de référence.

En ce qui concerne l'extraction des classes sémantiques ou concepts de l'application, nous avons utilisé l'algorithme des k-means. L'avantage de cet algorithme permet de traiter rapidement des données de taille importante, puisqu'il converge à une vitesse linéaire de l'ordre de O(n.k.t). Où n, k et t désignent respectivement le nombre de mots à classer, le nombre de classes sémantiques et le nombre d'itérations maximales. En plus il est simple à implémenter.

IV.4.4 Grammaire probabiliste

Une grammaire probabiliste: cette grammaire participe au choix des TSe adéquats à la description des mots constituant l'énoncé à interpréter. Cette grammaire permet de tenir

compte de plusieurs informations contextuelles en même temps. En plus elle ne considère que les TSe pertinents déjà utilisés pour la prédiction du TSe correspondant à un mot non pas encore interprété. Notre modèle permet de contraindre l'analyse sémantique des énoncés reconnus, en réduisant l'espace de recherche du décodage sémantique des énoncés. Ceci est réalisé en se reposant sur une estimation des probabilités d'interprétation d'un mot donné, sur les mots qui agissent sur son sens (en utilisant la notion d'information mutuelle moyenne), et sur l'utilisation de modèles de type POS tagging pour la détermination de chacun des traits de l'ensemble TSe. L'utilisation pour l'interprétation d'un mot donné, les mots qui agissent sémantiquement sur ce dernier, permet de surmonter les problèmes de l'oral spontané. Ceci a été prouvé à travers le formalisme des grammaires de cas de Fillmore. Quand aux modèles POS tagging, leur performance a été démontrée dans le domaine de l'analyse syntaxique. Ci-dessous l'équation exprimant la probabilité d'interprétation d'un mot Mi par le couple (Ci, TMi) en tenant compte du type de l'énoncé. On remarque que dans cette formule nous n'avons pas considéré le domaine de l'application puisqu'il est prédéfini à l'avance. Dans notre cas, il s'agit du domaine des renseignements ferroviaires. Les approximations et les assomptions que nous avons considérées pour l'obtention de ce modèle (décrit par la formule 1 suivante) sont détaillées dans [Zou05a].

$$P((Ci, TMi)/ Mi, NTj) = P(NTj/ Mr) \times P(Ci / NTj, Mi-1, CPi-1, CPi-2) \times P(TMi /Ci,TsePi-1) \quad (1)$$

On remarque bien que cette probabilité est calculée en fonction du produit de trois probabilités conditionnelles. La première probabilité P (NTj/ Mrk) permet d'identifier le type de l'énoncé, s'il s'agit d'une demande de réservation, d'annulation de billet, etc. Ceci en tenant compte des mots de références Mrk présents dans l'énoncé du locuteur. Cette probabilité est calculée comme suit :

$$P (NTj / Mrk) = N (NTj(E), Mrk) / N (Mrk) \quad (2)$$

Où N (NTj(E) / Mrk) est le nombre d'occurrence des mots Mrk dans les énoncés de type NTj. Et N (Mrk) est le nombre total de cooccurrences de Mrk dans un même énoncé.

La deuxième probabilité P(Ci / NTj, Mi-1, CPi-1, CPi-2) permet de déterminer la classe sémantique Ci à laquelle appartient le mot à interpréter Mi, en tenant compte du type de l'énoncé et des deux classes sémantiques pertinentes précédentes. Cette probabilité est calculée comme suit :

$$P(Ci / NTj, CPi-1, CPi-2) = N(NTj(E), Ci, CPi-1, CPi-2) / N(NTj(E), CPi-1, CPi-2) \quad (3)$$

Où CPi-1 et CPi-2 sont les classes sémantiques affectées aux mots ayant la plus grande affinité sémantique avec le mot Mi à interpréter. L'affinité sémantique entre 2 termes Mi et Mj donnés est calculé comme suit:

$$AFFsém (Mi, Mj) = IM (Mi, Mj) = \log [P(Mi / Mj) / P(Mi).P(Mj)]; \text{ avec } Mj \in Cdroit(Mi) \quad (4)$$

Et la troisième probabilité P(TMi / Ci, TsPi-1) permet de déterminer le trait micro sémantique TMi à attribuer à Mi, en tenant compte de la classe qui a été attribuée à ce mot et du Tse pertinent précédent. Cette probabilité est calculée comme suit :

$$P(TMi / Ci, TsePi-1) = N(Tsei, TsePi-1) / N(Ci, TsePi-1) \quad (5)$$

IV.5 Tests

Voici un exemple illustratif de l'extraction des mots ayant la plus grande affinité sémantique avec le mot à interpréter تونس dans l'énoncé suivant :

بم سعر الذهاب إلى تونس.

Suite à une comparaison des affinités sémantiques (voir figure 19) que possède le mot تونس avec chaque mot de son contexte droit, on remarque que les mots الذهاب et إلى ont la plus grande affinité sémantique avec تونس

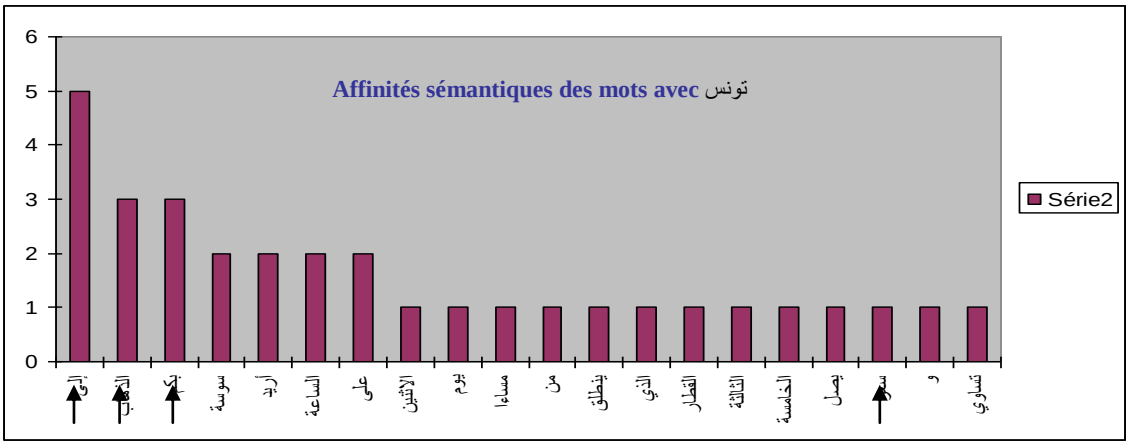


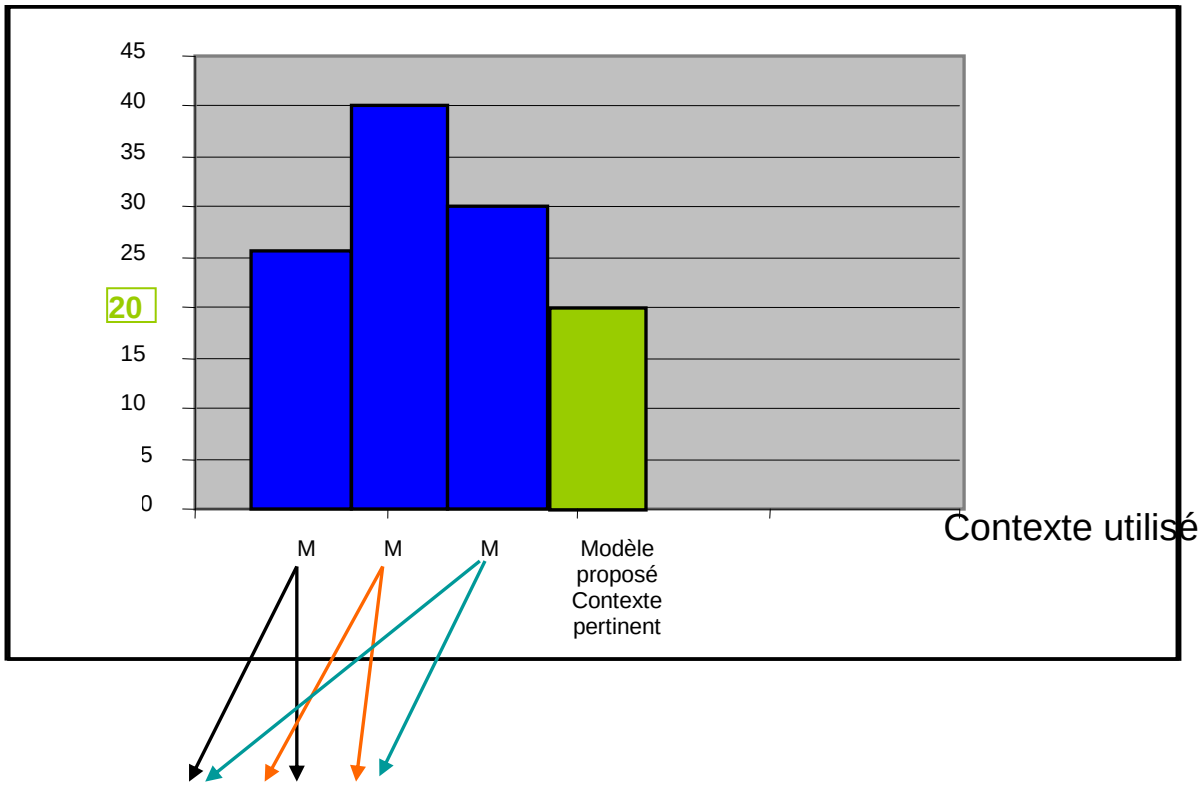
Figure 30 Vocabulaire de l'application.

Ainsi, on utilisera les classes Ci attribuées aux mots الذهاب et إلى pour la détermination de la classe sémantique à laquelle appartient تونس → CPi-1= مؤشر et CPi-2= حركة, et l'ensemble Tse = {CPi-1= مؤشر , TMPi = وجهة} pour la détermination du trait micro sémantique du mot تونس.

Pour montrer l'efficacité de notre système, nous avons testé et évalué sa performance. Les résultats trouvés sont encourageants.

Par exemple, les figures 20 et 21 suivantes montrent bien :

- L'influence de la considération du contexte pertinent sur le résultat d'interprétation, par rapport à des modèles considérant un historique fixe et déterminé à l'avance.
- l'influence de la considération de plusieurs types d'informations contextuelles et de la taille du contexte sur le résultat d'interprétation.



E = M Mi-1 Mi-2 Mi-3 Mi-4 Mi-5…

Figure 31 L'influence du contexte pertinent sur le résultat d'interprétation

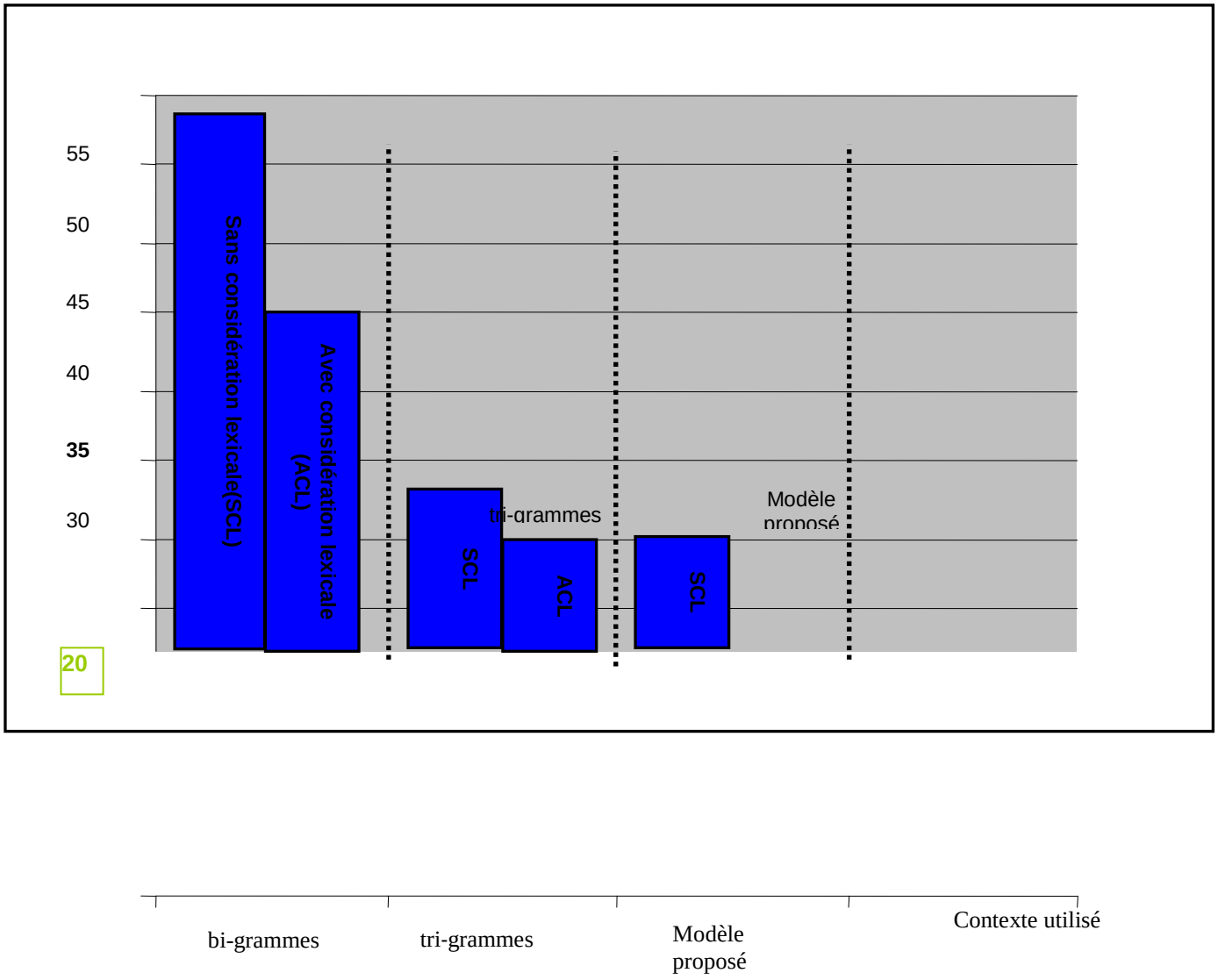


Figure 32 : Influence du contexte

VI.6 Discussion et perspectives

Nous avons utilisé un modèle probabiliste pour le décodage sémantique de la parole arabe spontanée. Ce modèle permet d'utiliser le contexte jugé pertinent, et d'intégrer différents types de données contextuelles pour déduire le sens d'un mot donné.

Afin de montrer l'efficacité de notre modèle, nous avons testé et évalué sa performance. Les résultats trouvés sont encourageants. Les taux d'erreurs du décodeur sémantique probabiliste et du générateur de formulaires (paires attribut/valeur) sont respectivement égaux à 29% et 25%.

Nous comptons améliorer les performances de notre système. Pour cela, nous prospectons de réaliser dans le prochain avenir.

- Une extension de la modélisation probabiliste proposée pour le décodage sémantique des énoncés. Cette extension permet d'introduire des données syntaxiques dans le modèle, afin d'améliorer sa performance. En effet en évaluant la qualité de notre système de

compréhension, en calculant le pourcentage des ensembles TSe incorrectement attribués, nous avons obtenu un taux d'erreur de l'ordre de 29%. Ce taux d'erreur est du au fait, que certains énoncés du corpus de test ont une structure syntaxique très complexe. Afin de remédier ce problème, certains systèmes combinent une analyse syntaxique profonde avec une analyse sélective tel que le système TINA de [Sen92]. D'autres systèmes utilisent les stratégies d'analyses du TAL robuste [Ant03]. En plus du problème de la complexité syntaxique de certains énoncés, l'absence de la voyellation dans la majorité des textes arabes modernes rend la compréhension automatique de cette langue encore plus difficile. Or la détermination de la voyellation correspondante à un mot (et par conséquent son sens), nécessite plusieurs niveaux de connaissances: morphologique, syntaxique, sémantique et pragmatique. [Deb02]. C'est pour cela, que nous prospectons d'améliorer la performance du système de compréhension, en intégrant des données syntaxiques dans le modèle probabiliste utilisé pour le décodage sémantique.

- Afin de rendre notre système plus robuste face aux erreurs engendrées par le module de reconnaissance de la parole, nous comptons s'intéresser à la compréhension des énoncés en présence des mots mal reconnus ou inconnus par le lexique de l'application. Ceci en rendant notre système capable en même temps d'identifier ce type de mots ainsi que de prédire la classe sémantique à laquelle appartient le mot inconnu ou mal reconnu pour pouvoir déterminer son sens. Nous rappelons que les mots inconnus sont dus à un apprentissage insuffisant, ou à la grande variabilité du langage. Alors que les mots mal reconnus peuvent être causés par un environnement bruité, ou par l'accent de l'utilisateur.

V. CONCLUSION

V.1. Bilan

Ce mémoire fait un bilan de plus d'une douzaine d'années de recherche sur le TALN (cas de l'arabe). Si le cadre de mes travaux a toujours été le traitement automatique de l'arabe, trois grands axes ont été explorés: la génération automatique de dictionnaires et d'outils d'analyse morphologiques, syntaxiques et sémantiques, la synthèse et la reconnaissance de la parole et la traduction de l'oral.

La génération de dictionnaires et d'outils d'analyse morphologiques et syntaxiques pour l'arabe est devenu une activité marginale avec encore quelques travaux en cours. Nous avons réalisé un dictionnaire linguistique **étiqueté** d'analyse, de dérivation et de conjugaison de la langue arabe. Ce dictionnaire contient plus que 264 milles mots et il est muni de 7 types d'étiquettes qui sont : **اسم الفاعل، اسم المفعول، اسم التفضيل، اسم المكان، اسم المرّة، اسم آلة، اسم الهيئة.**

De nombreuses méthodes et algorithmes ont vu le jour jusqu'à nos jours et par différentes équipes dans le monde arabe et notamment SAKHR en Égypte, RIADI et LARIS en Tunisie et d'autres équipes de recherche au Maroc, en Algérie et Syrie; et ailleurs en France (CLIPS, LIDILEM, IRIT,etc…).

Les travaux sur l'analyse sémantique de l'arabe n'apportent pas de solutions satisfaisantes: la complexité des méthodes, le manque de structuration et de collaboration entre les équipes de recherche, et entre linguistes arabes et informaticiens sont certainement responsables de l'inefficacité et de la rareté des solutions dans ce domaine.

Notre équipe dispose actuellement de plusieurs outils logiciels et d'un ensemble de dictionnaires pour l'analyse lexicale, syntaxique, sémantique et phonétique de l'arabe ; ces outils aident à générer un système de synthèse ou de reconnaissance de la parole arabe, en plus d'un prototype de traduction de l'arabe vers le français; l'équipe dispose aussi d'un système d'édition Braille multilingue.

V.2. Perspectives et évolution

Le système oreillodule comprend trois catégories principales de fonctionnalités :

- des fonctions de reconnaissance de la parole
- des fonctions de traduction linguistique

- des fonctions de synthèse de la parole.

Prises chacune à part, ces fonctions existent et il y a plusieurs laboratoires qui sont en train de développer des systèmes de traduction d'une langue vers une autre comme le CLIPS en France, d'autres des systèmes de reconnaissance de la parole, d'autres des systèmes de synthèse de la parole et enfin une troisième catégorie qu'on peut appelé « informatique pour les handicapés » visuels ou autres ;

Ces systèmes tournent sur des plates-formes contraignantes de par leur taille et leur puissance de calcul: la question qui se pose alors est comment surmonter ses contraintes matérielles et logicielles pour réaliser notre objectif: à savoir l'oreillodule?

Depuis la naissance du transistor, élément de base de l'électronique, plusieurs applications complexes ont été réalisées aujourd'hui grâce à l'évolution technologique qui a permis d'augmenter le nombre de transistors. Cette évolution va dans le sens de la réalisation d'applications temps réel : ces applications sont complexes et semblent être un défi aux concepteurs.

Aujourd'hui, selon l'architecture cible (les ASIC : Application Specific Integrated Circuits, les FPGA : Field Programmable Gate Array, les SoC : System on Chip), plusieurs solutions existent et répondent au marché de l'industrie électronique qui exige des produits logiciels portables, performants et moins coûteux. Nous pouvons citer comme produit, les systèmes de synthèse et de reconnaissance de la parole qui produisent une voix se rapprochant considérablement de la voix humaine. La plupart de ces produits servent notamment dans la télécommunication (la messagerie vocale, service de consultation de courrier électronique par téléphone, serveurs vocaux interactifs, livres et journaux parlants, service bancaire par téléphone, etc.…).

En même temps, la mise en œuvre des nouvelles technologies devient de plus en plus risqué en raison du niveau d'investissement requis, tandis que la complexité des systèmes conduit à une exigence de plus en plus forte sur les méthodes de conception.

Dés lors, les questions suivantes se posent à l'industrie : comment tirer le meilleur profit des technologies ? Comment s'ouvrir à de nouvelles applications ? Comment concevoir des produits toujours plus performants et innovants notamment sur le plan de la rapidité, de l'encombrement, de la robustesse vis-à-vis de perturbations et de la consommation électrique ?

Parmi ces applications, nous pouvons citer les machines parlantes, les traducteurs de poche et autres comme notre Oreillodule: objectif de nos travaux. C'est dans ce cadre que se situe l'évolution de nos travaux: il s'agit de réaliser un système embarqué, c'est àdire mettre tout le système sur un seul puce.

Un tel système est composé d'une couche logicielle utilisant plusieurs techniques gourmandes en mémoire et en puissance de calcul. On citera en exemple les fonctions d'accès aux bases de données, les traitements lexicaux, syntaxiques, sémantiques, pragmatiques et phonétiques.

V.3. Les défis

Une étude préliminaire conduit à une mise en valeur de différents problèmes et défis pouvant mettre en difficulté les différentes composantes logicielles dans ce projet. Parmi ces défis nous citons :

-L'exploration d'architecture à partir d'une spécification de haut niveau. Traditionnellement l'architecture des systèmes monopuces est définie par des spécialistes expérimentés. La conception des architectures dédiées à des applications particulières peut nécessiter des schémas de communication très sophistiqués mettant en œuvre des parties matérielles, des parties logicielles et des parties non numériques telles que les interfaces de composants optiques ou micromécaniques. La seule expérience des concepteurs devient insuffisante pour définir de telles architectures. Des outils permettant d'explorer un grand nombre de solutions deviennent cruciaux pour évaluer le coût de la solution à concevoir et

éviter les voies sans issue. L'architecture retenue devra permettre un développement en parallèle du flot logiciel et matériel.

-Définir des modèles et des techniques permettant de mettre en oeuvre à court, moyen ou long termes notre système finalisé reposant sur une forte composante langagière. Ainsi une coopération entre plusieurs équipes de recherche en informatique linguistique permettra l'étude des mécanismes fondamentaux de la communication en langue naturelle. Cette recherche s'effectue dans un contexte pluridisciplinaire alliant linguistique et informatique principalement. La réalisation de systèmes de dialogue effectif dans le cadre notamment de collaborations industrielles. Cette activité nous permet par ailleurs de disposer d'une plateforme d'expérimentation pour la validation des différents modèles que nous concevons. La définition d'outils et de méthodes génériques permettant d'étudier de façon fine des situations de dialogues réels, issus de la transcription d'expériences de simulation ou d'observations directes.

ANNEXE 1 : Histoire des dictionnaires dans le monde arabe

I. Les écoles :

Plusieurs dictionnaires arabes sont apparus durant les siècles précédents. Cette multitude n'a pas changé la raison de créer ces dictionnaires ; à savoir la protection du Coran contre les erreurs de prononciation et de compréhension, ainsi que de garder le trésor de la langue arabe contre les intrus.

L'histoire des dictionnaires de la langue arabes a connu plusieurs écoles lexicales et linguistiques, parmi ces grandes écoles, nous pouvons citer :

- **Ecole Elkalil : (مدرسة الخليل)**

C'est la première école connue dans l'histoire des dictionnaires de la langue arabe, et "Elkalil" est le premier linguiste qui a entamé la rédaction dictionnairique.

Le principe de cette école est de classifier les matières suivant les lettres de sortie et de diviser le dictionnaire en livres, dériver ces livres en des portes, ces portes contiennent les mots.

مدرسة الخليل

وقوام مدرسته ترتيب المواد على الحروف حسب مخارجها وتقسيم المعجم إلى كتب، وتفريع الكتب إلى أبواب بحسب الأبنية، وحشد الكلمات في الأبواب، وقَلْبُ الكلمة إلى مختلف الصيغ التي تأتي منها، مثل قوله في باب السين والميم مع الواو والألف والياء: سوم، وسم، سمو، مسو، موس . وقد سار بعض رواد التأليف المعجمي على نهج الخليل، فالتزمه الأزهري في "التهذيب" وابن عباد في "المحيط"، والقالي في "البارع." [1]

- **Ecole Abi abid : مدرسة أبي عبيد**

Cette école comporte les plus prestigieux maîtres de la langue arabe, c'est "أبي عُبيد القاسم بن سلام" "son principe est de rédiger le dictionnaire suivant le sens et le thème, en créant des articles pour les mots de même sens.

مدرسة أبي عبيد

وهي التي تنتسب إلى أحد أئمة اللغة والأدب أبي عُبيد القاسم بن سلام، وقواعدها بناء المعجم على المعاني و الموضوعات، وذلك بعقد أبواب وفصول للمسميات التي تتشابه في المعنى أو تتقارب و فَضْلُ أبي عبيد جمع كتابا مثل: خلق الإنسان، والنساء، واللباس، والطعام أشتات هذه الموضوعات والمعاني في كتاب كبير، يضم أكثر من ثلاثين والشراب، ... ومجموع ما تضم هذه الكتب الثلاثون سبعة عشر ألف حرف وأكثر . [1]

- **École Eljawhari : " مدرسة الجوهري"**

[1] **www.ajeeb.com**

Cette école est due au maître innovateur "Eljawhari" qui a créé le principe de classification des mots dans le dictionnaire suivant la dernière lettre au lieu de la première, puis de ranger les articles suivant les lettres de l'alphabet arabe.

مدرسة لجوهري

هذه المدرسة تنتسب إلى الإمام المجدد الجوهري الذي ابتكر في التأليف المعجمي منهجا قرّب اللغة إلى الباحثين. ومئات المعاجم و الكتب اللغوية مرتبة ترتيب الجوهري مما يدل على عِظم مدرسته. ونظام هذه المدرسة ترتيب المواد على حروف المعجم باعتبار آخر الكلمة بدلا من أولها، ثم النظر إلى ترتيب حروف الهجاء عند ترتيب الفصول، والأول سماه بابا، والثاني فصلا، فكلمة "بسط" يُبحث عنها في باب الطاء لأنها آخر حرف فيها، وتقع في فصل الباء لأنها مبدوءة بها.[2]

La dictionnairique (la science des dictionnaires) qui est une branche du traitement automatique du langage naturel, à été l'une des activités intéressantes pour les linguistes arabes, mais en leurs temps, ils traitaient les dictionnaires papiers. Ils ont commencé par regrouper les termes selon leurs domaines comme les livres de "elasmai" "كتب الأصمعي في النّخل والخيْل ", puis ils ont conçu le dictionnaire selon une méthode qui facilite la recherche des termes et leur sens. On croit que " الخليل بن أحمد الفراهيدي " est le premier qui a mais un dictionnaire linguistique arabe qui l'a appelé "كتاب العيْن ". Plusieurs linguistes ont apparus après et qui ont mis plusieurs dictionnaires, parmi ces dictionnaires nous citons:

date	auteur	dictionnaire
(ت 206 هـ)	لأبي عمرو الشيباني	معجم (الحروف)
(ت 244 هـ)	لابن السكِّيت	معجم (الألفاظ)
(ت 321 ه)	لابن دريد	(الجمهرة)
(ت 356 هـ)	لأبي علي القالي	(البارع)
(ت 370 هـ)	للأزهري	(تهذيب اللغة)
(ت 395 هـ)	لابن فارس	(و (المُجمل (اللّغة مقاييس)
(ت 400 هـ)	للجوهريُّ	(الصّحاح)
(ت 458 ه)	لابن سيده	(و (المخصّص (المُحكم)
(ت 711 هـ)	لابن منظور	(لسان العرب)
(ت 817 هـ)	للفيروز آبادي	(المحيط القاموس)
(ت 1205 ه)	للزّبيدي	(تاج العروس)

II. Organisation des dictionnaires (Microstructure):

[2] **www.lexicons.ajeeb.com**

1. Classification alphabétique (المعاجم ذات منهج القافية):

Le dictionnaire est sectionné en portes suivant les caractères de l'alphabet de l'arabe, puis au sein de chaque porte, les verbes sont rangés suivant leurs fins vocaliques, parmi les dictionnaires qui ont adopté cette classification, nous citons :

الصحاح " أو " تاج اللُّغَة العربية و صحاح اللُّغَة " لإسماعيل بن حماد الجوهري (ت 398 هـ) . و الذي يُعتبر أول " معجم صنف على نظام القافية . ثمّ " لسان العرب " لابن منظور الأفريقي (ت 711 هـ) . و منها " القاموس المحيط " لِمحمد بن يعقوب الفيروز آبادي (ت 817 هـ) ؛ و قد استطاع " القاموس " أن يزاحم المعاجم الأخرى في التداول ، و أن يحيط نفسه بالأدبيات الجدلية و التعليقات و الاختصارات و تُرجم إلى اللغتين الفارسيّة و التركيّة . و أيضاً معجم " تاج العروس من جواهر القاموس " للزبيدي (القرن الثامن عشر ميلادي) . و " التاج " مكّون من عشرة مجلدات . و كان أحد أهم مصادر المعجميات العربية في الدراسات العربية في أوربا [3].

2. Classification radical (المنهج الجذري):

Ce type de dictionnaire comporte 28 portes, selon l'ordre d'apparition des caractères de l'alphabet de la langue arabe, chaque caractère a une porte, puis les termes sont classifiés selon leur racine et par ordre d'apparition alphabétique.

Le premier dictionnaire écrit suivant cette classification est " أساس البلاغة " qui à était écrit en 538 H par "لجار الله محمد بن عمر الزمخشري", ce dictionnaire à connu un succès très important, pourtant qu'il ne contient pas beaucoup de termes par rapport aux autres dictionnaire, en ajoutant en plus des racines, leurs sens allégorique.

La majorité des dictionnaires de nos jours, ont adopté la méthode de classification radicale, car il est plus simple, en plus qu'il regroupe les termes qui sont issus d'une même racine. Parmi ces dictionnaires: "" " محيط المحيط " " ", المصباح المنير " لأحمد بن محمد الفيومي (ت 776 هـ) ", " مجمع اللُّغَة العربية بالقاهرة " réalisé par " لبطرس البستاني (ت 1883 م) " , " المعجم الوسيط ", " المنظمة العربية للتربية و الثقافة و العلوم " écrit par " المعجم العربي الأساسي ". Plusieurs dictionnaires aussi, ont été réorganisés et restructurés, comme "" القاموس المحيط " للفيروز آبادي " qui a été réécrit par " الطاهر أحمد الزواوي " en 1959, de même pour "" لسان العرب " لابن منظور " qui a été repris par " نديم مرعشلي و يوسف خياط " en 1970.

La problématique de la langue et des dictionnaires arabes, a été très développé et surtout pendant la deuxième moitié du 20éme siècle, parmi les spécialistes qui ont travaillé sur ces thèmes, nous pouvons citer " " إبراهيم أنيس ", " محمد المبارك ", " صبحي الصالح ", " إبراهيم السامرائي ", " مهدي المخزومي " et " مراد كامل ".

[3] **http://www.nisaba.net**

III. Les banques terminologiques arabes: (بنوك المصطلحات العربيّة)

Plusieurs sociétés arabes ont réalisés des banques terminologiques dans le but de résoudre quelques problèmes d'arabisation des sciences, nous citons la :

- **Banque maghrébine (**بنك مغربي**):**

Cette banque est réalisée en 1978 à Ribat (Maroc), dans le cadre d'une collaboration entre " المنظمة العربية للتربية والثقافة والعلوم" et "برنامج الأمم المتحدة للتنمية ومعد الدراسات والبحوث للتعريب ".[4]

- **Banque terminologique saoudienne (**البنك الآلي السعودي للمصطلحات**):**

Cette banque est réalisée en 1983 dans le centre national saoudien des sciences et des technologies (من قبل مدينة الملك عبد العزيز للعلوم والتقنية). Elle regroupe un nombre important de termes dans plusieurs domaines. Ces termes sont accrédités (مجامع اللغة العربية ومكتب تنسيق التعريب).

Banque jordanienne (بنك المصطلحات الأردني**) :** Cette banque est réalisée en 1985 par le centre jordanien de la langue arabe (مجمع اللغة العربية الأردني) et à pour objectif le stockage des termes scientifiques et techniques utile à la traduction et l'arabisation.

Dictionnaire Informatisé de l'Arabe Multilingue et Basé sur Corpus(DIINAR-MBC) : L'objectif général du projet est de contribuer à produire une boîte à outils pour les linguistes, les lexicographes et professionnels de la technologie de la langue. Dans le cadre du projet européen DIINAR-MBC (DIctionnaire INformatisé de l'ARabe,Multilingue et Basé sur Corpus) (Dichy et al., 1998), il a été réalisé un dictionnaire prototype. Ce prototype est constitué d'environ 8000 unités lexicales (verbes, noms et adjectifs). Il reprend, les informations morphologiques et syntaxiques antérieurement développées et stockées dans DIINAR.1 (Braham et al., 2002) et ajoute les définitions et les équivalents en français et en anglais. Sur la base de ce prototype, il a été établi un dictionnaire électronique pour apprenant exploité dans le cadre d'un environnement d'apprentissage (Zaafrani, 2002b). d'abord, la réalisation des outils de traitement automatique de l'arabe qui vont assister l'expert humain dans le choix du vocabulaire définitoire et des exemples du corpus textuel.

[4] **http://www.c4arab.com**

DicoBase™ Arabe : Le DicoBase™ Arabe de Linga est un dictionnaire de type base de données linguistiques pour la langue arabe utilisée dans les applications d'informatique linguistique. Le DicoBase™ Arabe est le résultat de plusieurs années de travail dans le domaine de la terminologie et de la lexicographie. L'équipe de Linga est composée de plusieurs Docteurs en littérature arabe et de terminologistes professionnels sous la direction du Professeur Leila-Guillemot. Les principales caractéristiques du DicoBase™ Arabe sont les suivantes :

Catégorie	**Nombre**
Termes	31 245
Dérivations	95 983
Expressions	10 723
Logique de structure de traits	845
Nombre total d'entrée	1 245 785

Encodage des données suivant le standards UNICODE ISO 10646

- **Dictionnaire multilingue ajeeb (arabe-français-anglais)**

c'est un dictionnaire multilingue (arabe-français-anglais) présenté par le site www. ajeeb.com

Figure A.11 dictionnaire ajeeb

1- les dictionnaire multilingues CIMOS : (Figure A.12)

Les dictionnaires existent en deux versions:

- Version Simple
- Version Client-Serveur

Il existe 4 types de dictionnaires:

- **Dictionnaire général** contenant des mots d'usage courant avec plus de 300 000 mots, phrases et verbes nominaux.
- **Dictionnaire spécifique** contenant des mots utilisés par des spécialistes et des experts dans un domaine précis.
- **Dictionnaire d'idiomes** contenant des expressions et des verbes nominaux.
- **Dictionnaire utilisateur** contenant des mots ajoutés ou mis à jour par l'utilisateur.

Caractéristiques principales:

- Capable de trouver toutes les formes fléchies d'un mot
- Recherche de verbes nominaux
- Intégration facile avec les applications multimédia
- Interface utilisateur multi-langue (Anglais, Français, Arabe...)

Sujets inclus que les dictionnaires spécifiques utilisent:

- Comptabilité 12 000 entrées
- Affaires 3 400 entrées
- Informatique 1 200 entrées
- Finance 1 980 entrées
- Médecine 25 000 entrées
- Militaire 4 000 entrées
- Termes techniques et scientifiques 60 000 entrées

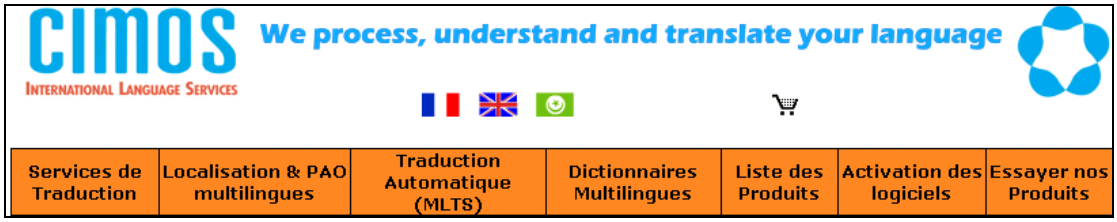


Figure A.12 CMOS

ANNEXE 2 : Etudes statistiques sur l'arabe

Dans ce qui suit, nous allons présenter une étude statistique concernant trois grands dictionnaires da la langue arabe, en classifiant les racines suivant leurs nature (trilitères, quadrilitères et quinquilitère) et suivant leurs occurrence d'apparitions :

- Dictionnaire " ASSIHAH d'EL-JAWHARI " : **الصحاح لابن الجوهري**
- Dictionnaire " LISSANE ELARAB " : **لســـان العــرب**
- Dictionnaire " TEJ ELAROUSSE " : **تـاج الـعـروس**

RACINE	NOMBRE	%
Trilitères	7597	63.42 %
quadrilitères	4081	34.07 %
Quinquilitères	300	2.51 %
Total	11978	100 %

Dictionnaire « الصحاح لابن الجوهري »

RACINES	NOMBRE	%
Trilitères	6538	70.50 %
Quadrilitères	2548	27.47 %
Quinquilitères	187	2.03 %
Total	9273	100%

Dictionnaire « تـاج الـعـروس »

RACINES	NOMBRE	%
Bilitères	21	0.37 %
Trilitères	4814	85.33 %
Quadrilitères	768	13.61 %
Quinquilitères	38	0.64 %
Total	5641	100 %

Dictionnaire « لســان العــرب »

Dans le tableau suivant, qui prend en compte plusieurs dictionnaires (الوسيط ,محيط المحيط ,المحيط , نجعة الرائد ,لسان العرب ,القاموس المحيط ,الغني) nous avons présenté des informations sur le nombre des mots, leurs occurrences dans ces dictionnaire(عـدد المشتـقات ,عدد المـواد,متوسط عدد المواد للـحرف ,عـدد الكلمـات بـالمـعجم)[5].

المعجم	متوسط عدد المواد للـحرف:	عـدد المـواد بالمـعـجم:	عـدد المشتـقات بالمعـجم:	عـدد الكلمـات بـالمـعجم:
المحيط	1.429	40.000	40.000	810.000
محيط المحيط	400	11.200	84.965	1.300.000
الوسيط	250	7000	30.000	450.000
الغني	1.071	30.000	195.000	2.000.000

[5] **www.ajeeb.com**

القاموس المحيط	390	11.000	70.000	733.000
لسان العرب	335	9.393	158.149	4.493.934
نجعة الرائد	---	142	5.629	119.176

Statistiques sur les dictionnaires arabes

ANNEXE 3 : Equipes et Laboratoires de recherche en TAL arabe dans le monde arabe :

Les premières recherches d'arabisation dans le monde ont pratiquement commencé au début des années 70. Les recherches sur le traitement automatique des langues et en particulier l'arabe ont commencé dès 1979. L'arabisation a été prise en charge dans un premier temps par les pays arabes qui avaient pour objectif la contribution à la promotion de la langue arabe pour la doter de moyens lui permettant d'être utilisée sans complexe dans les nouvelles technologies. Comment cela a-t-il été fait ? Quels moyens ont été mis en œuvre ? Cela suppose la mise en place d'une ingénierie de la langue arabe dans le contexte du multilinguisme qu'exige l'ouverture du monde arabe sur les autres cultures qui véhiculent actuellement le savoir scientifique et technologique. La démarche scientifique ne procède que par des constructions hypothétiques et relatives, par procédures de découverte, d'innovation et de décision en perpétuel changement et surtout par des recherches de données et des réalités objectives qui ne sont établies et valables que par rapport à des problématiques définies dans des périodes déterminées. Dans ce cadre, il est implicatif de lancer des projets de linguistique automatique en arabe en particulier, des systèmes de gestion de bases de données textuelles et multilingues, des systèmes de gestion de bases de données documentaires multilingues, des traitements statistiques de documents, d'interrogations de bases de données en langage naturel, d'arabisation des logiciels et des interfaces. Mais pour cela, il suffit l'existence de laboratoires d'ingénierie de l'arabe dans le contexte multilingue et, en particulier, en relation avec le français et l'anglais. Parmi ces laboratoires et unités de recherche nous citons :

1. RIADI

RIADI est le premier laboratoire en Tunisie qui s'intéressa au traitement automatique de la langue Arabe et ce depuis le début de l'année 1980. RIADI compte actuellement plus de 100 chercheurs travaillant dans les domaines du TALN, de la documentique et de la gestion électronique des documents, de l'ingénierie des connaissances et le génie logiciel.

Le laboratoire et particulièrement l'unité de Monastir, a publié plus de 150 articles scientifiques, thèses de doctorats, mémoires de Mastère et projets de fin d'études sur le traitement automatique de l'arabe.

Les travaux de RIADI ont été couronnés par plusieurs réalisations de logiciels comme les analyseurs morphosyntaxiques, les correcteurs, les dictionnaires, etc.

2- LARIS : Laboratoire de Recherche en Informatique de Sfax

Les thèmes de recherche de ce laboratoire sont :

1. *Traitement de langage naturel (analyse et vérification des textes)*
2. *Communication homme-machine en langage naturel*
3. *Documentique (condensation et résumé automatique)*
4. *Ingénierie (intégration et coopération)*

3- UTIC : Unité de recherche en technologie de l'information et de la communication créée en 2002 à l'Ecole Supérieure des Sciences et Techniques de Tunis. Elle regroupe actuellement environ une cinquantaine de chercheurs. Ses travaux de recherche sont focalisés essentiellement sur le e-Learning et l'accessibilité des personnes handicapées aux nouvelles technologies. Quelques projets sont en cours concernant la synthèse de la parole et la construction d'une base de données morphologiques[6].

4- Laboratoire Informatique et Traitement Automatique de l'Arabe(Maroc)

[6] **http://www.latl.unige.ch/uvf/partenaires.html**

Dirigé Pr. Yahia Hlal, LIT2A a été créé au Maroc officiellement en 1991 pour mener les recherches sur le traitement automatique des langues, et en particulier l'arabe, qui ont commencé à l'EMI dés 1979.

Le laboratoire a pour objectif la contribution à la promotion de la langue arabe pour la doter de moyens lui permettant d'être utilisée sans complexe dans la technologie informatique d'aujourd'hui.

Les thèmes abordés sont les suivants :

- Linguistique automatique et arabe en particulier
- Système d'aide au néologisme
- Système de gestion de Bases de données textuelles arabe et multilingue
- Système de gestion de Bases de données documentaires arabe et multilingue (Indexation automatique)
- Système de gestion de Bases de données juridiques arabe et multilingue
- Systèmes d'aide à la traduction depuis et vers l'arabe
- Système de détection et de correction des erreurs dans la saisie de textes arabes et latins
- Traitements statistiques de documents arabes et latins
- Interrogation de bases de données en langage naturel et arabe en particulier
- Systèmes EAO pour l'arabe
- Arabisation des logiciels
- Interfaces arabes pour la communication dans le cadre des inforoutes

5- Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (ALGERIE)

Le centre de recherche scientifique et technique pour le développement de la langue arabe CRSTDLA[7] a pour missions essentielles :

- la mise en œuvre de projets de recherche dans les domaines des sciences et techniques du langage appliqués à la langue arabe et aux langues d'enseignement en vue du développement de ces langues sur les plans didactique et technologique.

[7] **http://www.crstdla.edu.dz/index.htm**

Les recherches menées au sein de ce département ont pour objectif essentiel le traitement automatique de l'arabe en collaboration avec le département de linguistique arabe. Les travaux y afférents portent donc sur la formation de la théorie néo-khalilienne et la réalisation d'outils informatiques adéquats. Outre ces objectifs, les recherches visent également à:

1- codifier, classifier et normaliser l'usage réel du lexique arabe technique et non technique.

2- élaborer sur cette base des lexiques spécialisés pour l'enseignement et la recherche

3- élaborer une banque de données textuelles automatisée qui doit servir de source et de référence à tous les travaux futurs de lexicographie.

Ce dernier objectif qui constitue une priorité, constitue, en fait, un projet international supervisé par l'ALECSO et coordonné par le Centre.

1. laboratoire de phonètique et de traitement automatique de la parole (Algérie)

C'est une équipe qui s'intéresse au traitement de l'oral : synthèse et reconnaissance. Les domaines de recherche sont :

- **Phonétique de l'Arabe Standard**

1. Etude acoustico-articulatoire de la Haraka et du Sukun en Arabe Standard ;
2. Analyse acoustico-articulatoire des différents phonèmes de l'Arabe Standard.

- **Traitement Automatique de la parole**

1. **Synthèse de la parole** :
   - Synthèse par concaténation d'unités pré-stockées;
   - Synthèse par règles.
2. **Reconnaissance automatique de la parole** ;
3. **Codage de la parole.**

Équipe de recherche :

- Abbas Mourad , **a_mourad29@hotmail.com**
- Benbellil Khoudir, **kbenbellil@hotmail.com**
- Droua-Hamdani Ghania, **gh.droua@post.com**
- Ferrat Kamel, **kferrat@wissal.dz**

6- institut d'études et de recherche pour l'arabisation (IERA) (Maroc)

L'IERA est une institution universitaire de recherche scientifique fondamentale et appliquée qui a pour vocation de faire de l'arabe une langue moderne de travail, d'enseignement, des sciences et des techniques, dotée d'outils pédagogiques et technologiques adéquats. Ses objectifs sont :

- Situer l'arabe dans son environnement national naturel, en tant que langue officielle, et dans son environnement international, en tant qu'une des six langues internationales.
- Doter la langue arabe d'outils conceptuels, terminologiques et computationnels permettant le réaménagement de son statut et de ses fonctionnalités, en tant que langue de communication, de savoir, de science et de technologie.

Parmi les projets en cours, nous citons :

- Un système de réforme de la graphie arabe, dit ASV-Codar (Arabe Standard Voyellé Code Arabe) a été conçu. L'alphabet arabe a été introduit en informatique et plusieurs normes arabes et internationales ont été mises au point, en collaboration avec des instances arabes concernées (ASMO et ALESCO).
- Une première base de données lexicographique euro-arabe a été élaborée en collaboration avec l'UNESCO, l'ALESCO, le PNUD et l'Agence Spatiale Européenne.

- Réalisation de livres scolaires et d'outils pédagogiques du niveau de l'enseignement fondamental qui permettraient à l'apprenant de la langue arabe de lire, d'écrire, de compter et d'acquérir des connaissances de base en langue arabe.

- GENFO : Génération automatique des formes lexicales et conception d'une base de données lexicologique arabe intelligente, recensant toutes les formes morphologiques et morphosyntaxiques potentielles. Ce système est exploitable dans les domaines suivants :

    - Traduction assistée par ordinateur.

    - Constitution des bases de données lexicales et terminologiques générées automatiquement,
    - Exécution de didacticiels pour l'enseignement et l'apprentissage de la langue arabe.

- Exploitation du système GENFO comme outil de recherche pour les linguistes et les personnes concernées par les applications lexicographiques et didactiques.

- GENTERM : Génération automatique des termes. Ce système permet de créer une base de connaissances terminologiques multilingues arabes en ayant recours à des mécanismes de génération automatique.
- MULTILEX : Base de données lexicographiques multilingues (700000 entrées)

7- Université Abou Bekr Belkaid (Tlemcen Algérie)

Au sein de cette université, existe une équipe de recherche s'intéressant au TALN. Parmi ses thèmes de recherches nous citons :

- Reconnaissances automatique de l'écriture et de la parole arabe.
- Système de gestion documentaire arabe.
- Systèmes d'aide à la traduction depuis et vers l'arabe.

Parmi les projets de recherches exécutées ou en cours d'exécution nous citons :

Type	***Date d'agrément***	***Durée***	***Titre du projet***
ANDRU	06/02/2001	03 ans	*العلاج الآلي للغة العربية*
ANDRU	Mai 2000	03 ans	**العلاج الآلي للمنطوق العربي**
CNEPRU	2000	03 ans	**التعرف على الوحدات الدالة اللغوية**

Projets arabes

8- SAKHR:

SAKHR est un éditeur de logiciels orientés vers le public arabophone. L'équipe de recherche appartenant à cette entreprise Koweitienne est plutôt connue par ses produits que par ses publications ou interventions scientifiques. En effet, cette firme commerciale fait un nombre

impressionnant d'annonces de réalisations et de produits. Nous osons même dire qu'ils sont quasiment au nombre de problèmes que se posent les chercheurs dans le domaine. Nous n'avons pu acquérir ces produits et encore moins effectuer des tests de leurs performances.
Sachons simplement qu'aussi bien au niveau d la recherche " fondamentale " qu'au niveau des applications, la société SAKHR[8] déclare avoir des solutions en Analyse Morphologique, Dictionnaires Monolingues et Bilingues, Indexation de Documents, Correction Orthographique, etc.

Parmi les produits de cette société nous pouvons citer :

ArabDox : un système de gestion de document d'Arabic/English/French qui offre à des entreprises une solution intégrée pour contrôler des quantités croissantes de l'information non-structurée dans les documents. Le processus commence en saisissant l'information de toutes les sources (les papiers, microfilms, fax, E-mails, texte classe, des documents de HTML, des documents de bureau, etc.) Il décrit alors cette information avec des attributs de base de données, et la rend finalement aisément disponible et trouvable par un web browser.

Idrisi : La recherche intelligente de documents et d'information de l'Internet. Idrisi est un moteur de recherche bilingue (arabe, anglais) pour des documents, des bases de données d'entreprise (ODBC conforme), et sites Web.

Johaina : surveille les nouvelles concernant le Moyen-Orient. Il balaye des centaines de sites arabes, anglais et français pour chercher toutes les nouvelles informations visant le Moyen-Orient.

Ibsar (vision en anglais) : Il est basé sur le texte de Sakhr Arabic/English à la parole TTS et aux moteurs de ROC de reconnaissance optique des charactères de Sakhr. Ibsar permet aux utilisateurs aveugles de lire les livres et les documents imprimés aussi bien que les fichiers électroniques tout seuls, sans n'importe quelle aide. Il leur permet également d'écrire des textes en arabe et anglais et d'imprimer ces textes en Braille.

SET : est un système intégré qui contrôle le cycle complet d'entrer dans, de traiter et d'éditer les livres électroniques et de papier sur l'Internet et l'Intranet.

9- L'école syrienne (SYRIE)

[8] *www.sakhr.com*

Un collège de quatre enseignants et chercheurs de l'université syrienne de Damas Ont édité en 1996, un dictionnaire électronique des verbes de l'arabe. Un ensemble d'études statistiques assez instructives y sont données. L'ouvrage en question concerne uniquement les verbes arabes, mais y sont annoncés d'autres ouvrages s'intéressant aux autres mots de la langue (tels que les noms, les adjectifs, ….

Il est regrettable que les auteurs n'entreprennent pas de décrire la manière avec laquelle le dictionnaire a été construit ou alors très brièvement. Nous ne savons notamment pas si des outils automatiques ont pu être mis en œuvre.

10-CNRS (IDL) (France) :

Plusieurs travaux sont menés dans l'équipe de recherche U.R.A (Unité de Recherche Associé) dans le laboratoire IDL (Laboratoire Informatique Droit et Linguistique) au sein du CNRS (Centre Nationale de Recherche scientifique) et portant sur le Traitement Automatique du Langage Naturel et ses applications.

Sous la direction du Dr. Fathi DEBILI, directeur de recherche au CNRS, l'équipe à réalisé certains travaux sur le TALN et a permis de faire soutenir un ensemble de thèses et d'articles sur la langue arabe.

VI. BIBLIOGRAPHIE

[Ant03] J-Y. Antoine, J. Goulian et J. Villaneau. Quand le TAL robuste s'attaque au langage parlé: analyse incrémentale pour la compréhension de la parole spontanée. Proceedings of TALN, Batz-sur-Mer, 2003.

[Ben01] Ben Sassi S., Braham R., Belgith A. 2001. Neural speech synthesis system for Arabic language using celp algorithm, Proc. IEEE Conference on Computer Systems and Applications, 119-121.

[Bla75] R. Blachère, M. Gaudefroy-Demombynes. Grammairede l'arabe classique, Maisonneuve & Larose, Paris, 1975.

[Bou01] Boula de Mareuil Philippe, Célérier Philippe, Cesses Thierry, Fabre Serge, Jobin Carine, Le Meur Pierre-Yves, Obadia David, Soulage Benoît, Toen Jacques. 2001. Elan text to speech : un système multilingue de synthèse de la parole à partir du texte. Elan TTS Toulouse.

[Bou02] C. Bousquet-Vernhettes. Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine – Décodage conceptuel stochastique. Thèse de l'université de Toulouse III, 2002.

[Bou92] Bouaissi Amel, Redjeb Lilia. 1992. Contribution à la synthèse de la parole arabe. Mémoire de fin d'études. Institut supérieur de gestion.

[Cal89] Calliope. 1989. La parole et son traitement automatique. Masson. Paris.

[Cat01] R. Cattoni, M. Federico and A. Lavie. Robust analysis of spoken input combining statistical and knowledge-based information sources. Proceedings of ASRU, Trento, 2001.

[Cha03] N. Chaabene et L. Belguith. L'étiquetage morpho-syntaxique: comment lever l'ambiguïté dans les textes arabes non voyellés ?. Proceedings des 3ème journées scientifique des jeunes chercheurs en génie électrique et informatique, Mahdia, Tunisie, 2003.

[Cha95] CHANOD, J.-P et TAPANAINEN,P. (1995) : "Creating a tagset, lexicon and guesser for a french tagger " In Proceedings of EACL SIGDAT workshop on From Texts To Tags: Issues In Multilingual Language Analysis.

[Dut93] Dutoit Thierry. 1993. High quality text to speech synthesis of the french language. Thèse. Faculté polytechnique de Mons.

[Els02] Elshafei M., Al-Muhtaseb, H., Al-Gamdi M. 2002. Techniques for high quality Arabic speech synthesis, Information sciences, Vol.140, 255-267.

[Eme77] Emerard Frençoise. 1977. Les diphones et le traitement de la prosodie dans la synthèse de la parole. Bulletin de l'institut de phonétique de grenoble.

[Gha04] Ghannouchi Ahlem, Charfi Aicha. 2004. Un système de concatenation de fichiers sonores. Mémoire de fin d'études. Faculté des sciences de Monastir.

[Gha90] S. Ghazali, H. Habaili, M. Zrigui. Correspondance graphème phonème pour la synthèse de la parole arabe à partir du texte, IRSIT. Proceedings of Congrès dialogue homme machine, Tunis, 1990.

[Gue83] Guerti Mhania. 1983. Contribution à la synthèse de la parole par diphones en arabe standard. Thèse. Institut de Linguistique et de Phonétique. Alger.

[Gue98] Guerti Mhania. 1998. Le principe de la synthèse de la parole. Ecole Nationale Polytechniques Alger. JTEA98.

[Hab76] H. HABAILI, (1976). Contraintes de structure morphématique en Arabe, DEA en linguistique, Canada, université de Montréal.

[Had04] A. HADDAD, (2004). Un système de génération automatique de dictionnaires linguistiques et thématiques de la langue arabe. Mastère en informatique, Ecole Nationale des Sciences de l'informatique, TUNISIE.

[HAD05] B. Haddad, M. Yaseen. A Compositional Approach Towards Semantic Representation and Construction of ARABIC. Proceedings of LACL, 2005.

[Jam04] S. Jamoussi. Méthodes statistiques pour la compréhension automatique de la parole. Thèse de doctorat de l'université Henri Poincaré, 2004.

[Jep04] JEP-TALN 2004, Fès, Maroc.

[JOH97] B. AL-JOHAR, J. MCGREGOR. A Logical Meaning Representation for Arabic (LMRA), Proceedings of 15th National Computer Conference, 1997.

[Kni01] S. Knight, G. Gorell, M. Rayner, D. Milward, R. Koeling, I. Lewin. Comparing grammar-based and robust approaches to speech understanding: a case study. Proceedings of European conference on speech communication and technology, 2001.

[Koh98] T. Kohonen. Self-organisation and associative memory. Spring-Verlag, Berlin,1998.

[Kso03] Ksouri Haifa, Zouari Imed. 2003. Constitution d'un dictionnaire d'unités acoustiques de l'arabe. Mémoire de fin d'études. Faculté des sciences de Monastir.

[Lef00] F. Lefèvre. Estimation de probabilité non paramétrique pour la reconnaissance markovienne de la parole. Thèse de l'Université Pierre et Marie Curie, 2000.

[Lem00] Lemmety Sami. 2000. Review of speech synthesis technology. Thèse. Helsinki University of Technology.

[Man96] C. MANKAÏ Naanaa. Compréhension automatique de la langue arabe. Application: Le système Al Biruni, Thèse de doctorat de l'université de Tunis II, 1996.

[McQ67] J. McQueen. Some methods for classification and analysis of multivariate observations, Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[MEF01] K. Meftouh, M.T. Laskri. Generation of the Sense of a Sentence in Arabic Language with a Connectionist Approach, Proceedings of AICCSA, 2001.

[Mer95] MERIALDO B, (1995):"Modèles probabilistes et étiquetage automatique" TAL, volume 36, 1995.

[Min99] W. Minker. Compréhension automatique de la parole spontanée. L'Harmattan, Paris, 1999.

[Mou73] A. H. MOUSSA, (1973). Statistical study of Arabic roots in moijam arous. Koweit .

[Mou96] Moulines Eric, Cappe Olivier. 1996. Synthèse de la parole à partir du texte, Techniques de l'ingénieur. H1960 pp 7.

[OUE01] R. Ouersighni. A major offshoot of the Dinar-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts. Proceedings of ACL/EACL, 2001.

[ROS94] R. Rosenfeld. Adaptive statistical language modelling: A maximum entropy approach, Thèse de doctorat de l'université de Carnegie Mellon, 1994.

[Sai02] Saidane Tahar, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2002. La Transcription Orthographique-Phonétique de la Langue Arabe. Tahar 2ème Conférence internationale JTEA.

[Sai04] Saidane Tahar, Haddad Ahmed, Zrigui Mounir, Ben Ahmed Mohamed. 2004. Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones. JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, Maroc.

[Sai04] Saidane Tahar, Zrigui Mounir, Ben Ahmed Mohamed. 2004. Constitution d'un dictionnaire de polyphones pour un système de synthèse de la parole arabe. SETIT 2004. Tunisie.

[Sai04] Saidane Tahar, Zrigui Mounir, Pr Ben Ahmed Mohamed. 2004. La Transcription Orthographique-Phonétique de la Langue Arabe. RÉCITAL 2004, Fès, Maroc.

[Sai04] T.SAIDANE, A.HADDAD, M.ZRIGUI, Pr. M. BEN AHMED, (2004). Réalisation d'un système hybride de synthèse de la parole arabe utilisant un dictionnaire de polyphones

[Sai05] T. Saïdane, M. Zrigui, et M. Ben Ahmed. Arabic speech synthesis using a concatenation of polyphones: the results. Proceedings of Canadian Conference on AI, Montréal, Canada, 2005.

[Sen92] S. Seneff. Robust parsing for spoken language systems, Proceedings of ICASSP, 1992.

[Sil93] M. SILBERZTEIN, (1993).Dictionnaires électroniques et analyse automatique de textes (Le système INTEX). (Masson, Paris)

[Van99] G. Van Noord, G. Bouma, R. Koeling, M.J Nederhof . Robust grammatical analysis for spoken dialogue systems, Natural Language Engineering 5(1), 1999.

[Zaa04] R ZAAFRANI,(2004). Un dictionnaire électronique pour apprenant de l'arabe (langue seconde) basé sur corpus. JEP-TALN 2004, Fès, Maroc.

[Zou04] A. Zouaghi, M. Zrigui, et M. Ben Ahmed. Une structure sémantique pour l'interprétation des énoncés. Proceedings of JEP-TALN, Fès, Maroc, 2004.

[ZOU05] A. Zouaghi, M. Zrigui, et M. Ben Ahmed. Un étiqueteur sémantique des énoncés en langue arabe, Proceedings of RECITAL, 2005.

[Zou06] A. Zouaghi, M. Zrigui, et M. Ben Ahmed. L'influence du contexte sur la compréhension de la parole arabe spontanée. Proceedings of TALN, Louvain, Belgique, 2006.

[Zri90] Zrigui M., Ghazali S., Ben miled Z., Jemni M. 1990. Synthèse de l'Arabe standard à partir du texte par TD PSOLA, 18ème journée d'étude sur la parole. Belgique.

[Zri91] Zrigui M., Mili A, Jemni M. 1991. Vers un système automatique de synthèse de la parole arabe, Maghrebin symposium on programming and system, Alger. p 180-197.

[Zri87] M. Zrigui, M. Truquet, B. Causse. Towards an arabic new Braille. 9ième congrès international de l'informatique et la langue arabe, Ryadh, Arabie Saoudite du 06 au 09/10 1986

[Zri02] M. Zrigui, M. Ben Ahmed. Une écriture phonologique compacte de l'arabe. Actes du 2ième colloque international d'informatique linguistique, pp 32 à 48, Kuwait du 27 au 30/11/2002

[Zri02] M. Zrigui, A. Mili, M. Jemni. An expert system for graphem-phonem conversion in arabic. Symposium on information technology & applications. Arab school of science and technology. Damascus du 25 au 31 mai 1992.

[Zri92] M. Zrigui, M. Achour. Speech synthesis from text. 1ier symposuim international sur la linguistique computationnelle. Caire, Egypt, du 19 au 23 Juin 1992

[Zri92] M. Zrigui, M. Achour. Nour: a multilingual system for the blind. 3ième congrès international pour la linguistique computationnelle; Durham, Grande Bretagne du 11 au 15 déc. 1992