UNIVERSITÉ LumiÈre LYON II
THÈSE pour obtenir le grade de DOCTEUR DE L'UNIVERSITÉ LYON II
Discipline : Sciences de l'Information et de la communication

DÉVELOPPEMENT D'UN ENVIRONNEMENT INTERACTIF D'APPRENTISSAGE AVEC ORDINATEUR DE L'ARABE LANGUE ÉTRANGÈRE

Riadh ZAAFRANI
18 Janvier 2002
Directeurs de thèse : M. Mohamed HASSOUN et M. Joseph DICHY

Jury : M. Jean Pierre DESCLÉS Président M. Mohamed HASSOUN Directeur M. Joseph DICHY Co-directeur M. Abdelfattah BRAHAM Rapporteur M. Jean-Paul METZGER Rapporteur M. Evehard DITTERS Examinateur

Table des matières

[19] Ghenima (1998), (Brahem & Ghazeli, 1998) et (Dichy, Hassoun, Zaafrani, 2002 d).

[24] Ghenima (1998), (Brahem & Ghazeli, 1998) et (Dichy, Hassoun, Zaafrani, 2002 a).

[25] Cf. (Dichy, Hassoun, Mouelhi, Zaafrani, 2002).

[28] Cf. (Dichy, Hassoun, Zaafrani, 2002 b)

[49] Le travail sur le dictionnaire arabe-arabe a été effectué par M. Abdelfattah BRAHAM (Maître de conférences à l'université de la Manouba en Tunisie).

[57] Les principes pédagogiques qui seront présentés dans cette section, sont synthétisés à partir d'informations récupérées du site de l'unité (TECFA), active dans le domaine des technologies éducatives et qui fait partie de la Faculté de Psychologie et des Sciences de l'Education de l'Université de Genève.: http://tecfa.unige.ch/themes (dernière consultation - Octobre 2001).

A toutes celles, à tous ceux qui m'ont donné la force de créer cette trace.

REMERCIEMENTS

Je remercie Monsieur Mohamed Hassoun, professeur à l'ENSSIB et Monsieur Joseph Dichy, professeur à l'université Lumière Lyon II, qui m'ont accepté dans leur équipe de recherche et qui ont assuré la direction de ma thèse.

Je voudrais remercier Monsieur Jean Pierre DESCLÉS, professeur à l'université de Paris, qui m'a fait l'honneur de présider le jury de cette thèse.

Je remercie Monsieur Abdelfattah BRAHAM, maître de conférences chargé de recherches à la faculté de lettres de la Manouba (Tunisie) et Monsieur Jean-Paul METZGER, Professeur à Lyon III, qui ont accepté d'être rapporteurs de cette thèse.

Je remercie Monsieur Evehard DITTERS professeur à l'université de Nimègue (Pays-Bas), qui a bien voulu examiner cette thèse et m'a fait l'honneur de participer au jury.

Je remercie Monsieur Xavier Lelubre, maître de conférences à l'université Lyon II, Monsieur Zoubeir MOUELHI, doctorant à l'université Lyon II, Monsieur Ammar MEDFAI, professeur d'arabe à l'Institut des Langues vivantes de Tunis et Mademoiselle Nacira GARBOUT, chercheuse linguistique à l'IRSIT, pour leur précieuse collaboration et pour l'intérêt qu'ils ont porté à ce travail.

Je remercie enfin toutes les personnes qui m'ont permis de mener à terme, ce travail de recherches.

Introduction générale

« LECTEUR, pour vivre bien content, lisez pour apprendre à bien vivre, et ne perdez point votre temps, à chercher les fautes d'un livre ; Il n'en est point de si parfait, où vous ne puissez reprendre ; Il n'en est point de si mal fait, en qui vous ne puissiez apprendre » Jean de LA RIVIERE (1721)

Cette thèse concerne l'élaboration d'un environnement informatique d'aide à l'apprentissage lexical et grammatical de l'arabe[1] langue seconde ou étrangère « *AL-Mu*^{c} *aLLiM* ». Notre travail porte essentiellement sur trois axes :

- l'élaboration de ressources linguistiques et d'outils informatiques pour le traitement automatique de la langue arabe,
- leur utilisation pour construire les ressources du système,
- l'individualisation de l'apprentissage notamment par la gestion d'un modèle de l'apprenant.

Cette thèse est issue des travaux effectués sur le traitement automatique de la langue arabe au sein de notre groupe de recherche SAMIA « **S** ynthèse et **A** nalyse **M** orpho-syntaxiques **I** nformatisées de l' **A** rabe en vue d'une application en enseignement assisté par ordinateur » (SAMIA, 1984) (Lelubre, 1985) (Dichy, 1987) (Abu Al-Chay, 1988) (Dichy & Hassoun, 1989) (Dichy, 1993) (Lelubre, 1993), sous la responsabilité de M. Hassoun (École Nationale Supérieure des Sciences de l'Information et des

[1] Il s'agit de l'arabe littéraire, avec insistance sur la langue moderne (à l'exclusion des dialectes).

Bibliothèques : ENSSIB) pour les aspects informatiques et de J. Dichy (Université Lyon II) pour les aspects linguistiques.

A partir de 1993, l'équipe SAMIA a signé une convention de partenariat avec l'Institut Régional des Sciences de l'Informatique et des Télécommunications de Tunis (IRSIT- S. Ghazeli et A. Braham - Voir Annexe 2). Cette collaboration a abouti à la réalisation de la base de données lexicale de l'arabe DIINAR.1 « DIctionnaire INformatique de l'ARabe » (Hassoun, 1987) (Dichy, 1990) (Gader, 1996) (Dichy, 1997) (Ghenima, 1998) (Braham & Ghazeli, 1998).

De notre côté, nous avons débuté notre travail de recherches, en réalisant un prototype d'une base de données des formes verbales et déverbales de l'arabe en vue d'une application en EAO, dans le cadre de notre mémoire de maîtrise (Zaafrani & Ouersighni, 1993). Nous avons, par la suite, exploité un sous-lexique généré à partir de cette base de données, pour concevoir un premier environnement d'apprentissage dans le cadre de notre mémoire de DEA (Zaafrani, 1994).

Après la soutenance de ce mémoire, nous avons entamé cette thèse par une étude exploratoire du domaine de l'Apprentissage des Langues Assisté par Ordinateur (ALAO). Nous avons notamment constaté que l'**individualisation** de l'apprentissage (l'espace et le temps de l'apprentissage au choix de l'apprenant, un rythme de progression adapté, une participation intensifiée de l'apprenant, une évaluation personnalisée et en temps réel) reste le rôle le plus assigné à l'ordinateur dans les environnements de l'ALAO. Cette individualisation passe obligatoirement par la mise à la disposition des apprenants de ressources linguistiques informatisées propres à leur fournir les matériaux d'apprentissage et les outils d'aide et par l'articulation de ces moyens autour d'un programme pédagogique pertinent.

Nos premiers travaux ont relevé par conséquent de la construction matérielle de cette énorme masse d'informations linguistiques pour la langue arabe. Nous avons pris en charge la maintenance de la base de données lexicale (DIINAR.1) et la réalisation d'une série d'applications de Traitement Automatique de la Langue arabe. Notre participation aux travaux de construction et d'amélioration du dictionnaire DIINAR.1 recouvre trois phases d'activités distinctes :

- Dans la première phase, nous nous sommes familiarisés avec la conception linguistique et informatique de DIINAR.1 et nous l'avons fait émigrer vers un nouvel environnement informatique afin de faciliter son exploitation.
- Dans la deuxième phase, nous avons apporté quelques modifications à la structure de la base de données DIINAR.1 afin de répondre aux besoins des applications de TAL qui l'utilisent. Notre travail a consisté notamment, en l'association de nouveaux spécificateurs linguistiques aux entrées lexicales nominales et verbales et l'ajout de deux nouvelles parties à la base de données pour gérer les mots outils et les noms propres.
- Dans la troisième phase qui s'est déroulée dans le cadre du projet DIINAR-MBC (**DI** ctionnaire **IN** formatisé **Ar** abe - **M** ultilingue et **B** asé sur **C** orpus) (Voir Annexe 3), nous avons intégré de nouveaux spécificateurs de nature syntaxique et sémantique

sur un lexique réduit de DIINAR.1[2]. Nous nous sommes ainsi occupés de la réalisation informatique du nouveau dictionnaire prototype multilingue **PROLEMAA** (**Pro**totype de **Le**xique **M**ultilingue **à** partir de l'**A**rabe), qui associe aux entrées lexicales arabes du dictionnaire leurs homologues en français et en anglais avec leurs spécificateurs linguistiques respectifs.

La réalisation de ces ressources s'est effectuée en harmonie avec le développement d'applications de TAL de l'arabe. Nous avons d'abord construit un générateur automatique qui permet une production paramétrée du lexique et surtout un analyseur morpho-syntaxique fonctionnant à partir du lexique généré. Nous avons ensuite développé un certain nombre d'**applications** qui se basent sur les résultats retournés par l'analyseur :

- Un outil d'étiquetage morpho-syntaxique de textes bruts.
- Un concordanceur qui permet de sélectionner des exemples d'utilisation d'une unité lexicale ou d'une catégorie syntaxique dans un corpus textuel donné ou de générer des activités d'apprentissage.
- Un programme pour le calcul des fréquences des unités lexicales, qui a permis de sélectionner le lexique de PROLEMAA.

La partie centrale de notre travail a été de concevoir un environnement adapté à un apprentissage autonome du lexique et de la grammaire arabe pour des apprenants étrangers de niveaux différents. L'orientation de l'apprentissage vers le lexique s'appuie sur les récents travaux en psycholinguistique, qui ont montré qu'une bonne connaissance du lexique est fondamentale pour la réelle maîtrise d'une langue et permet d'améliorer les différentes compétences de l'apprenant : morphologiques, syntaxiques et sémantiques.

Les théories psycholinguistiques sur le lexique mental, ont indiqué aussi que la maîtrise lexicale passe par la structuration et la réflexion sur le vocabulaire et l'entraînement et l'exécution d'activités lexicales. Dès lors, nous avons conçu un environnement d'apprentissage qui fonctionne autour d'un schéma d'apprentissage en trois volets : exposition / compréhension de textes, rétention du lexique et maîtrise de la grammaire.

Pour traiter le volet de l'exposition / compréhension, nous avons circonscrit notre cadre d'étude sur quelques textes qui pourront être par la suite enrichis. En effet, la couverture de la langue entière impose la constitution d'un corpus de textes de très grosse taille ainsi qu'un dictionnaire couvrant les sens de tous les mots de ce corpus. Nous nous sommes ainsi contentés de quelques textes que nous avons étiquetés par des informations d'ordre morpho-syntaxiques et ajouter les différents sens des mots manquants dans le dictionnaire et leurs correspondants dans les langues cibles des apprenants.

Comme nous le montrerons dans cette thèse, un corpus textuel bien étiqueté permet de résoudre les problèmes d'accès lexical et de compréhension de textes. Notre travail a consisté dans ce volet, à faciliter le passage de l'apprenant du texte au dictionnaire et vice-versa et à profiter des possibilités de l'informatique pour lui permettre une navigation

[2] Le lexique choisi correspond aux lemmes les plus fréquents du corpus textuel sélectionné dans le cadre du

mieux organisée à l'intérieur du dictionnaire.

Pour ce qui est du deuxième volet de notre environnement qu'est la mémorisation du lexique, nous avons conçu une interface qui offre la possibilité à l'apprenant d'organiser son vocabulaire dans un *dictionnaire personnel*. Il est en effet admis que la quantité du travail sur le lexique facilite sa rétention. Le fait que l'apprenant structure lui-même son propre vocabulaire implique une organisation plus profonde de celui-ci. L'effort mental ainsi généré est bénéfique pour l'incorporation de nouvelles connaissances au sein des anciennes et favorise de ce fait la rétention du vocabulaire. En plus, nous nous appuyons sur le dictionnaire personnel de l'apprenant pour personnaliser les activités lexicales et grammaticales.

Le troisième volet de notre schéma concerne la maîtrise des règles grammaticales. Nous avons construit un module qui permet à l'enseignant d'organiser les leçons grammaticales et de définir des activités génériques. Pour cela, nous avons défini quelques maquettes d'activités adaptées au média informatique et aux ressources de notre environnement. Ces activités sont générées automatiquement à partir d'un *modèle de l'apprenant* qui permet de synthétiser toutes les interactions apprenant-environnement et de suivre l'évolution de l'apprenant.

Ce travail est composé de neuf chapitres au cours desquels nous décrivons progressivement l'environnement d'apprentissage réalisé. Après un premier chapitre retraçant l'historique du domaine de l'ALAO, nous présentons les différentes ressources de l'environnement.

Le deuxième chapitre détaille la conception et la réalisation informatique de la base de données lexicale DIINAR.1, dont les nouvelles parties ajoutées pour gérer les mots outils et les noms propres. Dans ce chapitre, nous décrivons aussi le travail de maintenance effectué sur cette base de données et qui a permis d'améliorer les performances des différentes applications de TAL attachées.

Le troisième chapitre présente les processus d'analyse des mots graphiques et de génération automatique du lexique DIINAR.

Le quatrième chapitre décrit le fonctionnement des différentes applications développées à partir de l'analyseur du mot graphique. Ces applications ont servi notamment à définir le dictionnaire PROLEMAA et à étiqueter les textes utilisés dans l'environnement d'apprentissage.

Après ces quatre premiers chapitres, consacrés à la définition des ressources et des outils du système, nous exposons dans le cinquième chapitre les principes linguistiques et les pratiques pédagogiques retenues qui nous ont guidé pour l'élaboration des différentes composantes de l'environnement d'apprentissage.

Le sixième chapitre décrit la conception et la réalisation du dictionnaire électronique pour apprenant de l'arabe langue étrangère. Nous montrons notamment en quoi ce dernier est différent du dictionnaire classique sur papier et comment il permet de résoudre le problème d'accès lexical.

Le septième chapitre traite de l'implémentation des activités lexicales et grammaticales. Nous expliquons d'abord les principes de conception d'une activité

informatisée et nous décrivons ensuite le processus de génération automatique de ces activités à partir des ressources du système (base de données lexicale, corpus de textes étiquetés, dictionnaire général, modèle de l'apprenant, dictionnaire personnel).

Le huitième chapitre décrit le modèle de l'apprenant. Nous déterminons à la fois les informations que doit contenir le modèle et le diagnostiqueur qui est l'ensemble des processus qui l'élaborent et le mettent à jour.

Enfin, le neuvième et dernier chapitre récapitule l'architecture de l'environnement « *AL-Mu*^{c} *aLLiM* » et décrit les modules qui n'ont pas été étudiés dans les chapitres précédents : le module de recherche et du choix du texte, le module de compréhension d'un texte, le module de gestion du dictionnaire personnel de l'apprenant et le module de l'enseignant.

Chapitre 1 L'Apprentissage des Langues Assisté par Ordinateur : Bilan et perspectives

" La société vers laquelle nous courons tous est celle de l'amusement, comme mode privilégié d'accès au monde. C'est aussi une des raisons pour lesquelles on est souvent tenté d'introduire les nouveaux médias à l'école : on se dit que grâce à ces nouvelles techniques, la transmission deviendra un jeu et que les enfants qui aiment l'écran aimeront sur l'écran ce qu'on leur enseigne. Seulement voir n'est pas savoir, savoir n'est pas penser." A.Finkielkraut (1995) Le nouvel observateur, supplément au N°1618, p. VIII

1.1 Introduction

Nous avons choisi d'introduire cette thèse par un chapitre sur l'historique de l'Apprentissage des Langues Assisté par Ordinateur (ALAO), afin de situer notre travail par rapport au nombre considérable de réalisations qui se font dans ce domaine. En effet, le champ de l'ALAO a été forgé par une multitude de théories, de disciplines et de réalisations touchant différentes compétences linguistiques, ce qui a rendu tout travail de circonscription du domaine compromis :

- L'ALAO est nourri par de nombreuses disciplines : intelligence artificielle (IA), linguistique computationnelle (i.e. Traitement Automatique des Langues naturelles (TAL), traduction assistée par ordinateur, traitement de corpus textuels, etc.), ingénierie didactique, interaction homme-machine, etc.
- L'ALAO met en jeu plusieurs théories : théorie comportementale, théorie constructiviste, théories en acquisition d'une langue seconde, psychologie cognitive, psycholinguistique, etc.
- Les *compétences en langues* sont constituées d'un ensemble complexe de compétences interconnectées (phonologiques, morphologiques, syntaxiques, lexicales, communicatives, sociolinguistiques, etc.)
- L'ALAO recouvre différents types d'activités : test de connaissances, renforcement, mémorisation, activités de découverte, compréhension, production, etc.

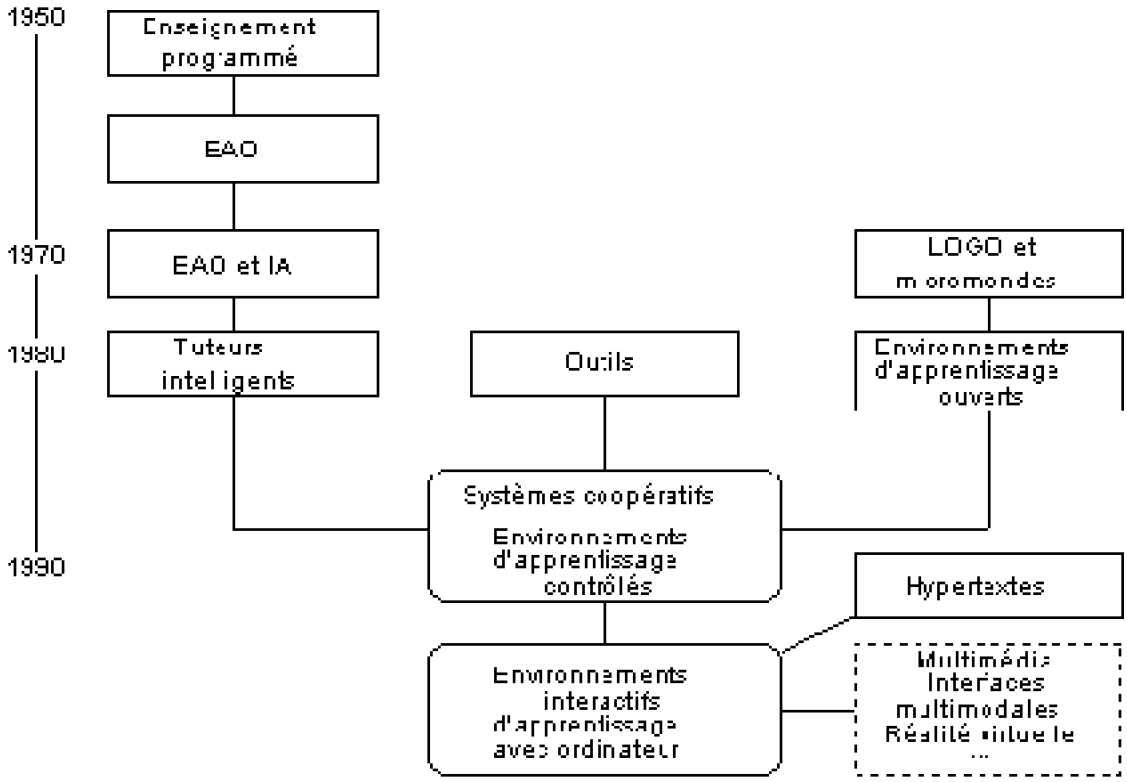
L'histoire de cette jeune discipline[3] peut être présentée de différentes manières selon l'angle où nous nous plaçons. Nous avons choisi d'aborder cette histoire à travers le rôle joué par la machine dans l'apprentissage, ce qui correspond à trois types de modèles d'apprentissage :

- Les modèles de type *transmissif* correspondant aux logiciels issus de l'enseignement programmé et aux tuteurs intelligents.
- Les modèles de type *découverte* correspondant aux micromondes et aux environnements d'apprentissage ouverts.
- Les modèles de type *exploratoire* correspondant aux systèmes hypermédias et aux environnements coopératifs.

Chaque modèle d'apprentissage correspond à une époque avec des théories, des pratiques et des technologies différentes. Dans cette rétrospective, nous ne chercherons pas à être exhaustif, mais à présenter, pour chaque compétence linguistique qui a été touchée un exemple de réalisation. La figure (1-1) ci-dessous (extraite de Bruillard, 1997, p. 24) donne un aperçu des principales étapes qui ont marqué l'histoire de l'informatique en éducation.

Dans ce schéma, la première colonne réfère à des courants qui ont privilégié l'enseignement, c'est-à-dire pour lesquels la machine est principalement amenée à jouer le rôle du maître (de l'enseignement programmé aux tuteurs intelligents). La troisième colonne rend compte de recherches dans lesquelles l'ordinateur est un moyen permettant aux apprenants d'effectuer constructions, explorations et découvertes. La colonne centrale correspond aux tentatives de synthèse de ces deux approches, essayant de concilier un certain guidage par la machine dans des environnements largement contrôlés par les apprenants. Bien entendu, un tel découpage est schématique et réducteur et ne correspond que très imparfaitement à ce qui s'est passé, les idées ayant traversé les différents courants.

[3] Au bout de plusieurs décennies d'existence, le champ de l'ALAO n'a toujours pas atteint le statut de discipline (Levy, 1997).



En parcourant les trois subdivisions de l'historique de l'ALAO, nous essaierons de dégager les apports et les limites des différents systèmes rencontrés et de définir les voies à travers lesquelles l'informatique pourra assister et/ou modéliser l'apprentissage des langues. En conclusion de ce chapitre, nous essayerons de synthétiser tous les concepts et les intuitions de ces années passées, afin de définir les bases d'un "Environnement interactif d'apprentissage par ordinateur de l'arabe langue seconde", tel que nous l'entendons.

1.2 Les modèles de type transmissif : l'ordinateur comme tuteur

Dans les années 50, sous l'impulsion de la cybernétique et de la psychologie comportementale, une technologie de l'enseignement s'est implantée, c'est-à-dire la mise en œuvre des méthodes scientifiques et des connaissances sur les processus d'enseignement en vue d'atteindre des buts éducatifs précis et contrôlables. Si au début l'ordinateur ne servait qu'à être un support à l'enseignement programmé, il s'est transformé plus tard en une machine adaptative à l'apprenant grâce notamment aux techniques de l'Intelligence artificielle.

1.2.1 Les didacticiels

Ce sont les classiques logiciels d'Enseignement Assisté par Ordinateur (EAO) qui mettent en situation, plus ou moins interactive, un apprenant et un problème à résoudre. La gamme des activités possibles est assez vaste mais chaque séquence est fermée par un choix de réponses restreint à celles qui sont interprétables par le programme.

Ces logiciels favorisent donc peu l'initiative de l'apprenant et l'apprentissage consiste pour le sujet à mémoriser certains concepts.

Cette manière de concevoir des didacticiels s'explique par le fait que l'EAO a été très fortement marqué par l'enseignement programmé[4] et a été inspiré des machines à enseigner cybernétiques et de la psychologie béhavioriste. Dans ces machines, l'apprentissage consistait à associer des conduites à d'autres conduites innées (réflexes) ou préalablement acquises. Cette association se faisait par un renforcement systématique. Quatre grands principes ont été mis en évidence par les chercheurs qui ont travaillé sur l'enseignement programmé (Bruillard, 1997) :

- Structuration de la matière à enseigner : la matière est décomposée en unités élémentaires, il faut fragmenter les difficultés suivant le principe des petits pas.
- Adaptation : la progression s'effectue par petites étapes et le rythme de progression est celui de l'élève. Un enseignement programmé doit être expérimenté jusqu'à ce qu'il "marche".
- Stimulation : participation active de l'élève, sollicité par des questions auxquelles il doit fournir une réponse effective, qu'elle soit construite ou uniquement choisie parmi plusieurs proposées. C'est le principe du conditionnement opérant mis en valeur par Skinner. Selon Skinner, l'apprenant est en mesure de répondre correctement aux questions qui lui sont posées, s'il suit convenablement le cours. Skinner plaide donc pour un schéma uniséquentiel. Cette conception oblige à supprimer toute difficulté pour qu'aucun obstacle insurmontable ne soit rencontré par l'apprenant.
- Contrôle et connaissance immédiate de la réponse. Un comportement nouveau s'acquiert plus rapidement s'il y a renforcement (ou feed-back) immédiat et, si possible positif.

Les relations entre l'apprenant et la machine seront de ce fait longtemps stéréotypées et les schémas assez réducteurs de l'EAO initial seront amenés plus tard à évoluer. Les sept situations suivantes représentent le travail interactif d'un apprenant au cours d'une séquence d'EAO idéale (Weidenfeld et alii, 97, pp. 108) :

- Information : le système interactif présente un ensemble d'informations ;
- Sollicitation : le système pose, une question ou propose un exercice. La nature de la question ou de l'exercice est largement conditionnée par la capacité d'analyse des réponses (4) ci-dessous.
- Résolution de problème : l'apprenant utilise les différentes fonctions mises à sa

[4] Par rapport à l'EAO, l'enseignement programmé constitue une méthode d'enseignement indépendante de tout support, de tout mode de présentation. Les supports utilisés sont des livres (manuel programmé ou livre brouillé), les fiches et les machines.

disposition au sein du didacticiel pour élaborer sa solution. La séquence d'élaboration de solution se termine :

– soit à l'initiative de l'apprenant qui « valide » sa réponse au problème posé.
– soit à l'initiative du système par le déclenchement d'un signal de surveillance.

· Analyse de réponse : le travail effectué par l'apprenant est comparé aux différentes attentes du système. Ce sera le programme d'analyse de réponse.
· Diagnostic : le résultat de la comparaison dépend lui aussi de la sophistication du programme de diagnostic consécutif à l'analyse précédente. Dans les cas très sophistiqués, il peut être utilisé pour comprendre la nature de l'erreur en vue de générer une explication (6, ci-dessous) ou de mettre à jour le *profil de l'apprenant* afin d'assurer un parcours personnalisé.
· Explication : le système adresse à l'apprenant un *commentaire textuel* sur son activité. Les commentaires sont généralement statiques et prédéfinis mais des systèmes plus sophistiqués (utilisant des techniques de l'intelligence artificielle) peuvent produire des commentaires individualisés adaptés à l'apprenant, voire exploiter le profil de l'apprenant pour redéfinir la tactique ou la stratégie la mieux adaptée pour la suite du travail de l'apprenant.
· Gestion du parcours pédagogique : si le didacticiel n'est pas terminé, on boucle à nouveau sur un autre cycle.

1.2.2 Les langages-auteur

L'arrivée des micro-ordinateurs dans les années 80 a totalement modifié la donne, puisqu'il devenait possible à tout un chacun de 'mettre en machine' ses propres projets pédagogiques, si modestes soient-ils. Pour permettre à des auteurs (non informaticiens) de concevoir et mettre au point un didacticiel, à l'aide de trames d'exercices préétablies, des langages de programmation appelés langages-auteur[5] ont vu le jour. A partir de la séquence d'EAO idéale (cf. § 1.2.1), nous pouvons aisément décrire les principales fonctionnalités offertes par les langages-auteur (Weidenfeld et alii, 97, pp. 109-111) :

· **Information :** affichage de textes et parfois de graphiques.
· **Sollicitation :** question générale textuelle, réponse à fournir au clavier (Activités à réponses ouvertes) ou choix parmi des réponses préétablis (Activités à réponses fermées).
· **Résolution de problème :** L'apprenant utilise les différents outils mises à sa disposition au sein du didacticiel (retour en arrière, aide en ligne, etc.) pour élaborer sa solution. En général, dans un système d'EAO ce n'est pas un **problème** qui est

[5] Parmi les systèmes-auteur les plus diffusés actuellement nous citons **Macromedia Authorware** basé sur des interactions très classiques mais qui permet aisément d'intégrer du multimédia.

posé mais un exercice dont la résolution ne requiert pas la mise en œuvre d'un **raisonnement**. L'enseignement de notions nécessitant la mise en œuvre de raisonnement, suppose que le système soit en mesure de résoudre lui-même le problème, ce qui nécessite des techniques plus élaborées issues de l'IA.

- **Analyse de réponse :** les réponses attendues dans un EAO classique sont pré-définies. Beaucoup de fonctions des langages auteur ont consisté en des traitements linguistiques destinés à prévenir les erreurs de frappe, fautes d'orthographe, etc. Mais peu d'analyses de réponses se sont avérées assez fiables pour éviter les deux écueils mortels de l'EAO :

  – accepter comme exacte une réponse erronée,
  – refuser une réponse exacte mais non prévue explicitement par le concepteur.

- **Diagnostic :** Le diagnostic s'effectue par référence à un ensemble d'erreurs types explicitées au moment de l'analyse pédagogique. Dans les versions les plus courantes les réponses types sont au nombre de trois :

  – les réponses **exactes** ou à tout le moins acceptables.
  – les erreurs **fécondes**, essentiellement prévisibles, pour lesquelles un commentaire adapté est susceptible d'un fort impact pédagogique.
  – les réponses **totalement erronées**, pour lesquelles tout commentaire est superflu.

- **Explication :** cette fonction est très liée au diagnostic. En général dans les systèmes d'EAO bien conçus, les explications sont prédéfinies par l'auteur, à partir d'une typologie des erreurs *intéressantes* (i.e. fécondes - Voir ci-dessus). Lorsque le système informatique est censé reproduire le raisonnement humain, il est tentant d'attendre de sa part une explication automatique des erreurs. Certains systèmes experts le réalisent mais les résultats sont généralement décevants.
- **Gestion du parcours pédagogique :** la réussite d'un apprenant dans un exercice permet de déduire son niveau de maîtrise des notions associées à cet exercice et fournit une image de ses connaissances à un instant donné. En théorie, on peut donc proposer le nouvel exercice le plus adapté à un apprenant donné à un instant donné. Ceci suppose que l'ensemble des moyens de s'approprier d'un domaine de connaissances donné soit bien explicité : Quels sont les notions qui constituent ce domaine ? Quels sont les prérequis entre ces notions ? Quels sont les parcours *naturels* permettant d'accéder à la connaissance d'une notion ? Etc. Ces questions relèvent d'une expertise pédagogique qui n'est généralement pas facile à produire.

1.2.3 Les limites de l'EAO

Certains des initiateurs de l'EAO, fascinés par l'ordinateur, ont pu clamer que *celui-ci allait remplacer les maîtres*.L'expérience a néanmoins montré que la présence d'un formateur

dans le processus est requise, au moins pour préconiser le didacticiel.

Plusieurs types de considérations ont limité l'usage de l'EAO et contribué à sa diversification et à son renouvellement.

- Le modèle pédagogique de l'EAO est centré autour du maître. En fait les dispositifs de l'EAO sont calqués sur un *modèle classe* : un maître enseigne des notions à des élèves ou leur pose un exercice ; puis il détermine leur prochaine activité. Il existe d'autres pratiques pédagogiques, plus centrées sur l'apprenant comme l'approche constructiviste ou l'approche communicative en didactique des langues (voir ci-dessus).
- La forme stéréotypée de l'interaction pédagogique EAO la prédispose inégalement aux différents domaines de connaissance. Ainsi l'EAO pourra être adapté à l'acquisition de connaissances factuelles (règles grammaticales) et de connaissances **procédurales simples**. Par contre l'EAO est inadapté à des **procédures complexes** ou à une **documentation importante**. Les méthodes de l'EAO ne permettent plus alors de rendre compte des **processus cognitifs** mis en œuvre et il faut avoir recours à d'autres approches.
- La frontière entre l'EAO traditionnel et d'autres approches n'est pas toujours nette. Ainsi certaines analyses de réponses prennent en compte plusieurs solutions et de ce fait rejoignent les approches faisant appel à l'IA. A contrario, la forme souvent très rudimentaire de l'analyse de réponse de maints langages auteurs n'autorise que des vérifications de connaissances assez sommaires.

Nous retrouvons toutes ces limites dans les didacticiels d'enseignement des langues, dont les auteurs conçoivent généralement des activités d'enseignement très conventionnelles autour d'**exercices structuraux** de **grammaire sans contexte**. Les interactions apprenant-système sont très limitées et visent surtout à sanctionner les connaissances de l'apprenant. Les didacticiels construits sont donc surtout des environnements de test des connaissances de l'apprenant, et sont abusivement qualifiés de support d'apprentissage (Chanier, 92).

1.2.4 EAO et intelligence artificielle : Apparition des tuteurs intelligents

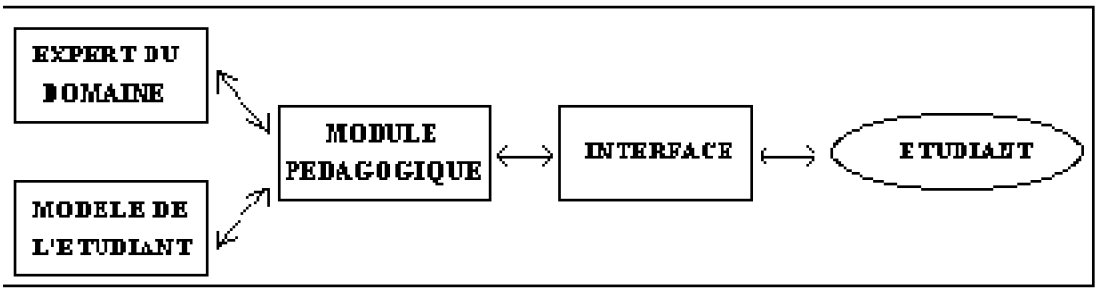
Bien que consacrant une certaine rupture avec la théorie comportementale, les recherches sur les tuteurs intelligents vont prolonger les travaux développés dans le cadre de l'enseignement programmé et l'EAO et essayer de tirer parti au mieux de cette machine adaptative qu'est l'ordinateur. Grâce aux techniques d'intelligence artificielle (IA), il devient possible de doter les machines de connaissances et de certaines capacités dont elle est en mesure de les utiliser. Les écueils dus à la rigidité des programmes (questions et réponses préenregistrées), à la représentation de l'élève trop rudimentaire et à la représentation de la matière beaucoup trop atomisée, devraient pouvoir être évités.

Ainsi, à la fin des années 1970, des systèmes informatiques dits Tuteurs Intelligents (TI) (WENGER, 87) destinés à l'enseignement et basés sur des techniques de

représentation de connaissances se sont développés. Ces systèmes sont dotés de trois types de connaissances :

- d'abord, à la différence d'un système d'EAO, le système connaît le contenu à enseigner dans la mesure où il est capable de résoudre les problèmes qu'il propose au sujet ou que le sujet lui propose, ceci ayant lieu la plupart du temps grâce à une base de connaissances dite experte,
- ensuite, le système est capable d'inférer l'état des connaissances du sujet en train de résoudre un problème donné. C'est la construction en ligne du "modèle" de l'apprenant à partir de sa performance,
- enfin, la stratégie pédagogique du tuteur est de s'adapter au sujet en produisant un guidage ou une action qui réduise la différence entre les connaissances de l'apprenant et celles du système.

Ces trois types de connaissances vont se traduire par un système composé de trois modules et d'une interface (figure 1-2). Les trois modules sont respectivement le **module expert** du domaine, le **module** diagnostic des connaissances **de l'apprenant** et le **module tutoriel** ou **pédagogique**. L'**interface**, à côté des trois modules, est centrale dans un tuteur intelligent. Elle doit faciliter l'interaction Apprenant-Machine. Si elle est mal conçue, l'efficacité des trois modules peut alors être altérée.



1.2.5 l'EIAO et l'enseignement des langues

L'Enseignement Intelligemment Assisté par Ordinateur (EIAO) apporta de nouvelles perspectives à l'ALAO. Mais la double complexité de la langue naturelle et d'une construction didactique visant à faire acquérir une langue étrangère n'a pas favorisé l'apparition de beaucoup de tuteurs intelligents dans ce domaine.

Les systèmes d'EIAO qui ont émergé sont essentiellement issus d'aspects "résolution de problème". Ces systèmes permettent par exemple d'acquérir la démarche optimale pour rechercher une panne. La compréhension des interactions causales entre divers phénomènes a aussi fait l'objet de quelques réalisations : "Pourquoi pleut-il ? Le climat de telle région sera-t-il favorable à la culture du riz et pourquoi ?".

Pendant un enseignement de langue, l'apprenant n'est pas confronté à un problème à résoudre à l'aide d'un algorithme à conscientiser et à maîtriser. Il ne se trouve pas non plus devant un réseau de relations causales qu'il faut démêler pour expliciter ou justifier

un phénomène. L'EIAO semblait donc ne pas convenir au domaine de l'ALAO.

L'accent mis sur le développement d'un savoir linguistique expert, concentra alors les travaux sur l'acquisition de compétences grammaticales autour d'énoncés détachés d'une réelle situation communicative[6].

Le degré d'individualisation offert par les logiciels resta cependant très limité, bien que ce point fût l'un des principaux sujets de controverse entre EAO et EIAO. Les modèles d'apprenants gérés automatiquement par les systèmes étaient sommaires et prenaient peu en compte les processus cognitifs mis en jeu dans l'acquisition d'une langue (Chanier, 92).

S'appuyant sur une critique facile des logiciels d'EAO existants, certains articles présentant des tuteurs intelligents (TI) laissaient entendre qu'il serait facile de doter les systèmes informatiques de larges aptitudes de compréhension du langage et de possibilités de conduite de dialogues très interactifs. Ces affirmations qui sous-estimaient largement les difficultés, en laissant croire prématurément à une intégration des tuteurs dans les cursus d'enseignement, furent battues en brèche par des chercheurs qui présentaient des logiciels d'EAO mieux conçus qu'auparavant (Chanier, 92).

1.3 les modèles de type "découverte" : l'ordinateur comme apprenant

Dès la fin des années soixante, un second courant s'écarte de l'idée consistant à faire de l'ordinateur un super-enseignant, en essayant de le promouvoir comme moyen d'expression et d'expérimentation pour les apprenants. Il ne s'agit plus que l'ordinateur programme l'apprenant mais de donner à l'apprenant la possibilité de programmer la machine. La cybernétique et l'enseignement programmé ont dégagé l'idée de contrôle, mais ce contrôle, au lieu d'être confié à une machine, peut être rendu à l'apprenant.

1.3.1 Les micromondes

Rendre le contrôle à l'apprenant pose la question d'un langage ou d'une forme de communication simple avec la machine. Les langages issus des recherches en IA, conçus pour faciliter l'expressivité afin d'explorer des problèmes complexes, s'avèrent de bons candidats. Il n'est pas question de se cantonner à l'apprentissage de la programmation à l'aide d'un langage spécifique, mais plutôt de faciliter des apprentissages par la programmation. Sous l'impulsion de Papert émerge alors la notion de *micromonde* (Papert, 1981).

Les micromondes sont, comme des langages de commande, des systèmes informatiques ouverts. L'apprenant peut explorer un domaine ou un dispositif avec un

[6] Voir quelques exemples de réalisations dans le livre de (Demaizière, 86, pp. 54-56).

minimum de contraintes de la part du système. Il lui permet d'agir sur des objets, dont le comportement respecte certaines contraintes de fidélité et de cohérence.

L'exemple le plus connu de "micromonde" est, sans aucun doute, la géométrie-tortue[7] de LOGO[8], dans lequel l'explorateur peut "apprendre" à un robot virtuel à réaliser toute sorte de tâches (dessins) en le programmant. L'objectif pédagogique assigné à ces environnements est souvent ambitieux. L'apprenant est censé apprendre à apprendre ; il se sert de l'environnement pour "réfléchir" ses connaissances et construire de nouveaux objets de savoir. Du point de vue informatique, la conception de ces systèmes est similaire à celle d'un langage de programmation de haut niveau. L'apprentissage de tels langages est de type "constructiviste" car il s'agit en effet pour l'apprenant de construire des objets de plus en plus complexes à partir de "schémas élémentaires" appelés "primitives" et d'une grammaire permettant de créer ces assemblages.

Outre LOGO, d'autres langages conçus comme des supports à l'apprentissage par la programmation ont émergé. C'est le cas des extensions de langages de programmation traditionnels, comme les extensions de LOGO et plus particulièrement celles à base d'objets, SMALLTALK, PROLOG... Ces langages informatiques, en eux-mêmes, n'ont pas de vertus éducatives particulières, mais ils respectent certaines contraintes ou intègrent certaines caractéristiques permettant de développer des activités qui, elles, sont intéressantes du point de vue éducatif. C'est bien dans des façons spécifiques d'exploiter les possibilités offertes par ces langages que l'on peut espérer développer des capacités souhaitées chez les apprenants.

1.3.2 Le LOGO et l'apprentissage des langues

Papert compare l'apprentissage des mathématiques tel qu'il le souhaite à l'apprentissage "sans douleur" d'une langue vivante en milieu naturel : "L'idée de 'parler mathématique' avec un ordinateur peut être élargie à celle d'apprendre les mathématiques en 'Mathématie', tout comme on apprendrait l'italien en Italie" (papert, op. cit., pp. 27).

Cette comparaison peut susciter l'illusion d'un passage facile à l'apprentissage d'une langue vivante grâce à Logo, et l'on entend beaucoup dire que les micromondes de Logo pourront s'appliquer en linguistique, en grammaire, etc. La transition, n'est cependant pas si simple, même si l'on considère que l'idée d'une initiative heuristique à laisser à l'apprenant par rapport à des méthodes de type "dressage répétitif", peut et doit être retenue (Demaizière, 1986, pp. 58).

Les exemples d'utilisation de Logo donnés pour la grammaire sont certes intéressants mais restent limités. Il s'agit d'exercices de construction automatique

[7] Afin de rendre l'ordinateur plus intéressant pour les enfants, Papert adjoint la fameuse tortue au LOGO. Il y voyait un moyen de capturer dans une forme computationnelle quelque chose de physique, analogue au fait de marcher et de dessiner.

[8] LOGO est un langage de programmation basé sur LISP, mis au point dans le courant des années soixante dix par Seymour Papert au MIT (*Massachusetts Institute of Technology*), dans le but d'aider les élèves à développer une compréhension plus profonde d'idées mathématiques jugées importantes.

d'énoncés par l'ordinateur à partir d'une structure syntaxique donnée par l'apprenant, et ce par un choix aléatoire dans des listes de mots mises en machine. De tels exercices font apparaître l'intérêt du classement des mots en catégories grammaticales. On n'a pas pour autant résolu des problèmes fondamentaux en langue, se rapportant à des concepts somme le temps, l'aspect, la modalité, les repérages situationnels, le point de vue de l'énonciateur, etc.

1.3.3 La simulation et l'enseignement des langues

L'ordinateur a été souvent utilisé pour simuler une expérience scientifique ou une situation impliquant diverses analyses et prises de décision (diagnostic médical, gestion d'entreprises, etc.). Ces programmes de simulation peuvent servir aussi comme un support d'apprentissage. L'apprenant se verra présenter divers résultats en fonction des demandes qu'il aura faites à la machine ou des données qu'il aura fournies. L'avantage est que le résultat peut être fourni immédiatement et sans risques et que l'apprenant peut ainsi faire de nombreux essais facilitant son apprentissage dans des conditions impossibles à remplir autrement. Une expérience demandant plusieurs heures (ou plusieurs années) ou impliquant des risques (explosion …) peut être réalisée en quelques secondes.

Dans le cadre de l'enseignement des langues, on constate, d'une part, que l'on parle beaucoup de simulation dans les milieux concernés et, d'autre part, que son domaine n'est pas spécialement touché, au niveau des réalisations existantes.

Cette absence s'explique par le fait que l'on ne voit pas d'emblée quelles activités de langue pourraient être simulées efficacement grâce à l'ordinateur. L'une des activités les plus attrayantes en classe de langue est celle de simuler le monde extérieur par des exercices de conversation en langue étrangère, mais cette activité est non automatisable par l'ordinateur[9].

On constate toutefois la présence d'applications simples permettant de simuler des règles purement morphologiques de différentes langues et de générer des conjugaisons, des pluriels, des déclinaisons, etc. Le problème de ces programmes est leur focalisation sur des algorithmisations de règles morphologiques qui sont loin d'être centrales pour l'apprentissage d'une langue et leur présentation est généralement pauvre et inadaptée aux apprenants (mots isolés à traiter).

1.3.4 Des environnements ouverts à la découverte guidée

Si le travail avec des environnements ouverts semble séduisant au plan éducatif (surtout pour la résolution de problèmes), les évaluations montrent que l'apprenant rencontre de sérieuses difficultés pour atteindre son but sans l'aide du système. En effet, par opposition aux tuteurs, la machine ne contrôle pas l'**adéquation** entre les **outils utilisés** et l'**objectif**

[9] Les traitements automatiques que l'on peut faire sur une langue naturelle ne permettent pas encore de simuler des conversations en langue naturelle.

que l'apprenant poursuit. Les recherches convergent sur l'idée de la nécessité d'une assistance durant l'activité qui sera assumée par le système lui-même.

La "boîte noire" cède la place à la "boîte de verre" : A un modèle déjà implanté ne laissant à l'utilisateur d'autres choix que de modifier la valeur de certains paramètres prévus à l'avance, vont se substituer des environnements ouverts, détaillant pas à pas leur fonctionnement et fournissant des modes d'accès ou de modification au modèle sous-jacent à la simulation.

Dans le domaine de l'apprentissage des langues, le système SWIM (ZOCK, 91) a adopté l'approche *boîte de verre* pour rendre le processus de production de phrases transparent. Produire du langage, consiste à effectuer une série de choix à différents niveaux (pragmatique, conceptuel, linguistique) interdépendants. L'une des difficultés que trouve l'apprenant dans une classe ou dans certains systèmes de simulation, est qu'il ne voit que l'entrée (scène visuelle, sens) et la sortie (phrases), ce qui se passe entre les deux n'est pas accessible à ses sens. De ce fait l'apprenant est réduit à observer des covariations entrées-sorties pour déduire les règles de la langue (approche boîte noire).

Afin de faciliter la compréhension du processus de la production de langage, le système SWIM autorise l'apprenant à construire explicitement ses hypothèses sur l'interlangue, à les tester, et conséquemment à mettre à jour ses propres règles. Le système guide l'apprenant dans sa tâche, en lui proposant lors de chaque étape des aides interactives. Même si aucune modélisation explicite de l'apprenant n'est visée, le système donne un aperçu de la façon dont sont apprises les structures de la langue cible.

D'autres applications intéressantes ont été produites pour aider les apprenants dans la traduction de textes (Farrington, 1984). A chaque point de sa progression, l'apprenant peut demander une liste de diverses possibilités envisageables pour continuer. Le système assure la correction et la compatibilité de ce qui est proposé avec ce qui a déjà été donné par l'apprenant.

1.4 les modèles de type exploratoire : l'ordinateur comme instrument

L'évolution récente des technologies de communication a amené la communauté ALAO à s'intéresser de façon approfondie à l'ingénierie de l'apprentissage multiforme et à distance. En quelque sorte, le problème posé aujourd'hui est de faire évoluer l'ALAO en prenant en compte une transformation radicale de l'informatique dont la référence technologique est moins l'ordinateur individuel que les systèmes permettant la communication et l'interaction, en temps réel ou en temps différé, entre des machines et des humains distribués dans l'espace.

1.4.1 Les nouvelles technologies de l'information et de la communication (TIC)

Nous désignerons par le terme générique NTIC l'ensemble des technologies informatiques qui permettent de représenter, capter, traiter et distribuer l'information sous toutes ses formes (symbolique ou analogique). Avec l'avènement d'Internet, la distribution de l'information se réalise de plus en plus, par des réseaux hétérogènes, difficilement localisables et qui génèrent plus ou moins spontanément et de manière collaborative un ensemble de services en constante évolution.

Le World Wide Web sur le réseau Internet est le meilleur exemple de ce type de dispositif. Si la technologie associée à ces environnements est relativement standard, l'usage qui en résulte est une expérience sans précédent dans le domaine de l'éducation.

Dans le domaine de l'apprentissage des langues vivantes étrangères, les NTIC ont favorisé l'émergence de nouveaux matériaux et de nouvelles ressources pédagogiques. Leur classification et leur évaluation ne sont pas toujours évidentes. Néanmoins nous pouvons dégager un certain usage des NTIC de la part des utilisateurs (enseignants ou apprenants), que nous avons regroupé en cinq éléments cruciaux :

1. Internet comme source d'informations : Ils peuvent être des sources " brutes " qui ne sont pas forcément créées pour l'usage pédagogique, mais qui peuvent être utilisées à fins pédagogiques dans l'apprentissage des langues étrangères. Ce sont les sites de presse et d'actualités, les grands journaux, et quelques magazines en accès libre sur Internet. Il y a aussi des émissions de radio et de télévision aux quatre coins du monde, plusieurs types d'extraits de vidéo et de cinéma, etc. Le site de la BBCL'adresse Internet du site de la BBCest : http://www.bbc.co.uk/arabic/fm.shtml (dernière consultation : septembre 2001). par exemple est un site multilingue très complet, qui propose pour les arabophones les informations en langue arabe et un enseignement à distance de l'anglais. Il y a aussi les sites touristiques concernant un pays, les sites de villes ou d'ambassades, les sites d'entreprises ou d'organisation publique et les sites d'agences de voyage qui offrent également des possibilités de fournir un bon matériel innovant à des fins pédagogiques. Les sources d'information peuvent être également les dictionnaires multilingues en ligne et les lexiques d'un domaine spécifique.
2. Internet comme source d'activités et de produits pédagogiques : Des banques d'activités sont mises à la disposition des utilisateurs du WEB par les instances publiques, par les équipes académiques, les établissements scolaires ou les organismes de recherche. De plus il existe un nombre illimité d'associations et de groupes mettant des activités pédagogiques et des idées utiles à l'apprentissage des langues étrangères.
3. Internet comme support des activités développées par les enseignants et les apprenants : L'enseignant peut égalementpublier ses propres sources éducatives. Il peut mettre en ligne ses exercices et matériaux utilisés pour sa classe afin d'être consultés à n'importe quel moment par ses apprenants ou par d'autres personnes. Le partage des idées et des ressources entre enseignants de divers pays est en effet possible par le biais d'Internet. Le WEB motive aussi les apprenants à publier leurs projets et leurs travaux et de les partager avec une audience mondiale.

4. Internet comme un outil de recherche : L'enseignant peut donner aux apprenants de trouver des informations sur une question précise ou effectuer des activités liées à des données qu'il faut rechercher sur le réseau. L'apprenant pour sa part, peut consulter les bases de données du réseau et faire des recherches sur des documents, des ressources pédagogiques ou des activités et des exercices spécifiques.

5. Internet comme un outil de communication : Les NTIC offrent de nouvelles possibilités de communication, notamment le courrier électronique (temps différé) et le chat (temps réel). Grâce à la convivialité des interfaces, la communication devient facile et efficace. L'enseignant et les apprenants peuvent communiquer à n'importe quel moment et à n'importe quel lieu. La technologie considérée dans ce cas comme contexte, permet l'idée de l'immersion totale dans un nouveau contexte linguistique, qui, grâce aux aspects sociaux et émotionnels (i.e. communication sur des sujets relatifs à des situations réelles de la vie), favorise l'apprentissage de la langue.

1.4.2 Apports de la navigation et du multimédia

Un des apports importants des nouvelles technologies est la fourniture d'informations sous forme hypermédia. Le fait que cette information soit purement textuelle (hypertexte) ou comporte également images, images animés ou son (multimédia) modifie les formes générales d'interaction aux différentes phases du cycle de l'apprenant dans l'utilisation d'un didacticiel (cf. § 1.2.1) et va permettre des interactions didactiques fondamentalement nouvelles :

- **Information :** les informations proposées sont accessibles par boutons ou icônes et affichées dans des fenêtres **sur l'initiative de l'apprenant**. Elles sont de n'importe quel type (texte, image, son, etc.) et leur contenu n'est plus limité. La recherche d'information peut être intégrée au didacticiel.
- **Sollicitation :** un ensemble d'informations visuelles et/ou sonores peut précéder la sollicitation. Mais un autre élément important est la diversification des modalités de saisie de réponse introduites. Par exemple un clic souris va pouvoir matérialiser le choix d'une portion de l'image ou le choix d'un arrêt sur image dans une séquence vidéo ou la détection d'un signal sonore convenu. Cette diversification rend l'interaction plus conviviale et elle simplifie l'analyse de réponse.
- **Résolution de problème :** la communication entre les divers processus présents sur le micro-ordinateur ou sur le réseau auquel l'apprenant est connecté lui donnent bien plus d'outils de résolution que lorsqu'il fallait les inclure dans le didacticiel, au prix d'un effort de programmation important.
- **Analyse de réponse :** les considérations précédentes montrent que la reconnaissance syntaxique des réponses est simplifiée.
- **Explication :** le système peut adresser à l'apprenant un commentaire hypertextuel qui permet, en laissant à l'apprenant le soin d'approfondir le sens en fonction de ses besoins, de résoudre partiellement les problèmes du niveau d'implicite inhérent à tout

message. Mais ce peut être aussi un commentaire sous forme vidéo qui illustre concrètement les conséquences d'un choix.

Dans le domaine de l'ALAO, c'est surtout dans le contexte de l'apprentissage de la prononciation que les apports du multimédia sont les plus spectaculaires. Beaucoup de logiciels permettent d'enregistrer la prononciation d'une phrase par l'apprenant et d'interagir sur cette prononciation en la comparant à celle d'un modèle et en mettant en évidence graphique les distorsions importantes (visualisation différentielle des signaux correspondants aux deux enregistrements).

1.4.3 Vers des Environnements Informatiques pour l'Apprentissage Humain (EIAH)

L'évolution récente des technologies de communication et l'intégration de plus en plus affirmée entre informatique et télécommunication d'une part, et la manifestation d'un grand besoin en matière d'apprentissage de la part des sociétés développées[10] d'autre part, a amené la communauté de l'ALAO à s'intéresser de façon approfondie à l'ingénierie d'enseignement à distance (EAD). En quelque sorte, le problème est aujourd'hui posé de faire évoluer l'ALAO en prenant en compte une transformation radicale de l'informatique dont la référence technologique est moins l'ordinateur individuel que les systèmes permettant la communication et l'interaction, en temps réel ou en temps différé, entre des machines et des humains distribués dans l'espace (Balacheff et al, 97).

Certains travaux actuels visent ainsi des Environnements Informatiques pour l'Apprentissage Humain (EIAH) dans lesquels coopèrent agents humains et agents artificiels, ce qui nécessite des conceptualisations et des stratégies différentes de celles de l'ordinateur individuel.

L'autonomie de l'apprenant est l'une des questions mises en jeu. Un grand nombre de travaux impliqués dans des dispositifs de formation à distance font le constat que l'apprenant ne peut être isolé et que son autonomie ne peut être totale (Balacheff et al, 97).

Il s'agit aujourd'hui de concevoir des systèmes coopératifs d'apprentissage qui intègrent comme acteurs des enseignants et des apprenants, et qui offrent de bonnes conditions d'interaction à travers le réseau entre agents humains et agents artificiels, ainsi que de bonnes conditions d'accès à des ressources d'apprentissage distribuées.

1.5 Conclusion

Cette synthèse de l'historique de l'ALAO, montre bien que l'on est passé d'une volonté

[10] En particulier la prise en compte de ce que la commission européenne a appelée "l'éducation et la formation tout au long de la vie"

d'optimisation des modes d'enseignement, à une centration sur la construction des connaissances. L'on est passé d'un intérêt pour les comportements à une focalisation sur les connaissances et la genèse de ces connaissances. De plus, la vision individualisée de l'apprentissage a cédé la place à une vision plus collective et plus collaborative de cet apprentissage. Avec l'évolution des interfaces et des possibilités d'action et d'exploration, l'ordinateur tend à devenir un poste de travail intégrant de multiples environnements, allant de simples correcteurs orthographiques à des environnements hautement interactifs donnant accès via Internet à de multiples ressources.

Sur le plan des réalisations, on constate néanmoins que le fait d'utiliser un ordinateur semble encore aujourd'hui suffire à justifier à priori n'importe quel contenu. Ce n'est pas parce qu'*un programme informatique permet de faire tel traitement linguistique qu'il est pour autant justifié de l'intégrer tel quel dans un enseignement*. On constate en effet qu'il y a un certain nombre de décalages entre l'explosion massive des applications utilisant le multimédia et l'hypertexte et les *besoins réels* des apprenants de ces applications.

Tel que nous essayerons de le définir tout au long de cette thèse, un "Environnement Interactif d'Apprentissage par ordinateur" (EIAO) d'une langue, doit regrouper différentes méthodes adaptées aux différentes compétences linguistiques ciblées. Le concept d'EIAO permet en effet de dépasser l'opposition simpliste et manichéenne entre les défenseurs de l'apprentissage par induction (résultant des seules activités exploratoires du sujet) et les partisans des tutoriels inspirés par l'enseignement programmé (apprentissage par enseignement). Un système d'apprentissage doit être capable de favoriser l'acquisition des concepts et des procédures, par l'utilisation de la méthode la mieux adaptée au domaine de connaissances ciblé.

Il apparaît donc nécessaire d'élaborer une stratégie d'intervention en fonction des compétences linguistiques visées pour un apprenant. L'apprentissage des compétences grammaticales pour une langue (habileté à produire et comprendre des énoncés correctement sur les plans phonologique, syntaxique et lexical) s'adaptent peut-être plus à un modèle d'enseignement tutoriel, que des compétences sociolinguistiques (habileté à utiliser le langage de façon appropriée dans un contexte socioculturel donné), ou communicatives (habileté à transmettre efficacement une information à un auditeur...).

Il s'agit de construire une "station de travail" que chaque apprenant pourrait l'adapter à ses besoins. Un bon EIAO est un système qui réalise la synthèse entre, les avantages de l'exploration libre et de la construction progressive des objets de connaissance d'une part, et les apports des activités et de l'aide fournie par les systèmes tutoriels d'autre part. L'idée centrale est de permettre à l'apprenant de transformer rapidement et efficacement ses expériences en connaissances organisées.

Un EIAO doit privilégier l'idée que la meilleure façon d'apprendre c'est de se trouver dans une situation (quasi) réelle de travail. Plutôt que de construire des logiciels orientés sur l'explicitation formelle des connaissances, nous pensons qu'il est préférable de concevoir des outils qui assisteraient l'apprenant efficacement dans la résolution des problèmes auxquels il doit faire face. Il apparaît en effet, que l'apprentissage incident de l'ensemble des connaissances nécessaires à la résolution d'une tâche pose moins de problème de motivation et d'attention, si l'intérêt pour la tâche est assuré à un niveau

élevé.

Un EIAO de langue est en résumé un logiciel qui permet d'entraîner un certain nombre d'automatismes et de compétences communicatives nécessaires à la maîtrise de la langue. Il doit prendre en compte les champs conceptuels du domaine et les définitions de termes pour donner à l'apprenant la possibilité de communiquer avec les autres apprenants et les spécialistes du domaine.

Dans la suite de cette thèse, nous construirons d'abord des ressources linguistiques et des outils informatiques de TAL arabe et nous verrons par la suite comment nous pourrions les articuler dans un EIAO de l'arabe langue seconde ou étrangère répondant aux objectifs définis ci-dessus.

Chapitre 2 Présentation de la base de données lexicale DIINAR.1

« Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher. » Antoine de SAINT-EXUPÉRY

2.1 Introduction

Les progrès technologiques ont permis l'accroissement des volumes de données stockés sur des supports magnétiques et la rapidité de leur traitement. C'est ainsi qu'on assiste à l'émergence de *dictionnaires électroniques* [11], qui semblent être très utiles, voire indispensables de par la quantité d'informations de natures diverses qu'ils contiennent aux applications linguistiques et plus particulièrement à l'élaboration d'activités d'apprentissage.

La construction d'un dictionnaire électronique fiable et utilisable par diverses

[11] Dans le domaine du TALN, on désigne sous ce terme des dictionnaires associés à des bases de données qui sont créés pour un traitement linguistique informatisé. De tels dictionnaires ne doivent pas être confondus avec **les dictionnaires usuels** mis sur support électronique (tels par exemple le Robert pour le français, ou le Longman pour l'anglais), que nous traiterons dans le sixième chapitre de cette thèse.

applications, implique des prises de décisions portant sur le contenu et le format des données, aussi bien au niveau linguistique qu'au niveau informatique : « A partir d'observations sur le fonctionnement des langues, la linguistique informatique construit des **concepts** et des **modèles**, qu'elle valide ensuite par une **simulation informatique** » (Desclés, 1989, p.14).

Nous introduirons ce chapitre par les fondements linguistiques, qui ont servi à la conception de la base de données lexicale DIINAR.1 (DIctionnaire INformatisé de l'ARabe). Nous détaillerons ensuite la réalisation informatique en énumérant les différents constituants du dictionnaire, dont les nouvelles parties des mots outils et des noms propres.

Cette base de données servira principalement à générer le lexique qui sera utilisé par l'analyseur. L'exploitation de ce dernier, entraîne souvent la détection de nouveaux mots absents du dictionnaire, qui doivent être codées avant d'être intégrés. La maintenance du dictionnaire constitue un moyen important pour l'amélioration des résultats d'analyse et doit par conséquent être traitée avec le plus grand soin. L'examen de cette tâche tout à fait particulière fera l'objet d'une présentation à la fin du chapitre.

2.2 Fondements théoriques

L'objectif des premiers travaux effectués par l'équipe SAMIA, était la définition d'un modèle linguistique pouvant être simulé sur ordinateur et utilisé dans le cadre d'un système d'enseignement assisté par ordinateur de la langue arabe (Dichy & Hassoun, 1989, pp 61). Sans traiter de façon exhaustive tous les phénomènes, la modélisation devrait permettre à un apprenant débutant d'étudier un maximum de régularités morphologiques afin qu'il puisse analyser et générer des mots correctement à partir de leurs traits linguistiques.

Le domaine qui a été pris en compte par l'équipe SAMIA était le mot graphique (mot séparé par deux blancs), et non pas la phrase ou le texte, bien qu'il soit généralement admis qu'une analyse au niveau de la phrase présente une démarche beaucoup plus puissante et moins génératrice d'ambiguïtés.

Ce choix s'explique par le fait que la structure particulière de cette unité en arabe rend nécessaire une étude approfondie des phénomènes qui se produisent à son niveau pour rendre possible l'analyse syntaxique (Jaccarini & Audebert, 1986). Le domaine du mot graphique arabe constitue en effet, un passage presque obligé de toutes les recherches dans le domaine du traitement automatique de l'arabe.

Dans cette section, nous détaillerons les fondements linguistiques des travaux SAMIA autour du mot graphique en arabe (Dichy, 1990) (Dichy & Hassoun, 1989), qui ont permis, par la suite, la conception d'un dictionnaire pour la synthèse et l'analyse automatique du mot graphique arabe (Hassoun, 1987) et la réalisation de la base de données lexicale DIINAR (Dichy, 1997) (Braham & Ghazeli, 1998) (Ghenima, 1998).

Nous présenterons d'abord le schéma de décomposition du mot graphique arabe tel

qu'il a été défini dans le cadre des travaux de SAMIA. A partir de cette décomposition et des besoins des traitements de synthèse et d'analyse morphologique, nous recenserons les principales règles de gestion des formants du mot graphique, les relations entre les unités lexicales et le contenu du "spécificateur" qui doit être associé au mot graphique arabe.

2.2.1 La décomposition du mot graphique en arabe

Dans le modèle SAMIA, le mot graphique[12] arabe est considéré comme une structure d'objet complexe contenant une suite de morphèmes[13]. Chaque mot graphique peut se décomposer en une suite ordonnée de : proclitique(s), préfixe, base, suffixe(s), enclitique(s).

La base du mot est, construite selon un procédé appelé *dérivation interne* c'est à dire selon des schèmes[14] qui sont des patrons syllabiques dans lesquels s'inscrivent les consonnes de la racine. La racine est un groupe de consonnes qui se présentent selon un ordre fixe et qui représente une notion définie : par exemple, la racine (ب،ت،ك : *KTB*) représente la notion d'écrire. Le schème est constitué de voyelles brèves ou longues, du redoublement d'une radicale et de consonnes affixées appartenant aux racines mono-consonantiques (Roman, 1990). Les schèmes ont un sens codifié par la grammaire, qui entre en combinaison avec celui des racines pour donner le sens du mot. Par exemple, l'agent d'action des verbes de forme (I) est construit théoriquement à partir du schème suivant : $R_1 âR_2 iR_3$. Le participe actif du verbe (يَكْتُبُ – كَتَبَ: *KaTaBa – YaKTuBu*)= («écrire»), est (كَاتِب : *KâTiB*) = («écrivain»). Ce participe actif est obtenu par l'inscription des consonnes de la racine (ب،ت،ك : *KTB*) dans le schème $R_1 âR_2 iR_3$.

Ce procédé s'applique pour la totalité des verbes, des dérivés nominaux immédiats (nom verbal, participe actif, etc.) et pour une partie importante des noms. Toutefois, un sous-ensemble important des noms, c'est le cas des mots empruntés aux autres langues et des noms propres, ne sont pas construits selon ce procédé. Ces noms correspondent à des pro-bases (Dichy, 1997).

Bases et pro-bases sont le noyau lexical (ou formants-noyau, Fn) du mot, les autres constituants sont des extensions (ou formants-extensions, Fe) qui peuvent s'ajouter au mot pour former un mot graphique maximal. Ce mécanisme appelé *dérivation externe* ou

[12] Le mot graphique en arabe est communément défini, d'une manière empirique, comme l'unité située dans l'écriture entre deux "blancs" ou entre deux séparateurs similaires tels que les signes de ponctuation ([.], [,], [?], [!], [(], [«], etc.). Le tiret qui pose des problèmes de délimitation pour les écritures latines est considéré aussi comme un blanc en arabe.

[13] Les morphèmes constitutifs de l'unité-mot sont appelés des *formants de mot*, c'est-à-dire, des signes linguistiques minimaux dont les relations de contextualisation sont limitées aux autres morphèmes inclus dans l'unité composée que constitue le mot dans sa manifestation graphique (Dichy, 1987).

[14] La grammaire arabe fait traditionnellement usage, pour cela, d'une convention qui consiste à faire appel à une racine tri-consonantique théorique (ف،ع،ل) = («faire»). On utilise également de nos jours d'autres représentations symboliques comme R_1R2R_3 (où « R » signifie consonne radicale et le chiffre en indice indique la position dans la racine).

suffixation, fait de l'arabe une langue agglutinante.

La première couche des formants-extensions est constituée par les préfixes et les suffixes qui forment avec la base du mot ce que nous appelons un *mot minimal* (Cohen, 1961/70).

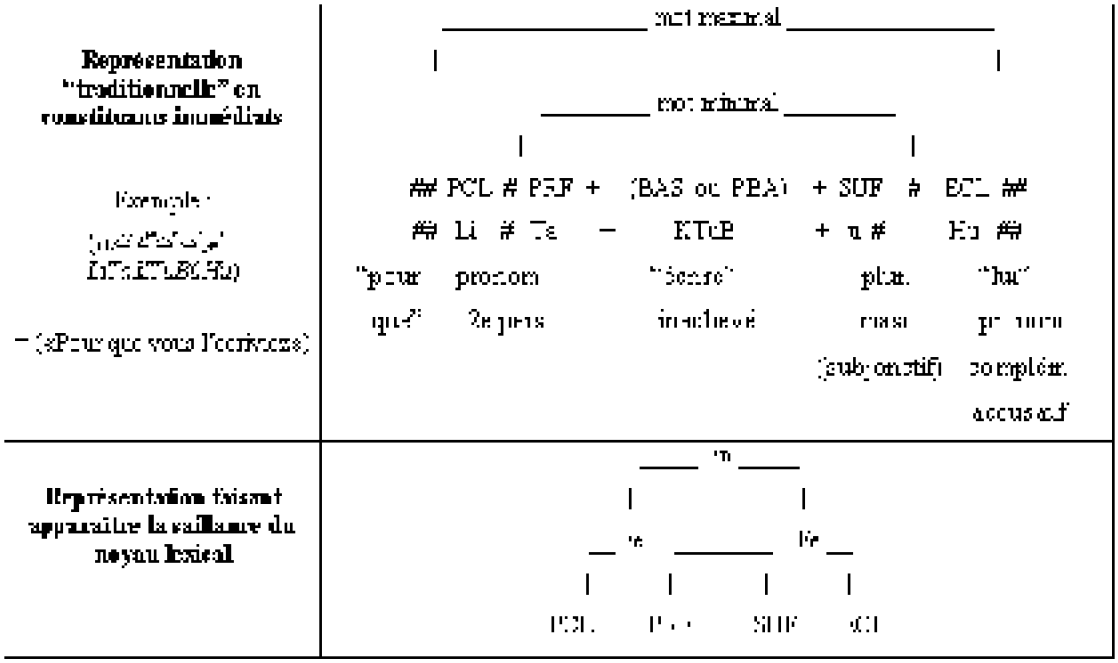
Le préfixe est un morphème verbal relatif à l'inaccompli (actif ou passif) placé avant la base. Pour les noms et les autres aspects de la conjugaison nous dirons que le préfixe est vide. Le suffixe est un morphème situé immédiatement après la base. La base peut être une base verbale suivie d'un suffixe verbal ou une base nominale suivie d'un suffixe nominal. Chaque suffixe peut être une combinaison de deux suffixes cas de (قبسن ال ءاي : *Yâ? ALNiSBa*) = («adjectifs de relation»). Comme pour les préfixes, mais dans des cas beaucoup plus rares, le suffixe peut être vide.

La seconde couche des formants-extensions est constituée par les proclitiques et les enclitiques, qui s'attachent au mot minimal par une seconde dérivation externe et forment le *mot maximal* (Cohen, 1961/70). Par proclitique, nous désignons les proclitiques simples (morphèmes à une lettre) et les proclitiques composés (morphèmes à plusieurs lettres). Les premiers sont des coordonnants, des conjonctions, des prépositions, etc. Les seconds sont obtenus par combinaison des premiers. L'article (ال : *?aL*) est également considéré comme proclitique simple, bien qu'il comporte deux lettres. L'enclitique est un pronom personnel complément qui peut être simple ou double. Il est attaché au mot qui le précède pour ne former qu'un seul mot.

En faisant usage des *frontières faible* et *forte* de morphème (respectivement # et +), on peut représenter le mot graphique en arabe ainsi (figure 2-1) :

15

[15] Pour la représentation en constituants immédiats (Desclés et al., 1983), lire "##" comme *frontière de mot*; le critère empirique permettant de distinguer la frontière "+" (pré- ou suffixation) de la frontière "#" (enclise : pro- ou enclitiques) est celui de la pause potentielle (Lyons 1978/90) : en l'absence du PRF ou du SUF auquel elle est liée par une frontière "+", la BAS - ou la PBA - ne peut constituer une forme libre minimale. En revanche, elle peut, de ce point de vue, "se passer" des ECL et des PCL.



2.2.2 Analyse des besoins d'un synthétiseur et d'un analyseur automatique des mots graphiques

Comme nous l'avons déjà annoncé dans l'introduction de ce chapitre, l'objectif initial de la modélisation linguistique de l'équipe SAMIA était la réalisation d'un synthétiseur et d'un analyseur du mot graphique arabe, qui devront être utilisés dans le cadre d'un EAO de l'arabe. Ces outils pourront être par la suite utilisés dans d'autres applications tels que la traduction automatique, la génération automatique de textes et la correction orthographique.

Afin d'aboutir au bon fonctionnement de ces outils, un recensement des différents éléments et règles linguistiques a été effectué à partir de la décomposition du mot graphique arabe ci-dessus. Ce travail a permis de dégager les besoins suivants du synthétiseur :

- les listes exhaustives des morphèmes du mot,
- les règles permettant leur agencement,
- les règles testant leur compatibilité,
- les règles de transformation morphologique et phonologique lors de leur assemblage,
- les relations de fléchage entre unités lexicales (Voir ci-dessous § 2.2.4).

L'analyse morphologique doit de son côté, non seulement identifier les différents morphèmes du mot à analyser[16] et les attester, mais doit aussi fournir un ensemble d'informations morphologiques et syntaxiques sur le mot à analyser. L'unité lexicale doit

[16] Le mot pourra être non vocalisé, partiellement vocalisé ou complètement vocalisé.

être définie (base, mot minimal ou mot maximal) ainsi que l'ensemble des informations linguistiques qui lui seront associées.

2.2.3 Les règles grammaticales des formants du mot

Une grammaire des formants du mot issue de la décomposition du mot graphique et permettant son traitement en synthèse et/ou en analyse, doit comporter deux sortes de relations contextuelles (Dichy, 1987) :

- La *relation d'ordre* qui régit la position des formants et qui est strictement prédictible. Par exemple, la préposition proclitique (لِ : *Li*) = (« pour ») est placée avant l'article (ال : *?aL*), et après les coordonnants (وَ : *Wa*) et (فَ : *Fa*).
- Les *relations de collocation morpho-phonologiques et syntaxiques*, gérant les incompatibilités et cooccurrences, ainsi que les modifications morpho-phonologiques entraînées par la présence d'un formant. L'article (ال : *?aL*), par exemple, est incompatible avec les marques casuelles de l'indétermination. Le pronom clitique (هُ : *Hu*) est réalisé (هِ : *Hi*) après la voyelle brève (ِ : *i*).

2.2.4 Définition de l'unité lexicale (Dichy, 1997)

Une question fondamentale se pose lorsqu'on entend réaliser un lexique pour les applications d'analyse et de génération : c'est celle du choix des unités lexicales (UL) qui constitueront les formes générées dans le dictionnaire à qui seront associées les propriétés syntaxiques et sémantiques.

Cette UL ne coïncide pas toujours avec le formant noyau (Fn). En effet, en ce qui concerne les noms (et à la différence des verbes), certains formants-extensions (Fe) sont susceptibles, lorsqu'ils sont associés à une base nominale, de se trouver pris avec elles dans un processus de lexicalisation. Un formant-extension sera dit lexicalisé (appelé formant-extension lexicalisé – Fel) lorsque l'unité <Fn, Fel> résultant de son association à un formant noyau constitue une unité du lexique indépendante. Une UL est donc constituée :

- Soit d'un Fn (on écrira UL = <Fn>) : c'est le cas des bases verbales ou des bases nominales dépourvues de Fe lexicalisé, par exemple : UL = <Fn = (بَتْكَم : *MaKTaB*)> =(« bureau »);
- Soit d'un ensemble UL = <Fn, Fel>, où Fel peut inclure plus d'un formant, et dont l'ordre séquentiel est assigné par la grammaire des formants du mot ; par exemple (ةَبَتْكَم: *MaKTaBa&*) = (« bibliothèque » ou « librairie »), s'analyse ainsi : UL = <Fn = (بَتْكَم : *MaKTaB*), Fel = (ةَ : *a&*)>, et constitue une UL distincte de (بَتْكَم : *MaKTaB*).

L'unité lexicale (UL) est définie aussi par ses relations avec d'autres UL_s. Ces relations ou ce *fléchage* entre deux ou plusieurs UL_s, peuvent être représentées sous cette forme : ULx□(UL_1, UL_2, …,UL_n).

Le fléchage reflète au départ une propriété morphologique "correspondant à la dérivation interne" : le passage du verbe au participe actif, d'une forme nominale au singulier à son correspondant au pluriel ("pluriel brisé "), etc., ce passage pouvant s'effectuer de manières classiquement présentées ainsi :

par « dérivation externe ». Par exemple, (مٍـلَـعُـم : *Mu c aLLim*) =(« maître d'école ») a pour pluriel masculin (نوُمٍـلَـعُـم : *Mu c aLLimûn*), construit par la suffixation de (نوُ : *ûn*) au masculin singulier.

par « dérivation interne ». Par exemple, (فيطٍل : *LaTîf*) =(« doux ») a pour pluriel (فاَطٍل : *LiTâf*), qui bien qu'ayant la même racine (ف،ط،ل : *LTF*), est construit selon un schème différent.

Le fléchage permet d'un autre côté de nuancer certains sens polysémiques. En effet, dans les cas de polysémie d'une forme correspondant à une entrée nominale du dictionnaire, des pluriels distincts sont susceptibles d'être associés à des sens différents. Par exemple, l'unité lexicale UL_1 = <ثيٍدَح : *Hadî T* > =(« discours rapporté, récit, conversation »), est reliée par un fléchage au pluriel UL_2 = <ثيٍداَحأَ : *?aHadî T* > et elle est distincte de l'unité lexicale UL'_1 = <ثيٍدَح : *Hadî T* > =(« moderne »), qui est plutôt reliée par un fléchage aux pluriels UL'_2 = <ثاَدٍح : *Hidâ T* > et UL'_2 = <ءاَثَدُح : *Huda T â?*>. Ce phénomène est observable aussi pour la relation entre les verbes de la forme simple et les noms de procès (ردالمصا).

Par conséquent, « On dit qu'il y a un fléchage entre deux UL (ou plus) lorsque l'une d'entre elles est supplétive de l'autre, c'est à dire lorsque l'une d'entre elles remplace l'autre dans les paradigmes morpho-syntaxiques d'une manière qui ne peut être prédite au sens strict par des règles opérant en génération, ou qui présente un caractère d'opacité en reconnaissance » (Dichy, 1997). Ainsi, la relation de fléchage n'est pas bijective.

Cette définition donne au concept de fléchage une extension plus large que celle de la « dérivation interne ». Deux types de cas, non compris dans cette dernière, sont en effet inclus (Dichy, 1997) :

- le fléchage entre une unité lexicale coïncidant avec le formant-noyau, et une UL comportant un noyau et un formant-extension lexicalisé, qu'on peut représenter par : UL_n = <Fn> □ UL_m = <Fn, Fel> dans un sens ou dans l'autre. Par exemple, (ةَبٍلُـع : *c uLBa&*) =(« boîte ») de structure UL = <Fn> (بٍلُـع : *c uLB*) + <Fel> = (ةَ : *a&*), a pour pluriel (بٍلَـعُـ : *c uLaB*) de structure UL = <Fn>.
- Le fléchage entre bases de racines différentes. Par exemple, (ةَأٍرَم : *MaR?a&*) =(« femme »), de racine (أ،ر،م : *MR?*), a pour pluriel (ءاَسٍن : *NiSâ?*) de racine (و،س،ن : *NSW*).

2.2.5 Les spécificateurs morpho-syntaxiques associés au mot

Le concept de *spécificateurs* relève de la linguistique informatique : associés aux entrées lexicales, ils gèrent les relations (grammaticales ou lexicales), qui lient ces unités aux

autres unités présentes dans le domaine d'extension considéré.

Ces spécificateurs peuvent être de différentes natures (morpho-phonologique, morpho-syntaxique, syntaxique, sémantique). On peut intégrer dans ces spécificateurs par exemple :

- la catégorie lexicale : la base du noyau lexical est nominale, verbale, ou un mot outil.
- la sous catégorie lexicale : par exemple le type du déverbal.
- les variables permettant de générer les formes fléchies de l'unité lexicale : le numéro du modèle de conjugaison pour les verbes ou le numéro du modèle de déclinaison pour les noms et les déverbaux.
- la catégorie syntaxique ou les structures syntaxiques des phrases associées au verbe

Nous reviendrons dans la description de la réalisation informatique sur le contenu de ces spécificateurs.

2.3 Réalisation informatique

Au cours de ces dernières décennies, les acteurs du domaine du TAL ont progressivement pris conscience de l'enjeu que constituent les dictionnaires électroniques et de l'importance des modalités d'implémentation. Le choix du formalisme de représentation des connaissances linguistiques, doit être pris en considération en priorité puisqu'il est un élément central et structurant dans la conception des applications de TAL. Ainsi, le contenu, le volume et le format du dictionnaire dépendent du formalisme choisi et favorisent des traitements linguistiques par rapport à d'autres.

Un dictionnaire électronique, sans pour autant être exhaustif, doit contenir le maximum de connaissances sur la langue étudiée. Selon le type d'application visé, ces connaissances seront de nature différente (phonologique, morphologique, syntaxique, sémantique, pragmatique, etc.) et de niveaux de structuration et de granularité différents. Si l'on souhaite être indépendant des traitements qui seront associés ou envisagés, on doit modéliser toutes ces informations linguistiques dans le même dictionnaire.

Les difficultés liées à la gestion de cette énorme quantité d'informations ont poussé les spécialistes à recourir aux techniques informatiques de représentation des connaissances, dont les bases de données de type relationnel, qui sont plus connus dans le domaine du TAL sous l'appellation de base de données lexicale.

L'organisation des données avec un système de gestion des bases de données (SGBD) assure une quasi parfaite indépendance entre données et programmes, mais surtout permet l'évolution du dictionnaire à travers le temps. En effet, les SGBD permettent d'étendre la couverture de la base de données lexicale à des niveaux de traitements non prévus au début de l'implémentation sans que cela n'affecte la structure globale de la base ni les programmes déjà réalisés, à condition que ces derniers soient bien conçus au départ.

Initialement, les données linguistiques issues de la conception linguistique ont été implémentées dans une base de données multi-fichiers DIINAR.1 qui fonctionnait sous le système d'exploitation MS-DOS (Gader 96) et (Ghenima 98). Cette première version a permis de saisir environ 20000 entrées verbales et 39000 entrées nominales. Cette base avait l'inconvénient d'être dépendante des programmes qui la gèrent et ne pouvait être exploitée qu'à travers une connaissance approfondie du code source.

Nous avons alors pris la responsabilité de faire émigrer DIINAR.1 dans une nouvelle base de données relationnelle, indépendante des langages de requêtes et de programmation qui pourraient l'exploiter. Cette émigration[17] a considérablement facilité notre travail de maintenance de la base et son exploitation. Nous avons pu générer des sous-lexiques adaptés aux différents besoins des autres membres de notre équipe. Par exemple, Pour les besoins de l'analyseur syntaxique, nous avons pu générer un lexique adapté sous forme d'une liste de prédicats.

Les principaux critères d'évaluation d'une base de données lexicale sont d'une part, la couverture, c'est à dire le nombre d'entrées, et d'autre part, l'exactitude et la précision des informations linguistiques. Le premier critère étant presque atteint avec DIINAR.1, nous nous sommes plutôt employés à mettre à jour la base pour atteindre le second objectif.

Dans cette section, nous décrirons grossièrement[18] les parties de la base relatives à la gestion des verbes, des déverbaux et des noms de DIINAR.2. Les deux nouvelles parties de la base de données, correspondants aux mots outils et aux noms propres, seront plus détaillées et décrites dans des sections indépendantes puisqu'elles n'ont jamais fait l'objet d'une présentation.

2.3.1 Rappel sur les bases de données

Plusieurs méthodes peuvent être utilisées pour la conception et la réalisation d'une base de données. Indépendamment de la méthode, le modèle conceptuel de données (MCD) constitue le point de départ et la partie fondamentale de toute conception d'une base de données. Il permet de mettre en lumière les caractéristiques essentielles du système d'information observé. Cependant, il n'est pas directement utilisable par une machine, mais c'est un mode de représentation intermédiaire entre la réalité observée et la machine avec son logiciel qu'il soit un SGF (Système de Gestion de Fichiers) comme c'était le cas pour DIINAR.1 ou un SGBD (Système de Gestion de Bases de Données) comme c'est le cas pour cette nouvelle base de données.

Une fois que le MCD a été défini, nous devons choisir le modèle d'implémentation de la base. Pour cette réalisation, nous avons opté pour le modèle relationnel qui est le plus récent et le plus utilisé.

[17] Le travail d'émigration a impliqué des modifications au niveau de la structure de la base de données et le développement de programmes informatiques pour l'automatisation du processus.

[18] Pour plus de détails sur la réalisation de DIINAR.1, nous renvoyons le lecteur à la thèse de Ghenima (1998).

Nous illustrons l'organisation des données dans DIINAR.1 par des Modèles Logiques de Données (M.L.D) qui permettent une représentation simplifiée et fidèle de l'implémentation de la base de données dans la machine.Dans ces schémas les informations s'articulent autour de trois concepts principaux : la propriété, l'entité et la relation.

- **Entité :** Une entité est un objet ou un concept manipulé doté d'une existence propre, identifiable et d'intérêt pour l'application. Elle est représentée par un rectangle sur le schéma du MLD.
- **Relation :** C'est une association définie entre N entités. Chaque occurrence de la relation doit-être liée à une occurrence de chacun des objets qui la composent. Elle est représentée par un rectangle arrondi sur le schéma du MLD.
- **Propriété** (ou attribut) **:** C'est une information élémentaire qui a un sens en lui-même et qui caractérise soit une entité soit une relation.
- Identifiant :

  – D'un objet : Propriété particulière de l'entité choisie de telle manière qu'à chaque valeur prise par cette propriété corresponde une et une seule occurrence de cette entité.

  – D'une relation : Concaténation des identifiants des entités qui participent à la relation.

Toutes les propriétés soulignées sur les MLD constituent des identifiants.

- **Cardinalité :** Les cardinalités d'une relation indiquent pour chaque couple Entité/relation, le nombre minimum et maximum d'occurrences de la relation pouvant exister pour une occurrence de l'entité.

2.3.2 Modélisation des verbes[19]

La partie verbale couvre environ 20.000 entrées. Une entrée verbale est définie par la donnée d'une racine et d'un schème verbal. Chaque entrée est représentée par sa forme conjuguée à l'accompli et à l'inaccompli à la 3ème personne du singulier masculin.

[19] **Ghenima (1998), (Brahem & Ghazeli, 1998) et (Dichy, Hassoun, Zaafrani, 2002 d).**

A chaque entrée verbale, est associé un spécificateur morpho-syntaxique incluant :

a) le numéro du modèle de conjugaison du verbe [20] : Un verbe conjugué en arabe se compose d'un préfixe, d'une base et d'un suffixe. Il varie en fonction du **paradigme de conjugaison** [21] et en fonction du **pronom de conjugaison** qui dépend de la ***personne*** (1ère personne (المتكلّم = celui qui parle), 2ème personne (المخاطب = celui à qui on s'adresse), 3ème personne (الغائب = l'absent)), du ***genre*** (masculin ou féminin) et du

[20] Ce travail reprend les modèles de conjugaison de l'arabe définis dans (Abu al-chay, 1988) et établis dans (Ammar & Dichy, 1999).

[21] Seulement 9 paradigmes restent aujourd'hui vivants : les 4 paradigmes de la voix active = (المعلوم) : l'accompli = (الماضي), l'inaccompli indicatif = (المضارع المرفوع), l'inaccompli apocopé = (المضارع المجزوم) et l'inaccompli subjonctif = (المضارع المنصوب) ; les 4 paradigmes correspondants à la voix passive = (المجهول) et l'impératif = (الأمر). Les verbes d'état durable n'admettant pas de voix "passive" en raison d'une impossibilité sémantique. D'autres paradigmes sont sortis de l'usage comme l'inaccompli énergique = (المؤكّد).

nombre (singulier, duel ou pluriel). Les paradigmes de conjugaison des verbes issus d'une racine anomale[22] comportent plusieurs réalisations de la base verbale, chaque réalisation étant compatible avec un sous-ensemble déterminé de la liste des suffixes. Par exemple, les verbes ayant une racine redoublée comme (شَدَّ - يَشُدُّ : *S aDDa-Ya S uDDu*) = (« attacher ») admettent à l'achevé 2 bases (Figure 2-2) : $Base_1$ = (شَدَ : *S aDaD*), utilisée par exemple dans les formes conjuguées (شَدَتْ : *S aDaDTu*) et (شَدَانْ : *S aDaDNa*) et $Base_2$ = (شَدَّ : *S aDD*), utilisée par exemple dans les formes conjuguées (شَدَّتْ : *S aDDaT*) et (شَدَّا : *S aDDâ*).

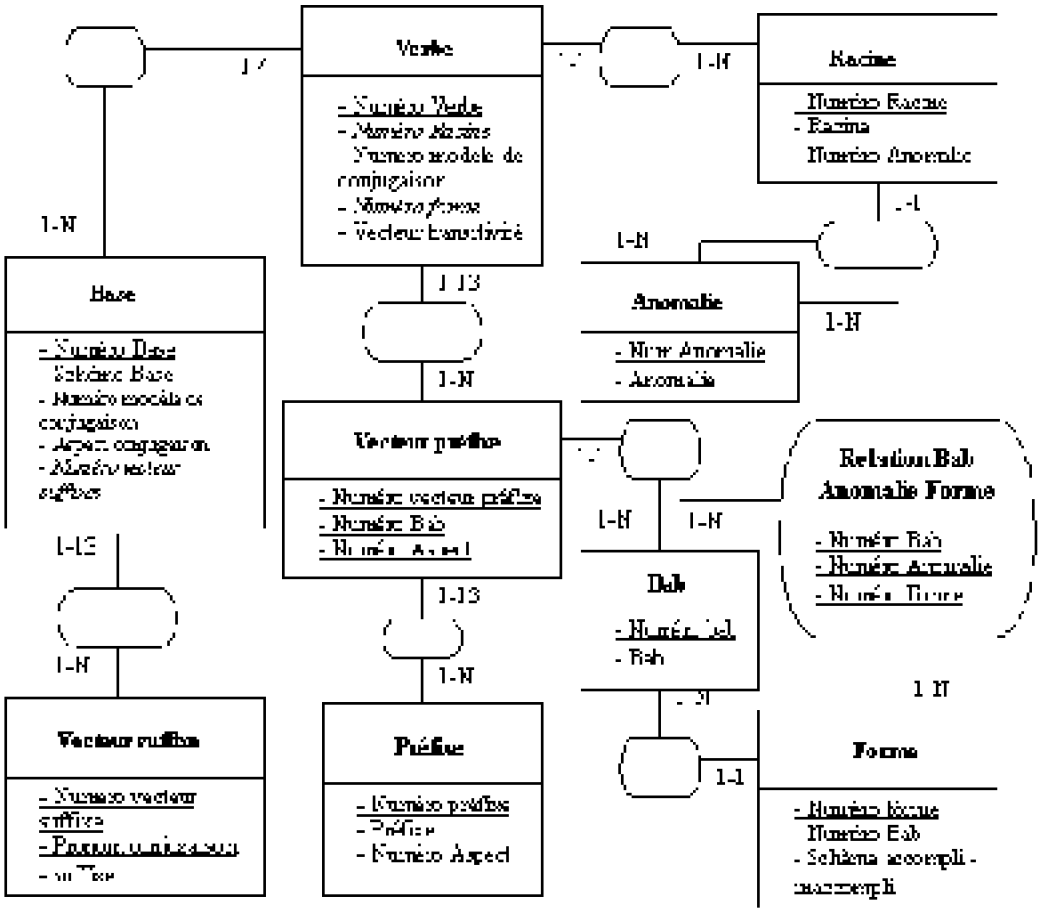
Chaque modèle de conjugaison admet un ensemble de schèmes de bases verbales[23]. Chaque schème d'une base est relié à une liste des préfixes (vecteur préfixes) et une liste de suffixes (vecteur préfixes) permettant d'obtenir les différentes formes conjuguées. La figure (2-3) ci-dessous, illustre les principales entités et relations de la base de données permettant la conjugaison des verbes.

b) la liste des dérivés nominaux immédiats : Un certain nombre de déverbaux très réguliers, dont l'inventaire est fini et systématiquement mentionnés dans les dictionnaires papiers ont été retenus (figure 2-4) :

- **Le nom de procès ou le nom verbal** = (المصدر) :A mi-chemin entre le nom et le verbe, il désigne le procès du verbe et signifie le « fait de… » ou l'« action de… ».
- **Le participe actif ou le nom d'agent =** (اسم الفاعل) : désigne celui qui fait l'action du verbe.
- **Le participe passif ou le nom de patient =** (اسم المفعول) : dérivé exclusivement des verbes transitifs, désigne celui qui subit l'action du verbe ou parfois le résultat de l'action.
- **L'adjectif analogue =** (الصفة المشبّهة) : ne peut être dérivé qu'à partir des verbes simples trilitères ou quadrilitères.
- **Le nom de temps et de lieu** = (اسم الزمان أو المكان) : exprime soit l'endroit, soit le moment où une action a lieu, et parfois les deux en même temps.

[22] On distingue grossièrement trois types de racines anomales : la racine redoublée = (المضاعف) dans laquelle la 2ème et la 3ème consonnes radicales sont identiques, la racine hamzée = (المهموز), dont l'une des consonnes radicales est une hamza (ء : *?*) et la racine défectueuse = (المعتل) dont l'une des consonnes radicales est (و : *w*) et/ou (ي : *y*).

[23] Un paradigme de conjugaison donné peut avoir jusqu'à quatre bases de conjugaison différentes.



Le sens du verbe est en relation avec un ensemble de déverbaux. Les déclinaisons de chaque déverbal sont obtenues par la donnée d'un numéro de modèle de déclinaison qui est associé dès le départ au schème du déverbal. Chaque déverbal peut être décliné suivant : le mode (indéterminé, déterminé par l'article, déterminé par annexion), le cas (nominatif, accusatif, génitif), le genre (masculin, féminin) et le nombre (singulier, duel, pluriel). 19 modèles de déclinaison de déverbaux sont employés. Chaque modèle permet de générer au plus 54 formes fléchies pour le même déverbal.

c) le spécificateur morpho-syntaxique : Ce spécificateur inclut des informations de nature morphologique et syntaxique (figure 2-4). Ces informations permettent :

- de déterminer si le verbe peut être transitif aux humains et/ou aux non humains, avec ou sans mot outil (متعدّي إلى العاقل / غير العاقل، قادر أو غير قادر),
- si le verbe accepte la levée de la règle qui lui interdit un pronom enclitique de la première ou de la deuxième personne si son sujet est de la même personne (فعل من أفعال القلوب),
- si le verbe admet un double complément d'objet, c'est à dire s'il accepte les doubles enclitiques (متعدّ إلى مفعولين).

2.3.3 Modélisation des noms[24]

La partie nominale couvre environ 29000 entrées et 10000 formes de pluriel interne. Une

entrée nominale peut être :

- soit un nom au singulier. Dans le cas où le pluriel est obtenu par une dérivation interne, il constitue une entrée séparée. Ces pluriels sont dits des “pluriels brisés” =(جمع التكسير) comme (كِلاَب : KilâB) =(« chiens ») pluriel de (كَلْب : KaLB) =(« chien »);
- soit un adjectif (non déverbal) au masculin singulier. Lorsque le féminin ou le pluriel de ces adjectifs ne sont pas réguliers, ils sont intégrés dans la base comme des nouvelles entrées distinctes. Par exemple, (زَرْقَاء : *ZaRQâ?*) =(« bleue ») féminin de (أَزْرَق : *?aZRaQ*) =(« bleu » constitue une entrée lexicale séparée.

A chaque entrée nominale, est associé un spécificateur morpho-syntaxique permettent de définir notamment (Voir figure 2-5) :

- si l’entrée est formée par un noyau simple ou si l’entrée est obtenue par une suffixation. Dans ce dernier cas, le proclitique (لْ) et/ou le(s) suffixe(s) doivent être déterminés.
- le genre (masculin, féminin) et le nombre (singulier, pluriel, collectif).
- si l’entrée constitue un humain ou un non humain.
- le modèle de déclinaison (11 modèles sont recensés dans DIINAR). Chaque modèle permet de générer 9 formes déclinées selon le mode (indéterminé, déterminé par annexion, déterminé par l'article) et le cas (nominatif, accusatif, génitif) de déclinaison.
- si l’entrée est un nom propre.
- si l’entrée admet des duels (i.e. suffixes انِ, نْيَ), des pluriels (i.e. suffixes اتُ, ونَ, ينَ), un féminin (i.e. suffixe ةَ) ou un relatif (i.e. suffixe يّ) obtenus par suffixation.
- Si l’entrée admet un masculin, un féminin, un pluriel brisé ou un relatif obtenus par une dérivation interne.
- Si l’entrée peut être définie avec l’article (لْ).

24 **Ghenima (1998), (Brahem & Ghazeli, 1998) et (Dichy, Hassoun, Zaafrani, 2002 a).**

2.4 Modélisation des mots outils

La classe des mots outils est un fourre-tout, qui regroupe toutes les parties restantes du discours qui ne peuvent pas être classées dans la catégorie des verbes ou celle des noms, tels que les conjonctions, les prépositions et les pronoms. Les mots outils sont très fréquents dans les textes et la plupart de leurs éléments ont une répercussion syntaxique importante, soit qu'ils engagent la globalité de la phrase, soit qu'ils aient une incidence sur l'environnement immédiat.

Dans les systèmes d'analyse développés jusque-là par les membres de notre équipe, la reconnaissance de ces mots outils se faisait par la consultation directe d'une liste regroupant une bonne partie de leurs éléments et dépourvue d'informations linguistiques. Ainsi l'analyse ne permettait pas de décortiquer la structure morphologique des mots

outils composés et de récupérer des informations utiles à l'analyse syntaxique. Dans le cadre de l'environnement d'apprentissage « *AL-Mu* ^{c} *aLLiM* », les informations associées aux mots outils sont d'un grand apport à l'apprenant surtout lors d'un travail de compréhension, puisqu'elles lui permettent d'avoir des points de repère pour l'analyse des phrases.

Toutes ces considérations, nous ont poussé à concevoir une nouvelle base de données des mots outils que nous détaillerons dans cette section. Nous définirons d'abord la structure du mot outil qui va permettre son analyse et sa synthèse automatique et les critères syntaxiques et sémantiques qui sont associés aux entrées du dictionnaire. Nous présenterons ensuite le schéma général "entités-relations" de la base de données et nous montrerons en fin de section quelques interfaces reflétant les fonctionnalités les plus importantes du système réalisé.

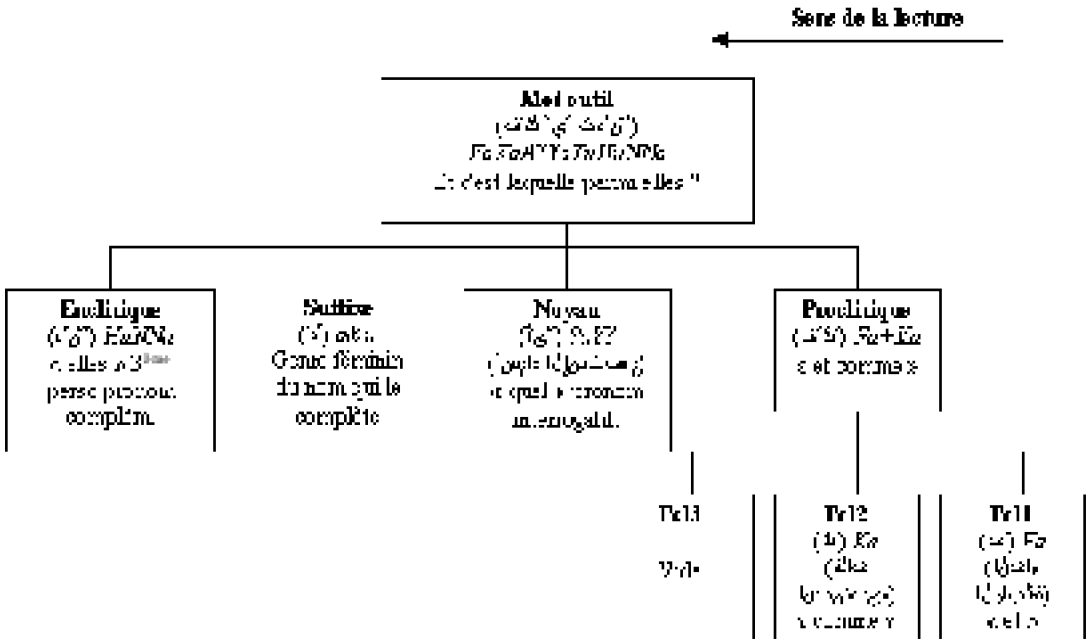
2.4.1 Conception linguistique[25]

A partir d'un premier travail de recensement des mots outils, nous avons dégagé une structure commune pour tous les mots outils. Chaque occurrence peut se décomposer en une suite ordonnée de : un à trois proclitique(s) (قبالسوا), d'un noyau (ةادألا), d'un suffixe (قحاللا) et d'un enclitique (قحلملا ريمضلا). A travers l'exemple de (فَكَأَ ّيَتَ ُهَنّ : *#Fa+Ka+AYYaTu+HuNNa#*), la figure (6-2) ci-dessous illustre ces différents morphèmes.

La modélisation linguistique, a ainsi adopté une structuration autour du noyau qui constitue l'entrée du dictionnaire, auquel pourrait s'ajouter l'ensemble des extensions cités ci-dessus. Afin de faciliter la recherche et la consultation du dictionnaire, nous avons tout d'abord organisé les entrées du dictionnaire. Nous avons, en effet, divisé cette importante masse de particules hétéroclites en trois grandes classes : les **particules simples** (فورح يناعملا), les **particules nominales** (تاودألا ءامسألا) et les **particules verbales** (لاعفألا ىضاملا).

Chaque classe a été à son tour subdivisée en des sous-classes plus fines selon des critères syntaxiques et sémantiques. La classe des particules simples (فورح يناعملا), a été divisée en trois sous-classes selon l'incidence syntaxique du mot outil sur son voisinage : les particules régissant uniquement les noms (ءامسألاب ةصّتخملا فورحلا), les particules régissant uniquement les verbes (لاعفألاب ةصّتخملا فورحلا) et les particules régissant les noms et les verbes (كةرتشملا فورحلا).

[25] **Cf. (Dichy, Hassoun, Mouelhi, Zaafrani, 2002).**



La sous-classe des *particules régissant uniquement les noms* regroupe les sous-classes des (حروف التنبيه), (حروف الجرّ), (حروف القسم), (حروف الاستثناء), (حروف النداء), (الحروف المشبّهة بالفعل), (حرف المفاجأة), (حرف التفصيل), (حرف التوجّع والإشفاق), (حرف التشبيه) et (الحروف المشبّهة بليس).

La sous-classe des *particules régissant uniquement les verbes* regroupe les sous-classes (حروف النصب), (حروف المصدريّة), (حروف الجزم), (حروف الشرط), (حروف التحضيض والتندي), (حروف الاستقبال), (حرف التوقّع), (حروف العَرْض).

La sous-classe des *particules régissant les noms et les verbes* regroupe quant à elle les sous-classes (حروف العطف), (حرف الاستفهام), (حرف التفسير), (حرف الاستفتاح), (حروف النفي), (حروف الجواب), (حروف التعليل), (حروف الصلة), (حروف التمنّي), (حروف التوكيد).

La classe des particules nominales (الأسماء الأدوات) a été divisée en trois sous-classes : (أسماء الإشارة), (الظروف), (أسماء الاستفهام), (أسماء الشرط), (الأسماء الموصولة), (أسماء الأفعال), (أسماء الأصوات), (الكنايات), (الضمائر).

La classe des particules verbales (الأفعال الجامدة) a été divisée aussi en trois sous-classes : (أفعال جامدة على صيغة الماضي), (أفعال جامدة على صيغة المضارع), (أفعال جامدة على صيغة الأمر).

Nous avons ensuite défini les propriétés morphologiques et syntaxiques qui ont été associées aux entrées du dictionnaire. Bien qu'uniquement une petite partie des mots outils (principalement les démonstratifs) diffèrent en genre et en nombre, nous avons retenu ces deux propriétés pour l'ensemble des mots outils. Pour ceux n'ayant pas de genre et/ou de nombre, nous avons ajouté les possibilités «sans genre» = (لا جنس) et «sans nombre» = (لا عدد). Nous avons aussi retenu le trait qui détermine si le mot outil est figé (أداة مبنيّة) ou pas, c'est à dire s'il s'accorde avec les prépositions (حروف الجرّ) ou pas. Lorsque le mot outil n'est pas figé, comme par exemple l'interrogatif (أيُّ : *?aYYu*) = («quel»), il s'accorde avec les prépositions (i.e. "بِ : *Bi*", "لِ : *Li*") et se réalise avec forme

différente (i.e. "ـَيَأ : *?aYYi*"). Par contre, les mots outils figés (i.e. le démonstratif "اذَه : *HaDâ*") ne s'accordent pas avec les prépositions et gardent leur forme initiale.

Les autres traits retenus sont d'ordre syntaxique et sont relatifs au voisinage immédiat du mot outil. Ces traits signalent, d'une part, les mots outils qui marquent obligatoirement le début de la phrase (i.e. les interrogatifs "اَلَأ : *?aLâ*" et "اَمَأ : *?aMâ*") et ceux qui peuvent se trouver au début de la phrase, et d'autre part, les catégories des mots qui peuvent suivre le mot outil : un nom, un verbe, un autre mot outil, une phrase débutant par "أن المصدرية" = *?aN* ou un signe de ponctuation marquant la fin d'une phrase (i.e. "مْعَن : *Na c aM*", "ىلَب : *BaLaÿ*"). Ces informations permettent de diminuer considérablement les ambiguïtés syntaxiques lors d'une analyse automatique.

Nous avons enfin recensé les affixes qui peuvent être suffixés au noyau du mot outil. Nous avons inventorié 21 suffixes et 40 pronoms compléments qui peuvent être suffixés aux mots outils. Les pronoms compléments ont été répartis sur trois vecteurs : Le premier vecteur contient la liste des *pronoms compléments des verbes* = (الضمائر المختصة الفعال), le second vecteur correspond à la liste des *pronoms compléments des noms* = (الضمائر المختصة بالإسم) et le dernier vecteur correspond à *la liste des pronoms compléments non-humains* = (الضمائر غير العاقل). Les deux premiers contiennent 18 pronoms correspondant aux différentes combinaisons du nombre (singulier, duel, pluriel), du genre (masculin, féminin) et de la personne (1ère, 2ème, 3ème personne), alors que le dernier vecteur ne contient que les 4 pronoms relatifs à la 3ème personne du singulier et du duel (au masculin et au féminin).

Il est à signaler que certains mots outils sont compatibles avec des pronoms appartenant à deux vecteurs différents. Nous avons finalement répertorié la liste des proclitiques[26] qui peuvent être associés aux mots outils. Nous avons obtenu une liste de 15 morphèmes, qui peuvent se combiner entre eux pour former des proclitiques composés (doubles ou triples).

2.4.2 Modélisation de la base de données des mots outils

La base de données des mots outils comprend 8 entités et 2 relations[27] (figure 2-7) :

- Entité "catégories" : Cette entité contient les trois grandes classes des mots outils : les **particules simples** = (الحروف العاملة), les **particules nominales** = (الأسماء الأدوات) et les **particules verbales** = (الأفعال الجامدة). Chaque classe est identifiée par un numéro.

[26] Chaque proclitique est lui-même un mot outil simple formé d'un seul caractère.

[27] Sur la première colonne des tableaux de description des entités, les symboles (2) et (ind) signifient que les champs correspondants sont respectivement une clé et un index de l'entité.

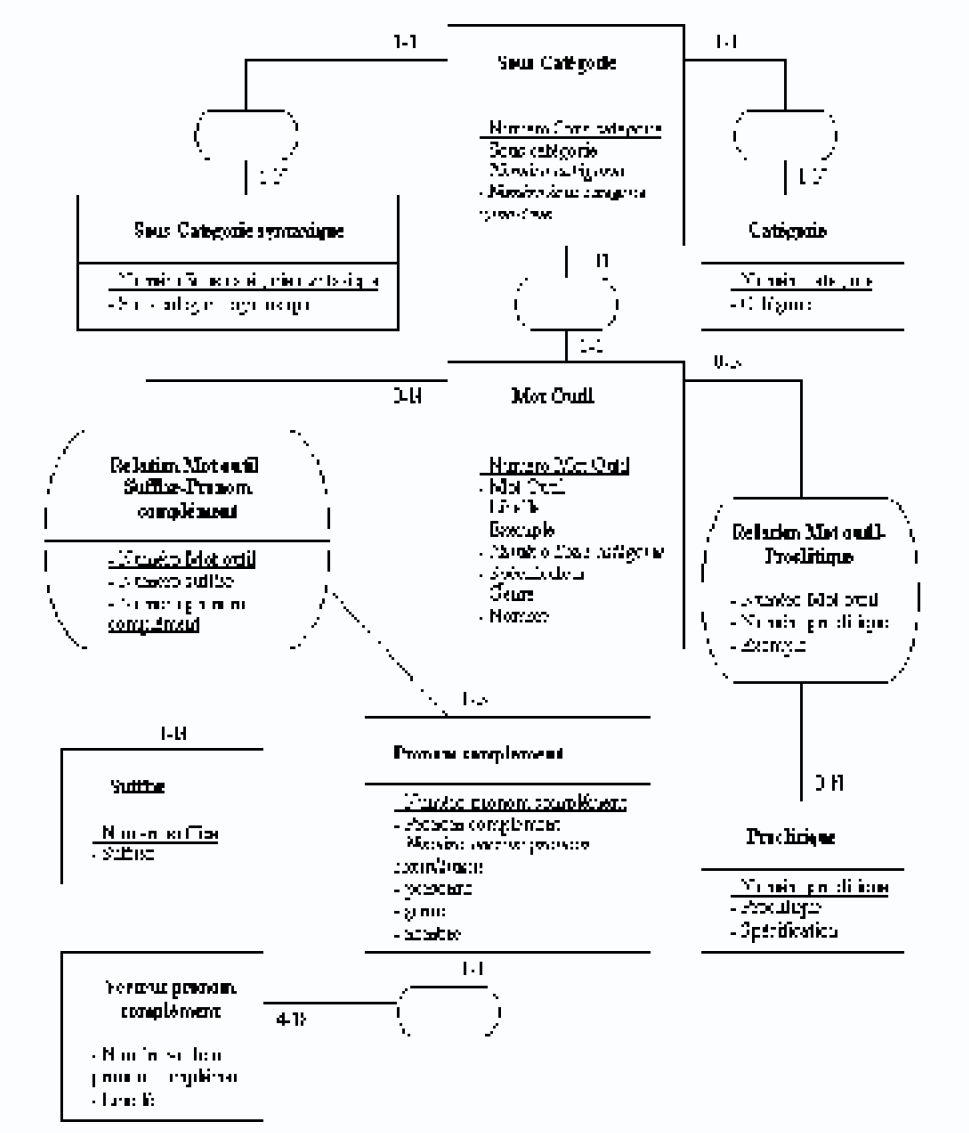
	Nom du champ	**Type des données**	**Description**
🔑	Ncatégorie	Numérique(Octet)	Numéro de la catégorie du mot outil
Ind	Catégorie	Texte(20)	Le libellé de la catégorie du mot outil

· Entité "sous-catégories" : Cette entité contient les 41 classes sémantiques des mots outils (cf. § 2.4.1). Chaque occurrence est relative à une grande catégorie et à une sous-catégorie syntaxique.

	Nom du champ	**Type des données**	**Description**
🔑	NSousCatégorie	Numérique(Entier)	Numéro de la sous-catégorie du mot outil
Ind	SousCatégorie	Texte(30)	Le libellé de la sous-catégorie du mot outil
Ind	NCatégorie	Numérique(Octet)	Numéro de la catégorie du mot outil
Ind	NsousCatégorieSyntaxique	Numérique(Octet)	Numéro de la sous-catégorie du mot outil

· Entité "Sous-catégories syntaxiques" : Cette entité contient les 3 sous-catégories syntaxiques relatives aux particules simples (cf. § 2.4.1).

	Nom du champ	**Type des données**	**Description**
🔑	NsousCatégorieSyntaxique	Numérique(Octet)	Numéro de la sous-catégorie syntaxique du mot outil
Ind	SousCatégorieSyntaxique	Texte(20)	Le libellé de la sous-catégorie syntaxique du mot outil



- Entité "mots outils : Cette entité contient actuellement environ 442 noyaux de mots outils. Chaque occurrence est relative à une sous-catégorie sémantique. A chaque occurrence est associé un spécificateur permettant d'identifier les propriétés syntaxiques du mot outil ainsi que deux autres propriétés permettant de définir son genre et son nombre.

	Nom du champ	Type des données	Description
🔑	NMot Outil	Numérique (Entier)	Numéro du mot outil
Ind	MotOutil	Texte(15)	Le noyau du mot outil
	Libellé	Texte(30)	Libellé du mot outil
	Exemple	Texte(100)	Exemple général employant le mot outil
Ind	Nsous Catégorie	Numérique (Octet)	Numéro de la sous-catégorie sémantique du mot outil
	Spécificateur	Texte(1)	Bit(1) : Si le mot outil est figé ou pas, bit(2) : S'il peut débuter la phrase, bit(3) : S'il débute obligatoirement la phrase, bit(4) : s'il peut être suivi par un nom, bit(5) : s'il peut être suivi par un verbe, bit(6) : s'il peut être suivi par un autre mot outil, bit(7) : s'il peut être suivi par (أن المصدرية) *?n*, bit(8) : s'il peut clôturer une phrase
	Genre	Texte(1)	1 : Masculin, 2 : Féminin, 3 : Commun, 4 : Sans genre
	Nombre	Texte(1)	1 : Singulier, 2 : Duel, 3 : Pluriel, 4 : Sans Nombre

· Entité "proclitiques" : Cette entité contient les 2954 combinaisons possibles des proclitiques. Chaque identifiant est obtenu par la concaténation de 3 identifiants de proclitiques simples (00-15).

	Nom du champ	Type des données	Description
🔑	Nproclitique	Texte(6)	Numéro du proclitique du mot outil
Ind	Proclitique	Texte(15)	Le proclitique du mot outil
	Spécification	Texte(100)	La spécification du proclitique du mot outil

· Entité "suffixes" : Cette entité contient les 21 suffixes possibles.

	Nom du champ	Type des données	Description
🔑	NSuffixe	Texte(2)	Numéro du suffixe du mot outil
	Suffixe	Texte(10)	Le suffixe du mot outil

· Entité "vecteurs des pronoms compléments" : Cette entité contient les 3 vecteurs pronoms compléments possibles (cf. § 2.4.1).

	Nom du champ	**Type des données**	**Description**
🔑	NVecteur pronom	Numérique(Octet)	Numéro du vecteur pronom complément
	Libellé francais	Texte(70)	Spécification du vecteur des pronoms compléments en français
	Libellé arabe	Texte(50)	Spécification du vecteur des pronoms compléments en arabe

· Entité "pronoms compléments" : Cette entité contient les 40 pronoms compléments possibles. Chaque pronom complément est relatif à un vecteur. A chaque occurrence est associé les propriétés permettant de définir la personne, le genre et le nombre du pronom complément.

	Nom du champ	**Type des données**	**Description**
🔑	Npronom complément	Numérique(Octet)	Numéro du pronom complément
	Pronom complément	Texte(10)	Le pronom complément
	NVecteur pronom	Numérique(Octet)	Numéro du vecteur pronom complément
	Personne	Texte(20)	La personne relative au pronom complément
	Genre	Texte(20)	Le genre de la personne relative au pronom complément
	Nombre	Texte(20)	Le nombre de la personne relative au pronom complément
	Libéllé	Texte(20)	Libéllé du pronom complément

· Relation "Proclitique-Mot outil" : Cette entité joue le rôle de relation entre les entités "Mot outil" et "Proclitique".

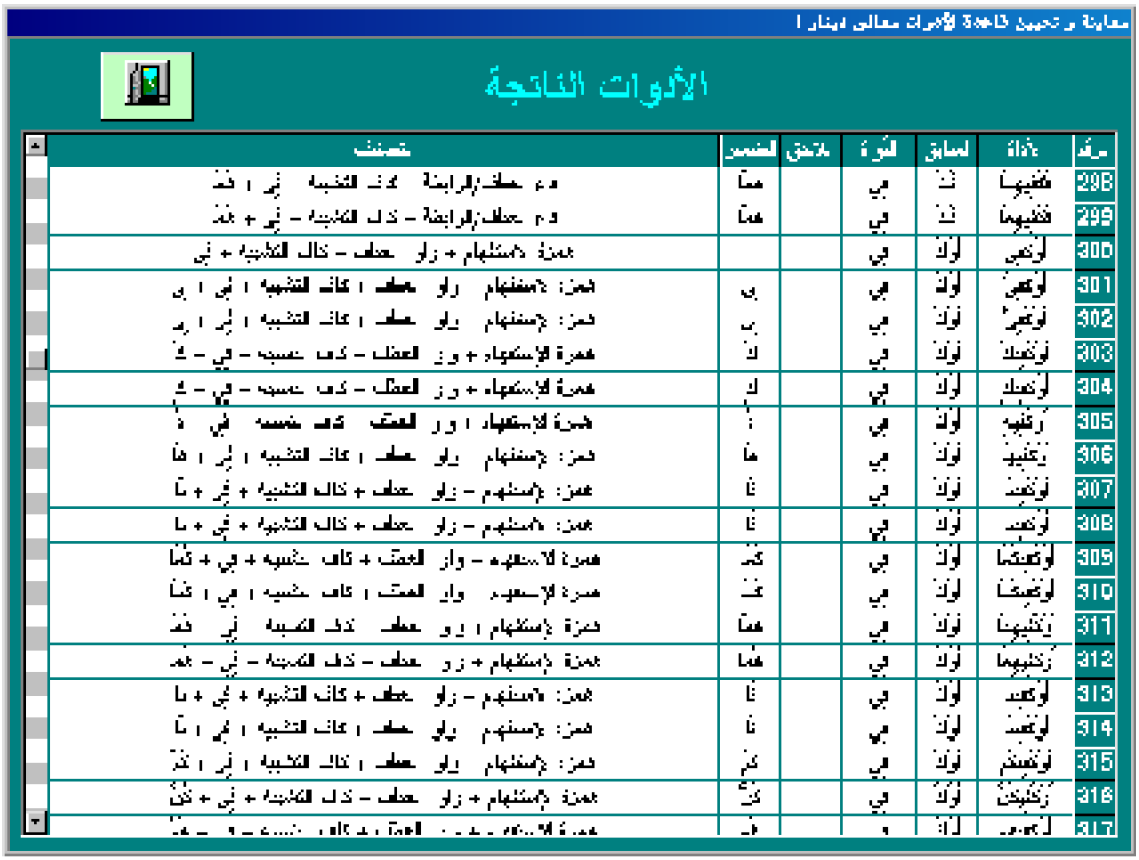
	Nom du champ	**Type des données**	**Description**
🔑	NMotOutil	Numérique (Entier)	Numéro du mot outil
🔑	Nproclitique	Texte(6)	Numéro du proclitique
	Exemple	Texte(150)	Exemple d'utilisation du proclitique avec le mot outil dans une phrase
	Interrogation	Texte(1)	1 : Si l'expert n'est pas sûr de la relation, 0 : sinon

Relation "Mot outil-Suffixe-Pronom complément" : Cette entité joue le rôle de relation entre les entités "Mot outil", "Suffixe" et "Pronom complément".

	Nom du champ	Type des données	Description
🔑	NMotOutil	Numérique (Entier)	Numéro du mot outil
🔑	NSuffixe	Texte(2)	Numéro du suffixe
🔑	NPronom complément	Numérique (Octet)	Numéro du proclitique

2.4.3 Présentation des interfaces de saisie et de mise à jour

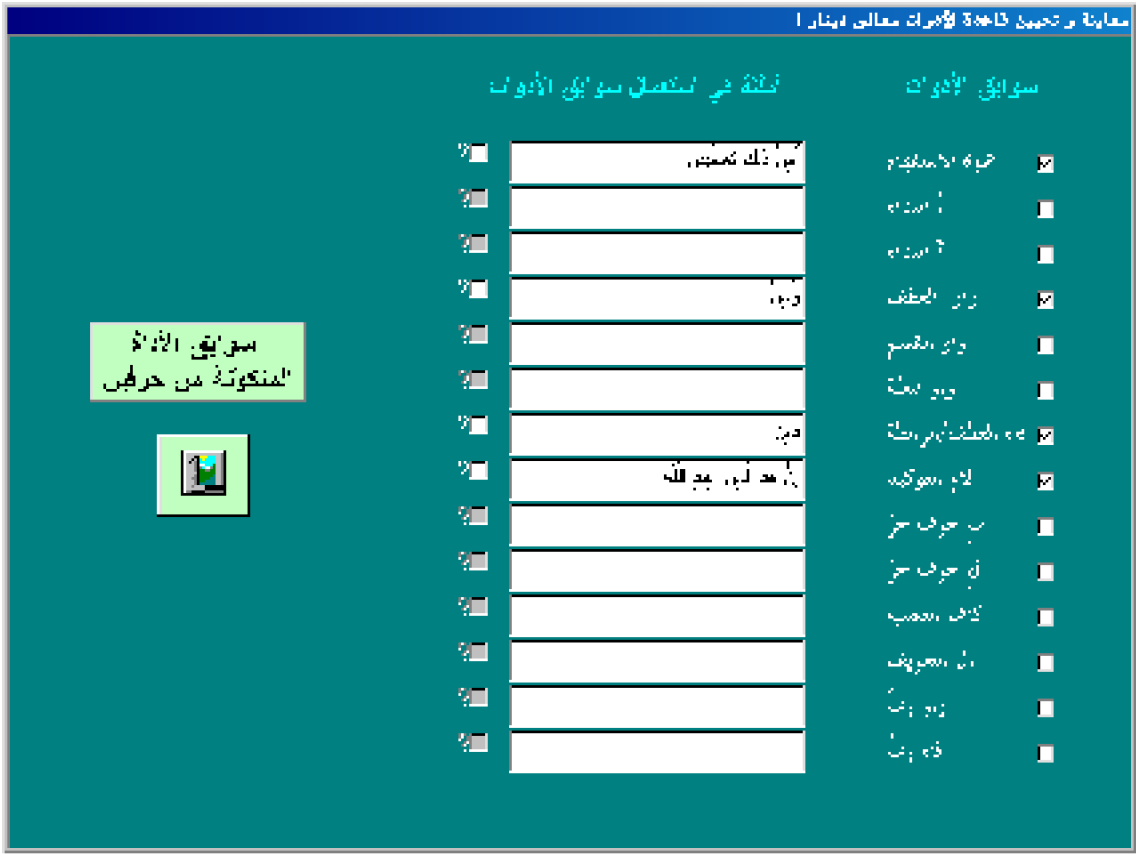
Afin de faciliter la saisie et la mise et à jour des données, nous avons conçu et réalisé un certain nombre d'interfaces qui fonctionnent autour de l'organisation de la base de données des mots outils.



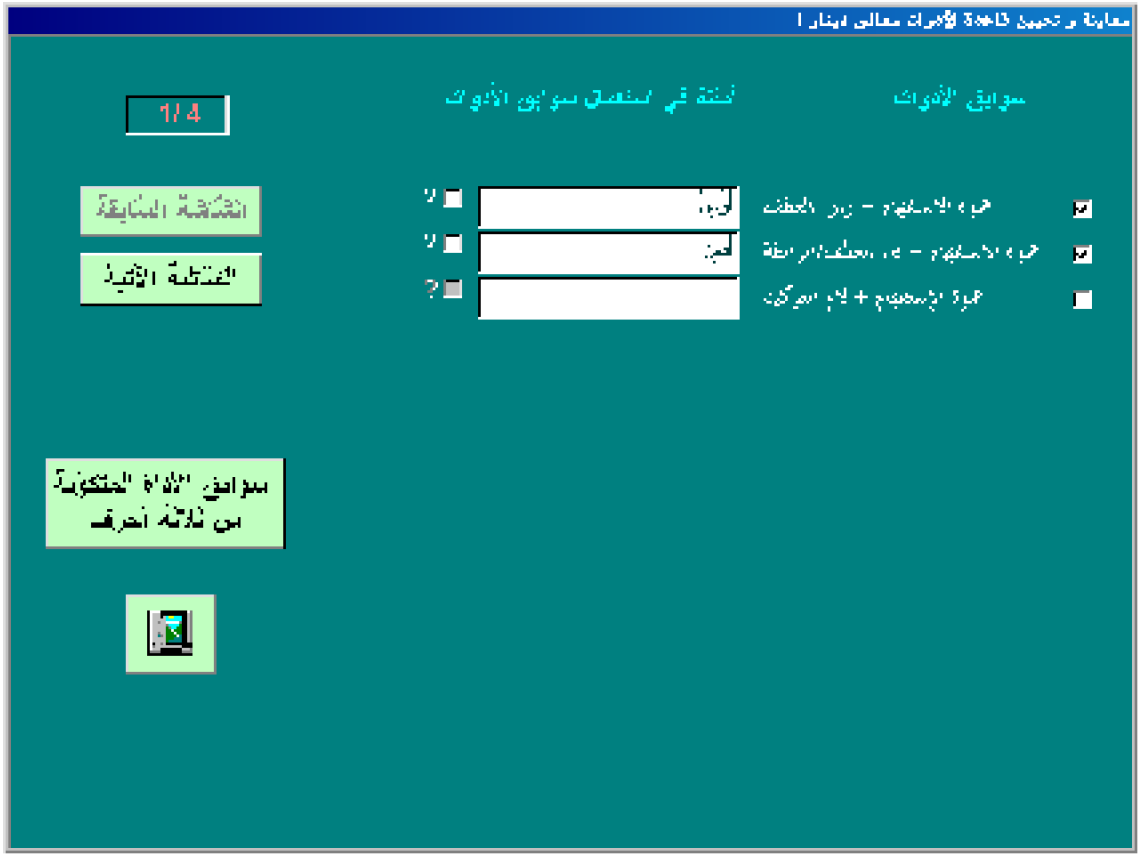
Dans ce paragraphe, nous n'allons pas revenir sur la réalisation de toutes ces interfaces, mais nous allons juste revenir sur quelques fonctionnalités qui ont été utiles à

l'expert linguiste lors de l'élaboration de ce dictionnaire :

- Génération instantanée des formes fléchies des mots outils : Cette fonctionnalité permet de générer automatiquement toutes les formes fléchies à partir du noyau du mot outil. Elle permet à l'utilisateur du système de vérifier l'incidence du choix des affixes sur les résultats de la génération et de pallier éventuellement les erreurs par l'ajout de nouvelles règles morphologiques. Cette fonctionnalité nous a permis, en effet, de corriger certaines formes fléchies erronées par l'ajout de nouvelles règles morphologiques. Par exemple, les pronoms compléments (هُمْ : *HuM*), (هُنَّ : *HuNNa*), (هُمَا : *HuMâ*), (هُ : *Hu*) se transforment respectivement en (هِمْ : *HiM*), (هِنَّ : *HiNNA*), (هِمَا : *HiMâ*), (هِ : *Hi*) lorsqu'ils sont précédés par la voyelle (ـِ : *i*) ou par la voyelle longue (ـِي : *î*) : Règle « harmonique vocalique ». La figure (2-8) ci-dessus, qui illustre l'interface de consultation des formes fléchies, montre par exemple qu'à partir de la suffixation du pronom (هُمَا : *HuMâ*) à la préposition (فِي : *Fi*) nous obtenons la forme correcte (فِيهِمَا : *FîHiMâ*) (ligne 298) au lieu de (فِيهُمَا* : *FiHuMâ*). Sur cette même figure, nous pouvons remarquer également qu'à partir de l'application d'une autre règle morphologique, nous obtenons (فِيَّ : *FiYYa*) à partir de la concaténation de (فِي : *Fî*) et du pronom personnel (ـِي : *iY*) (ligne 301) au lieu de (فِي يِ* : *FîiY*).



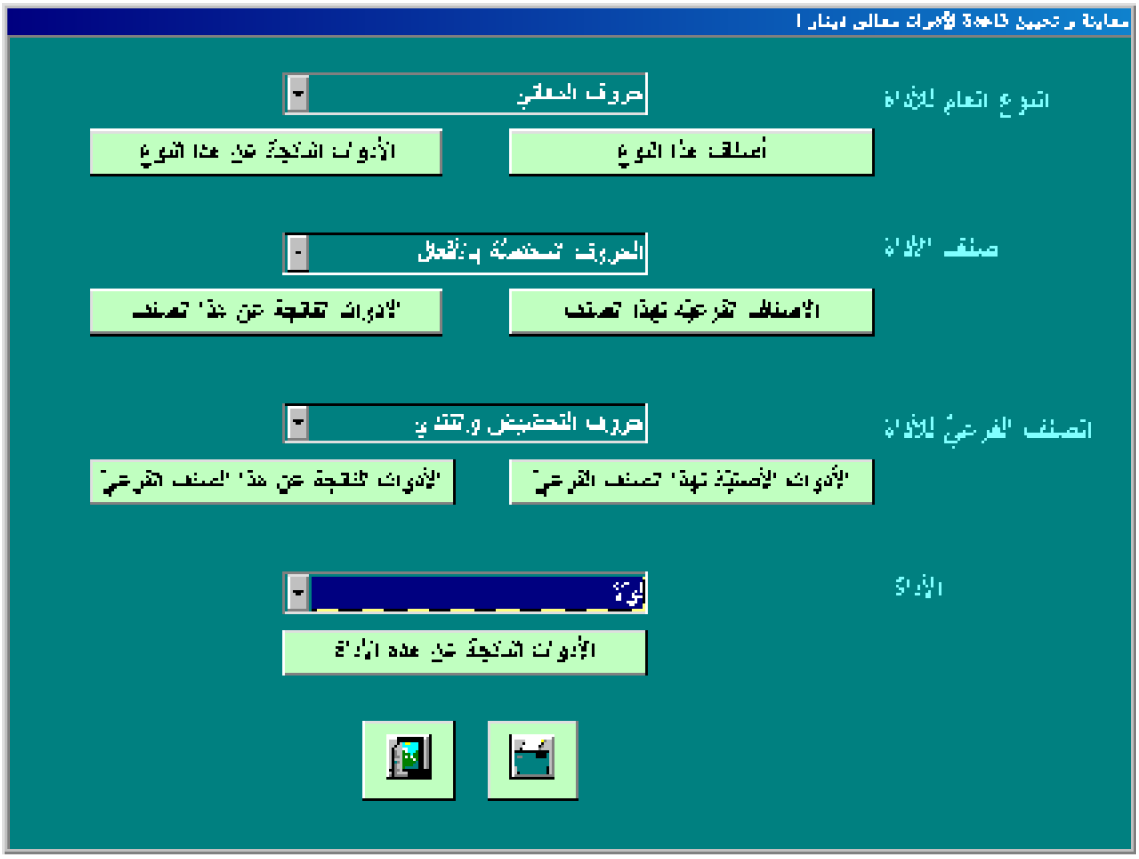
- Génération automatique des proclitiques composés (possibles) : Cette deuxième fonctionnalité assiste l'expert lors de la saisie des proclitiques composés. La figure (2-9), illustre les quatre proclitiques simples qui sont compatibles avec le mot outil (مِنْ : *MiN*). Ces proclitiques peuvent se combiner entre eux pour former des proclitiques doubles ou triples. Cette fonctionnalité permet ainsi de limiter les propositions de choix aux seules combinaisons de ces quatre proclitiques, soit au total quatre écrans au lieu de 15, chacun correspondant aux combinaisons d'un proclitique avec les trois autres.



Cette fonctionnalité est accessible à partir de l'écran de saisie des proclitiques simples, par un simple clic sur le bouton (Proclitiques doubles : سوابق الأداة المتكوّنة من حرفين). La figure (2-10) montre le premier écran relatif aux proclitique doubles commençant par (أَ : *?a*) = (همزة الإستفهام). L'utilisateur peut naviguer entre les différents écrans à l'aide des boutons (Ecran suivant : الشاشة الآتية) et (Ecran précédent : الشاشة السابقة). Il peut passer directement aux proclitiques triples, par un nouveau clic sur le bouton (Proclitiques triples : سوابق الأداة المتكوّنة من ثلاثة أحرف). A partir des choix effectués sur l'écran de la figure (2-10) (i.e. 2 proclitiques doubles sélectionnés), 8 nouveaux écrans de choix de proclitiques triples seront proposés à l'expert (4 écrans pour chaque proclitique double qui sera combiné avec les 4 proclitiques simples déjà sélectionnés). Cette fonctionnalité a ainsi permis d'éviter à l'expert d'avoir à parcourir les 2954 combinaisons de proclitiques composés possibles lors de chaque saisie d'un mot outil.

· Impression des listes des mots outils par catégorie : La dernière fonctionnalité permet

l'impression papier des mots outils relatifs à une catégorie, à une sous-catégorie sémantique ou à une sous-catégorie syntaxique. L'impression papier, l'outil de travail préféré des linguistes, permet d'effectuer un contrôle minutieux des résultats de génération. Nous avons prévu pour cet effet une interface indépendante, qui permet d'imprimer aussi les différentes catégories des mots outils utilisées (figure 2-11).



2.5 Modélisation des noms propres[28]

Les échecs des systèmes d'analyse automatique des langues sont dus, en grande partie, à l'absence de noms propres du dictionnaire. Afin de minimiser ces échecs, nous avons

[28] Cf. (Dichy, Hassoun, Zaafrani, 2002 b)

conçu et réalisé une nouvelle base de données des noms propres que nous détaillerons dans cette section. Nous décrirons d'abord la conception de cette base. Nous présenterons ensuite son schéma "Entités-Relations" et nous décrirons enfin les entités de cette base.

2.5.1 Conception de la base de données

Pour réaliser cette base de données des noms propres, nous nous sommes basés sur la modélisation des noms (cf. § 2.3.3) qui présente beaucoup de traits communs avec cette nouvelle base de données dont notamment la partie relative aux modèles de déclinaison.

Nous avons tout d'abord défini les entrées du dictionnaire. Nous étions confrontés à deux problèmes de choix où il fallait trancher. Le premier problème est dû aux multiples transcriptions d'un même nom propre « Les noms propres n'ont, dit-on, ni orthographe ni prononciation ». Faut-il définir une seule entrée principale et la lier aux autres transcriptions ou utiliser plusieurs entrées différentes. Nous avons opté pour ce dernier choix parce que les différentes transcriptions peuvent se décliner différemment et le fait de lier deux transcriptions relatives au même nom propre n'apporte aucun apport au processus d'analyse. Le second problème concerne les noms propres composés. Faut-il avoir une entrée pour chaque élément du nom propre composé et les associer par une relation ou bien les considérer comme un tout constituant une seule entrée. Nous avons opté pour ce dernier choix puisque les composants ne constituent pas forcément des noms propres. Par exemple, « الدّين : *aDDîN* » de « عالم الدّين : *Lâ? ADDîN* » ne constitue pas un nom propre.

Afin notamment de faciliter la recherche et la navigation dans le dictionnaire, nous avons ensuite essayé d'organiser les entrées du dictionnaire dans différentes catégories selon des critères sémantiques. Nous les avons ainsi divisés en trois principales divisons : les **noms des lieux** (اسم مكان), les **noms des personnes** (اسم شخص) et les **autres noms propres** (اسم آخر), qui étaient à leur tour distribués sur d'autres subdivisons plus fines.

Dans cette première réalisation, la catégorie des "noms des lieux" regroupe les subdivisions *pays* = (بلدان), *capitales* = (عواصم), *grandes villes* = (مدن مشهورة), *monuments historiques* = (أماكن تاريخيّة), *fleuves* = (أنهار), *montagnes* = (جبال), *lacs* = (بحيرات), *continents* = (قارّات) et *océans* = (بحار). Celle des "noms des personnes" regroupe les *noms courants* = (اسم علم متداول), les *surnoms* = (لقب كنية متداول), les *personnalités* = (شخصيّات) et les *établissements* = (مؤسّسات). Enfin, la subdivision "autres noms propres" regroupe quant à elle les *jours de semaine* = (أيّام السبوع), les *mois* = (الشّهور), les *nombres cardinaux* = (الأعداد الأصليّة), les *nombres ordinaux* = (الأعداد الترتيبيّة), les *raccourcis* = (اختصارات), les *nations et populations* = (شعوب وقبائل), les *dynasties* = (سلالة) et les *noms propres non caractérisés* = (غير مبوّب).

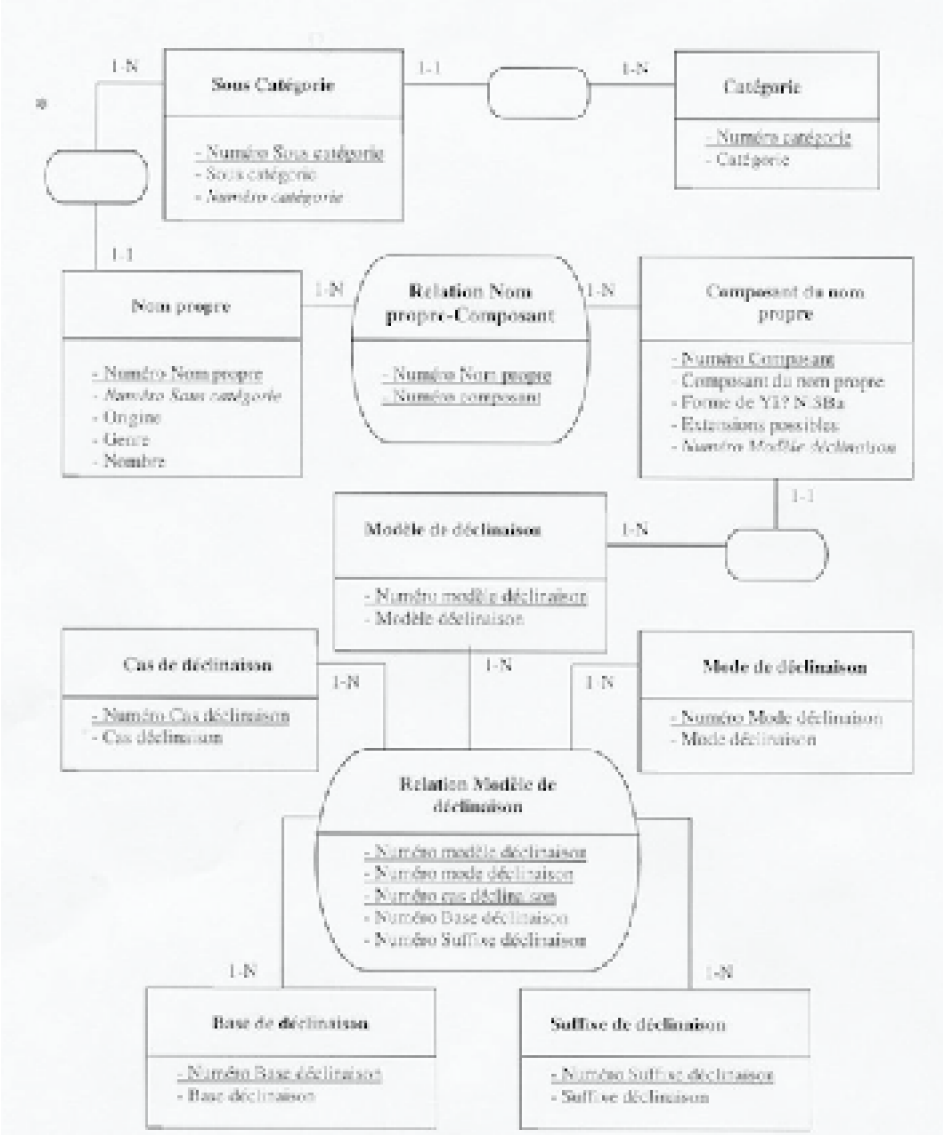
Nous avons enfin défini et structuré les entités de la base de données à partir des règles linguistiques y relatives. Nous avons constaté qu'un nombre important de noms propres composés partagent le même composant, comme par exemple les prénoms composés de (عبد : *BD*) (=«adorateur») et de l'un des 99 attributs de dieu = (اسم الله

الحسنى). Afin d'éviter la redondance du même composant dans plusieurs entrées différentes, nous avons créé deux entités indépendantes : Une première entité contenant uniquement les indices des noms propres et une seconde entité contenant les composants qui sont reliés par une entité relation définie par la concaténation de leur clé (figure 2-12). Cette décomposition a l'avantage d'accélérer la saisie, puisqu'elle évite de ressaisir les propriétés des composants déjà saisis.

Nous avons, d'autre part, associé trois propriétés aux entrées du dictionnaires qu'elles soient simples ou composées : son origine (arabe / étrangère), son genre (masculin, féminin) et son nombre (singulier, duel, pluriel). Afin de pouvoir générer les formes fléchies à partir du (des) composant(s) du nom propre, nous avons aussi retenu les informations suivantes :

- Les différents affixes acceptés par les composants : l'article (ٱلـ : *al*), les suffixes du duel (ـانِ : *âNi*), (ـيْنَ : *aYNi*), les suffixes du pluriel (ـاتُ : *âT*), (ـونَ : *ûN*), (ـينِ : *îN*), le suffixe du féminin (ـةَ : *a&*) et le suffixe de l'adjectif de relation (ـيّ : *iYY*).
- Le noyau (base) de l'adjectif de relation lorsqu'il est différent de celui du nom propre.
- Le modèle de déclinaison pour déterminer les suffixes du nom propre selon le mode (indéterminé, déterminé par annexion, déterminé par l'article) et le cas (nominatif, accusatif, génitif). Nous avons ajouté aux 11 modèles de déclinaison préalablement définis pour les noms, 8 nouveaux modèles pour tenir compte des spécificités des noms propres dont notamment l'absence du mode "déterminé par l'article".

2.5.2 Présentation du Schéma du MLD



2.5.3 Présentation des entités du dictionnaire

La base de données des noms propres comprend 9 entités et 2 relations (figure 2-12) :

· Entité "catégories" : Cette entité regroupe les trois principales divisions des noms propres les **noms des lieux** = (أسماء الأماكن), les **noms des personnes** = (أسماء الأشخاص) et les **autres noms propres** = (أسماء أخرى). Chaque classe est identifiée par un numéro de catégorie.

	Nom du champ	**Type des données**	**Description**
🔑	NumCatégorie	Numérique (Octet)	Numéro de la catégorie du nom propre
Ind	Catégorie	Texte (20)	Le libellé de la catégorie du nom propre

· Entité "sous-catégorie" : Cette entité contient les 21 subdivisions des noms propres (cf. § 2.5.1). Chaque occurrence est relative à une catégorie.

	Nom du champ	**Type des données**	**Description**
🔑	NumSousCatégorieNP	Numérique(Octet)	Numéro de la sous-catégorie
Ind	SousCatégorieNP	Texte(20)	Le libellé de la sous-catégorie
Ind	NumCatégorieNP	Numérique(Octet)	Numéro de la catégorie

· Entité "Nom propre" : Cette entité contient les entrées du dictionnaire. Chaque entrée appartient à une division particulière de noms propres et est définie par son origine, son genre et son nombre.

	Nom du champ	**Type des données**	**Description**
🔑	NumNomPropre	Numérique(Entier)	Numéro du nom propre
Ind	NumSousCatégorieNP	Numérique(Octet)	Numéro de la sous-catégorie
	Origine	Texte(1)	1 : Arabe, 2 : Non arabe, 3 : général
	Genre	Texte(1)	1 : Masculin, 2 : Féminin
	Nombre	Texte(1)	1 : Singulier, 2 : Duel, 3 : Pluriel

· Entité "Composant du nom propre" : Cette entité contient les éléments de chaque nom propre, qu'il soit simple (un seul élément) ou composé (plusieurs éléments). A Chaque composant sont associés les différents affixes acceptés, la base du nom de relation et le numéro de son modèle de déclinaison.

	Nom du champ	**Type des données**	**Description**
🔑	NumComposant	Numérique(Entier)	Numéro du composant du nom propre
	NomComposant	Texte(30)	Le composant du nom propre
	FormeNPAvecYa	Texte(20)	La base de l'adjectif nom de relation lorsqu'elle est différente du composant
	Affixes_possibles	Texte(1)	Bit(1) : (ـِيّ : *iYY*), bit(2) : (ـَة : *a&*), bit(3) : (ـاَت : *âT*), bit(4) : (ـُون : *ûN*), bit(5) : (ـِين : *îN*), bit(6) : (ـاَن : *âN*), bit(7) : (ـَيْن : *aYN*), bit(8) : (ـاَل : *aL*).
	NModèle de déclinaison	Numérique(Octet)	Le numéro du modèle de déclinaison

· Relation "Noms Propre-Composant" : Cette entité joue le rôle de relation entre les entités "Nom propre" et "Composant".

	Nom du champ	Type des données	Description
🔑	NumNomPropre	Numérique(Entier)	Numéro du nom propre
🔑	NumComposant	Numérique(Entier)	Numéro du composant du nom propre

· Entité "Modèle de déclinaison" : Cette entité contient les 19 modèles de déclinaison.

	Nom du champ	Type des données	Description
🔑	NModèle de déclinaison	Numérique(Octet)	Le numéro du modèle de déclinaison
	Modèle de déclinaison	Texte(20)	Le libellé du modèle de déclinaison

· Entité "Mode de déclinaison" : Cette entité contient les 3 modes de déclinaison (indéterminé, déterminé par annexion, déterminé par l'article).

	Nom du champ	Type des données	Description
🔑	NMode	Numérique(Octet)	Le numéro du mode de déclinaison
	Mode	Texte(20)	Le libellé du mode de déclinaison

· Entité "Cas de déclinaison" : Cette entité contient les 3 modes de déclinaison (nominatif, accusatif, génitif).

	Nom du champ	Type des données	Description
🔑	NCas	Numérique(Octet)	Le numéro du cas de déclinaison
	Cas	Texte(20)	Le libellé du cas de déclinaison

· Relation "Modèle de déclinaison" : Cette entité permet à partir d'un modèle, d'un mode et d'un cas de déclinaison de déterminer la base et le suffixe de déclinaison.

	Nom du champ	Type des données	Description
🔑	NModèle de déclinaison	Numérique(Octet)	Le numéro du modèle de déclinaison
🔑	NMode	Numérique(Octet)	Le numéro du mode de déclinaison
🔑	NCas	Numérique(Octet)	Le numéro du cas de déclinaison
	NBase de déclinaison	Numérique(Octet)	Le numéro de la base de déclinaison
	NSuffixe de déclinaison	Numérique(Octet)	Le numéro du suffixe de déclinaison

· Entité "Base de déclinaison" : Cette entité contient les 25 schèmes des bases de déclinaison.

	Nom du champ	Type des données	Description
🔑	NBase de déclinaison	Numérique(Octet)	Le numéro de base de déclinaison
	Base de déclinaison	Texte(20)	La base de déclinaison

· Entité "Suffixe de déclinaison" : Cette entité contient les 13 suffixes de déclinaison.

	Nom du champ	Type des données	Description
🔑	NSuffixe de déclinaison	Numérique(Octet)	Le numéro de suffixe de déclinaison
	Suffixe de déclinaison	Texte(20)	La suffixe de déclinaison

2.6 Maintenance de la base de données lexicale DIINAR.1

Dans un système d'analyse linguistique utilisant un dictionnaire, le traitement des échecs n'est autre qu'un problème de maintenance du dictionnaire. En effet, ces échecs sont dus, soit à une simple absence des mots du dictionnaire, soit à une insuffisance des traits linguistiques. Ainsi, les modifications que nous apportons à la structure et au contenu du dictionnaire DIINAR permettent de pallier les échecs d'analyse et d'améliorer les performances des applications.

L'enrichissement de la base de données lexicale par des modules de gestion des mots outils et des noms propres, a constitué un premier apport au travail de maintenance de la base. Nous avons, d'autre part, apporté quelques modifications à la structure de la base de données verbale dont notamment l'association d'un schéma syntaxique à chaque entrée verbale. Ces schémas sont certes rudimentaires et ne permettent pas à eux seuls d'obtenir une analyse syntaxique performante, mais ils permettent de diminuer le nombre d'ambiguïtés lors de l'analyse syntaxique[29].

Chaque schéma syntaxique admet un sujet et peut avoir jusqu'à trois compléments directs ou indirects[30] lorsque le verbe est transitif. Chaque argument de ce schéma est défini par un ensemble de traits reflétant sa nature (nom simple – syntagme nominal, etc.) et ses propriétés catégorielles (humain / non humain / concret / abstrait / animé / non animé).

[29] Ces informations vont notamment pouvoir être intégrées dans l'analyseur syntaxique LARUSA, qui fonctionne à partir d'une grammaire de format AGFL (Ditters, 1992) obtenue à partir d'un lexique généré de la base de données lexicale DIINAR.1 (ouersighni, 2001) et (ouersighni, 1998).

[30] Lorsque le complément est indirect, la préposition est spécifiée parmi l'une de ces huit prépositions : نَيْ‌بَ – مْ‌نِ – لِ‌ىَ – نْ‌عَ – لِ – بِ – فِي – لَ‌ىَ.

Le travail de maintenance a consisté aussi en la mise à jour des données de la base. Nous avons ainsi associé aux différentes applications une base de données "corbeille" qui permet d'enregistrer toutes les chaînes de caractères non analysées par le système avec leur contexte. Pour le traitement de ces échecs, nous avons mis à la disposition de l'expert linguiste un module qui parcourt la corbeille et qui lui permettra d'accéder aux différentes interfaces de DIINAR afin d'effectuer les mises à jour nécessaires.

2.7 Conclusion

Comme nous l'avons introduit au début de ce chapitre, l'objectif de la construction du dictionnaire électronique DIINAR est double : la génération du lexique qui permettra une analyse robuste du mot graphique arabe (cf. chapitre 3) et l'élaboration d'activités d'apprentissage (cf. chapitre 7).

Nous avons par conséquent essayé d'inclure un maximum d'informations de natures diverses qui permettent de diminuer les ambiguïtés d'analyse et d'améliorer les diverses compétences linguistiques de l'apprenant. C'est dans ce sens que nous avons pris la responsabilité de la maintenance du dictionnaire DIINAR.1, que nous avons enrichi par de nouveaux spécificateurs (i.e. bases de données verbale et nominale) et par de nouvelles parties pour gérer les mots outils et les noms propres.

Pour pouvoir gérer l'énorme quantité d'informations du dictionnaire, on a eu recourt aux techniques des bases de données, plus particulièrement le modèle relationnel, d'où le terme *base de données lexicale*. Dans le chapitre suivant, nous allons expliquer comment cette base de données va être utilisée pour générer le lexique qui sera à son tour utilisé par l'analyseur des mots graphiques.

Chapitre 3 L'analyseur morpho-syntaxique des mots graphiques

« Le malheur des humains vient de ce que trop d'entre eux n'ont jamais compris que les mots ne sont que des outils à leur disposition, et que la seule présence d'un mot dans le dictionnaire (le mot "vivant" par exemple) ne signifie pas que ce mot se rapporte forcément à quelque chose de défini dans le monde réel. »
Richard DAWKINS

3.1 Introduction

La majorité des applications qui mettent en jeu du texte, notamment celles relevant de la syntaxe, verraient leurs performances s'améliorer par l'intégration de meilleurs systèmes d'analyse morphologique.

Un analyseur morphologique est communément défini, comme un programme qui permet de reconnaître un même mot sous les diverses formes qu'il peut prendre dans les phrases. Pour chaque forme trouvée, il doit isoler ses différents éléments et déduire les traits morphologiques et syntaxiques hors contexte qui leurs sont associés.

Théoriquement, on peut s'en passer de l'analyseur morphologique en construisant un lexique contenant l'ensemble des formes fléchies et en associant à chaque forme ses traits. Cette méthode permet un accès direct à tous les mots à traiter. Toutefois, il est clair que ces données ne peuvent être fournies « à la main » et qu'on doit les générer à partir d'un dictionnaire des formes canoniques et des règles de flexion. Par conséquent, l'analyseur doit être de toute façon considéré. Dans le cas où l'on souhaiterait utiliser un lexique limité aux formes canoniques, il faudra mettre en œuvre des mécanismes pour ramener les formes fléchies aux formes stockées.

Deux points permettent de préciser les différences entre les deux méthodes : le **stockage des traits** et la **vitesse de reconnaissance** G. Sabah (1989, pp.22). Un dictionnaire comportant toutes les formes possibles implique que les traits soient attachés à toutes les formes fléchies d'un même mot alors qu'un dictionnaire réduit aux formes canoniques permet de ne les stocker qu'une seule fois. Le gain de place mémoire est compensé par la perte de temps au moment où est cherchée la forme canonique.

De nos jours, l'augmentation des mémoires d'ordinateur fait préférer généralement la méthode du **dictionnaire des formes fléchies**. Cependant, ce choix nous semble non raisonnable pour la langue arabe et ce pour au moins deux raisons : l'arabe est une langue fortement agglutinée et les textes peuvent être non voyellés, partiellement voyellés ou complètement voyellés. Si l'on désire analyser la totalité des formes fléchies par la simple consultation d'un dictionnaire, il faudra générer un lexique dont la taille est de l'ordre de centaines de millions de mots.

Toutes ces considérations, notamment celles relatives aux propriétés morphologiques de l'arabe, nous ont amené à adopter une approche qui se situe à mi-chemin entre la simple consultation d'un dictionnaire de toutes les formes fléchies et une analyse complètement formelle.

Nous essaierons dans ce chapitre d'expliquer la méthode d'analyse que nous adopterons en mentionnant les différentes ressources (lexiques) qui seront utilisées et la façon dont ils sont générés. D'abord, nous examinerons les principales méthodes d'analyse morphologique et nous verrons si elles peuvent s'appliquer à la langue arabe. Nous décrirons ensuite à travers des exemples, les différentes étapes du processus d'analyse retenu. Nous décrirons enfin l'environnement informatique qui permet de génération des différents lexiques utilisés par l'analyseur à partir de la base de données lexicale DIINAR.

3.2 Les méthodes d'analyse morphologique

Dans cette section, nous nous intéressons aux **techniques informatiques de reconnaissance des formes fléchies**. On distingue les analyseurs procéduraux (où les connaissances sont données sous la forme de programmes) des analyseurs déclaratifs (où les connaissances sont données de façon indépendante de leur emploi).

3.2.1 Les analyseurs procéduraux

Un analyseur *procédural* s'appuie sur l'analyse de la séquence d'actions à effectuer pour conjuguer un verbe (synthèse) ou pour retrouver la forme canonique à partir d'une forme fléchie (analyse), afin de déduire les règles permettant de reproduire ces processus. On trouvera une description plus détaillée de ce type d'algorithme et de l'exemple que nous reprenons ci-dessous dans G. Sabah (1989, pp. 25-28).

Ce processus a été employé par WINOGRAD en 1972 pour fabriquer un programme de reconnaissance d'une partie des formes fléchies de l'anglais (SHRDLU). Ce programme procède par une série d'actions sur la forme fléchie afin de retrouver la forme canonique : des suppressions de fin de mots (comme *n't* pour la négation, *ing*, *ed* ou *en* pour les verbes, *est* et *er* pour les superlatifs et les comparatifs, etc.) et des ajouts de certaines lettres à la fin du mot (comme ajouter *an* après avoir ôté *en* de *men*, pour analyser le pluriel).

L'algorithme d'analyse a été implémenté sous forme d'un automate d'états finis. Il utilise une fonction *FIN* qui teste si la fin du mot correspond à l'une des chaînes (n't, ing, etc.), une fonction *OTER* qui précise ce qu'il faut supprimer à la fin du mot et une fonction *AJOUTER* qui permet de remplacer les lettres supprimées. Lorsque l'automate aboutit à l'état final, une recherche dans le dictionnaire est effectuée pour voir si la base (= ce qui reste de la chaîne) existe. Par exemple, la forme fléchie *plied* est analysée en ôtant le suffixe *ed* et en remplaçant la lettre *i* par la lettre *y* ce qui permet de retrouver sa forme canonique *ply*.

Un tel mécanisme est généralement insuffisant. Quand une base est trouvée dans le dictionnaire, il convient de vérifier qu'elle est en accord avec le suffixe que l'on a ôté. La liste des traits morpho-syntaxiques attachés à la base doivent par conséquent, permettre de vérifier la compatibilité entre la base et le suffixe.

Pour le traitement des exceptions (comme pour passer de *sang* à *sing* ou de *won't* à *will*), l'ajout des règles ralentit la vitesse de reconnaissance. Les algorithmes dans ce cas, doivent inclure deux entrées indépendantes. De ce fait, ces programmes s'appliquent dans des domaines très restreints de la langue où la morphologie est très régulière.

3.2.2 Les analyseurs déclaratifs

Un analyseur *déclaratif* comme son nom l'indique reçoit de manière déclarative les données de la langue. Le programme est général et la mise à jour se fait plutôt au niveau des données. Un exemple d'analyseur déclaratif est celui de PITRAT dont on trouvera une description dans G. Sabah (1989, pp. 25-36) et qu'on détaillera dans cette section.

Constatant que la méthode de WINOGRAD, devient très complexe si on étend le domaine de la langue ou on passe à une autre langue, PITRAT a conçu un même analyseur général qui a été testé sur une dizaine de langues dont l'arabe.

Les données utilisées dans son programme sont constituées de trois types d'informations :

- Le fichier des mots. Chaque entrée est formée du nom du mot, du nom de son modèle de conjugaison et de la suite ordonnée des bases qui seront utilisées pour générer les formes fléchies. Par exemple l'entrée (TENIR, **VENIR**, TIEN, TEN, TIENN, TIN, TÎN) indique que le verbe tenir suit la conjugaison qui s'appelle venir et utilisera les cinq racines *tien*, …
- Le fichier des terminaisons. Sous un nom de terminaison, précisant la partie de la conjugaison considérée, on indique la liste des suffixes utiles. Par exemple, la ligne (**VIP** ; S, S, T, ONS, EZ, ENT) désigne les suffixes de certains verbes à l'indicatif présent. Ce fichier a une relation étroite avec le précédent puisque les terminaisons vont dépendre de la façon dont on aura défini les bases.
- Le fichier des conjugaisons. Il indique les relations entre les bases et les terminaisons. Une entrée de ce fichier correspond à un nom de modèle de conjugaison (par exemple VENIR). La conjugaison est divisée en groupes identifiés par un nom de groupe, et dans chacun d'eux on indique le nom de terminaison et la séquence des numéros de bases correspondantes. Par exemple, l'entrée (Venir, IP, VIP, 1, 1, 1, 2, 2, 3 ; …) signifie que pour conjuguer un verbe à l'indicatif présent (IP), on utilise les terminaisons de VIP avec la première base (tien) pour les trois personnes du singulier, la deuxième base (ten) pour la première et la deuxième personne du pluriel et la troisième base (tienn) pour la troisième personne du pluriel.

Le processus d'analyse procède par des découpages successifs du mot en deux suites de chaînes de caractères. On regarde alors si la suite correspondant à la fin du mot est présente dans les terminaisons. Si elle y est, on recense les couples (nom de terminaison, rang) qu'on note (T, r) qui sont associés à la suite trouvée. Par exemple, si on analyse *tenez*, l'analyseur détectera la terminaison *ez* et lui trouvera associé le seul couple (VIP, 5), indiquant que cette finale est la cinquième de conjugaison des verbes à l'indicatif présent. On vérifie alors si le début du mot (*ten* pour l'exemple) correspond à une racine connue. L'analyseur déterminera alors les couples (Mot, Base) notés (M, B) formés du nom du mot de la base et de son rang dans le fichier des mots. Dans cet exemple, à partir de *ten* on trouvera (*Tenir*, 2), indiquant que la forme canonique du mot peut être *TENIR* et que la base *TEN* est la troisième dans le fichier des mots.

Une analyse est retenue lorsque la conjugaison associée au mot **M (TENIR)** dans le fichier des mots contient un groupe qui est associé à la suite de terminaisons trouvée **T (VIP)**. On vérifie alors que le numéro de base de rang **r (5)** est bien **R (2)** dans le fichier des conjugaisons, ce qui est le cas pour notre exemple (Venir, IP, VIP, 1, 1, 1, 2, **2** , 3 ;…).

En continuant systématiquement le processus même après avoir trouvé une analyse, on construit toutes les analyses possibles.

De façon symétrique, à partir d'un mot M, d'un groupe G et d'un numéro n dans ce groupe, les données permettent de construire la forme fléchie : à partir de (*Tenir*, *VIP*, *5*), on doit obtenir la forme *Tenez*, deuxième personne du pluriel de l'indicatif présent du verbe *tenir*.

3.2.3 Conclusion

Il nous paraît évident que représenter les données morphologiques d'une manière déclarative, est le moyen le plus adéquat pour développer et maintenir des algorithmes d'analyse et de génération morphologique, même s'ils sont parfois moins performants que les analyseurs procéduraux.

Le programme de PITRAT décrit ci-dessus est restreint aux seules formes base+suffixe. Pour analyser un mot graphique en arabe, on doit par conséquent l'étendre pour intégrer les autres composantes du mot graphique arabe (proclitique(s), préfixe et enclitique(s)).

Néanmoins, le seul recours aux listes des morphèmes ne permet pas de reconnaître certaines formes de l'arabe écrit. En effet, lors du processus de synthèse, certains morphèmes sont assimilés ou changent de forme graphique lorsqu'ils sont associés à d'autres morphèmes.

L'exemple le plus fréquent est celui de la hamza instable qui change de forme lorsqu'elle est suivie de certains morphèmes. Par exemple, la forme (يَنآرْقَ : *YaQR#Ni*) « ils (les deux) lisent » est générée à partir du préfixe (يَ : *Ya*), de la base (أرْق : *Qra?*) et du suffixe (ان : *âNi*). Cette forme devrait normalement être générée en (يَنااَرْقَ : *YaQRa?âNi*)*, mais par application d'une règle de transformation, la hamza (أ : *?*) se transforme en (آ : *#*) lorsqu'elle est suivie de (ا: *â*). C'est uniquement la forme erronée (يَنااَرْقَ : *YaQRa?âNi*)* qui pourrait être directement analysée par la simple consultation des listes des préfixes, bases et suffixes.

Pour résoudre ce problème, on peut, à la manière de WINOGRAD, utiliser une procédure de remplacement automatique des (آ : *#*) par des (اَأ : *?â*). Malheureusement, l'application de cette procédure ne pourrait pas être généralisée à toutes les formes. Par exemple, la forme (يَناَفَلآتَ : *YaTa#LaFâNi*) « ils se lient d'amitié ou d'amour» qui est générée à partir du préfixe : (يَ : *Ya*),de la base (فَلآتَ : *Ta#LaF*) et du suffixe (ان : *âNi*), ne doit pas être transformée en (يَناَفَلااَتَ : *YaTa?âLaFâNi*)* auquel cas la base ne pourrait pas être analysée. D'autres formes comme (ُلآكْ : *#KuKu*) « je mange » qui est générée à partir du préfixe (أَ : *?a*), la base (ُلأْكْ : *?KuL*) et le suffixe (ُ : *u*), doit subir une nouvelle règle de conversion qui transforme la graphie (آ : *#*) en (ُأَأْ : *?a?*), afin de pouvoir retrouver le préfixe et la base dans les listes des morphèmes du dictionnaire.

On rencontre souvent ce genre de problème surtout lorsque la racine du mot graphique est malade (لّتعم) ou irrégulière (فعاضم وأ زومهم). De par la fréquence de ces mots dans les textes arabes, toutes ces règles d'exceptions doivent être recensées et être minutieusement traitées. Dans la section suivante, nous présenterons l'analyseur du mot graphique arabe que nous avons développé en montrant comment nous avons pu contourner les problèmes cités ci-dessus.

3.3 Processus d'analyse

3.3.1 Introduction

Comme nous l'avons déjà signalé dans l'introduction de ce chapitre, l'analyse d'un texte sur le plan morphologique est l'opération qui consiste à vérifier l'appartenance à la langue de chacun de ses mots, à déterminer leurs constituants (morphèmes) et à donner pour chacun ses traits linguistiques hors contexte.

L'analyseur doit distinguer deux catégories de mots. Les premiers sont ceux qui sont directement accessibles dans le lexique et les seconds sont ceux qui sont obtenus par flexion et/ou par composition. Dans cette section, on présentera une méthode d'analyse commune permettant de reconnaître tous les mots graphiques arabes (non voyellés, partiellement voyellés ou complètement voyellés).

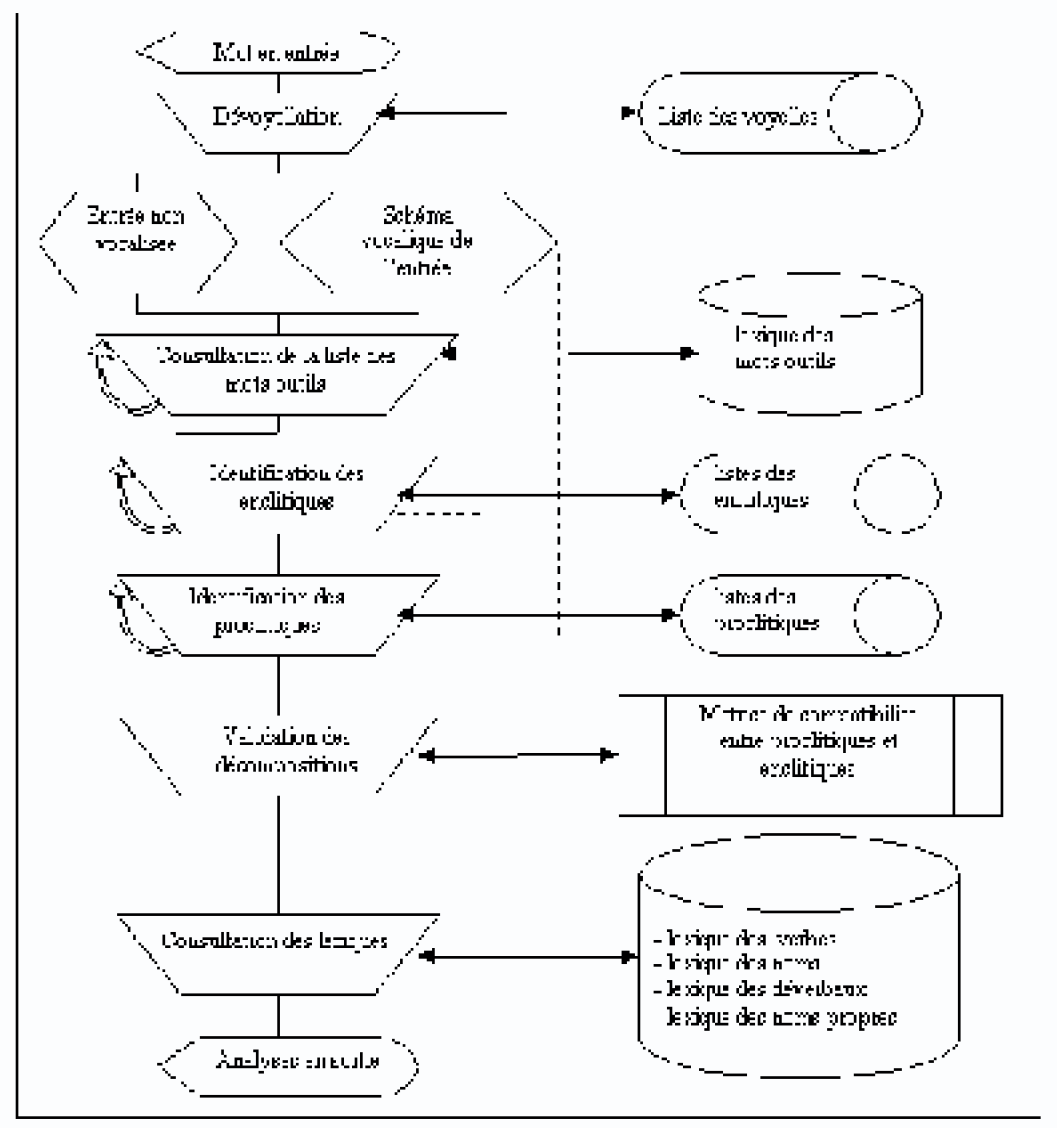
La méthode d'analyse proposée reprend *grosso modo* le même cheminement que celles qui ont permis, dans le cadre des travaux de notre équipe, la réalisation d'un correcteur orthographique de l'arabe (Gader 96) et d'un système de voyellation de textes arabes écrits (Ghenima 98). Le même processus d'analyse a été généralement utilisé par la majorité des analyseurs morphologiques de la langue arabe (Ben hamadou, 1991) avec parfois des techniques informatiques différentes comme par exemple l'utilisation d'un automate d'états finis (Beeslay, 1996). Nous avons toutefois apporté quelques modifications dont notamment celle d'utiliser le mot minimal au lieu du noyau comme entrée pour la consultation du lexique. Nous avons ainsi évité les problèmes de transformation morphologique lors de l'assemblage des constituants du mot minimal qui seront traités par le générateur du lexique.

Le processus d'analyse s'effectue en plusieurs étapes. Dans chaque étape, un mécanisme de filtrage permet d'éliminer, de l'ensemble des solutions candidates trouvées, celles qui sont non valides. La figure (3-1) illustre l'ordonnancement de ces étapes avec les ressources associées. Dans cette section, nous allons nous intéresser à une présentation de chaque étape et à la définition des algorithmes et structures de données permettant leur mise en œuvre.

3.3.2 Dévoyellation de l'entrée

Afin de pouvoir traiter les formes vocalisées, partiellement vocalisées et complètement vocalisées, nous identifions tout d'abord la forme non vocalisée (par élimination des voyelles) et le schéma vocalique de l'entrée. Cette étape ne nécessite que la liste des voyelles arabes. Par exemple, la procédure de dévoyellation de la chaîne (يَقْرَؤُونَ : *YaQRaÜûNa)* « ils lisent » fournit la forme non vocalisée (يقرؤون : *YQRÜWN*) et le schéma vocalique[31] (´???´?´?´? : ?a??a??u?a).

[31] Dans le schéma vocalique, chaque consonne est remplacée par un point d'interrogation à son emplacement dans le mot.



3.3.3 Consultation de la liste des mots outils

Étant donné l'importante fréquence des mots outils dans les textes arabes, le lexique regroupant toutes leurs formes fléchies des mots outils (Voir ci-dessous son processus de génération § 3.4) est consulté au début du processus d'analyse. Chaque entrée de ce lexique est formée par la forme non vocalisée du mot outil accompagnée de son schéma vocalique.

La recherche dans le lexique se faisant avec la forme non voyellée, en cas de succès, les schémas vocaliques des deux mots sont comparés. L'algorithme d'appartenance du schéma vocalique à celui accompagnant l'entrée du lexique est alors le suivant :

Arguments :

- Ch1 : le schéma vocalique du mot à analyser.
- Ch2 : le schéma vocalique du mot associée à l'entrée du lexique.
- ChNV : le mot non vocalisé.

Fonction Appartient (Ch1, Ch2, ChNV)

Début

Appartient □ Vrai *Par défaut, la schéma vocalique du mot appartient à celui du lexique*

i □ 1, j □ 1 *initialisation des compteurs permettant de parcourir les 2 schémas vocaliques*

CptCons □ 0 *initialisation du compteur permettant de parcourir le mot non vocalisé*

ChaîneDesVoyelles " ً ٍ ٌ ّ ْ ٓ ٔ ٕ "

Tant-que i <= longueur(Ch1) And j <= longueur (Ch2)

Parcourir les deux chaînes jusqu'au bout

c1 □ Sous-chaîne (Ch1, i, 1) *Sélection du caractère de la position i de la chaîne Ch1*

c2 □ Sous-chaîne (Ch2, j, 1) *Sélection du caractère de la position j de la chaîne Ch2*

Si c1 = c2 **alors** *Si les deux caractères courants sont égaux*

Si c1 = "?" **alors** CptCons = CptCons + 1 **FinSi** *il s'agit de deux consonnes*

i □ i + 1, j □ j + 1 *Passer aux deux caractères suivants dans les deux chaînes*

Sinon *(c1 ≠ c2)*

Si c1 = "?" And Appartient(c2, ChaîneDesVoyelles) <> 0 alors

Le caractère courant de la 1 ère chaîne est une consonne et celui de la seconde est une voyelle

j □ j + 1 *Avancer au caractère suivant de la seconde chaîne*

Sinon *(c1 ≠ c2)*

Si c2 = "?" And Appartient(c1, ChaîneDesVoyelles) <> 0 alors

Le caractère courant de la 2 ème chaîne est une consonne et celui de la 1 ère est une voyelle

i □ i + 1 *Avancer au caractère suivant de la première chaîne*

Sinon (c1 ≠ c2) *Il s'agit de deux voyelles différentes*

Appartient □ Faux, Quitter la fonction

FinSi

Fin Fonction

Cet algorithme permet donc d'analyser les mots non vocalisés, partiellement vocalisés ou complètement vocalisés. Si le processus réussit, le résultat de cette opération est retenu pour figurer directement en sortie de l'analyseur.

Par exemple, supposons que le mot à analyser est (وَلَكَ ّنَأ : *WaLaKa?aNNaKa)* dont le schéma vocalique est (َ??َ??َ?? : ?a???a??a?) et la forme non vocalisée est (وكّنأل : *WLK?NNK*). L'interrogation du lexique des mots outils avec la forme non vocalisée fournit deux entrées comme possibles solutions dont les schémas vocaliques sont (ِ?َ??َ?َ?َ? : ?a?a?a?a??a?i) et (َ?َ??َ?َ?َ? : ?a?a?a?a??a?a). C'est uniquement cette dernière entrée qui sera retenue comme solution possible par l'algorithme *Appartenir* puisque la dernière voyelle de la première solution ne coïncide pas avec celle du mot à analyser.

3.3.4 Identification des enclitiques

Cette étape consiste à repérer les différents enclitiques du mot à analyser par la consultation de la liste des enclitiques. Cette liste est formée par tous les enclitiques non voyellés (y compris l'enclitique vide) auxquels sont associés leur schéma vocalique. Le processus consiste à comparer successivement les fins du mot à analyser non voyellé (dont la taille ne doit pas dépasser l'enclitique le plus long de la liste) avec les éléments de la liste des enclitiques. A la fin de cette étape, on retiendra les enclitiques dont les formes vocalisées coïncident avec les fins du mot à analyser avec les restes du mot à analyser. L'algorithme de recherche des enclitiques se présente ainsi :

Arguments :

- MotAAnalyser : le mot à analyser.
- MotAAnalyserNV : le mot à analyser non voyellé.
- ChVoyMotAAnalyser : le schéma vocalique du mot à analyser.

Procédure RechercheEnclitiques (MotAAnalyser,MotAAnalyserNV,ChVoyMotAAnalyser)

Début

(On détermine la taille de l'enclitique le plus long possible du mot à analyser)

Si Longueur(MotAAnalyserNV)<5 **Alors**

(5 étant la longueur de l'enclitique le plus long de la liste)

Longueur_Maximale_Enclitique Longueur(MotAAnalyserNV)

Sinon: Longueur_Maximale_Enclitique 5

Fin Si

Pour i = 0 **jusqu'à** Longueur_Maximale_Enclitique

EclNVMotAAnalyser EnlèveFin (MotAAnalyserNV,i) *Extraction de l'enclitique*

Si Appartient(EclNVMotAAnalyser, listeEnclitiques) **alors**

Retenir (EclNVMotAAnalyser)

DebMotNVAAnalyser EnlèveFin(MotAAnalyserNV,long(EclNVMotAAnalyser))

RechercheProclitiques(DebMotAAnalyser,

DebMotAAnalyserNV,ChVoyDebMotAAnalyser)

FinSi

Fin Pour

Fin Procédure

3.3.5 Identification des proclitiques

Pour chaque enclitique retenu, une boucle permet de récupérer dans cette étape ses différents proclitiques possibles à partir du début du mot à analyser (ce qui reste du mot à analyser, une fois qu'on enlève l'enclitique en cours).

Cette opération fait appel à la liste des proclitiques non voyellés (y compris le proclitique vide) auxquels sont associés leur schéma vocalique. Le processus consiste à comparer successivement les débuts du mot (dont la taille ne doit pas dépasser le proclitique le plus long de la liste) avec les éléments de la liste des proclitiques. A la fin de cette étape, on retiendra pour chaque enclitique la liste des proclitiques possibles avec les restes du mot à analyser correspondants.

L'algorithme d'identification des proclitiques est similaire à celui des enclitiques. La différence tient uniquement au fait que l'analyse du mot est effectuée de gauche à droite (c'est-à-dire à partir du premier caractère) au lieu de droite à gauche en raison de la représentation inversée des proclitiques.

3.3.6 Validation des décompositions

Une fois que l'identification des enclitiques et des proclitiques terminée, on constitue tous les couples (Pcl, Ecl) possibles. Parmi ces couples, il s'agit de ne retenir que ceux qui sont attestés par la langue. Le mécanisme de filtrage des couples (Pcl, Ecl) valides utilise une matrice de compatibilité booléenne [$C(Pcl_i, Ecl_i)$].

Les informations contenues dans cette matrice se réduisent à un spécificateur qui renseigne sur la validité du couple (Pcl, Ecl) et sur le type présumé du mot minimal (verbe, nom, déverbal), ce qui permet d'orienter la recherche dans le lexique :

- **Si** $C(Pcl_i, Ecl_i) = 0$ **alors** les segments Pcl_i et Ecl_i identifiés ne peuvent pas constituer un mot graphique candidat.
- **Si** $C(Pcl_i, Ecl_i) = 1$ **alors** les segments Pcl_i et Ecl_i identifiés peuvent constituer un mot graphique candidat en les associant à un verbe.
- **Si** $C(Pcl_i, Ecl_i) = 2$ **alors** les segments Pcl_i et Ecl_i identifiés peuvent constituer un mot graphique candidat en les associant à un nom ou à un déverbal.
- **Si** $C(Pcl_i, Ecl_i) = 3$ **alors** les segments Pcl_i et Ecl_i identifiés peuvent constituer un mot graphique candidat indifféremment avec un nom ou un déverbal ou un verbe

Au terme de cette phase, seules les décompositions qui présentent des proclitiques et des enclitiques compatibles sont retenues avec une indication sur le type présumé du mot

minimal avec lequel ils sont compatibles.

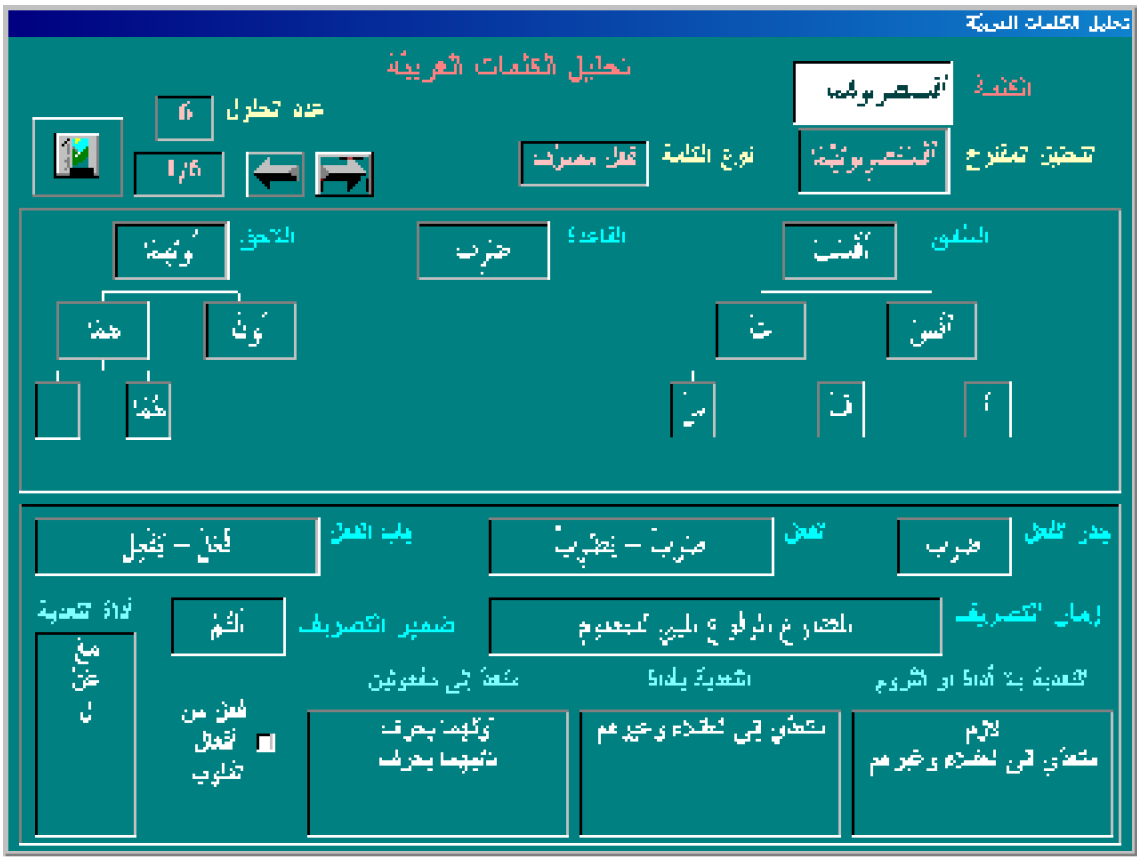
3.3.7 Consultation des lexiques

Pour toutes les décompositions validées, le lexique doit attester de l'existence du mot minimal et de sa compatibilité avec le reste de la décomposition.

Une dernière vérification de la compatibilité du mot minimal (verbe) et ses enclitiques à ce niveau permet d'éliminer les solutions suivantes :

- Les solutions à **double enclitique** dont le mot minimal est un verbe passif sont automatiquement rejetées (i.e. *اَهَكُتجِّوُز). Les formes à double enclitique incluant un verbe conjugué dans une voix active sont aussi rejetées, à l'exception de celles qui incluent un verbe transitif direct à double compléments (فرشابم ّدعتم لعف ىلإ ةادأ نودب نيلوعفم)
- Les solutions à **un seul enclitique** dont les verbes ne sont pas transitifs directs sont rejetées. Si la forme contient un verbe passif, ce dernier doit être obligatoirement transitif direct à double complément pour retenir la solution. Par exemple : (حَمَكَ + هُ= il le juge) est acceptable tandis que (حُمِكَ + هُ* = il est jugé lui) est inacceptable.

Une fois cette vérification effectuée, les lexiques relatifs aux formes conjuguées, aux noms, aux déverbaux et aux noms propres seront successivement consultés. Si le mot minimal est trouvé dans l'un de ces lexiques, toutes les informations lexicales le concernant sont réunies et sont retournées en sortie de l'analyseur. La figure (3-2) ci dessous, qui est une copie d'écran de l'interface d'expérimentation de l'analyseur montre un exemple d'analyse du mot graphique "امهنوبرضتسفأ". Les informations fournies en sortie d'analyse de la première solution de cet exemple, sont relatives aux solutions verbales.



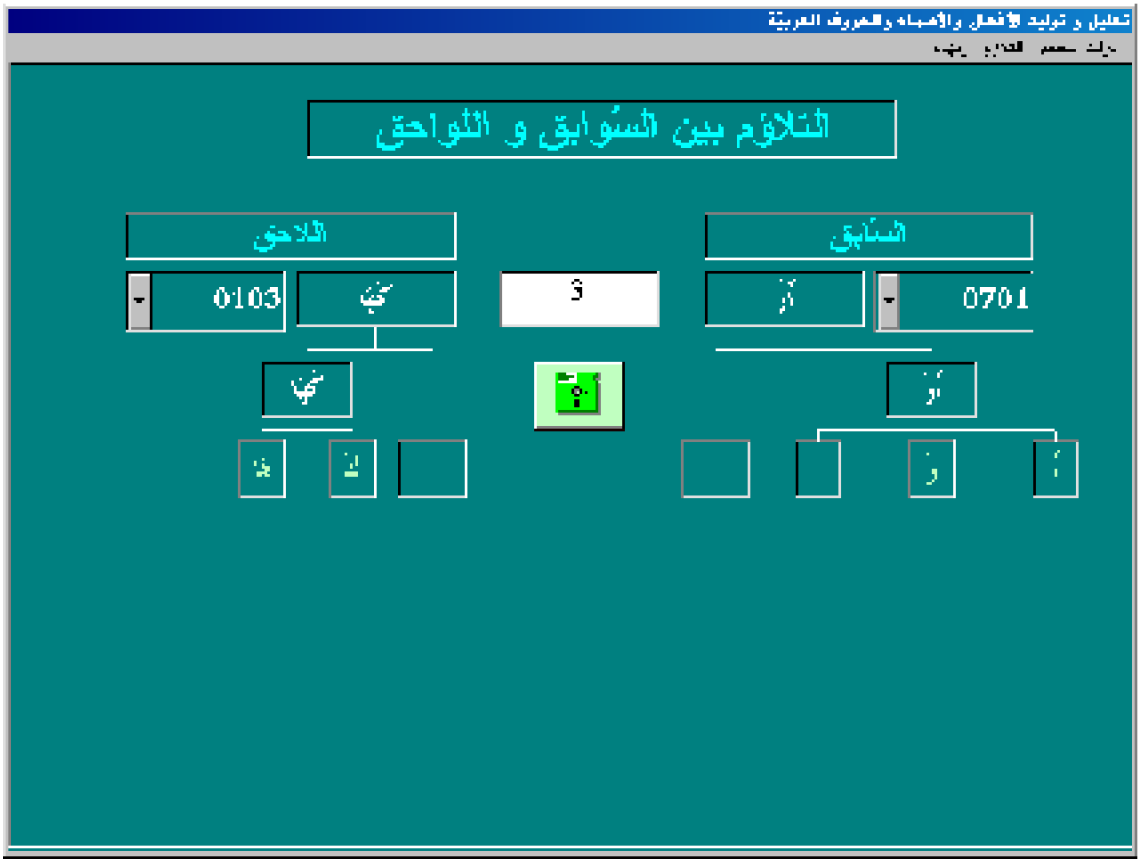
3.4 Génération des lexiques

3.4.1 Présentation du lexique

La modélisation de DIINAR.1, a permis d'organiser un maximum d'informations de natures diverses en différentes bases de données selon la nature de l'unité lexicale traitée (cf. Chapitre 2). Au bout de la chaîne d'analyse des mots graphiques (cf. § 3.3.7), l'analyseur a besoin d'accéder directement au lexique pour attester les solutions trouvées et de récupérer les diverses informations y relatives afin de les retourner en sortie d'analyse. Un lexique permettant de reconnaître rapidement toutes les formes fléchies et leurs différentes propriétés doit par conséquent être généré à partir de DIINAR.

Dans cette section, nous présenterons l'environnement qui permet de générer ce lexique. Cet environnement permet la régénération du même lexique pour tenir compte d'éventuelles modifications et de générer d'autres formes du lexique adaptées à d'autres analyseurs. En effet, aucun dictionnaire, qu'il soit électronique ou non, n'est bien entendu complet. La réalisation du dictionnaire DIINAR n'est pas terminée, et on peut se demander s'il est possible qu'elle le soit un jour. Cette base de données lexicale augmente sans cesse et est mise à jour régulièrement. Disposer d'un tel environnement permet d'inclure facilement les mises à jour et les informations ajoutées dans le lexique. D'autre part, d'autres analyseurs utilisent des lexiques générés à partir de DIINAR. Ces lexiques ont des contenus et des formes différentes du notre. L'environnement de génération permet de répondre à ces différents besoins sans toucher à la structure de la base de données lexicale.

L'environnement de génération du lexique permet aussi de mettre à jour la matrice de compatibilité Proclitiques - Enclitiques exploitée par l'analyseur (figure 3-3).



Dans cette section, nous allons d'abord définir les procédures et les règles morphologiques qui permettant l'obtention de formes acceptables par la langue arabe et ensuite présenter successivement les différents sous-lexique produits générés à partir de DIINAR.1.

3.4.2 Description des règles de génération

Lors du processus de génération du mot graphique arabe, les morphèmes subissent généralement des changements dans leur forme graphique lors de leur assemblage, pour des raisons morphologiques et/ou phonétiques (Baccouche, 1992) (Al-Hakkak & Neyreneuf, 1996) (Ben hamadou, 1991). Dans cette section, nous présentons les principaux phénomènes linguistiques qui ont été traités dans le programme de génération.

a) Les phénomènes de combinaison phonétique

Parmi les phénomènes de combinaison phonétique on distingue principalement l'assimilation et l'accommodation. D'autres phénomènes ont été directement traités dans les schèmes des morphèmes.

- L'accommodation : Ce phénomène se manifeste quand les points d'articulation des phonèmes consécutifs sont identiques mais l'un d'eux est sonore alors que l'autre est sourd ; auquel cas la consonne sourde se transforme en son correspondant sonore. La dérivation interne ne tient pas compte de ce phénomène et permet d'obtenir des formes erronées comme (امعتدا* : *?DT* ^{C} *M*) qui se réalise et s'écrit (امعدّا : *?DD* ^{C} *M*) ou (ارهتزا* : *?ZTHR*) qui se réalise et s'écrit (ارهزا : *?ZDHR*) : phénomène de *voisement*, (ابرتضا* : *? **D** TRB*) qui se réalise et s'écrit (ابرطضا : *? **D** TRB*) ou (احلتصا* : *? **S** TR **H***) qui se réalise et s'écrit (احلطصا : *? **ST** R **H***) : phénomène d'*emphatisation*, etc. Ces phénomènes sont généralement relatifs à la forme verbale VII : (افتعل - يفتعل).
- L'assimilation : C'est la tendance de deux phonèmes, ayant des traits pertinents voisins, à devenir identiques. Par exemple, (مُتَدْدَم* : *MaDDaT*) qui se réalise et s'écrit (مُتَدَم : *MaDDaT*) : phénomène de *chute* ou d'*assimilation graphique*. Ces phénomènes sont généralement relatifs à la forme verbale VII : (افتعل). Pour tenir compte de ces phénomènes, nous avons ajouté une règle de substitution de graphèmes.

Pour tenir compte de ces phénomènes, nous avons ajouté des règles de substitution qui se placent juste après la procédure de dérivation.

b) Les phénomènes de transcription

La transcription de certains archi-graphèmes pose certains problèmes en arabe. En effet, leur transcription est régie par des règles spécifiques qui dépendent de leur contexte immédiat dans le mot.

- Transcription de la HAMZA : La *hamza* prend diverses formes selon sa position, sa voyellation et la voyellation de la consonne qui la précède.

– En début de mot, ce sera toujours un "أ". Si la voyelle est une *kasra /i/*, la *hamza* se placera sous le *Alif* : "إ", sinon au-dessus : "مياه ارب ِإ" / "نيَأ" / "نذُأ".

– A la fin d'un mot, le support dépend de la voyelle qui la précède. Ce sera "أ" si cette voyelle est une *fatha /a/*, "ئ" si c'est une *kasra /i/*, "ؤ" si c'est une ***d** amma /u/* et "ء" (aucun support) si c'est un soukoun /Ø voyelle/ porté par une consonne ou une lettre de prolongation : "عودة" / "ءَرْم" / "ءام" / "ئِرب" / "ؤُرجي" / "أَدبي" / "ءُوّبتلا" / "لِئُس" / "لِئاَسم" / "لَءاَسَت" / "لُؤاَسَت" / "لَأْسم" / "لوُؤْسم" / "لاَؤُس".

· Transcription du 'Alif Mamdûda' : Quand deux *'Alif'* se suivent (اأ ou أأ), on écrit le 2ème sur le 1er et la *hamza* disparaît. C'est le 'Alif Mamdûda' (قدودمم فلأ) : (آ).

3.4.3 Génération du lexique

a) Génération des formes conjuguées

Ce lexique est formé par l'ensemble des formes conjuguées obtenues à partir des verbes de DIINAR et des neuf paradigmes de conjugaison vivants de la langue (cf. § 2.3.2). Néanmoins, pour des restrictions sémantiques quelques formes conjuguées sont exclues du lexique :

· Les formes conjuguées au passif et à l'impératif des verbes exclusivement intransitifs (ceux décrivant une qualité ou un état durable : les verbes de forme I : "لُعْفَ - لَعُفَ ُ" dans leur totalité et une partie des "لَعْفَ - لَعِفَ"). Par exemple : la forme "تُرْبِكْ" (j'ai été grandi) du verbe "رُبكْ - رَبُكَ" (grandir) est sémantiquement inacceptable.

· Si les verbes sont exclusivement intransitifs ou bien transitifs indirects (admettent un complément d'objet précédé d'une préposition), la seule forme conjuguée au passif qu'on retient est celle de la troisième personne du masculin singulier. Par exemple : le verbe "نم كِحَض" = (« rire de ») admet la seule forme "امنم كُحْضُي ال ٌروُمُأ" = (« des choses dont on ne rit pas ») / "منم كِحُض نالف" = (« on a rit de lui ») / "كِحُض قنالف اهيلع" = (« on la lui a posé un lapin »).

· Si les verbes sont exclusivement transitifs à des non-humains, les seules formes conjuguées qu'on retient au passif, sont celles de la troisième personne du singulier et du duel. Par exemple : le verbe "أَرَق" (lire) accepte la forme "اَئِرُق ناسرّدلا نإذه" (=Ces deux leçons-là ont été lues).

Chaque entrée de ce lexique est constituée par la forme conjuguée non vocalisée et de

son schéma vocalique. A chaque entrée est associé un spécificateur permettant de retrouver toutes les informations enregistrées dans DIINAR dont le verbe, la racine, le pronom de conjugaison et l'aspect de conjugaison.

b) Génération des formes nominales

Ce lexique est formé par l'ensemble des bases nominales obtenues à partir de la base de données nominale DIINAR.1.

Chaque entrée de ce lexique est constituée par la base nominale non vocalisée et de son schéma vocalique. A chaque entrée est associé un spécificateur permettant de spécifier le genre et le nombre de l'entrée, si elle accepte l'article (ٱلْ : *?aL*) et ses différents suffixes et déclinaisons possibles.

Aux bases nominales qui se terminent par un *"Tâ? final"* (souvent, marque du féminin) ou un *"?LiF final"*, on génère une deuxième base dans laquelle on substitue le *"Tâ?"* fermé : (ة) par un *"Tâ?"* ouvert : (ت) et le *"?LiF final"* qui prend l'une de ces deux formes : (ى, ا) par un *"Yâ?"* (ي). A cette seconde base nominale sont associés les suffixes du duel (ـَانِ, ـَيْنِ), ceux du pluriel (ـَاتُ, ـُونَ, ـِينَ) et celui du relatif (ـِيّ). Par exemple : la base (مَرَضِيَة) devient pour le duel (مَرَضِيَت) : (مَرَضِيَتَا + ـَانِ= مَرَضِيَتَانِ) ou (مَرَضِيَت + ـَيْنِ = مَرَضِيَتَيْنِ).

Aux bases nominales qui se terminent par une *HaMZa*, on génère aussi d'autres bases pour tenir compte des règles d'écriture de la Hamza.

c) Génération des déverbaux

Ce lexique est formé par l'ensemble des déverbaux obtenus à partir de la base de données verbale DIINAR.1.

Chaque entrée de ce lexique est constituée par la base non vocalisée du déverbal et son schéma vocalique. A chaque entrée (au singulier masculin) est associé un spécificateur permettant de définir la catégorie du déverbal, le verbe, la racine du verbe et les différentes déclinaisons.

d) Génération des mots outils

Ce lexique est formé par l'ensemble des formes fléchies obtenues à partir de la base de données des mots outils de DIINAR.1. Ce lexique est consulté au début du processus d'analyse.

Chaque entrée de ce lexique est constituée par la forme fléchie du mot outil non vocalisée et de son schéma vocalique. A chaque entrée est associé un spécificateur permettant de décrire les proclitiques, le noyau et les pronoms compléments qui forment le mot outil. Ce spécificateur permet aussi de spécifier les propriétés morphologiques et syntaxiques qui ont été associées au mot outil (cf. § 2.4).

e) Génération des noms propres

Ce lexique est formé par l'ensemble des bases nominales obtenues à partir de la base de

données des noms propres de DIINAR.1.

Chaque entrée de ce lexique est constituée par la base nominale non vocalisée et de son schéma vocalique. A chaque entrée est associé un spécificateur permettant de spécifier la catégorie sémantique, le genre, le nombre et les différents suffixes et déclinaisons possibles du nom propre. Les mêmes règles de génération utilisées pour les noms sont aussi appliquées aux noms propres.

3.5 Conclusion

Nous avons présenté dans ce chapitre, une méthode d'analyse des mots graphiques arabes non vocalisés, partiellement vocalisés ou complètement vocalisés. Nous avons adapté une approche déclarative qui se base sur les constituants du mot graphique arabe et sur le lexique généré à partir de la base de données lexicale DIINAR.1. Le programme de génération du lexique tient compte de l'ensemble des phénomènes morphologiques et phonétiques et ne permet d'obtenir que des formes attestées par la langue.

Néanmoins, nous serons souvent appelés à effectuer des mises à jour et des corrections dans DIINAR pour tenir compte par exemple de mots absents ou de mots mal saisis. Par le biais de ce programme, le lexique pourrait être automatiquement régénéré.

Dans le chapitre suivant, nous allons décrire des applications de TAL arabe, qui fonctionnent à partir des résultats retournés par l'analyseur, et qui vont nous permettre de construire les autres ressources linguistiques de l'environnement d'apprentissage « *AL-Mu*^{c} *aLLiM* ».

Ces applications vont nous permettre de tester l'analyseur, de détecter les incohérences et d'apporter éventuellement des correctifs à DIINAR et au processus de génération automatique.

Chapitre 4 Élaboration d'applications pour le traitement automatique des textes arabes

« Les obstacles sont les signes ambigus devant lesquels les uns désespèrent, les autres comprennent qu'il y a quelque chose à comprendre. Mais il en est qui ne les voient même pas... » Paul VALÉRY

4.1 Introduction

Dans ce chapitre, nous décrirons le fonctionnement de trois applications développées à partir de l'analyseur des mots graphiques arabes (cf. chapitre 3). Ces applications ont été réalisées essentiellement pour assister le processus de construction des ressources de l'environnement « *AL-Mu*^{c} *aLLiM* ». Elles pourraient néanmoins servir ultérieurement à effectuer des études quantitatives et statistiques des faits langagiers de la langue à partir de différents corpus textuels.

Le titre de ce chapitre « traitements automatiques des textes » évoque généralement la conception d'applications capables de **traiter** de façon **automatique** des données linguistiques, c'est à dire des données exprimées dans une **langue naturelle**. Les

traitements linguistiques complètement automatisés recourent généralement à des techniques statistiques ou probabilistes, mais ne permettent pas encore d'obtenir des résultats totalement corrects.

Dans le domaine de l'EIAO des langues, il est inadmissible d'utiliser des données qui peuvent être erronées. C'est pour cela, qu'une partie des applications développées sont semi-automatiques, c'est à dire qui font intervenir l'être humain dans l'exécution de certaines tâches lorsque les résultats sont équivoques.

D'autre part, le terme « **texte »** est souvent employé pour désigner un texte écrit. Il peut cependant désigner l'un quelconque des types des données linguistiques (textes écrits, dialogues, phrases, mots isolés, etc.) et prendre différentes formes (fichier texte, base de données, etc.). Les applications que nous avons réalisées, se baseront sur des **objets d'entrée particuliers** qui sont obtenus après une première phase d'indexation des textes bruts.

Ceci étant dit, nous entamerons ce chapitre par une description des différentes étapes du processus d'étiquetage semi-automatique des textes arabes bruts : segmentation, lemmatisation et association d'informations linguistiques aux lemmes. Nous verrons ensuite qu'à partir d'un texte étiqueté, nous pourrons réaliser des traitements complètement automatisés et obtenir des résultats précis. Nous nous intéresserons particulièrement à deux applications qui permettront la construction des principales ressources d'« *AL-Mu*^{c} *aLLiM* » : la recherche de concordances et le calcul de fréquences des mots ou des unités lexicales.

4.2 Étiquetage semi-automatique de textes arabes

Depuis les années soixante, les corpus textuels mis sur support électronique ont toujours existé pour des langues comme l'anglais ou le français[32]. La nouveauté réside dans l'enrichissement de ces corpus et le développement d'outils appropriés à leur traitement. D'abord, les corpus ne sont plus des suites de mots « nus », c'est-à-dire de simples chaînes de caractères, mais ils sont étiquetés (ou annotés ou encore enrichis). Nous entendons par-là l'ajout d'information, de quelque nature qu'elle soit : morphologique, syntaxique, sémantique, etc. Ensuite, les outils d'interrogation de ces corpus enrichis ainsi que les outils d'annotation proprement dits (étiqueteurs, analyseurs syntaxiques, etc.) se répandent.

Les corpus étiquetés sont principalement utilisés en analyse linguistique (cf., Habert & Nazarenko & Salem, 1997). Selon le type d'annotations effectuées, ils mettent en évidence des régularités qui échappent à l'observation « à l'œil nu ». Ils sont devenus désormais des outils indispensables à toute théorisation linguistique.

Nous débuterons cette section par une définition du processus d'étiquetage de texte

[32] En France, un fonds de quelque 160 millions de mots a été patiemment constitué à l'Institut National de la Langue Française (INaLF – CNRS) depuis les années soixante et constitue une base textuelle désormais accessible en ligne : *Frantext*.

et par un exemple de phrase arabe étiquetée. Nous aborderons ensuite les problèmes liés à la définition de l'unité de segmentation avant de décrire les processus de segmentation et de lemmatisation.

4.2.1 Définition et exemple

Nous définissons le processus d'**étiquetage de textes** comme l'ensemble des opérations qui permettent de passer d'un texte brut, exempt d'informations linguistiques, à une séquence d'**unités élémentaires** lexicales (les lemmes) assorties d'étiquettes morpho-syntaxiques[33]. Cette définition implique successivement le choix de l'unité élémentaire de segmentation, le processus de segmentation lui-même, la lemmatisation des unités et l'association des informations linguistiques aux lemmes.

Prenons l'exemple de la segmentation de la phrase suivante : (سيذهب محمد علي إلى المسجد) = (« Mohamed Ali ira à la mosquée. »), qui pourra aboutir à la suite d'unités (U_1, U2, U_3, U_4, U_5, U_6), dans laquelle chaque U_i correspond à une unité répertoriée, définie par un ensemble d'informations linguistiques :

- **U_1 : segment : (س)**
  - forme lemmatisée : (س)
  - informations morpho-syntaxiques : marque du futur

- **U_2 : segment : (يذهب)**
  - forme lemmatisée : (ذَهَبَ - يَذْهَبُ)
  - informations morpho-syntaxiques : verbe, inaccompli, indicatif, 3ème personne, singulier, masculin, constructions : intransitif ; transitif avec un complément introduit par la préposition « إلى » ; transitif avec deux compléments, le premier par « بِ », le deuxième par « إلى » ; etc.
  - informations sémantiques : sujet : humain/non humain, concret, animé objet : humain/non humain, concret, animé/ non animé

- **U_3 : segment : (محمد علي)**
  - forme lemmatisée : محمد علي
  - informations morpho-syntaxiques : nom propre composé, masculin, singulier
  - informations sémantiques : humain, concret, animé.

- **U_4 : segment : (إلى)**

[33] On parle aussi de **corpus arborés,** lorsque les unités élémentaires du texte sont munies d'arbres syntaxiques.

– forme lemmatisée : (ﺇﻟﻰ)
– informations morpho-syntaxiques : préposition, suivie d'un nom, d'un mot outil ou d'une phrase commençant par « أن ».

· **U_5 : segment : (ﺍﻟﻤﺴﺠﺪ)**

– forme lemmatisée : مسجد
– informations morpho-syntaxiques : nom, masculin, singulier, défini
– informations sémantiques : non humain, concret, non animé.

· **U_6 : segment : (.)**

– forme lemmatisée : .
– informations morpho-syntaxiques : ponctuation, délimiteur de phrases.

4.2.2 L'unité de segmentation

Dans un processus de segmentation automatique, le choix des unités doit obéir à deux impératifs : la segmentation ne doit pas être trop difficile à effectuer, et les unités doivent être suffisamment cohérentes et significatives pour faciliter les traitements ultérieurs (Fuchs, 93).

Cette double contrainte se heurte à une série de phénomènes linguistiques[34] : amalgames, flexions, dérivations, compositions, etc., qui conduisent à des obstacles lors de l'automatisation du processus.

Certains mots résultent, on le sait d'une séquence ou d'un **amalgame** de deux unités existantes. C'est le cas par exemple de ces expressions adverbiales (ذئنيح / ذئنع / ذئتقو / ذِئَتَعاس) « = à ce moment-là » qui sont formées à partir de la particule (ذإ) = (« car, puisque ») et d'une deuxième particule désignant le temps. Faut-il alors rétablir les deux unités (qui jouent chacune un rôle syntaxique spécifique) pour faciliter l'écriture des règles ultérieures ou les laisser dans une unique unité ? Dans la plupart des cas, il convient de répondre par la négative, les unités « composées » ayant un autre sens – au moins en partie – que celui de chacune des unités qui la composent.

Par ailleurs, on sait déjà que nombre de mots connaissent des phénomènes de **flexion** ou **dérivation externe**. Du fait de leur facilité de traitement due au caractère fermé de l'inventaire des affixes, leur analyse ne pose pas de problèmes. La question qui nous intéresse ici est celle du **statut** des désinences : dans la plupart des systèmes elles sont traitées comme une série d'attributs (de temps, de mode, de nombre, etc.) qui sont ajoutés à la forme lemmatisée, mais on peut aussi les considérer comme des unités

[34] Nous nous contentons des phénomènes propres à la langue arabe. Certains phénomènes comme celui d'élision, ne sera pas cité (cf. Haddar, 2000).

morphologiques à part entière et traiter la forme fléchie comme une concaténation de plusieurs unités.

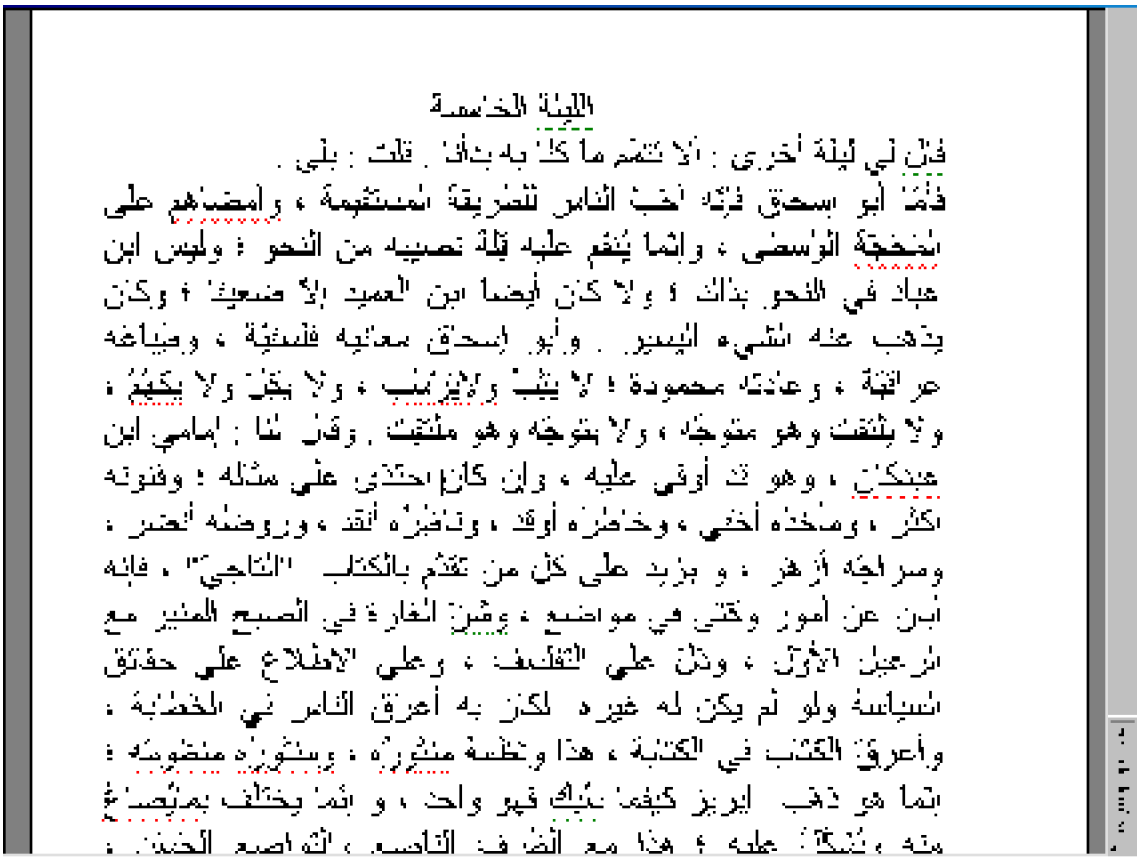
A ces difficultés, s'ajoutent celles relatives aux unités distribuées sur des séquences de plusieurs mots typographiques. C'est le cas des **unités discontinues**, comme les négations (لَنْ...أَدبًا / لَمْ...قُطُ) = (« ne…jamais ») ou (ال...سوى / ال...يغر) = (« ne…que, seulement »). Doit-on les traiter comme une unité unique discontinue ou comme plusieurs unités, sachant que chaque mot peut se rencontrer en emploi autonome ?

Le même ordre de difficulté se retrouve dans le cas des **mots composés** (لوسر دّمحم اللّه صلّى اللّه عليه وسلّم) (واحدٌ وعشرون), des **locutions** et des **formes figées**. Les mots peuvent être associés comme ils peuvent rester autonomes.

De toutes ces considérations il ressort que le **mot minimal**, tel qu'il a été défini lors de l'analyse des mots graphiques, ne constitue pas un mauvais point de départ. La réalisation informatique sera ainsi facilitée puisque nous aurons recours principalement aux résultats de l'analyseur. Chaque mot minimal correspond à une unité du lexique et les informations correspondantes sont directement récupérées. Il faudrait néanmoins ajouter un traitement spécifique pour les unités discontinues.

4.2.3 Le processus de segmentation des textes

Au-delà de la diversité des choix théoriques possibles, la tâche informatique de segmentation du texte en unités morphologiques nécessite la mise en place d'algorithmes qui obéissent *grosso modo* au même principe. On part d'un texte écrit, on repère les mots graphiques et on essaie, soit de les découper soit de les associer à d'autres mots voisins, de façon à ce que chaque segment correspond à une unité répertoriée dans le système (entrée du lexique).



الليلة الخامسة

قال لي ليلة أخرى : ألا تتمم ما كنا به بدأنا . قلت : بلى .
فأما أبو إسحق فإنه أخذ الناس للطريقة المستقيمة ، وأمضاهم على المحجة الوسطى ، وإنما ينقم عليه قلة تصنيفه من النحو ؛ وليس ابن عباد في النحو بذاك ؛ ولا كان أيضا ابن العميد إلا ضعيفا ؛ وكان يذهب عنه الشيء اليسير . وأبو إسحق معانيه فلسفية ، وعباراته عراقية ، وعادته محمودة ؛ لا يثب ولا يزئبر ، ولا يخل ولا يكهف ، ولا يلتفت وهو متوجه ، ولا يتوجه وهو ملتفت . وقال لنا : إمامي ابن عبدكان ، وهو ممن أوفى عليه ، وإن كان احتذى على مثاله ؛ وفنونه أكثر ، ومذاهبه أخفى ، وخاطره أوقد ، وتقاطره أنقد ، وروضه أنضر ، ومراجه أزهر ، و يزيد على كل من تقدم بالكتاب "التاجي" ، فإنه أبان عن أمور وكتى في مواضع ، وشبه الغارة في الصبح المنير مع الرعيل الأول ، ودل على التقدم ، وعلى الاطلاع على حقائق السياسة ولو لم يكن له غيره لكان به أعرق الناس في الخطابة ، وأعرق الكتاب في الكتابة ، هذا وخطبة منثوره ، ومنثوره منظومه ؛ إنما هو ذهب إبريز كيفما تريك فهو واحد ، و إنما يختلف بمصوغ منه ، تشكل عليه ؛ هذا مع الظرف الناصع ، والفاكهة الحلوة ،

Le processus de segmentation semi-automatique d'un texte est composé de trois phases :

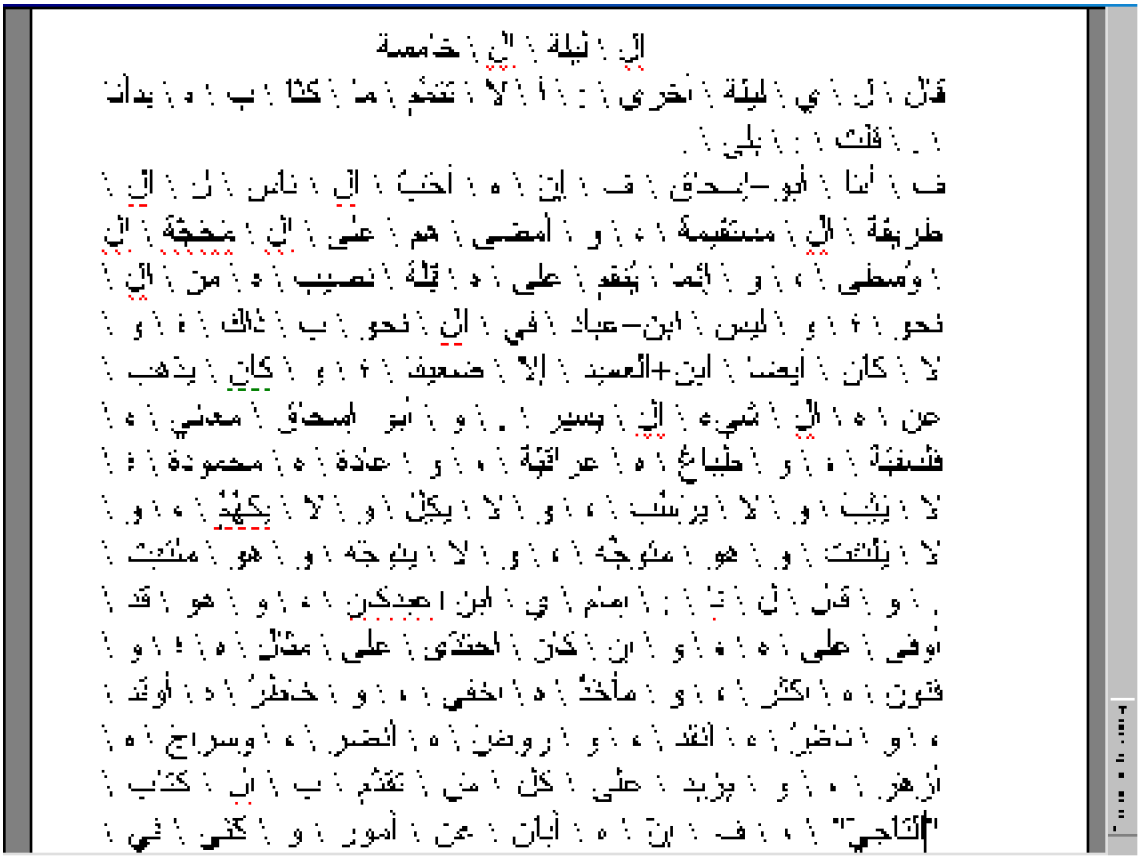
a) Repérage des mots graphiques : Le repérage des mots graphiques dans les textes arabes n'est pas délicat[35]. Les mots sont séparés par des blancs ou par des signes de ponctuation.

b) Analyse des mots graphiques : Une fois que la liste des mots graphiques a été répertoriée, on segmente chaque mot en ses différentes unités (proclitique(s) + mot minimal + enclitique(s)). Si le mot graphique présente plusieurs solutions de segmentation, nous demandons l'avis de l'expert linguiste qui supervise le processus.

c) Traitement des unités discontinues : Bien que les unités complexes (noms composés, locutions, formes figées, etc.) ne nous semblent pas très nombreuses dans les

[35] Dans les langues latines, un certain nombre de caractères fonctionnent tantôt comme séparateurs de mots tantôt comme composants de mots. C'est le cas du trait d'union ou de l'apostrophe.

textes arabes[36], on est obligé de les traiter à chaque cycle du processus d'étiquetage. Chaque unité répertoriée fait l'objet d'une recherche parmi les entrées de la liste des unités complexes. Si elle y figure, on vérifie la présence des éléments de l'unité complexe dans son contexte avoisinant. Si le processus de recherche réussit, l'unité est répertoriée comme une unité complexe. A la fin de cette étape, nous obtenons une suite d'unités séparées par des barres obliques. La figure (4-2), est obtenue à partir de la segmentation du texte de la figure (4-1).



4.2.4 Lemmatisation des unités segmentées

Le processus de lemmatisation consiste à regrouper toutes les unités segmentées sous une forme unique : le lemme. Lorsque l'analyseur propose plusieurs solutions pour la

[36] Les unités complexes occupent une place importante en français. On estime au cinquième d'un texte la surface qu'elles couvrent.

même unité segmentée, l'expert doit trancher pour l'une des solutions proposées. Par exemple, l'analyseur propose pour l'unité « لاق » trois solutions différentes :

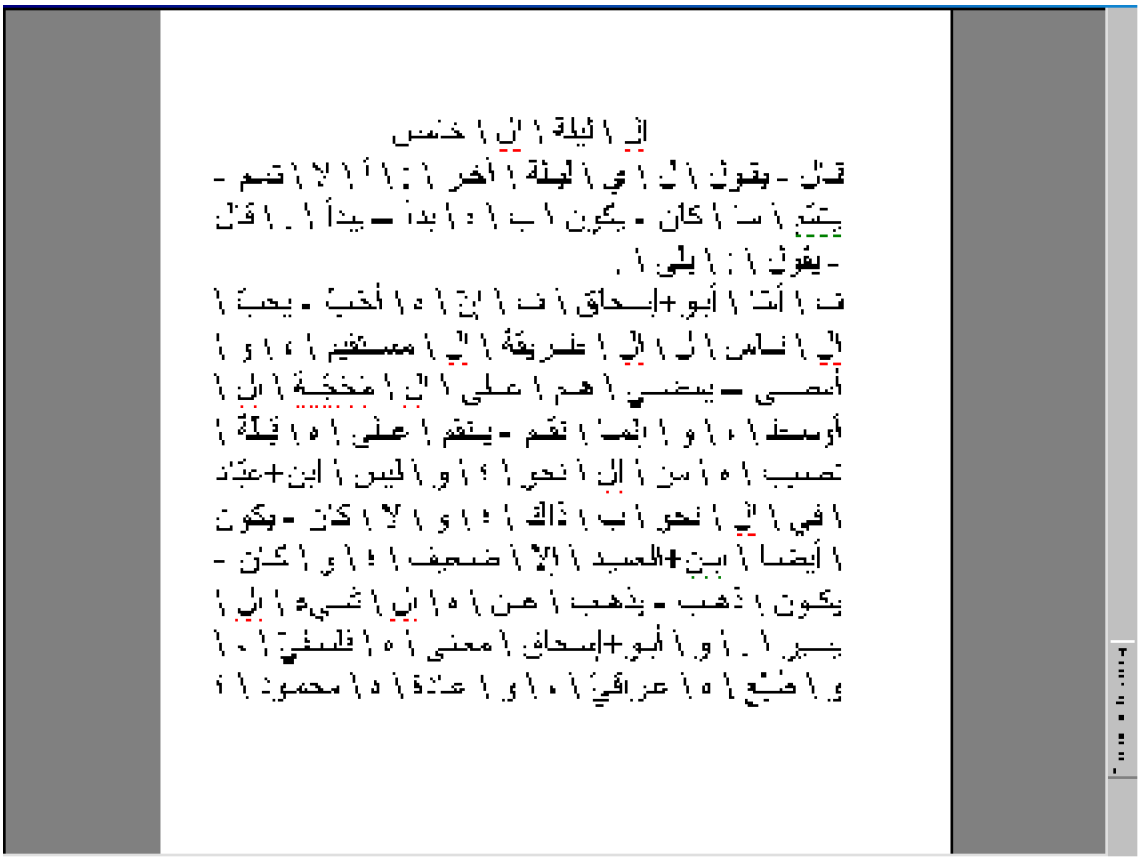
- « لاق » : forme nominale correspondant au lemme : (singulier « لاَق » - pluriel « لاَيِق ») = (« le dire »).
- « لاق » : forme conjuguée correspondant au lemme verbal « لاَق - يُلِيقَ » = (« faire la sieste »)
- « لاق » : forme conjuguée correspondant à un lemme verbal « لاَق - يُولُوقَ » = (« dire »).

L'expert, en se référant au contexte de la phrase « **لاق** لي ليلة أخرى » = (« Dans une autre nuit, il m'**a dit** ») doit évidemment choisir la dernière solution (figure 4-3).

Ainsi, chaque unité segmentée est remplacée par son lemme :

- Un verbe est remplacé par sa forme à l'accompli et à l'inaccompli à la 3ème personne du singulier masculin.
- Un déverbal est remplacé par sa forme au singulier masculin ou féminin.
- Un nom est remplacé par sa forme qui est généralement au masculin singulier, à laquelle nous ajoutons parfois le pluriel pour nuancer le sens.
- Un mot outil fléchi est ramené à son noyau au masculin singulier.

La figure (4-4) ci-dessous montre le même texte (figure 4-1) après le processus de lemmatisation.



4.2.5 Association des informations aux lemmes

Une fois que le texte a été lemmatisé, il ne reste qu'à associer à chaque lemme l'ensemble des informations morphologiques et syntaxiques correspondantes à partir de DIINAR[37] :

- Au lemme verbal est associé, l'aspect et le mode de conjugaison, le pronom de conjugaison, le verbe et le schéma syntaxique.
- Au lemme déverbal est associé, le type du déverbal, le verbe, le genre, le nombre, le mode de déclinaison et le cas de déclinaison.
- Au lemme nominal est associé, le trait humain/non humain, le trait nom propre/nom

[37] Il est à remarquer que cette étape se déroule en même temps que la précédente étape.

commun, le genre, le nombre, le mode de déclinaison et le cas de déclinaison.

- Au lemme d'un mot outil est associé le genre et le nombre.

4.3 Recherche automatique des concordances

Ces dernières années ont vu l'émergence concomitante de corpus de textes sous forme informatisée et de programmes informatiques permettant la recherche rapide des mots d'une langue donnée dans ces corpus. Ces programmes appelés concordanceurs doivent non seulement être capables de rechercher rapidement dans un vaste ensemble de textes un élément donné (morphème, mot ou expression), mais aussi de fournir tous les contextes des occurrences trouvées[38].

Ce genre d'applications est très utile pour l'apprentissage des langues. L'enseignant peut mettre à la disposition de ses élèves un nouveau type d'apprentissage du vocabulaire et de la grammaire s'appuyant sur des données authentiques. Nous essaierons au début de cette section de donner des exemples d'utilisation du concordanceur dans le domaine d'apprentissage des langues. Nous présenterons ensuite les principales fonctionnalités qu'un concordanceur doit assurer et nous verrons comment elles doivent être adaptées à la morphologie et aux textes arabes. Nous ajouterons quelques nouvelles fonctionnalités spécifiques aux concordanceurs fonctionnant dans des environnements d'apprentissage. Nous présenterons enfin la réalisation du concordanceur que nous utiliserons dans notre environnement d'apprentissage.

4.3.1 Les applications pédagogiques du concordanceur

Les concordanceurs sont très utiles lorsqu'un apprenant désire approfondir ses connaissances sur un mot donné. L'apprenant peut, soit explorer directement les résultats retournés par le concordanceur[39], soit s'exercer sur des activités générées automatiquement.

L'intérêt du premier type d'utilisation est que les phrases obtenues sont extraites de textes authentiques et reflètent véritablement l'usage de la langue. Ce ne sont pas des exemples construits par des spécialistes (traditionnellement, la préparation des compilations de concordances se réalisait par l'enseignant à la main qui construisait lui-même ses phrases). Avec cet outil, l'apprenant a la possibilité d'assumer lui-même le contrôle du processus d'apprentissage et devient le « linguiste », essayant d'identifier, de classifier et de dégager les régularités dans le comportement syntaxique, collocationnel et sémantique du vocable, ce que Tim Johns (1991) appelle *Data Driven Learning*, c'est à dire une exploration conduite par les données. D'après Johns, il est possible d'observer

[38] Pour la langue arabe, R. Abbes a entamé une thèse de doctorat sur le sujet (Abbes, 99), (Abbes & Hassoun, 99).

[39] Les résultats des concordances peuvent être directement visualisés sur écran ou imprimés sur papier.

de nombreux phénomènes linguistiques comme par exemple les structures grammaticales (constructions différentes pour les verbes anglais convince et persuade), les hyperonymes (révélés par l'emploi de such as : industries such as steelmaking) ou les adverbes (différence d'utilisation entre however et nevertheless). L'intérêt des concordanceurs est d'autant plus évident que ces informations ne sont pas toujours décrites par les dictionnaires.

Le concordanceur permet aussi la génération automatique d'activités lexicales ou grammaticales qui peuvent être adaptées au profil de l'apprenant. Différentes formes d'activités peuvent être imaginées à l'aide des concordanceurs[40] (cf. chapitre 7) : Exercices à trous, questions à choix multiples, etc. Dans le chapitre 4 de leur livre, Tribble et Jones (1997) proposent de nombreuses activités en anglais à exploiter en classe (individuellement, en binôme ou en sous-groupe), selon un principe de difficulté croissante pour l'élève. Les auteurs fournissent de nombreux exemples, expliquent les objectifs pédagogiques visés, le déroulement prévu et les consignes à donner pour chaque activité. La première consiste à remplacer le mot-cible par un mot inventé, puis à demander aux apprenants de rétablir la vérité. Ceci permet à l'apprenant de comprendre de lui-même ce qu'est une concordance. Les suivantes montrent comment travailler la syntaxe (travail sur les post-positions anglaises), la sémantique et le lexique (différences entre "look", "see" et "watch"), et l'idiomaticité (concordances sur "foot", "mout", etc). Joseph Rézeau (1997) propose aussi dans son article une série d'activités très intéressante dans le domaine de l'étude des locutions figées et semi-figées, des mots composés, des réseaux sémantiques et des mots instrumentaux pour le français (Voir aussi, (Johns, 1988)).

4.3.2 Un concordanceur pour l'EIAO de l'arabe

Un concordanceur doit être capable de parcourir une vaste base textuelle et de fournir toutes les phrases contenant le mot recherché. Le résultat est souvent affiché en format KWIC (Key Word In Context) ou, en français, MCC (Mot-Clef en Contexte), c'est à dire que le mot-clé est affiché au milieu de l'écran ligne par ligne et est entouré par son contexte gauche et droit. La régularité de cette disposition permet de mettre en évidence les caractéristiques du mot. En principe, un concordanceur doit pouvoir trier les contextes par ordre alphabétique ce qui permet d'étudier les collocations ou les schémas syntaxiques.

Pour retrouver les concordances, le programme de recherche (KWIC) parcourt le texte avec une ligne de longueur fixe et à chaque fois il compare le mot du milieu avec la forme graphique du mot recherché. Cette technique permet généralement de retrouver toutes les concordances de mots latins puisque ces derniers ne connaissent pas de variation morphologique importante. Par contre, pour la langue arabe cette technique ne donne pas des résultats satisfaisants.

A travers une série d'exemples de requêtes choisis[41], nous allons essayer d'expliquer

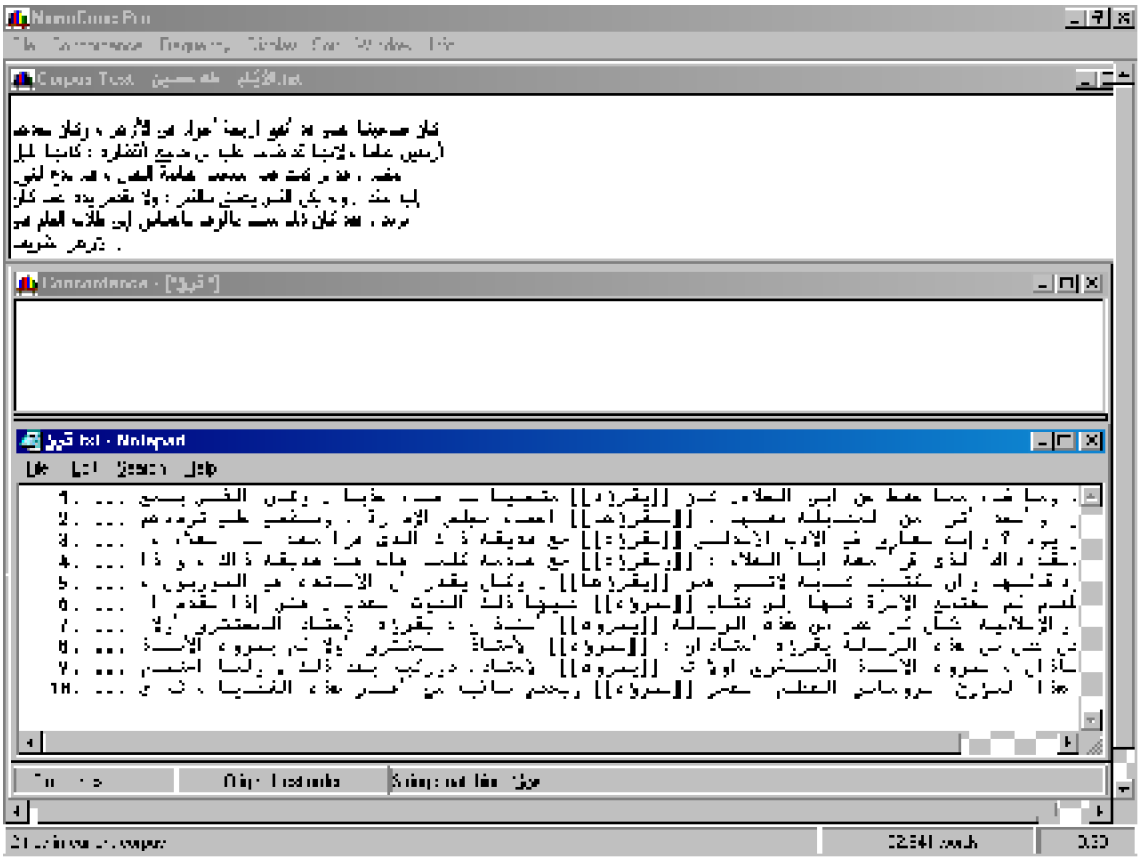
[40] Le site de Joseph Rézeau offre des exercices et des discussions basés sur le travail des concordances. Grande richesse de liens. Consulté en mai 2001 : http://www.uhb.fr/campus/joseph.rezeau/welcome.htm

pourquoi les concordanceurs KWIC ne répondent pas aux besoins d'un environnement d'apprentissage de la langue arabe, malgré les nombreuses fonctionnalités qu'ils offrent.

- **Requête 1 :** Nous avons lancé une première requête avec le mot graphique **(أرق)** sur une partie du livre (ما ّيألا) de (نيس ح مط), en espérant obtenir les contextes de toutes les formes conjuguées du verbe (أَرْقَ - يَ ْقرُ) = (« lire »). Le concordanceur nous a retourné 19 résultats seulement d'où sont absentes des formes conjugués simples comme (أرقي) (figure 4-4).
- **Requête 2 :** Pour essayer de retrouver les concordances manquantes, nous avons employé deux jokers (*), un avant et un après le mot : (*أرق*). Les jokers, qui sont très utilisés dans les environnements informatiques, permettent d'élargir la recherche à des chaînes de caractères non prévus explicitement par l'utilisateur. Il existe deux types de jokers : le caractère graphique (*) qui permet de remplacer un ou plusieurs caractères et le caractère graphique (?) qui remplace un seul caractère. Le résultat est nettement meilleur, puisque nous avons obtenu 87 résultats au lieu de 19 sur la même portion du livre (ما ّيألا) (figure 4-5), avec des formes comme « تأرق » ou « أرقي ». Néanmoins, cette fonctionnalité ne permet pas de retrouver les formes conjuguées qui emploient d'autres graphies de la HAMZA (autres que la graphie « أ »).

[41] Pour effectuer ces requêtes, nous avons utilisé « MonoConc Pro » - Version 1.0 (Build 208) – Michael Barlow – Athelstan 1996 – 1999 (info@athel.com)

- **Requête 3 :** Pour pallier ces inconvénients, on doit utiliser une requête complexe intégrant toutes les formes graphiques des formes conjuguées du verbe (أَرَقَ - أُرْقَ‌ي). On pourra recourir aux opérateurs booléens (OR, AND, NOT), pour construire de telles requêtes. Ainsi, on doit lancer la requête suivante : (*أرق*) OR (*ؤرق*) OR (*ىرق*) OR (*آرق*). La requête (*ؤرق*) à elle seule permet d'avoir dix nouveaux résultats (figure 4-6).



Evidement, nous ne pouvons pas produire automatiquement ce type de requête pour proposer par exemple des activités d'apprentissage des formes conjuguées du verbe (أَرْقَ - يَأْرَقُ).

Abstraction faite aux besoins de l'environnement d'apprentissage, l'utilisation de ces options sur des textes arabes ne résout pas complètement le problème du taux de silence[42] qui reste malgré tout présent. Ceci est dû pour deux raisons :

- La première raison est d'ordre morphologique. Certaines formes dérivées utilisent des noyaux différents de leur lemme (i.e. suppression ou insertion de consonnes dans le mot). Les formes conjuguées en arabe peuvent en effet être obtenues à partir de plusieurs bases de conjugaison[43]. Par exemple, le verbe « يولَ - ىَوَلَ » = («pivoter »), possède trois bases de conjugaison différentes «ول», «يَوَلَ » et «ىَوَلَ»

[42] On peut définir le taux de silence comme la proportion des mots dérivés du mot recherché qui n'ont pas été trouvés par le concordanceur.

pour la conjugaison de l'accompli. Si par exemple, nous lançons une requête avec la première base dans le mot à rechercher suivie d'un joker, nous obtenons toutes les formes conjuguées mais avec un taux de bruit très élevé (le résultat inclura les phrases contenant par exemple les prépositions «وْلَ », «الْوَلَ », etc.). Par contre, si nous mettons la deuxième ou la troisième base comme objet de requête, nous exclurons les résultats correspondants aux formes conjuguées obtenues avec les deux autres bases. On retrouve ce même problème avec les noms. Les pluriels brisés présentent généralement des noyaux différents de leur singulier. Par exemple, les singuliers qui sont construits avec le schème (فَعْلٌ) comme (فَرْحٌ) = (« mariage ») possèdent des pluriels avec un "alif" supplémentaire (فَارْحٌ). Il existe même des pluriels dont les consonnes sont complètement différentes de celles de leur singulier comme (نِسَاءٌ) « femmes » qui a pour singulier (اِمْرَأَةٌ).

- La deuxième raison du taux de silence pourrait être la nature même du corpus textuel utilisé. Dans le cas où le corpus serait formé par des textes partiellement voyellés (c'est généralement le cas pour les textes de l'environnement d'apprentissage), les taux de silence et de bruit seront très importants. Etant donné que les concordanceurs (KWIC) ne cherchent que les formes dont la voyellation est exactement la même que celle du mot-clé, ils vont passer sous silence toutes les autres formes. L'utilisation des jokers à la place des voyelles engendrera par contre un taux de bruit très élevé puisque les voyelles, ayant leur propre code (ASCII ou UNICODE), sont considérées comme des consonnes par les concordanceurs.

Pour toutes ces considérations, un concordanceur fonctionnant sur des textes arabes doit absolument assurer une fonction d'analyse qui ramène le mot recherché à son lemme (cf. § 4.3.3). Il doit d'autre part, assurer quelques fonctionnalités supplémentaires, pour répondre aux attentes des environnements d'apprentissage :

- La première fonctionnalité concerne le contexte d'apparition du mot recherché (le format d'affichage KWIC n'affiche généralement pas la totalité de la phrase). La restriction de l'affichage du contexte du mot à la seule phrase l'incluant est nécessaire pour les apprenants débutants surtout quand il s'agit d'un texte sans ponctuation.
- La seconde fonctionnalité concerne la génération automatique d'activités grammaticales sur un aspect morphologique ou syntaxique particulier (cf. § 4.3.2). Le concordanceur doit être en mesure de retrouver les phrases contenant des mots vérifiant les propriétés morpho-syntaxiques de ces activités.

4.3.3 Réalisation du concordanceur

Il existe deux types de concordanceurs : les "streaming concordancers" et les "indexing concordancers"[44] (Tribble C. & Jones G, 1997). Le premier lit les textes ligne par ligne et crée une concordance en temps réel alors que l'autre constitue un index durable qui

[43] Pour un seul aspect/mode de conjugaison, un certain nombre de verbes peuvent avoir jusqu'à quatre bases de conjugaison différentes (cf. § 2.3.2)

pourra servir par la suite de base à diverses opérations d'extraction des données.

Avec toutes les contraintes que nous nous sommes définies dans la section précédente, seuls les concordanceurs du second type peuvent répondre à ces besoins. Le processus de recherche d'une concordance dans une base textuelle sera plus rapide et les résultats obtenus plus précis. Nous avons par conséquent construit un « indexing concordancer » fonctionnant en deux phases distinctes :

a-) Indexation des textes

Cette phase n'est autre que le processus d'*étiquetage des textes arabes* que nous avons présenté au début de ce chapitre. Chaque texte du corpus est segmenté, lemmatisé et enrichi par des informations linguistiques associées aux lemmes.

Le résultat de l'indexation[45] est une suite ordonnée d'enregistrements. Chaque enregistrement est formé du mot, de son lemme, de ses différentes propriétés morpho-syntaxiques et de son contexte d'apparition (le processus de segmentation permettant de bien cerner les frontières des phrases). Les différents champs du fichier résultant constituent des index à partir desquels, on pourrait effectuer des requêtes. Ainsi, on pourra générer des activités appropriées à partir des résultats de requêtes formulées avec une propriété morphologique ou syntaxique par exemple.

b-) Recherche de concordances

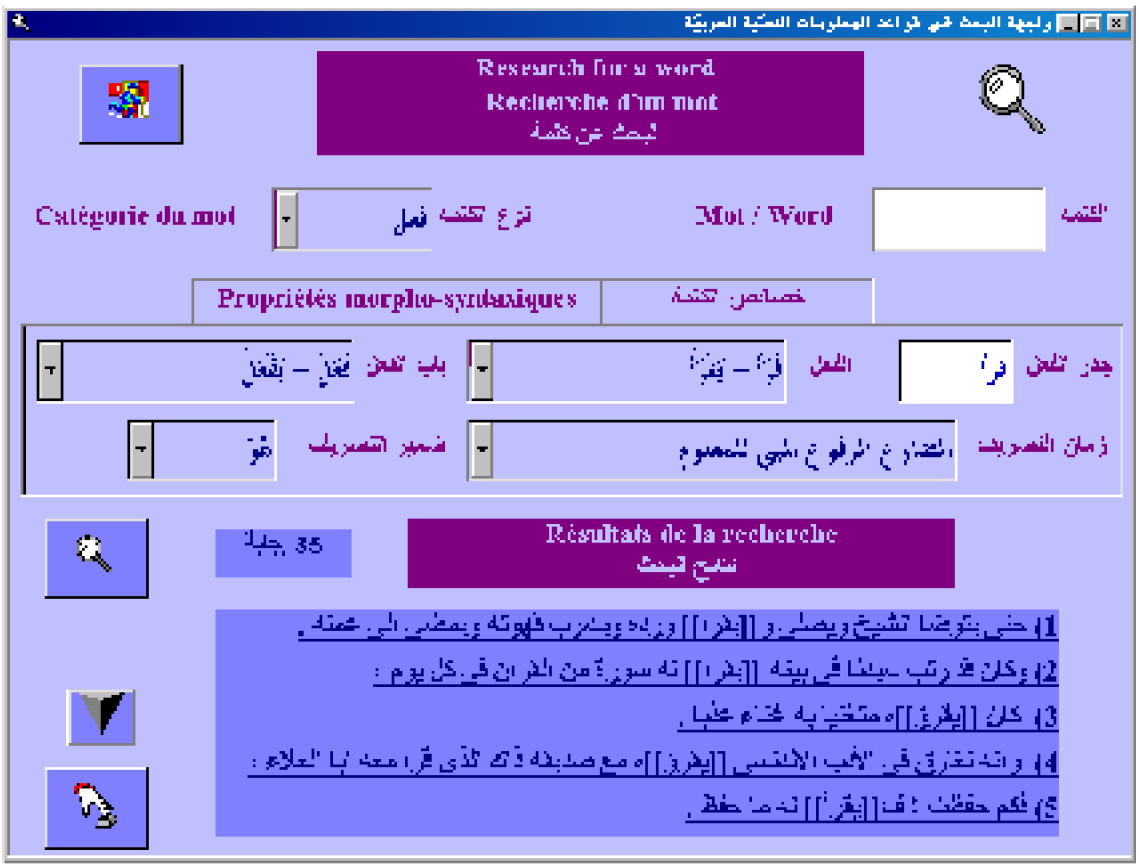
L'indexation du corpus textuel étant effectuée, l'utilisateur peut accéder directement à tous les contextes d'un mot ou d'une propriété donnée à partir d'une interface réalisée à cet effet (figure 4-7). Cette interface offre à l'utilisateur plusieurs possibilités de recherche :

- L'utilisateur peut effectuer une recherche classique en lançant une requête avec le mot recherché (champ de saisie : « الكلمة » = « mot »). La requête générera seulement les concordances du mot c'est à dire les phrases contenant le mot entre deux blancs. Le programme utilisera pour cela le champ index correspondant à la forme graphique du mot.
- L'utilisateur peut raffiner les résultats en sélectionnant la catégorie du mot à chercher (verbe, nom, déverbal, mot outil, nom propre) dans la liste « نوع الكلمة » = (« catégorie du mot »).
- L'utilisateur peut enfin accéder aux phrases contenant des mots ayant une certaine propriété morphologique ou syntaxique. Par exemple, l'utilisateur peut lancer des requêtes avec une racine d'un verbe, une forme d'un verbe, un aspect de conjugaison ou un pronom. Le concordanceur utilisera dans ce cas le champ d'index de la propriété indiquée par l'utilisateur.

[44] A notre connaissance, ces types n'ont pas de traduction en français. Une traduction mot à mot donne *"concordanceur en ruban"* pour le premier type et *"concordanceur indexeur"* pour le second.

[45] La forme du résultat est un fichier indexé mais aurait pu être une entité dans une base de données.

- L'utilisateur peut construire enfin des requêtes en combinant plusieurs propriétés. Le concordanceur utilisera dans ce cas les champs d'index concernés. La figure (4-7) montre un exemple de recherche élaboré. La requête permet de générer uniquement les concordances des formes conjuguées à partir du verbe "أَرَقَ - يَأْرَقُ" à la troisième personne du singulier masculin et à l'inaccompli indicatif. Nous pouvons remarquer que contrairement aux concordanceurs (KWIC), les contextes des différentes formes graphiques de ce verbe comme ("ؤرقي") sont automatiquement générés.



4.3.4 Conclusion

Dans cette section, nous avons réalisé un concordanceur adapté aux besoins d'un environnement d'apprentissage informatisé de la langue arabe. Le processus de recherche de concordances s'effectue à partir de textes préalablement segmentés et

étiquetés par un processus semi-automatique.

L'interface développée permet d'effectuer trois types de recherche. La première recherche classique servira notamment aux utilisateurs qui s'intéressent uniquement aux occurrences relatives à la forme graphique du mot. La seconde recherche permet d'obtenir les contextes des différentes formes dérivées d'un mot donné. Cette forme de recherche est mieux adaptée à la langue arabe puisqu'elle permet de générer des formes qui ne sont pas explicitement définies par l'apprenant. La dernière recherche permet d'avoir les contextes des occurrences de mots définis par leurs propriétés morpho-syntaxiques. L'interface offre une grande liberté à l'utilisateur pour s'essayer à diverses combinaisons de recherche inédites. Ce type de recherche nous servira particulièrement pour la génération automatique d'activités grammaticales pour notre environnement d'apprentissage.

4.4 Quantifier les faits langagiers

4.4.1 Introduction

A partir d'un corpus préalablement étiqueté, nous pourrons à l'aide d'un simple outil informatique calculer le nombre d'occurrences d'unités textuelles qui correspondent à un patron donné (mot, lemme, catégorie morpho-syntaxique, etc.). Etant donné que la construction de ce corpus va durer dans le temps, nous avons pensé à créer un outil de calcul fonctionnant sur des textes bruts.

Dans le cadre du projet DIINAR-MBC (cf. Annexe 3), nous disposions en effet d'un important corpus textuel brut et il fallait sélectionner les unités lexicales les plus fréquentes pour définir un prototype de dictionnaire électronique multilingue (cf. chapitre 6). Pour répondre à ce besoin, nous avons réalisé un programme de calcul de fréquences que nous présenterons dans cette section. Cette description sera précédée par quelques exemples d'applications qui pourront exploiter les résultats retournés par ce programme.

4.4.2 A quoi servent les calculs de fréquences ?

L'indication de fréquence est très utile notamment pour la construction de dictionnaires basés sur des corpus textuels. En effet, l'utilisateur trouve souvent des difficultés pour décider du sens à donner au mot recherché parmi ceux proposés par le dictionnaire. La fréquence du mot dans un corpus textuel peut constituer alors un élément d'information très utile qui peut guider l'apprenant à mieux choisir son mot.

Le dictionnaire anglais COBUILD applique systématiquement une échelle de 0 à 5 diamants (pour les plus fréquents). L'apprenant est ainsi informé de l'utilisation réelle d'une entrée et peut en mesurer, par exemple, son côté démodé ou au contraire dans le vent.

Pour un environnement d'apprentissage d'une langue donnée, l'étude quantitative des faits langagiers d'un corpus textuel permet par exemple de proposer aux apprenants un vocabulaire de base et des textes appropriés à leur niveau.

4.4.3 Description du programme

Il s'agit d'un outil permettant, à partir d'un texte brut donné, de générer dans un fichier la fréquence de tous les mots rencontrés. Ce programme se base bien entendu sur les résultats retournés par l'analyseur des mots graphiques, qui sont souvent équivoques. Plusieurs cas peuvent en effet se présenter :

- L'analyseur échoue dans son analyse : le mot graphique doit faire partie de la liste des mots non reconnus. Si le mot graphique ne figure pas déjà sur cette liste, il fait l'objet d'une nouvelle entrée sinon sa fréquence est incrémentée d'une unité. Cette liste est formée généralement par les noms propres et les mots absents du dictionnaire ;
- L'analyseur retourne une ou plusieurs solutions d'analyse ayant le même mot minimal et la même catégorie grammaticale : le mot minimal sans ses proclitiques et ses enclitiques est enregistré dans le fichier avec une incrémentation de la catégorie grammaticale correspondante (verbe, nom ou mot outil).
- L'analyseur retourne plusieurs solutions d'analyse ayant des catégories grammaticales distinctes : Tous les mots minimaux de ces solutions sont enregistrés avec une incrémentation la fréquence des différentes catégories grammaticales correspondantes aux analyses. Par exemple : le mot graphique « يفـل » a quatre analyses possibles :

  – Le mot outil يف de catégorie حرف الجر précédé de مال التوكيد.
  – Le mot outil يف de catégorie حرف التعليل précédé de مال التوكيد.
  – La forme conjuguée « يُفـْلَ » du verbe « اَفـَلَ - يَفـْلُ » à la 3ème personne du singulier à l'accompli passif.
  – La forme conjuguée «يِفـْلَ » du verbe « فـَلَ - يُفـْلِ » à la 2ème personne du singulier féminin à l'impératif.

Pour le mot graphique « يفـل », on doit donc ajouter deux enregistrements au fichier des fréquences : Le premier avec le mot « يف » qui est considéré comme un mot outil et le second avec le mot « يفـل » qui est considéré comme un verbe. Si le même « يفـل » était entouré par d'autres proclitiques ou enclitiques, on aurait pu avoir d'autres solutions différentes.

4.5 Conclusion

Dans ce chapitre, nous avons présenté les principales applications qui vont nous permettre de construire au fur et à mesure les ressources de l'environnement « *AL-Mu*^{c} *aLLiM* » (textes étiquetés par des informations morphologiques et syntaxiques). Les données de ces ressources doivent être correctes et précises, et c'est pour cela que les applications développées font appel à l'intervention d'une expertise humaine lorsque les résultats sont équivoques.

Le programme de recherche de concordances est particulièrement important, puisqu'il va permettre de générer automatiquement des activités adaptées au profil de l'apprenant. Le programme de calcul des fréquences, ne permet pas pour l'instant de produire des résultats fiables puisque nous n'avons pas encore à notre disposition un corpus textuel complètement étiqueté. On pourra néanmoins faire fonctionner ce programme sur les textes de l'environnement afin de pouvoir proposer des textes adaptés au profil des apprenants.

Avant de passer à la mise en œuvre de ces ressources, nous essaierons dans le chapitre suivant, de dégager quelques principes linguistiques et pratiques pédagogiques relatifs à l'apprentissage d'une langue seconde qui guideront l'élaboration du dictionnaire (chapitre 6) et l'architecture de l'environnement « *AL-Mu*^{c} *aLLiM* ».

Chapitre 5 Bases linguistiques et pratiques pédagogiques retenues pour la conception de l'environnement d'apprentissage « *AL-Mu c aLLiM* »

« Excellent de ne pas trouver le mot juste - cela y peut prouver qu'on envisage bien un fait mental, et non une ombre du dictionnaire. » Paul VALÉRY

5.1 Introduction

Les principes linguistiques et les pratiques pédagogiques sont multiples, divers et mêmes antagonistes. Nous avons pensé qu'il est judicieux de procéder à une détermination des principes linguistiques et des pratiques pédagogiques qui soient en conformité avec :

- Le produit DIINAR avec toutes ses composantes (DIINAR est lui même sous-tendu par une vision linguistique).
- L'emploi de la machine dans l'enseignement
- Les caractéristiques spécifiques à la langue arabe.

Le tout doit être conçu dans un cadre le plus harmonieux possible. Il s'agit donc, de choisir les principes pédagogiques et de spécifier les compétences linguistiques visées, orientant de ce fait la constitution et l'organisation des ressources du système.

L'environnement « *AL-MucaLLiM* » reprend certaines des idées développées dans le système « ALEXIA » (Chanier & Fouqueré & Issac, 1995), (Issac, 1997), (Selva, 1999), (Chanier & Selva 2000) et le système « Lexica » (Goodfellow, 1994), tout en apportant une contribution sur certains des constituants.

Nous montrerons tout d'abord l'intérêt d'orienter l'apprentissage d'une langue vers le lexique. Nous examinerons ensuite le résultat des travaux en psycholinguistique sur le lexique mental et nous étudierons le processus de l'acquisition lexicale qui se définit par l'intégration de nouvelles connaissances au sein des anciennes et les moyens qui le favorisent.

Nous présenterons enfin le protocole expérimental mis en place à l'université Lyon 2 et qui nous a permis de peaufiner l'architecture de l'environnement d'apprentissage « *AL-Mu* ^{c} *aLLiM* ».

5.2 Le lexique dans l'apprentissage de la langue arabe

Le *vocabulaire* n'a pas toujours occupé la même place en didactique des langues étrangères. Si maintenant aussi bien les enseignants que les apprenants considèrent qu'une bonne connaissance du lexique est fondamentale pour la réelle maîtrise d'une langue, il n'en a pas été toujours ainsi. On distingue trois époques *différentes dans l'histoire de la didactique des langues* (Issac, 1997, pp 18) :

- La première période, celle des méthodes dites traditionnelles. De telles méthodes prenaient comme support des textes poétiques ou littéraires et leur objectif était d'amener l'apprenant à un niveau leur permettant de lire ces textes. Cet objectif a pour corollaire l'acquisition d'un vocabulaire riche. Donc un apprenant devait acquérir un vocabulaire important afin de maîtriser des textes de langage écrit, par opposition au langage oral, et de style littéraire, par opposition à tout autre style.
- La deuxième période, a vu l'apparition des *méthodes structurales* (MAO : Méthodes audio-orales et MAV : méthodes audiovisuelles) et des *laboratoires de langue*. Les apprenants doivent surtout acquérir les mécanismes de base (comment doit-on structurer une phrase ?) et se contenter d'un vocabulaire minimal ne *parasitant* pas les apprentissages plus fondamentaux. Ainsi, dans les années 70, les ouvrages dits *classiques* servant de référence sont, pour leur plus grande part, orientés vers l'apprentissage des structures grammaticales, et de plus les évaluations se faisaient presque exclusivement sur des critères grammaticaux. Cette période coïncide avec les recherches des « lexiques de base » tel que le français fondamental.
- Depuis le milieu des années 80, non seulement les apprenants manifestent leur désir de combler un *déficit lexical* qui les pénalise dans leur lecture mais de plus en plus de

chercheurs (en didactique mais aussi en linguistique et en psycholinguistique) attribuent aux mots un statut plus important.

L'apprentissage orienté vers le lexique permet en effet, d'améliorer les différentes compétences de l'apprenant : compétences linguistiques (morphologiques, syntaxiques et sémantiques) et compétences communicatives.

D'un point de vue sémantique, on se rend compte que la signification d'un mot n'est pas figée, c'est à dire que le mot a une propension naturelle à la polysémie (Boogards, 1994) surtout si nous prenons compte des « situations de communication ». C'est le contexte qui précise les différents sens du mot, comme le montrent les exemples suivants relatifs au mot «رْأَس » :

« وَضع الولد يدو على سأر س » = (« Le garçon a mis sa main sur sa tête. »)

« تعتبر البلاد التونسيّة على رأس بلدان القارّة الإفريقيّة في ميدان... » = (« La Tunisie est considérée comme le premier pays du continent africain en matière…»).

« رأس القضيّة هو... » = (« Le nœud du problème est … »).

« رأس العصابة هو... » = (« Le chef de la bande est … »).

Sur le plan morphologique, les schèmes des mots arabes contiennent des informations de type syntaxique et sémantique. Par exemple, les verbes ayant pour schème «لُعْفَي - لَعُفَ» sont intransitifs et dénotent un « qualificatif » de type durable comme « رُبْكَي – رَبُكَ » = (« être âgé » / « être grand ») ou « رُغْصَي - رَغُصَ » = (« être petit » / « être peu »).

D'autre part, syntaxe et lexique ne peuvent être considérés indépendamment. Il est possible à un apprenant débutant d'utiliser correctement un mot (dans sa morphologie) sans utiliser la syntaxe qui lui est propre. La syntaxe deviendrait par contre nécessaire dès qu'il s'agit d'introduire le dit mot dans une phrase dans une situation de communication.

En plus, le sens du mot peut changer selon la structure syntaxique et la nature des arguments de la phrase dans laquelle il est employé, comme le montrent les exemples suivants du verbe «ذُخْأَي - ذَخَأَ» = (« prendre ») :

1. «ذَخَأَ فلان الشيئ » = (« Un tel saisit quelque chose »)
2. «ذَخَأَ فلان فلانا إلى مكان » = (« Un tel conduit / dirige un tel à un endroit »)
3. «ذَخَأَ فلان الشيء باب » = (« Un tel croit / adopte quelque chose »)
4. «ذَخَأَ فلان فلانا على ان » = (« Un tel reproche à un tel que / de… »)

La connaissance d'un mot fait donc intervenir de nombreux savoirs : c'est savoir dans quel contexte il est utilisé, c'est appréhender les limitations de son usage selon les variations de fonctions ou de situations, c'est, enfin, connaître ses comportements morphologiques, syntaxiques et sémantiques. Nous verrons au chapitre 6 que le dictionnaire pour apprenant doit inclure toutes ces informations et que chacune d'elles doit faire l'objet d'activités spécifiques (voir chapitre 7).

Un dernier argument en faveur d'un apprentissage orienté vers l'acquisition lexicale

est que, justement, celle-ci se fait à un taux très faible en langue seconde. Bogaards (Bogaards, 1994) rappelle que le nombre de mots réutilisables en production est seulement de 1500 après cinq ou six ans d'apprentissage scolaire, au niveau de la production.

5.3 Le lexique mental

Les recherches menées en psycholinguistique nous éclairent sur la manière dont les mots pourraient être stockés dans la mémoire de chaque individu. Ces données sont à la base de tous les environnements d'apprentissage du lexique, et de tous les systèmes cherchant à représenter des processus mentaux : la notion d'hypertexte est aussi basée sur ce principe, et c'est pourquoi elle est souvent utilisée en environnements d'apprentissage (Issac, 1997).

Les recherches en psycholinguistique montrent que les mots ne sont pas disposés au hasard, sans aucun lien entre eux, dans notre mémoire. Le nombre considérable de mots dont dispose chaque être humain suppose un classement performant et systématique. Un empilement en vrac dans la mémoire ne pourrait pas expliquer les performances étonnantes de chacun en matière de vitesse de reconnaissance et de production des mots.

Il ne faut pas penser non plus que les mots d'une langue sont organisés par ordre alphabétique comme pour un dictionnaire et qu'elles couvrent la réalité d'une manière régulière. Si le rangement des items lexicaux était alphabétique, on s'attendrait à ce que les lapsus fassent apparaître des mots proches alphabétiquement de ceux qui devraient normalement être produits. Or ce n'est pratiquement jamais le cas.

Mais la différence entre dictionnaire et lexique mental ne s'arrête pas là. Elle est bien plus profonde. En effet, on peut constater que les quantités d'informations de part et d'autre ne sont pas comparables. Le lexique mental contient de loin bien plus d'information que tout dictionnaire. Une foule de détails ne sont pas considérés car les dictionnaires sont inévitablement limités et ne peuvent pas contenir tous les détails possibles sur chaque mot. Hudson (1984, cité par Selva (1999)) remarque : « Il n'y a pas de limite à la quantité d'information détaillée… qui peut être associée à un item lexical. Les dictionnaires existants, même les plus gros, ne peuvent spécifier les items lexicaux que de manière incomplète. »

Les résultats des recherches montrent que les mots du lexique sont plutôt proches des pièces d'un « puzzle » qui s'emboîtent les unes dans les autres et qui se conçoivent les unes par rapport aux autres. Les choses ne sont pas si simples car il peut y avoir plusieurs mots pour exprimer une même notion tandis que d'autres concepts ne sont pas lexicalisés. Il y a parfois recouvrement de sens lorsque plusieurs mots ont un ou plusieurs traits en commun.

Beaucoup de modèles essayant d'expliquer ces cohabitations des mots dans le lexique mental ont été proposés, mais l'ensemble converge vers deux grands types de

théories (Selva, 1999). Il y a d'une part les « atomic globule theories » et d'autre part les « cobweb theories ». Les premières affirment que les mots sont construits à partir d'un ensemble commun d'« atomes de sens » (en fait de primitives sémantiques) et que les mots reliés possèdent plusieurs atomes en commun. Les secondes considèrent que si les mots sont reliés entre eux, c'est à cause de l'existence de liens créés par les locuteurs. D'un côté, les mots sont vus comme un assemblage de morceaux élémentaires, de l'autre ils sont considérés à part entière avec leurs caractéristiques et formant un réseau (théories des toiles verbales). Même si le consensus n'est pas total, les chercheurs se tournent désormais davantage vers la deuxième type de théories, car l'association de mots dans la mémoire a pu être mise en évidence tandis qu'aucune expérimentation n'a montré de façon concluante l'existence des primitives sémantiques.

La théorie des toiles verbales (Aitchison, 1987, pp. 72-85), considère le lexique mental comme un vaste réseau, une toile verbale, dans lequel les nœuds sont les items lexicaux reliés entre eux par des chemins. A partir des réponses données aux tests d'associations, on établit que les liens peuvent être principalement de quatre types, classés par fréquence de réponse, les plus courants en premier :

- entre mots co-occurrents, qui sont les mots apparaissant le plus souvent dans les réponses aux tests d'association. Ils appartiennent aux mêmes champs sémantiques avec le même niveau de détail (sel et poivre ; papillon et mite ; rouge, blanc, bleu, etc.),
- entre les membres d'une collocation, apparaissant souvent ensemble dans des expressions plus ou moins figées (eau et salé ; bleu et marine ; etc.),
- entre hyponyme et hyperonyme (papillon et insecte ; rouge et couleur ; etc.),
- entre synonymes, plus rarement (léopard et panthère).

Bogaards (1994, pp. 71), fait remarquer que ce ne sont pas véritablement les mots qui sont liés entre eux mais leurs lexies, c'est-à-dire des éléments ayant une unité certaine au niveau sémantique. Ainsi, les toiles verbales sont organisées selon des critères exclusivement sémantiques. Les mots sont principalement rangés en champs sémantiques et liés entre eux par des relations plus ou moins fortes suivant leur nature.

L'étude des lapsus montre aussi que très fréquemment un mot est remplacé par un autre de même catégorie grammaticale. Sémantique et syntaxe sont donc indissociables.

Pour finir, un dernier résultat concernant la morphologie et la dérivation : il semble que les mots soient stockés comme un tout à part entière et non pas décomposés en affixes et bases et recomposés lors de la compréhension ou de la production du discours.

Néanmoins, pour les mots décomposables ou fléchis d'une manière régulière, il semble que les marques de flexion ne sont pas stockées avec le mot mais ajoutées dans le feu du discours. Voyons maintenant les conséquences de ces résultats dans le processus de l'apprentissage lexical.

5.4 L'acquisition lexicale

Dans cette section, nous passerons en revue d'abord les principales caractéristiques du processus d'acquisition lexicale avant de s'attarder sur l'apprentissage à partir d'un contexte écrit. A la fin de cette section, nous réfléchirons aux moyens que nous pourrons mettre en œuvre afin d'aider les apprenants dans cette tâche (usage d'activités lexicales et d'un dictionnaire personnel).

5.4.1 Processus d'apprentissage lexical

Le vocabulaire comporte deux aspects, qui correspondent à deux niveaux de traitement. Le premier concerne la forme des mots (composante phonétique et graphique) tandis que le deuxième a rapport à leur sens (aspect sémantique). Ces aspects ne mettent pas en jeu les mêmes capacités cognitives. Le premier est en effet plus superficiel et intervient en premier dans l'acquisition d'une langue. Par contre, l'acquisition se produit par l'intégration d'indices sémantiques dans les réseaux du lexique mental. Lors de l'apprentissage d'une langue seconde, les deux aspects sont présents dès le départ, mais si l'intégration des mots nouveaux se produit plutôt suivant des critères de forme au début, elle laisse place peu à peu à des associations plus profondes de nature sémantique au fur et à mesure que la compétence se développe.

C'est la tâche d'acquisition qui fixe le niveau de traitement. Une tâche de répétition, qui ne fait intervenir que la forme des mots ne produira qu'une fixation superficielle dans la mémoire à long terme. Au contraire, une tâche de raisonnement ou de comparaison détaillée agira plus profondément et impliquera l'intégration du mot dans divers réseaux mentaux de l'apprenant. L'acquisition se produit lorsque les connaissances nouvelles véhiculées se fondent et s'associent aux anciennes, et cela principalement à un niveau sémantique. Plus la tâche initiale est complexe, plus l'enregistrement dans la mémoire qui en découle sera riche, détaillé, et précis. L'enregistrement d'un mot n'est pas un phénomène ponctuel et définitif, mis en place une fois pour toute. Il doit être réactualisé pour subsister. Or plus la trace mémorielle est riche et précise, plus elle a de chances d'être retrouvée, réutilisée et, par ce fait même, renforcée (Bogaards, 1994, pp. 93).

D'un autre côté, l'intention d'apprendre ne mène pas forcément au meilleur résultat et que les tâches significatives, c'est-à-dire qui signifient quelque chose pour l'apprenant et où celui-ci est impliqué personnellement, provoquent un apprentissage bien plus efficace.

L'acquisition lexicale est un processus graduel et lent. C'est sous l'effet de la **répétition** et de la **manipulation mentale du vocabulaire** que les associations se mettent en place à des rythmes divers. Pour fixer l'item lexical dans la mémoire, certains chercheurs (Oxford et Crookall, (1990, cités par Selva (1999)) préconisent la révision structurée. Il s'agit de se doter d'un « planning » de révision, sachant qu'un mot nouveau doit être vu 6 à 10 fois avant d'être mémorisé (les mots sont revus dans le temps à intervalles de durée croissante).

Toutefois, l'acquisition lexicale n'est pas uniquement affaire de répétition, même si celle-ci finit toujours par produire un effet. D'autres facteurs entrent en jeu tels que la motivation personnelle et les besoins individuels. En outre, chaque apprenant dispose dans son propre lexique mental de divers réseaux, qui ne sont pas tous structurés et employés de la même manière. La création d'associations ou la constitution de « toiles verbales » est donc une entreprise hautement individualisée. « Chacun construit, au cours de son histoire personnelle et avec ses accents individuels, le lexique qui lui convient, avec les connotations et les images qui sont propres à chaque individu et au contexte socioculturel où il vit » (Bogaards, 1994, pp. 97).

5.4.2 Compréhension de texte et acquisition du lexique

L'un des moyens naturels d'exposition à de nouveaux mots est la lecture et l'écoute. L'inférence du sens des mots à partir du contexte environnant est un des moyens naturels pour apprendre le vocabulaire. A ce titre, la compréhension joue un rôle capital dans l'acquisition du lexique.

Selon Tréville et Duquette (1996), « le processus de la compréhension peut se définir comme l'interaction entre les connaissances antérieures et les connaissances nouvelles. Il y a compréhension quand l'individu peut rendre significatif l'apport langagier (connaissances nouvelles), c'est-à-dire quand il peut établir un lien entre l'acquis récent (vocabulaire et règles lexicales par exemple) et l'acquis déjà ancré dans la mémoire à long terme (mémoire sémantique). »

De ce fait, lors de la confrontation au texte, la compréhension n'est possible que si une certaine proportion d'éléments lexicaux est connue. Le handicap le plus significatif est par conséquent un vocabulaire insuffisant, bien que la compréhension est tributaire d'autres facteurs tels que, pour les textes écrits, reconnaître le type du texte et sa structure, trouver l'idée principale d'un paragraphe, etc.

Des recherches ont été menées afin de déterminer quantitativement la nature du seuil lexical de compréhension. D'après Laufer (1991, cité par Selva (1999)), les apprenants doivent connaître environ 3 000 familles lexicales (c'est-à-dire le vocable lui-même, accompagné de ses dérivés) pour réussir avec le minimum requis (un résultat de 56 %) au test de lecture de leur institution. On obtient ensuite une progression linéaire de 7 % pour chaque millier de familles de mots supplémentaires connues, c'est-à-dire qu'en connaître 4 000 amène un résultat de 63 %, 5 000, 70 %, etc. jusqu'à un niveau où la progression s'estompe peu à peu.

Le choix des textes proposés aux apprenants doit être par conséquent bien ciblé. La fréquence des mots nouveaux doit être minime et les activités lexicales doivent être orientées vers le vocabulaire de base pour permettre aux apprenants débutants de réussir au plus vite la compréhension des textes.

5.4.3 Intérêt des activités lexicales

La présence d'activités lexicales dans l'apprentissage du vocabulaire se justifie par de

meilleurs résultats des apprenants à des tests d'évaluation de compétences lexicales par rapport à d'autres processus d'apprentissage telle l'exposition à de nouveaux mots par la lecture seule de textes.

Paribakht et Wesche (1997) ont mesuré les incidences pour l'apprentissage, d'une part, du processus de lecture seule, et d'autre part, du processus de lecture suivi d'activités lexicales sur 38 jeunes adultes apprenants d'anglais langue étrangère de niveau intermédiaire et ayant des langues maternelles différentes (français, arabe, chinois, etc.). Diverses catégories d'activités ont été proposées aux apprenants :

- **Attention sélective :** activités visant à s'assurer que les étudiants repéraient certains mots-cibles (extraits par exemple d'une liste)
- **Reconnaissance :** activités visant à s'assurer que les étudiants reconnaissaient les mots-cible et leur sens (connaissance partielle des mots-cible)
- **Manipulation :** activités impliquant des connaissances morphologiques sur les mots (dérivés syntaxiques, construction de mots à partir d'affixes et de bases)
- **Interprétation :** activités impliquant l'analyse du sens des mots vis-à-vis du contexte (collocations, synonymes, etc.)
- **Production :** activités demandant aux étudiants de produire des phrases contenant les mots-cible dans des contextes appropriés.

Le temps passé dans les activités lexicales était compensé dans l'autre groupe par un supplément de lecture. Ils ont permis de vérifier les hypothèses suivantes :

- Les étudiants possèdent une meilleure connaissance des mots-cible après la séance de lecture suivie d'exercices mais aussi après la séance de lecture seule (ceci pour vérifier que la lecture était utile à l'apprentissage)
- Pour un temps donné et égal dans les deux cas, les gains en apprentissage étaient plus grands pour la lecture suivie d'exercices que pour la lecture seule.
- Les gains en vocabulaire étaient à la fois quantitatifs (plus de mots connus à la fin) et qualitatifs (meilleure connaissance des mots, mesurée par l'application d'une échelle de connaissance spécifique)
- Les gains dans le cas de la lecture avec exercices concernaient davantage les mots pleins (verbes, noms) que les mots grammaticaux.

L'expérimentation a aussi montré que les étudiants avaient une opinion favorable des activités, pensant que celles-ci amélioraient leurs compétences lexicales.

5.4.4 Intérêt du dictionnaire personnel

Les travaux de Goodfellow (1995) ont montré l'importance d'un module de dictionnaire personnel visant à noter et à organiser le vocabulaire en partie connu. Ce dictionnaire doit permettre à l'apprenant de sélectionner des mots dans un texte, de saisir leurs diverses propriétés linguistiques et sémantiques, de les regrouper suivant des caractéristiques

communes et de visualiser les groupes constitués.

Cette démarche constructiviste est très intéressante de point de vue pédagogique puisqu'elle permet à l'apprenant de bien travailler les unités lexicales et ainsi de mieux les retenir. De plus, les items construits pourraient être utilisés par le système d'apprentissage pour générer des activités lexicales personnalisées.

5.5 Protocole expérimental

Après avoir défini les fondements théoriques de notre démarche et les ressources informatiques disponibles, il fallait articuler le tout autour d'un programme pédagogique pertinent. Pour cela, nous avons mis en place une expérience d'enseignement de l'arabe langue seconde à l'université lumière (Lyon 2)[46], qui s'est déroulée sur deux phases : Une première phase classique d'enseignement dans une classe et une seconde phase autonome qui était assurée par l'environnement informatique d'apprentissage.

Le public des étudiants était constitué d'une dizaine d'étudiants en troisième année universitaire en section Langues Étrangères Appliquées (LEA). Ces étudiants avaient un niveau intermédiaire : Ils disposaient de connaissances grammaticales assez importantes mais non encore assises : ils ont acquis les conjugaisons, les formes dérivées et l'essentiel des règles de syntaxe. L'objectif de ce cours était de leur permettre de retenir un nombre important de familles de mots pendant un semestre, d'une telle manière qu'ils pourront les utiliser lors de leurs productions écrites.

La démarche de l'enseignant en classe, consistait à faire des études de textes bien ciblés. Après la lecture à haute voix du texte, il s'arrête au niveau de chaque phrase, puis au niveau des mots qu'il juge difficiles ou intéressants. Il effectue tout d'abord une analyse morphologique et syntaxique du mot graphique. Si le mot est obtenu à partir d'un processus de dérivation régulier, l'enseignant revient sur ce processus pour expliquer le ou les sens du mot.

Il présente ensuite un schéma organisé autour de la racine du mot, incluant un ensemble d'unités lexicales qui peuvent être associées au mot traité. Les mots doivent être en effet appris avec toutes les informations qui permettent leur réemploi : On peut citer pour les noms les formes au singulier, au pluriel, au féminin, l'adjectif de relation, etc. ; pour les verbes : les formes à l'accompli, à l'inaccompli, le masdar, le « régime » des compléments : avec ou sans préposition structure syntaxique de la complétive, etc. A chacune de ces unités est associé une signification et sa traduction. Les apprenants notent ces informations sous forme de fiches ou chacune d'elles correspond à un mot nouveau et comprend l'ensemble de ces informations et l'analyse de l'exemple traité.

[46] Ce cours a été assuré par le professeur J. Dichy, qui s'est impliqué personnellement et a adapté son contenu pour les besoins de l'expérience. Mr Ammar MEDFAI, qui avait une très longue expérience dans l'enseignement de l'arabe langue seconde et qui poursuivait un DESS en nouvelles technologies éducatives, nous a fortement aidé notamment pour la conception des activités de l'environnement.

Pour pouvoir mémoriser ces mots, les étudiants passent en revue l'ensemble des fiches construites sur des périodes de plus en plus espacées. Les tests de mémorisation, consistent à se faire deviner le sens du mot à partir de sa traduction ou le contraire.

Dans la plupart des cas, cette méthode s'avère efficace, même si elle exige une motivation de la part des apprenants pour mener le processus de mémorisation du lexique jusqu'au bout.

Malheureusement, le volume horaire n'était pas assez suffisant pour pouvoir passer en revue plus que deux ou trois textes par semestre, ce qui correspond à un vocabulaire très limité. D'autre part, la présence du professeur est obligatoire pour pouvoir nuancer les sens des mots afin qu'ils puissent être correctement mémorisés. Enfin, l'absence d'entraînement et d'exécution d'activités lexicales qui favorisent l'inférence par les apprenants ne favorise pas la maîtrise lexicale.

Par conséquent, notre tâche consistait dans la réalisation d'un environnement d'apprentissage qui pourra compléter le travail entamé par le professeur en classe. Cet environnement qui est basé sur les principes pédagogiques cités dans ce chapitre, sera détaillé au fur et à mesure dans les chapitres restants de cette thèse.

5.6 Conclusion

Nous avons souligné dans ce chapitre la position de l'apprentissage du lexique dans l'apprentissage d'une langue seconde. Nous avons ensuite étudié les mécanismes de l'apprentissage lexical à partir des études psycholinguistiques sur la structure du lexique mental. Nous avons vu que le lexique mental semble composé de lexies reliées entre elles par des liens de nature sémantique et contextuelle.

Nous avons examiné par la suite le processus d'apprentissage lexical qui peut se définir comme l'incorporation de nouvelles informations lexicales dans les anciennes. Cette incorporation est fonction du niveau de traitement du vocabulaire, un traitement en profondeur sur le sens des mots favorisant l'apprentissage. Ce dernier n'est pas instantané mais se déroule dans le temps à travers lequel les facteurs répétition et révision du vocabulaire jouent un rôle important. La présence d'activités lexicales et du dictionnaire personnel sont ainsi des moyens qui favorisent la rétention du lexique.

Nous avons enfin présenté la mise en œuvre expérimentale et défini les composantes du système en tenant compte de tous ces principes.

Nous décrirons au fur et à mesure ces différents composants, en commençant par la réalisation du dictionnaire électronique qui inclura les diverses informations citées ci-dessus et dont la clarté des définitions, la simplicité de l'accès (i.e. par l'utilisation de textes étiquetés par des informations de type sémantique) et la mise en valeur par des liens hypermédias des relations entre unités lexicales, permettront à l'apprenant de s'en passer de la présence du professeur et de bien travailler son lexique mental.

Dans le chapitre sept seront présentés les différentes activités qui sont proposées

aux apprenants et qui sont générées automatiquement à partir du dictionnaire. Ces activités qui faciliteront la mémorisation du vocabulaire, sont adaptées au niveau de chaque apprenant.

C'est en fait le modèle de l'apprenant, qui sera décrit dans le chapitre huit, qui permettra à partir de l'enregistrement du comportement de l'apprenant l'individualisation de l'apprentissage.

Dans le chapitre neuf, nous présenterons enfin l'architecture du système d'apprentissage réalisé et nous décrirons notamment le module du dictionnaire personnel qui permettra à chaque apprenant d'organiser le nouveau vocabulaire et qui jouera en quelque sorte le rôle des fiches de définition des mots.

Chapitre 6 Vers un dictionnaire électronique pour apprenant de l'arabe langue seconde

« Un dictionnaire, c'est tout l'univers par ordre alphabétique » Anatole France

6.1 Introduction

Le processus de compréhension de textes arabes présente d'énormes difficultés surtout pour les apprenants débutants. La principale difficulté consiste à repérer des unités de sens dans un flux d'information qui peut sembler assez flou. Le texte est en effet représenté dans une *graphie* d'où sont absentes les voyelles brèves et dont le découpage en « mots » n'apparaît pas toujours évident. En dépit de la ponctuation (souvent défectueuse), il ne se présente pas clairement où commence la phrase et où elle finit.

L'apprenant se jette alors dans le dictionnaire, espérant que la signification d'un mot le mettra sur la voie. Mais du fait de la présence de mots-outils dans les mots graphiques, il est parfois incapable de dégager la « bonne » racine[47], et même s'il le réussit, il se perd

[47] Il est à rappeler que les racines constituent les entrées des dictionnaires arabes.

dans les différentes significations du mot à chercher, ce qui nuit véritablement au processus de lecture. Le dictionnaire papier reste cependant, son seul recours face aux textes incompréhensibles en l'absence du professeur.

Pour remédier à cet état de fait, nous tenterons dans ce chapitre de jeter les fondements d'un dictionnaire électronique adapté aux apprenants de l'arabe langue seconde et qui leur offre un accès simple et rapide aux sens des mots lus dans les textes.

Nous décrirons au début de ce chapitre, la réalisation du dictionnaire *PROLEMAA* (PROtotype de LExique Multilingue A partir de l'Arabe) à laquelle nous avons participé dans le cadre du projet DIINAR-MBC (Voir Annexe 3). Ce prototype ne constitue qu'un protocole d'expérimentation et n'a pas été conçu pour l'environnement d'apprentissage. Néanmoins, nous essayerons sur la base de ce prototype, d'établir un dictionnaire électronique pour apprenant que nous exploiterons dans l'environnement « *AL-Mu*^{c} *aLLiM* ».

Notre démarche, consiste donc à déterminer les informations manquantes au prototype PROLEMAA et éventuellement modifier sa structure pour les intégrer. Nous explorerons pour cet effet les résultats de quelques récentes études sur l'usage des différents types de dictionnaires utilisés par les apprenants : d'abord les dictionnaires classiques sur papier, ensuite les dictionnaires pédagogiques et enfin les dictionnaires électroniques.

6.2 Description du dictionnaire PROLEMAA

PROLEMAA (PROtotype de LExique Multilingue A partir de l'Arabe) est avant tout un prototype[48] qui se base sur le travail déjà accumulé sur la base de données lexicale DIINAR.1. Il ne constitue pas une nouvelle base de données indépendante, mais une extension de DIINAR.1. Ce prototype a été réalisé dans le cadre du projet DIINAR-MBC, et a vu la participation de nos collègues et de nos professeurs de l'IRSIT, de LYON 2 et de l'ENSSIB.

Il est composé de trois parties qui sont la partie arabe – arabe, la partie arabe – français et la partie arabe – anglais. La partie arabe – arabe étant réalisée en amont, sur la base d'un choix d'entrées lexicales (quelques milliers) établi sur un corpus de textes scolaires tunisiens. Le choix de ces derniers a relevé de l'outil que nous avons élaboré pour le calcul des fréquences et qui a fait l'objet d'une présentation (cf. § 4.4).

PROLEMAA arabe - français, et *PROLEMAA arabe – anglais*, arrivent en fin de chaîne, après la préparation du lexique arabe et s'appuient, dans le cadre de cette expérimentation, sur la prise en considération des données bilingues fournies par les ouvrages lexicographiques existants, prises comme autant d'éléments de départ pour la constitution d'un futur dictionnaire multilingue arabe – français / arabe - anglais.

[48] Prototype et non produit fini : il s'agit d'une expérimentation, sur la base de laquelle devrait pouvoir être réalisé par la suite un dictionnaire multilingue performant et complet.

6.2.1 PROLEMAA arabe – arabe[49]

Les unités lexicales arabes traitées dans PROLEMAA (environ 8000 unités) sont constituées de trois listes : une pour les verbes, une seconde pour les noms et une dernière pour les adjectifs. Chaque liste renferme des unités simples formées d'un seul mot, et constitue un sous-ensemble des unités lexicales de DIINAR.1. *PROLEMAA arabe - arabe* reprend, en effet, des informations inhérentes aux unités lexicales antérieurement développées et stockées dans DIINAR.

A-) Traitement des verbes

Il s'agissait d'abord de choisir les informations linguistiques qui vont être associées aux entrées du dictionnaire. En commun accord avec les participants au projet, nous avons décidé d'associer les indications suivantes à chaque entrée verbale du dictionnaire :

- Sa racine consonantique
- Sa forme conjuguée (accompli / inaccompli)
- Son (ses) nom(s) de procès
- Ses autres déverbaux
- La liste des arguments.
- Le ou les sens de chaque verbe. Ces sens ont été examinés, à partir de plusieurs dictionnaires avec une mention particulière pour le dictionnaire de l'Académie arabe (?aL-WaSî**T**)[50], pour établir le **schéma d'arguments** [51] dans chacune des acceptions. Le choix du schéma d'arguments va de pair avec la définition.

Afin de faciliter et d'accélérer la saisie de ces listes d'unités, nous avons défini des tableaux pour effectuer la saisie. Les logiciels de traitement de texte, disposent en effet de nombreuses fonctionnalités qui s'accommodent bien avec ce travail de saisie. Nous avons défini un premier tableau de 9 colonnes (figure 6-1), pour saisir les indications citées ci-dessus. Chaque colonne de ce tableau correspond à une indication et chaque ligne correspond à une entrée du dictionnaire.

49 **Le travail sur le dictionnaire arabe-arabe a été effectué par M. Abdelfattah BRAHAM (Maître de conférences à l'université de la Manouba en Tunisie).**

50 Les définitions puisées dans le dictionnaire (?aL-WaSî**T**) ont été utilisées à titre d'essai. Nous comptons rédiger - ultérieurement - les définitions des unités lexicales retenues avec un "Vocabulaire Arabe Fondamental" en fonction des usages attestés dans un corpus étendu à construire.

51 Le schéma à arguments a été défini dans le cadre de la maintenance du dictionnaire (cf. § 2.5).

Définition	Arg3	Arg2	Arg1	Arg0	Autres déverbaux	Nom de procès	Verbe	Racine

Chaque schéma d'arguments d'un verbe admet au plus quatre arguments : un argument sujet (Argument 0) et trois arguments compléments (Argument 1 à 3). Nous avons défini une syntaxe rigoureuse pour la saisie du sujet et de chaque argument : [préposition + Argument d'ordre 1 [+ (و) + Argument d'ordre 2] … [+ (و) +Argument d'ordre n]]. Le respect de cette syntaxe, nous permet la récupération automatique des informations sur les arguments :

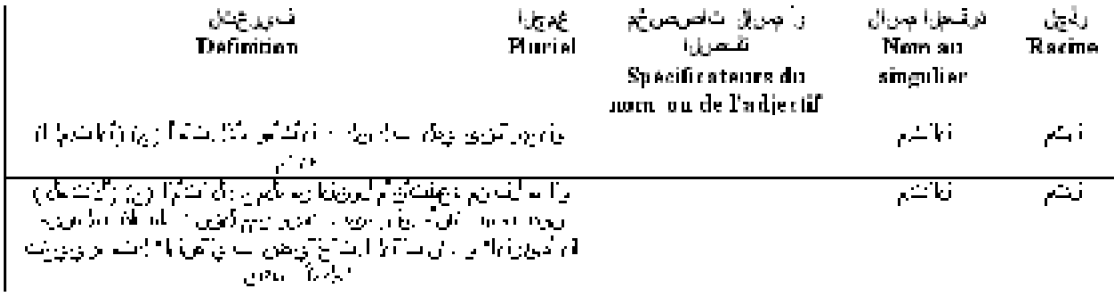
- La catégorie syntaxique parmi : syntagme nominal, syntagme prépositionnel, etc.
- Les propriétés sémantiques : (+ ou - [humain]), (+ ou - [animé]), concret / abstrait.
- L'outil de transition lorsque le régime du verbe est indirect.

Dans plusieurs cas, les informations sur ces schémas d'arguments ne figurent pas dans les dictionnaires arabes, ce qui a amené l'expert linguiste à les déterminer en fonction des sens indiqués et en fonction des utilisations possibles. L'expert a essayé de respecter l'ordre d'occurrence des arguments sur l'axe syntagmatique, tel qu'il apparaît dans le dictionnaire ou dans les usages, et au tant que possible, de distinguer les compléments "obligatoires" des compléments "facultatifs".

B-) Traitement des noms

Les informations concernant les unités lexicales nominales ont été également saisies sous la forme d'un tableau de quatre colonnes (figure 6-2). Nous avons décidé de retenir pour chaque entrée nominale du dictionnaire les indications suivantes :

- Sa racine consonantique
- Sa forme au singulier (ou au pluriel)
- Ses spécificateurs : Des codes sont utilisés pour spécifier les propriétés de l'unité lexicale. Ce spécificateur peut prendre la valeur (صم) pour désigner un nom verbal lexicalisé, la valeur (ج) pour désigner un nom au pluriel et la valeur (ؤم) pour désigner un nom au féminin. Lorsque ce spécificateur n'a pas de valeur, le nom est au singulier masculin.
- Sa forme au pluriel (ou au singulier)
- Sa définition



Définition Definition	Pluriel	Spécificateurs du nom ou de l'adjectif Specificateurs du nom ou de l'adjectif	Nom au singulier Nom au singulier	Racine Racine

D'autres informations sont récupérées automatiquement à partir du dictionnaire DIINAR.1 et sont directement associées à ces unités nominales :

- Les spécificateurs relatifs au nombre et au genre,
- Le spécificateur d'ordre sémantique ayant des implications syntaxiques : trait [± humain]
- Le collectif correspondant,
- L'adjectif de relation, le féminin, le duel et le pluriel régulier,
- Le schéma casuel ou de déclinaison.

C-) Traitement des adjectifs

La syntaxe arabe classique ne distingue pas dans ses "classes des mots" un statut particulier pour les adjectifs. Les "déverbaux" (à l'exception du nom de procès et des noms de lieu et temps) peuvent être considérés, selon les emplois, comme des adjectifs. Les exigences de PROLEMAA arabe- français et arabe - anglais ont nécessité une distinction de ces unités lexicales adjectivales du reste des noms.

Le même tableau de saisie des noms (figure 6-2) a servi pour la définition de ces unités adjectivales. C'est le champ "spécificateurs du nom ou de l'adjectif" qui permet de les distinguer : La valeur prise par ce spécificateur, nous permet de distinguer les adjectifs de relation (سن), les formes ressemblantes (فصف) et les participes actifs ou passifs (شم).

D'autres informations concernant les unités lexicales adjectivales sont récupérées automatiquement de DIINAR et leurs sont directement associées :

- Le chaînage avec le verbe ou le nom dont est dérivé l'adjectif
- Les formes de l'adjectif au singulier, au féminin et au pluriel
- Le schéma casuel ou de déclinaison.

6.2.2 PROLEMAA arabe – français et arabe – anglais

Sur la base du choix des entrées lexicales établi dans PROLEMAA arabe – arabe, nous avons contribué à la réalisation de deux dictionnaires prototypes bilingues *PROLEMAA arabe – français et PROLEMAA arabe – anglais*, qui sont respectivement dans le sens arabe -> français et arabe -> anglais. Il s'agit d'associer à chaque unité lexicale arabe un

ensemble d'équivalents sémantiques dans la langue cible.

Les unités lexicales arabes traitées sont toutes des unités formées d'un seul mot. Les équivalents en langue cible sont susceptibles d'être :

- Des unités lexicales formées d'un seul mot ;
- Des unités lexicales complexes, formées de plusieurs mots ;
- Des paraphrases, c'est-à-dire des syntagmes explicatifs non lexicalisés, lorsqu'il n'existe pas d'équivalent lexicalisé (arabe - français) ou des correspondants approximatifs, des collocations ou des spécificateurs sémantiques (arabe - anglais).

Comme pour les unités lexicales en arabe, nous avons associé à chacun de ces équivalents, un ensemble de paramètres et de données morphologiques, syntaxiques et sémantiques. Ces informations sont incomplètes (à titre d'exemple, le type de conjugaison des verbes français et anglais n'a pas été indiqué) car il s'agit avant tout que chaque équivalent en langue cible soit bien identifié. L'établissement d'une base de données lexicale en français ou en anglais étant pour l'instant hors de nos objectifs..

La recherche des équivalents français et anglais a été effectuée[52] à partir d'outils lexicographiques monolingues et bilingues existants[53].

La saisie des données bilingues a été effectuée sous forme de tableaux. Chaque unité lexicale arabe incluse dans PROLEMAA arabe – arabe se présente dans une ligne indépendante du tableau. Elle est suivie par l'ensemble des équivalents sémantiques dans la langue cible, chacun d'eux étant décrit dans une ligne indépendante. Nous décrirons dans ce qui suit, les formats des différents tableaux de saisie utilisés :

A-) Traitement des verbes

Les indications suivantes ont été retenues pour chaque équivalent **français** d'une unité verbale du dictionnaire (figure 6-3):

1. La forme du verbe à l'infinitif ;
2. Le type sémantique (+ ou - [humain]) du sujet (ou Argument 0) accepté par le verbe ;
3. Le régime (direct ou indirect, précédé d'un mot outil), suivi du type sémantique (+ ou - [humain]) du premier complément (ou Argument 1) accepté le cas échéant par le verbe ;
4. Le régime (direct ou indirect, précédé d'un mot outil), suivi du type sémantique (+ ou - [humain]) du deuxième complément (ou Argument 2) accepté le cas échéant par le

[52] Le travail sur le dictionnaire arabe-français a été effectué par Mr. Xavier LELUBRE (Maître de conférences à l'université LYON 2) tandis que le travail sur le dictionnaire arabe – anglais a été effectué par Melle. Nacira GARBOUT (chercheuse linguistique à l'IRSIT).

[53] Pour les dictionnaires monolingues en priorité le *?aL-Mu c JaM ?aL-WaSî **T*** et *?aL-Mu c JaM ?aL- c aRabî ?aL-?aSâSî* de l'ALECSO, et pour les dictionnaires bilingues le *Dictionary of Written Arabic* de Hans Wehr, mais d'autres dictionnaires ont été également consultés.

verbe;

5. Le régime (direct ou indirect, précédé d'un mot outil), suivi du type sémantique (+ ou -[humain]) du troisième complément (ou Argument 3) accepté le cas échéant par le verbe ;

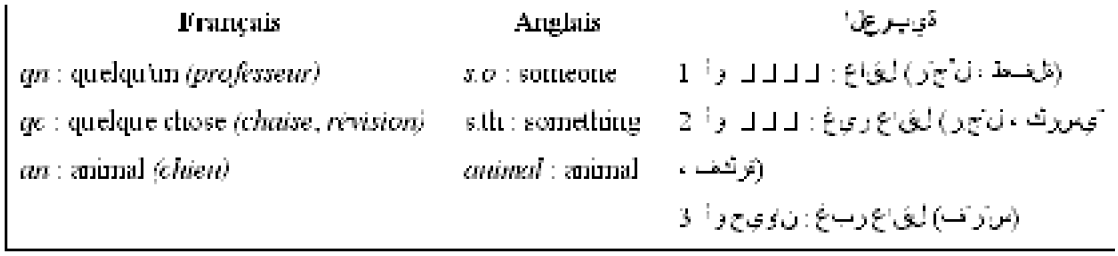
6. L'adjectif éventuellement lié au verbe ;

7. Le nom de procès du verbe, le cas échéant ;

8. Le champ "paraphrase" peut être : soit une paraphrase de l'unité lexicale source s'il n'existe pas d'équivalent lexicalisé en français ; soit un complément d'ordre sémantique, afin de caractériser de ce point de vue l'unité lexicale française, par indication soit du domaine sémantique concerné (noté entre crochets), soit d'un synonyme ou d'un parasynonyme, en cas d'ambiguïté sémantique de cette unité lexicale, soit d'un nom auquel l'adjectif est susceptible d'être associé.

Pour les équivalents **anglais**, nous avons retenu les mêmes indications (figure 6-4). Le champ "paraphrase" désigne dans ce cas : soit un équivalent approximatif, soit des expressions fixes soit des syntagmes nominaux ou verbaux associés au verbe anglais ou à son dérivé adjectival.

Les arguments des verbes correspondent à des schémas syntaxiques et sémantiques et respectent une certaine convention (figure 6-5). Afin de pouvoir automatiser le processus d'injection de ces tableaux dans la base, les valeurs des arguments, tant en français qu'en anglais, sont saisies en respectant cette règle : les mots-outils (prépositions, coordonnants) sont suivis d'une barre oblique (/) lorsqu'ils constituent des compléments indirects.



B-) Traitement des noms

Les indications suivantes ont été retenues pour chaque équivalent **français** d'une unité nominale du dictionnaire (figure 6-6) :

1. La forme de l'unité lexicale si elle existe, au singulier (si elle est lexicalisée au singulier) ;
2. Le spécificateur relatif au genre ;
3. Le spécificateur relatif au nombre : Du point de vue de sa lexicalisation, l'unité lexicale est susceptible soit d'être indifférente au nombre (elle accepte le singulier ou le pluriel) soit d'être lexicalisée à l'une ou à l'autre des valeurs de ce paramètre ;
4. L'adjectif de relation associé, s'il existe ;
5. La forme au pluriel, si celui-ci existe ;
6. Le champ "paraphrase", qui peut être : soit une paraphrase de l'unité lexicale source s'il n'existe pas d'équivalent lexicalisé en français, soit un complément d'ordre sémantique. Ce dernier permet de caractériser l'unité lexicale française, par indication soit du domaine sémantique concerné (noté entre crochets), soit d'un synonyme ou d'un parasynonyme, soit encore de l'emploi au propre ou au figuré, en cas d'ambiguïté sémantique de cette unité lexicale.

Pour les équivalents **anglais**, nous avons retenu les mêmes indications (figure 6-7).

Les valeurs des spécificateurs anglais du genre et du nombre sont différentes de celles des spécificateurs français : une unité lexicale en anglais peut être un nom comptable ou un nom non dénombrable ou les deux à la fois, ou encore un nom au pluriel et elle peut avoir un genre neutre. Le tableau de la figure (6-8) illustre toutes les valeurs utilisées dans PROLEMAA pour ces spécificateurs dans les trois langues. De manière générale la valeur 0 d'un paramètre est réservée à ce qui correspond à la valeur "libre", selon contexte, et non pas à une valeur bloquée a priori.

Genre ([illegible])	Nombre ([illegible])	Humain / non humain ([illegible])
0 : masculin ou féminin (selon le sexe) 1 : masculin 2 : féminin 3 : neutre 4 : masculin ou féminin	0 : libre, selon contexte 1 : lexicalisé au singulier 2 : lexicalisé au pluriel 3 : lexicalisé au duel	1: humain 2: non humain
طفل ، [illegible] 1 [illegible] 2 [illegible] 4 0 *élève* 1 *homme*, [illegible] 2 *femme* 4 [illegible]	طفل 0 [illegible] 1 [illegible] 2 [illegible] 3 0 *homme élève* 1 *informatique* 2 *ciseaux*	[illegible] 1 [illegible] 2 [illegible]
0 [illegible] 1 [illegible] 2 [illegible] 3 table (it : neutre)	0 [illegible] 1 [illegible] 2 [illegible]	

C-) Traitement des adjectifs

Les indications suivantes ont été retenues pour chaque équivalent **français** d'une unité adjectivale du dictionnaire (figure 6-9) :

1. La forme de l'adjectif au singulier ;
2. Le chaînage avec le verbe ou le nom dont est dérivé l'adjectif ;
3. Le type d'accord en genre et en nombre auquel est soumis l'adjectif (figure 6-10) ;
4. La forme au féminin, si celui-ci existe ;
5. La forme au masculin pluriel, si celui-ci existe ;
6. Le champ "paraphrase", qui peut être : soit une paraphrase de l'unité lexicale source s'il n'existe pas d'équivalent lexicalisé en français ; soit un complément d'ordre sémantique, afin de caractériser de ce point de vue l'unité lexicale française, par

indication soit du domaine sémantique concerné (noté entre crochets), soit d'un synonyme ou d'un parasynonyme, en cas d'ambiguïté sémantique de cette unité lexicale, soit d'un nom auquel l'adjectif est susceptible d'être associé.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
équivalent français économique	origine de la dérivation	accord avec le nom 0	féminin e	pluriel masculin -	paraphrase	
social		0	e	s		
[illegible]		1	e			

Pour les équivalents **anglais** des adjectifs, seulement trois indications ont été retenues (figure 6-10) :

La forme de l'adjectif ; 1.

Le chaînage avec le verbe à l’infinitif ou le nom au singulier dont est dérivé l'adjectif ; 2.

Le champ "paraphrase", qui peut être soit un équivalent approximatif, soit des traits sémantiques ou encore une collocation relative à l’unité lexicale de base. 3.

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
anglais équivalent automatic instrumental mechanical	base instrument machine	paraphrase [illegible] [illegible] [illegible]				

Le tableau de la figure (6-11) illustre les valeurs des spécificateurs relatifs aux adjectifs utilisés dans PROLEMAA.

Accord	Type de l'adjectif [illegible]	Origine de dérivation
0 : selon contexte : grand, nécessaire 1 : toujours au masculin singulier : marron, bleu clair 2 : toujours au féminin (singulier ou pluriel, selon contexte) : enceinte 3 : hors accord, l'unité lexicale adjectivale équivalente étant un syntagme adjectival : de bonne humeur	[illegible] 1 [illegible] 2 [illegible] 3 [illegible] 4 [illegible] 5 [illegible] 6 [illegible] 7	Si le nom associé est un nom de remplacement, il est mis entre parenthèses. villageois <-- village urbain <-- (ville)

6.2.3 Modélisation informatique du dictionnaire

Conçues comme des extensions de DIINAR.1, les différentes listes de PROLEMAA doivent être intégrées dans les différentes parties de ce dictionnaire. Nous avons par conséquent créé de nouvelles entités et apporté quelques modifications à DIINAR.1, afin d'incorporer ces nouvelles informations monolingues et multilingues. Dans cette section, nous décrirons successivement la nouvelle organisation de la partie verbale, nominale et adjectivale de la base de données.

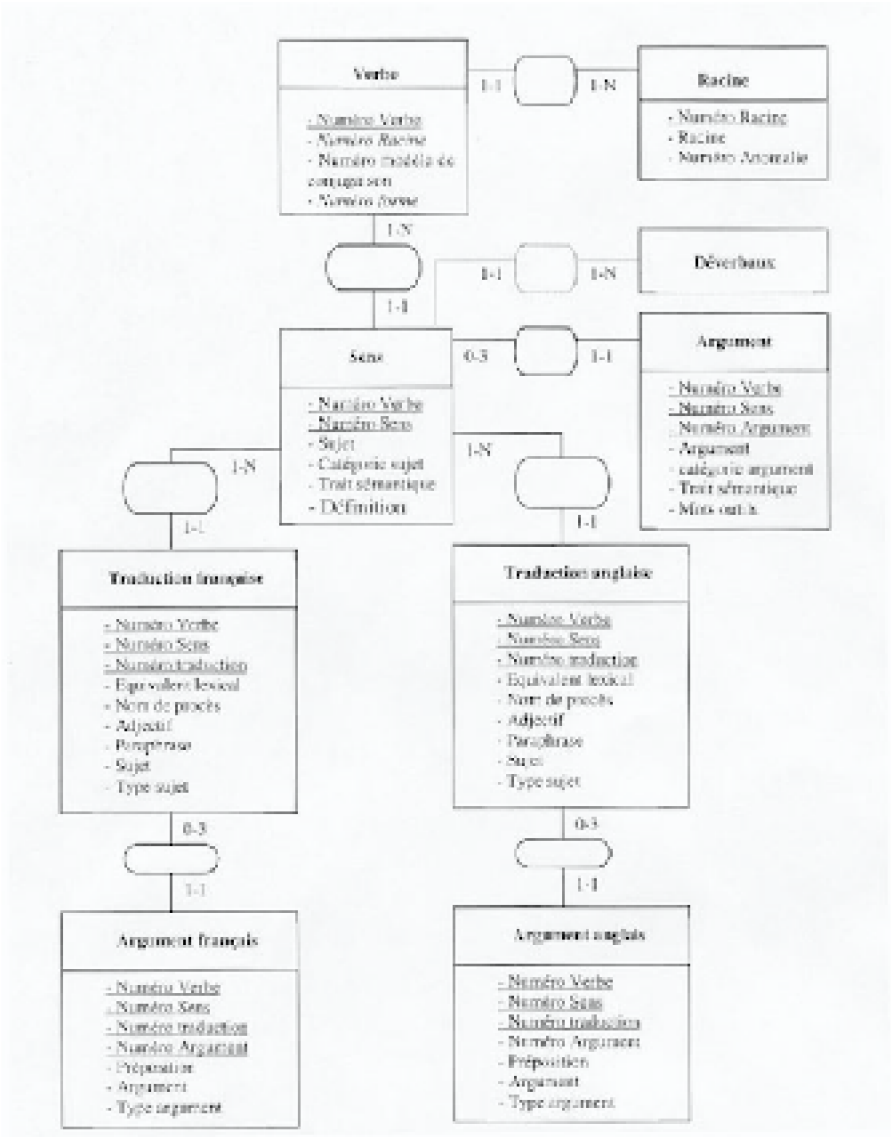
A-) Modélisation des unités verbales

Dans cette nouvelle modélisation de la base de données verbale, nous avons gardé toutes les informations de DIINAR.1 dont notamment celles qui permettent la génération des conjugaisons des verbes et des déverbaux qui sont représentés par des pointillés sur la figure (6-12) ci-dessous.

Par contre, nous avons ajouté quelques nouvelles entités pour gérer les significations des verbes, leurs schémas syntaxiques et leurs équivalents en français et en anglais. L'entité "Sens" renferme désormais la ou les significations de chaque verbe. Cette entité inclut aussi les informations relatives au sujet du schéma syntaxique du verbe : Sa catégorie et ses propriétés sémantiques : ([± humain], [± animé], concret / abstrait). Les informations sur les arguments compléments sont définies dans une entité indépendante "Argument" afin de minimiser la taille de la base. Cette séparation permet en effet, d'éviter que les champs d'arguments soient vides lorsque le verbe est intransitif ou n'admet pas trois d'arguments.

L'entité "Argument" inclut les informations suivantes : La catégorie de l'argument, ses propriétés sémantiques : ([± humain], [± animé], concret / abstrait) et un vecteur "Mot outil" correspondant à la préposition lorsque le complément est indirect.

Chacune des traductions en français et en anglais est introduite dans deux entités : Une entité "Traduction" et une entité "Argument". La première entité permet d'enregistrer l'équivalent lexical, les informations sur le sujet, le nom de procès, l'adjectif du verbe et la paraphrase qui remplace l'équivalent lexical. La seconde entité inclut les informations sur les différents arguments du schéma syntaxique de l'équivalent du verbe.



B-) Modélisation des unités nominales

Dans la partie nominale de la base de données, nous avons lié les noms verbaux lexicalisés à leurs verbes d'origine. Ce lien permettra, dans le cadre d'un système de navigation dans le dictionnaire, de passer de l'unité nominale à l'unité verbale origine de dérivation du nom.

La principale entité de cette base est l'entité "Nom" qui permet de définir les trois traits linguistiques retenus dans la phase de conception : le trait humain, le genre et le nombre. Les sens de chaque unité nominale sont introduits dans l'entité "Sens du nom".

Sur le schéma du MLD (figure 6-13), la cardinalité (0-N) du côté de l'entité "nom" dans la relation entre cette dernière et l'entité "Sens du nom" s'explique par le fait que certains noms n'ont pas reçu de signification lors de la saisie, soit qu'ils sont des pluriels de noms qui ont déjà reçu leur(s) signification(s), soit qu'ils sont des noms verbaux lexicalisés et leurs significations sont déduites à partir de leur verbe origine.

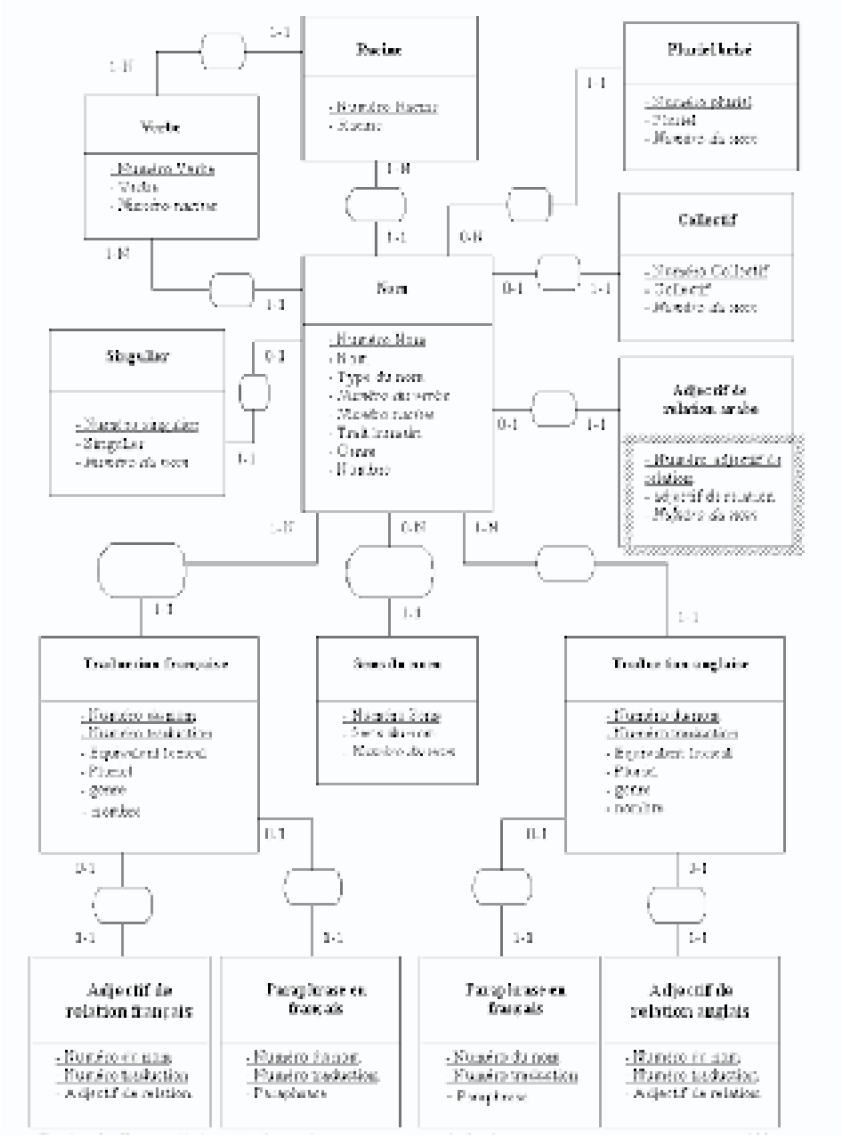
La figure (6-13), montre bien que les liens de l'entité "nom" avec ses formes dérivées (le singulier, le pluriel brisé, le collectif, l'adjectif nom de relation) sont bien conservés.

Pour introduire l'équivalent en français ou en anglais, nous avons créé une entité "Traduction", qui regroupe les informations définis dans les tableaux de saisie : l'équivalent lexical, le pluriel, le genre et le nombre. Dans un souci de minimiser la taille de la base, nous avons placé les paraphrases et les adjectifs de relation, qui ne sont pas automatiquement introduites avec l'équivalent lexical, dans deux autres entités indépendantes.

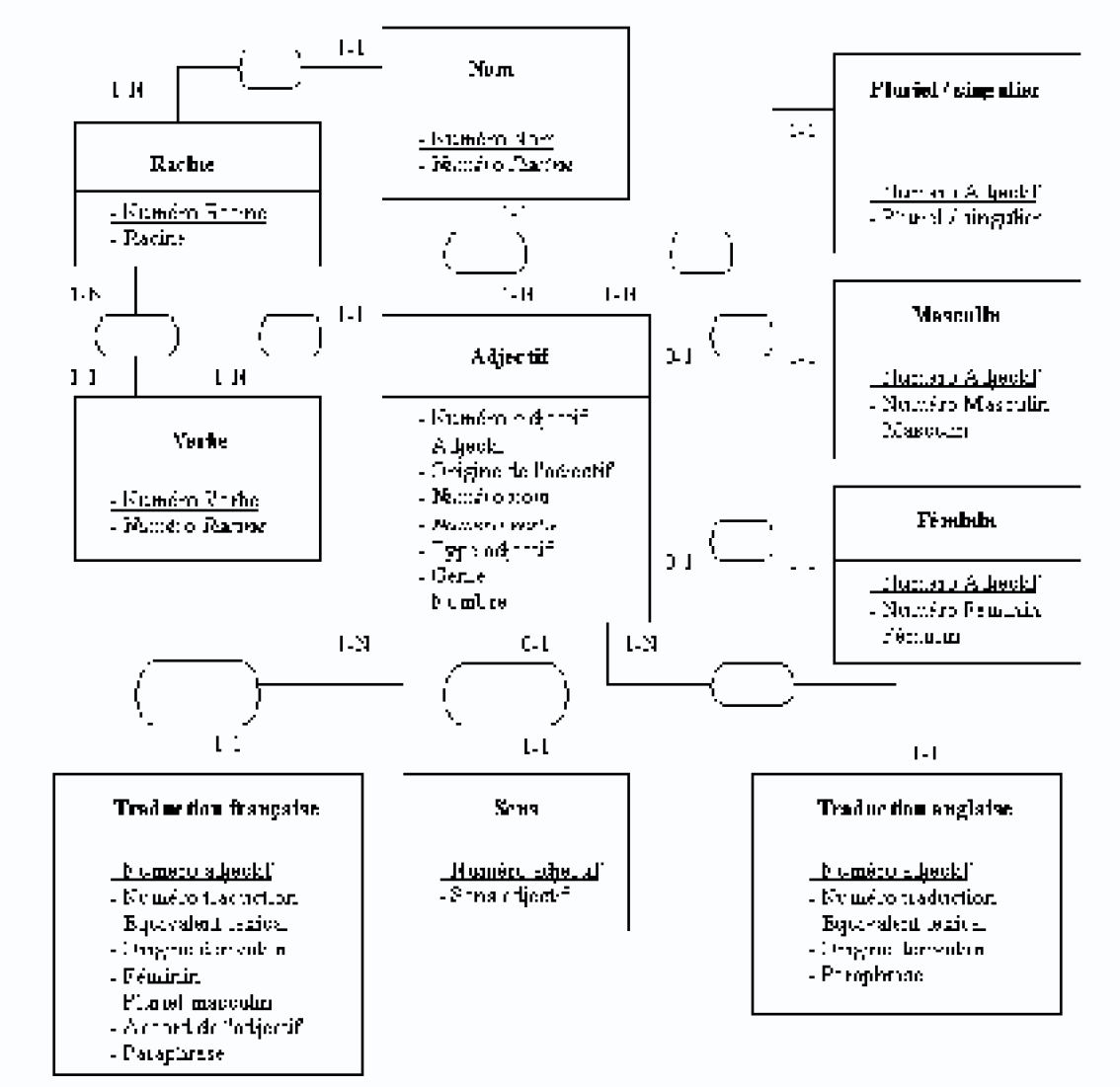
C-) Modélisation des unités adjectivales

Comme nous l'avons déjà fait remarquer, les adjectifs en arabe sont des noms à valeur qualitative souvent utilisés comme des substantifs. Dans DIINAR.1, ils n'ont pas été distingués des autres éléments du dictionnaire. Ils étaient soit des déverbaux et faisaient partie des unités verbales, soit des substantifs et faisaient partie des unités nominales.

Les exigences de la base de données multilingue ont nécessité la distinction de ces unités adjectivales des autres unités. Pour répondre à ces exigences, nous avons créé une nouvelle entité "Adjectif" dans la base de données pour gérer ces unités tout en gardant leurs liens avec les unités nominales et verbales, afin notamment de pouvoir présenter dans le dictionnaire le sens de l'adjectif. La signification de l'adjectif est en règle générale déduite à partir du sens du verbe qui lui est associé sauf dans de rares cas.



L'entité "adjectif" renferme par conséquent, l'identifiant du verbe ou du nom qui nous permettent d'accéder aux autres parties de la base et de récupérer toutes les informations sur l'origine de l'adjectif. Cette entité contient aussi les indications sur le genre, le nombre et la classe à laquelle appartient l'adjectif parmi ces huit classes : nom adjectif (صفة مشبهة), : participe actif (اسم فاعل), participe passif (اسم مفعول), intensif et nom de métier (صفة المبالغة), couleur et difformité (صفة للون), élatif (أفعل التفضيل), adjectif-nom de relation (نسبة) et substantif (إسم). Cette entité est enfin en relation avec les entités "Masculin", "féminin", "singulier" et "pluriel", qui permettent d'avoir les différentes formes dérivées de l'adjectif.



Nous avons prévu aussi une entité "Sens" pour enregistrer toutes les informations jugées utiles par l'expert linguiste et qui permettant une meilleure description de l'adjectif (colonne *définition* dans le tableau de saisie). Pour les équivalents français et anglais nous avons créé respectivement deux entités "Traduction française" et "Traduction anglaise", qui regroupent les informations définies dans les tableaux de saisie (figure 6-14).

6.2.4 Injection des données dans la base

Parallèlement à la saisie des données du dictionnaire PROLEMAA dans les tableaux, nous avons procédé à l'élaboration d'un programme d'injection automatique de cette grande variété d'informations dans la base de données. Cette méthode a l'avantage d'alléger la rude tâche de saisie entre plusieurs personnes et d'accélérer le travail de saisie.

Nous avons ainsi réalisé un programme d'injection automatique qui parcourt les différentes lignes des tableaux en repérant les unités lexicales correspondantes dans DIINAR.1, en gérant les erreurs de saisie et en synchronisant l'injection. En effet, l'introduction des données monolingues doit se réaliser en amont et l'injection des unités nominales doit précéder celle des unités verbales (les informations linguistiques associées aux arguments des schémas syntaxiques des verbes sont récupérées automatiquement à partir de la partie nominale de la base). A l'aide de ce programme, nous avons réussi à injecter environ 10.000 unités lexicales, dont plus que la moitié de ces unités ont reçu leurs équivalents en français et en anglais.

Cette démarche devrait être poursuivie pour l'achèvement du dictionnaire. Quelques problèmes liés aux formats des tableaux doivent être cependant préalablement résolus afin que ce processus réussisse à 100%. Nous reviendrons dans cette section sur ces difficultés et sur d'autres aspects intéressants de ce programme.

A-) Injection des unités lexicales arabes

Comme nous l'avons déjà décrit, les unités monolingues du dictionnaire PROLEMAA ont été saisies dans deux tableaux différents : un premier tableau pour les unités verbales et un second pour les unités nominales et adjectivales. L'injection de ce dernier tableau doit précéder le premier, puisque les unités nominales constituent les sujets et les arguments des schémas syntaxiques associés aux unités verbales.

Le tableau de saisie des noms et des adjectifs comprend quatre colonnes (figure 6-2) et respecte certaines conventions. C'est la valeur prise par le spécificateur de la troisième colonne qui nous permet de distinguer les propriétés du nom ou de l'adjectif en cours :

- Lorsque cette valeur est "ردصم", l'unité lexicale est un **nom verbal lexicalisé**, qui figure dans la partie verbale de DIINAR. Le verbe associé est identifié dans DIINAR.1 à partir des indications sur sa racine (première colonne du tableau) et la relation "Verbe-Nom verbal" dans la base. Les informations relatives au verbe sont ainsi récupérées et associées aux nouvelles informations sémantiques.
- Lorsque le spécificateur n'a pas de valeur, l'unité lexicale est un **substantif**. Les informations morphologiques relatives à cette unité et ses relations avec les autres unités lexicales sont directement récupérées à partir de la partie nominale de DIINAR.1 et associées aux nouvelles informations sémantiques. Deux cas particuliers peuvent aussi se présenter : Lorsque la valeur du spécificateur est "ثنؤم", il s'agit d'un nom féminin et lorsque cette valeur est "عمج", il s'agit d'un nom pluriel. Dans ces deux cas, la signification de l'unité est déduite à partir de la signification du nom masculin singulier qui est préalablement injecté.
- Lorsque cette valeur est "قبسن", l'unité lexicale est un **adjectif de relation**. Les informations linguistiques relatives à cette unité sont directement récupérées à partir de la partie nominale de DIINAR.1 et associées aux nouvelles informations sémantiques.
- Lorsque cette valeur est "ققص", l'unité lexicale est une **forme ressemblante** (ققص ةهّبشم), qui figure dans la partie verbale de DIINAR. Son verbe est identifié à partir

des informations sur sa racine et la relation "Verbe - Forme ressemblante" dans DIINAR.1. Les informations sur son verbe sont ainsi récupérées et associées à sa signification.

- Lorsque le spécificateur est "مشتق", l'unité lexicale est un participe actif (اسم فاعل) ou un participe passif (اسم مفعول), qui figure dans la partie verbale de DIINAR. Son verbe est identifié à partir des informations sur sa racine et les relations "Verbe – Participe actif" et "Verbe – Participe passif"dans DIINAR. L'information sur l'identité du verbe est ainsi récupérée et associée à sa signification.

Malheureusement, le programme n'a pas réussi à identifier une partie de ces unités dans DIINAR.1. Les unités non reconnues étaient mal orthographiées ou ne figuraient pas dans DIINAR.1 (les adjectifs de l'intensif (أفعال المبالغة) par exemple n'ont pas été traités dans DIINAR.1). Ces unités ont été, saisies en aval à l'aide de l'interface de saisie et de mise à jour des données de PROLEMAA.

Le tableau de saisie des unités verbales comprend quant à lui neuf colonnes (figure 6-1), correspondant aux informations permettant d'identifier l'unité lexicale verbale en cours : la racine, la forme conjuguée, la liste des noms verbaux, la liste des autres déverbaux, le sujet et les arguments et la signification. C'est la signification du verbe qui constitue désormais l'unité lexicale en entrée et non plus le verbe avec l'ensemble de ses significations comme c'était le cas dans DIINAR.1.

Le programme d'injection automatique doit repérer le verbe dans la base, à partir des indications sur sa racine et sa forme conjuguée, et associer les déverbaux directement aux significations correspondantes du verbe. Il doit ensuite identifier le sujet et les arguments du schéma syntaxique du verbe à partir de la partie nominale de la base. Il doit enfin associer la signification du verbe à l'unité en entrée.

Dans les tableaux de saisie, la syntaxe avec laquelle étaient définis les arguments n'était pas toujours respectée[54]. Cette syntaxe qui était au départ du projet insuffisamment définie, nous a aussi posé des difficultés pour dégager les propriétés de quelques arguments (i.e. les arguments indirects qui sont précédés par les prépositions (بِ / لِ) ne peuvent pas être distingués des arguments simples commençant par les mêmes caractères). Nous étions obligés à chaque fois d'ôter ces prépositions et de vérifier si l'argument correspondait à un argument valide dans la base de données nominale ou pas. Le même problème a été rencontré avec le coordonnant (و : *Wa*) où pour les arguments composés de plusieurs éléments, nous ne pouvions pas différencier un argument précédé par ce coordinateur et un argument composé dont le second membre commence par un (و : *Wa*).

Ce genre de problème a été évité par la suite pour les derniers tableaux de saisie. Nous avons défini une syntaxe qui exige le placement d'une barre oblique (/) entre les mots-outils et les arguments indirects et une virgule (,) entre les éléments des arguments composés.

[54] Dans un schéma qui devait respecter cette syntaxe : [préposition + Argument d'ordre 1 [+ (و) + Argument d'ordre 2] … [+ (و) +Argument d'ordre n]], nous trouvons parfois des (وٓ) *?aW* à la place de (وَ) *Wa* pour séparer les différentes classes de l'argument.

B-) Injection des unités lexicales multilingues

La partie multilingue de PROLEMAA comprend six tableaux : trois pour chaque langue contenant les correspondants des unités verbales, nominales et adjectivales. Tous ces tableaux ont été saisis de la même façon : une ligne pour la signification arabe qui est suivie par une ou plusieurs lignes correspondant à leurs traductions en langue cible.

Le programme devait d'abord, repérer la signification de l'unité lexicale arabe (i.e. entrée du dictionnaire) dans la base et lui associer ensuite ses traductions. Puisque les saisies des données monolingues et multilingues se sont déroulées en même temps, quelques informations manquaient aux tableaux multilingues (i.e. les significations monolingues). Pour les unités verbales, nous nous sommes appuyés sur les informations relatives aux arguments des schémas syntaxiques pour déterminer les significations des verbes. Par contre, pour les noms et des adjectifs nous nous sommes contentés de l'introduction des correspondants des unités qui présentaient une seule signification. Les autres unités ont été entrées manuellement dans le dictionnaire.

C-) Conclusion et perspectives

Cette expérience nous a permis d'initier une méthode qui permet de faire participer plusieurs personnes, chacun selon ses compétences, à l'élaboration du dictionnaire. Toutes les incohérences de saisie seront détectées ultérieurement par le programme et traitées manuellement en aval.

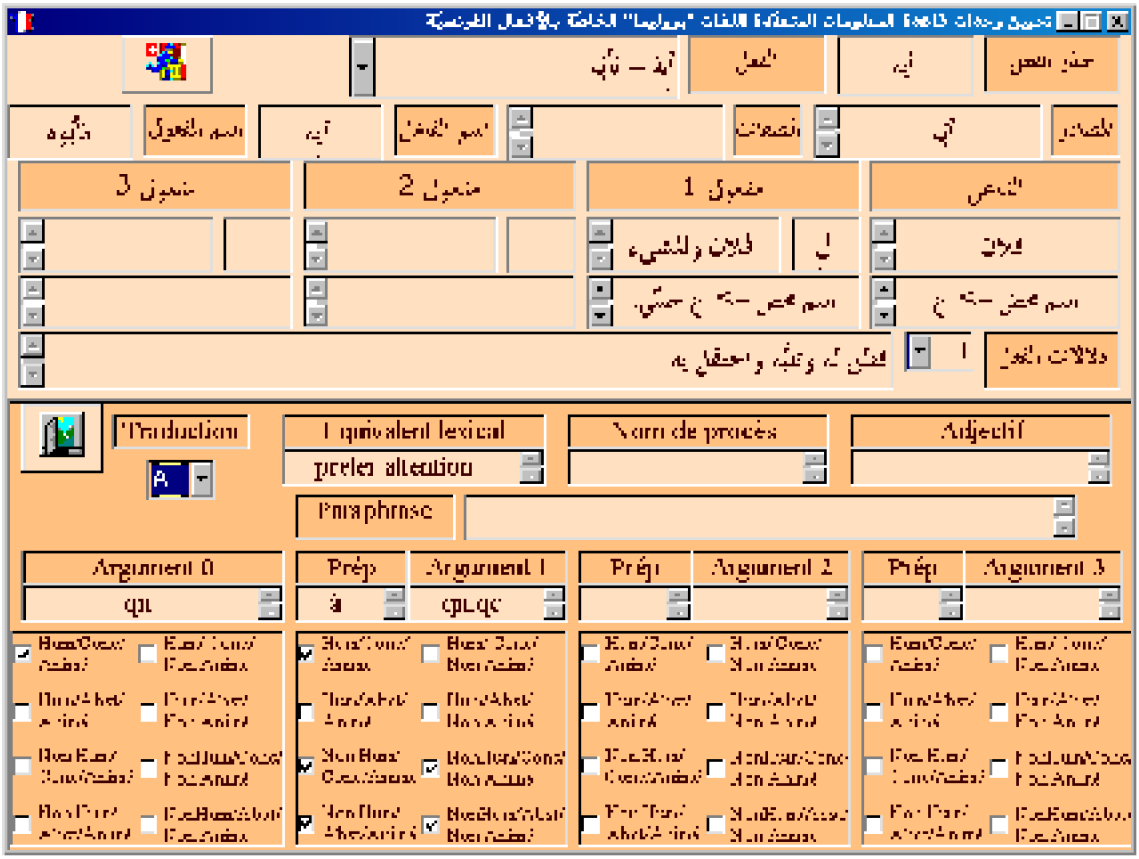
Cette méthode a l'avantage de pouvoir régénérer indéfiniment le dictionnaire. On pourrait par exemple revenir sur les significations des unités lexicales pour uniformiser le vocabulaire utilisé ou pour les redéfinir à partir d'un corpus textuel donné. On pourrait aussi, à partir de la même masse d'informations saisies, générer de nouveaux dictionnaires à thèmes par la sélection de sous listes spécifiques.

L'élaboration d'un dictionnaire comme PROLEMAA est par conséquent une opération complexe et de longue haleine et sollicite de nos jours l'aide de la machine par la réalisation d'outils informatiques comme le programme d'injection automatique qui facilitent sa maintenance.

6.2.5 Présentation des interfaces de consultation et de mise à jour

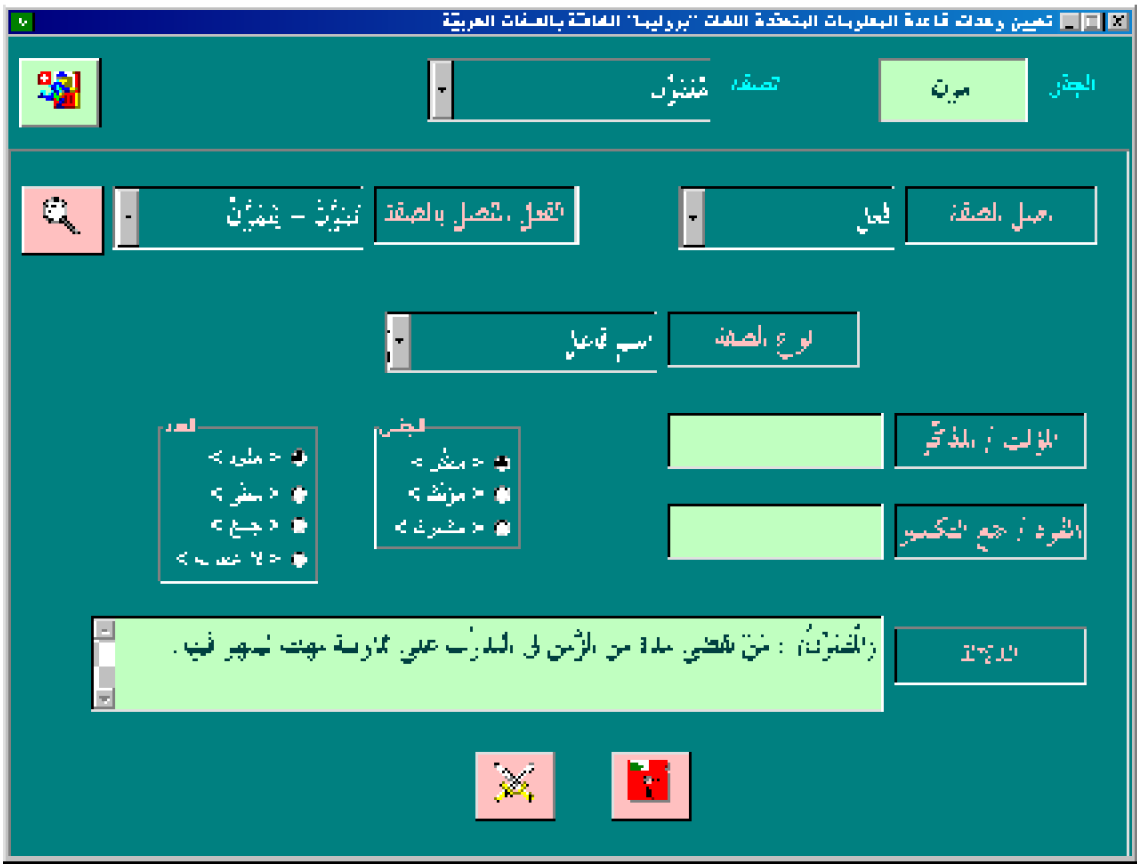
Pour la manipulation de PROLEMMA, nous avons procédé à la construction de nouvelles interfaces ergonomiques, qui évitent au maximum à l'utilisateur d'avoir à entrer du texte, avec les erreurs de frappe que cela peut entraîner. Ces interfaces vont servir à la correction des erreurs dues à la procédure d'injection automatique des données et surtout à la navigation de l'apprenant dans le dictionnaire dans le cadre de l'environnement d'apprentissage.

Comme PROLEMAA est une extension de DIINAR.1, nous avons conçu des interfaces, qui mettent en valeur les liens entre les données ajoutées (i.e. notamment celles de nature sémantique) et celles déjà existantes.



La figure (6-15) montre par exemple, l'écran de saisie et de consultation des unités lexicales verbales multilingues, où nous avons réussi à intégrer sur un même écran les informations relatives aux correspondants en langue cible avec celles de l'unité lexicale en arabe. Cette présentation est très utile aux utilisateurs du dictionnaire (i.e. les apprenants) puisqu'elle leur permet de mieux distinguer les informations pertinentes tant en compréhension de texte qu'en production.

Nous avons d'autre part, conçu pour la première fois une interface pour la consultation et la mise à jour des adjectifs arabes (figure 6–16). La catégorie des adjectifs n'était pas traitée dans DIINAR.1, comme une catégorie distincte des noms, conformément à la tradition grammaticale arabe. Cette interface permet de spécifier l'origine et la catégorie de l'adjectif et surtout d'accéder directement au sens de l'unité lexicale verbale lorsque l'adjectif est dérivé à partir d'un verbe (symbole « loupe » dans l'interface ci-dessous). Le sens de l'adjectif en arabe étant dans la majorité des cas déduit à partir de l'unité lexicale source.



6.3 L'usage du dictionnaire papier

Les études sur l'utilisation du dictionnaire papier (Bogaards, 1988) montrent qu'une très forte proportion déclare se servir du dictionnaire au moins une fois par semaine. Il faut cependant faire la différence entre le dictionnaire monolingue et le dictionnaire bilingue. Le second est le plus utilisé (au moins une fois par semaine pour 97 % contre 60 % pour le premier) mais la fréquence diminue avec le niveau de l'apprenant : plus celui-ci s'élève et plus l'emploi du monolingue augmente.

Malgré une utilisation importante du dictionnaire papier par les apprenants, notamment au cours de la lecture d'un texte dans une langue étrangère ou lors d'une traduction, son utilité n'est pas toujours évidente. Selon Bogaards (1995), plusieurs

expériences ont montré que le dictionnaire ne semblait pas améliorer la compréhension des textes d'une manière significative. Pour cela, il avance plusieurs raisons :

- Les apprenants n'aiment pas utiliser un dictionnaire. Ils le considèrent comme une étape obligée et contraignante qui les détourne de leur lecture.
- Ils ne savent pas utiliser un dictionnaire. Ils ont des difficultés à repérer l'information pertinente et acceptent la moindre indication qui va dans le sens de leur hypothèse initiale de manière à abréger l'épreuve. De plus, pour les monolingues, ils sont souvent dans l'obligation d'aller consulter d'autres entrées pour comprendre la première, soit par référence explicite, soit parce que la première définition contient des mots peu ou mal connus. Ils ont alors toutes les chances de perdre le fil du texte.
- Le dictionnaire nuit au processus de lecture : des expériences montrent que des étudiants utilisant un dictionnaire mettaient souvent plus de temps à terminer leur tâche, sans pour autant obtenir de meilleurs résultats. Plus un apprenant met de temps à chercher une information, moins il a de chance de la trouver.

Face à ce constat, Bogaards en déduit qu'il faut, d'une part, avoir un niveau de connaissance avancé sur la langue pour pouvoir profiter des informations contenues dans les dictionnaires, et d'autre part, avoir une bonne dose de ténacité et de courage.

6.3.1 Nature des informations recherchées

Dans quelles situations les apprenants utilisent le dictionnaire ? D'après Béjoint (1981, cité par (Selva, 1999)) il est plus utilisé pour une tâche écrite que pour une tâche orale, et que l'encodage prime sur le décodage. Cette préférence du dictionnaire pour les tâches écrites semble normale vu qu'il est lui-même sous forme écrite et que sa consultation demande un certain temps, temps dont on dispose rarement dans une conversation ou une écoute, au contraire d'une lecture ou d'une traduction.

Que cherche-t-on dans le dictionnaire ? D'après Bogaards (1988), c'est le sens et les définitions des mots qui sont les premières préoccupations. Puis viennent les synonymes et l'orthographe. Ensuite les informations grammaticales et l'emploi des mots en contexte. Suit la prononciation. Ces trois derniers éléments sont surtout demandés par les apprenants étrangers, moins par les locuteurs natifs. Enfin, l'étymologie, les niveaux de langues ou les homonymes ne sont presque jamais demandés. Peut-être l'intérêt de ces informations n'a-t-il pas suffisamment sauté aux yeux des apprenants.

Atkins et Varantola (1997, cités par (Selva, 1999)) ont étudié la façon dont est utilisé le dictionnaire pour une aide en traduction. Comme on pouvait s'y attendre, la traduction d'un mot que ce soit dans le sens L1-L2 ou L2-L1 (43 % et 59 % des consultations) est de loin la première information recherchée. La vérification du mot que les utilisateurs pressentaient comme traduction correcte vient en seconde position avec un tiers des consultations. Loin derrière viennent les informations concernant les collocations (11 % et 4 %) et encore plus loin, avec 4 %, les informations grammaticales. Les autres types d'informations représentent 5 %. Si l'on prend en compte le niveau des utilisateurs, on constate que plus le niveau est faible et plus le dictionnaire est utilisé pour trouver une

traduction, tandis que la proportion de vérification baisse.

6.3.2 Bilingue ou monolingue

Un point qui semble sûr, c'est que la majorité des étudiants préfèrent le bilingue au monolingue, même si ce dernier semble beaucoup plus bénéfique (Selva, 99, pp. 50). Ce fait peut s'expliquer par le fait que les monolingues utilisés ne sont pas des dictionnaires pour apprenant (cf. § 6.4) : Les définitions dans ces dictionnaires sont véritablement rédigées pour des natifs et aucun effort n'étant entrepris pour simplifier le vocabulaire définitoire. Cependant, les études montrent que la proportion d'utilisation du monolingue était plus importante au fur et à mesure que le niveau augmentait. Il faut donc avoir un certain niveau de langue pour tirer profit des informations du monolingue.

Il existe cependant un troisième type de dictionnaire, le **semi-bilingue**, sur lequel très peu d'études ont été menées à ce jour et qui est un terrain ouvert à l'investigation car les résultats obtenus par ce type de dictionnaire sont prometteurs. Les dictionnaires semi-bilingues sont un mélange des deux précédents dans le sens où, à la suite de la définition monolingue pour apprenant en langue seconde, se trouve la traduction du sens de l'unité lexicale considérée. L'utilisateur est donc assuré d'avoir compris la définition en langue étrangère, ce qui est confortable, à la condition toutefois de la lire et de faire l'effort de la comprendre, la tentation étant grande de ne regarder que la traduction.

Dans une étude menée Laufer et Hadar (1997, cités par (Selva, 1999)), qui avait pour but d'évaluer les performances des trois types de dictionnaires, monolingue pour apprenants, bilingue et semi-bilingue, les meilleurs résultats, autant en compréhension qu'en production sont obtenus avec le dictionnaire semi-bilingue. Ce dictionnaire rassemblant les deux types d'informations contenus dans les autres dictionnaires, les utilisateurs sont à même, d'une part de trouver à chaque fois les renseignements nécessaires, et d'autre part, de pouvoir les exploiter.

6.3.3 Conclusion

Il ressort de ces études que le dictionnaire papier est un outil très utile, quelle que soit sa nature et quel que soit le niveau de connaissances de l'apprenant. Cependant, chaque type de dictionnaire présente des forces et des faiblesses que nous récapitulons en ces points :

- Le monolingue nuit parfois à la compréhension : Comme toutes les informations sont données dans la langue étrangère, il y a parfois des problèmes de compréhension. C'est le cas avec les définitions compliquées ou qui bouclent. Même lorsque le vocabulaire est contrôlé et simplifié, comme c'est le cas dans les dictionnaires pour apprenants, il n'est pas toujours facile de saisir le sens précis des mots. Le problème ne vient toutefois pas toujours des dictionnaires : certains mots sont en eux-mêmes difficiles à comprendre ou à conceptualiser.
- Le monolingue ne permet pas l'accès aux mots inconnus. Comment, dans les tâches

productives, un apprenant peut-il trouver le mot qu'il lui faut mais qu'il ne connaît pas ?

- Le point fort du monolingue est le grand nombre d'informations différentes disponibles « authentiques » car exprimées dans la langue cible. Il est ainsi possible de voir le comportement réel des unités lexicales dans les définitions et les exemples ainsi que la nature des actants, qui sont reliés aussi bien syntaxiquement que sémantiquement et qui sont, de ce fait, nécessaires pour la maîtrise de l'unité lexicale.
- Le point fort du bilingue est la compréhension, mais il est inadéquat quand l'unité lexicale est très polysémique ou quand l'apprenant ne connaît pas le référent dans sa langue maternelle.

Il ressort donc que les meilleurs résultats, autant en compréhension qu'en production sont obtenus avec un dictionnaire riche rassemblant le maximum d'informations. Les utilisateurs sont à même, d'une part de trouver à chaque fois les renseignements nécessaires, et d'autre part, de pouvoir les exploiter. Dans la dernière décennie, de nouveaux dictionnaires pédagogiques ont ainsi vu le jour, pour pallier tous ces inconvénients dont les principaux apports seront décrits dans la section suivante.

6.4 Les dictionnaires pédagogiques

Le dictionnaire pédagogique ou dictionnaire pour apprenant est un monolingue destiné aux personnes apprenant une langue étrangère. Il se différencie sur un certain nombre de points du monolingue pour natif. Davantage de précisions sont données sur les entrées du dictionnaire. En principe le natif est déjà au fait de ces informations qui n'ont pas besoin d'être mentionnées dans les monolingues normaux. Cependant, pour des étrangers, elles sont des indices précieux sur le maniement de la langue.

6.4.1 Les définitions

Le point le plus significatif et le plus visible est l'utilisation d'un vocabulaire définitoire contrôlé pour décrire les entrées dans les définitions. Il est nécessaire en effet de définir et d'expliquer des mots inconnus avec des mots simples, que l'apprenant est susceptible de connaître déjà. Car s'il doit parcourir d'autres articles, l'efficacité du dictionnaire s'estompe et son utilité disparaît. Ce procédé a cependant l'inconvénient, dans certains cas, d'alourdir de façon notable les définitions.

D'autre part, certains dictionnaires pédagogiques comme le COBUILD ont adopté un format standard sous forme de phrase pour les définitions[55]. L'avantage est de pouvoir indiquer de manière transparente et naturelle, sans avoir recours à un métalangage ou à des codes, un certain nombre d'informations comme les structures grammaticales des

[55] Il est à signaler que les définitions entrées dans PROLEMAA ont été standardisées sous forme de phrases qui font clairement apparaître les arguments des schémas syntaxiques des entrées lexicales.

verbes, les actants, les contextes d'emploi ou les collocations. Ce format a par contre l'inconvénient de ne pas être toujours directement adaptable au contexte du mot en question que l'on cherche à comprendre dans le texte.

6.4.2 L'utilisation des exemples

Avec l'utilisation de corpus dans l'élaboration des dictionnaires, des exemples authentiques sont très utilisés pour servir principalement d'illustration à l'entrée définie. La conséquence, est l'obtention d'exemples contenant un nombre important de mots hors du vocabulaire contrôlé des définitions.

D'autres dictionnaires procurant un rôle plus pédagogique aux exemples, utilisent des exemples inventés, qui permettent de suppléer la définition et de montrer les propriétés grammaticales et collocationnelles de l'entrée. De ce fait, les exemples tendent à être plus stéréotypés.

Laufer (1992, cité par Selva (1999)) a étudié les rôles respectifs des différents types d'exemples et de leur influence dans la compréhension. Ces résultats montrent que l'exemple seul ne suffit pas pour comprendre, mais que, dans ce cas de figure, les exemples inventés conduisent à de meilleurs résultats, tant en compréhension qu'en production.

6.4.3 Les illustrations

L'illustration est très utile dans les dictionnaires pour apprenants, puisqu'elle supplée la définition lorsque les moyens purement linguistiques sont insuffisants pour expliquer un mot ou produisent une définition trop lourde. L'illustration permet surtout de minimiser l'un des problèmes cruciaux de l'apprenant, lorsqu'il veut exprimer une idée avec un mot qu'il ne connaît pas.

Les informations d'ordre graphique sont quasiment toujours des illustrations concernant des objets concrets. On peut ainsi connaître différents types ou différentes parties d'un thème donné.

6.4.4 Les renvois

Les dictionnaires pédagogiques utilisent des renvois analogiques vers d'autres mots lorsqu'ils estiment qu'une comparaison est nécessaire pour bien saisir les nuances. Les mots auxquels on renvoie sont des synonymes, antonymes, mots composés ou présentant des similarités morphologiques.

Même si l'apprenant a trouvé le mot qui est censé exprimer son idée, il peut s'interroger sur sa justesse. En d'autres termes, il lui faut, dans un premier temps, examiner les différentes possibilités qui se présentent pour démasquer les candidats, puis comprendre les différences, parfois subtiles, entre eux pour pouvoir retenir le plus adéquat.

Ces renvois doivent être repérables et la nature de la relation doit être très claire. A côté de ces renvois, les dictionnaires utilisent souvent des notes d'usage qui discutent de manière contrastive des points précis ou des difficultés auxquelles peuvent être confrontés les apprenants.

6.4.5 Les informations grammaticales

Les informations grammaticales occupent un espace relativement important dans les dictionnaires pour apprenants. Les apprenants de langue étrangère ont en effet besoin, plus que les natifs, d'être dûment renseignés sur ce point afin de s'exprimer correctement.

Chaque dictionnaire a son système de notation, plus ou moins clair, faisant plus ou moins appel à des notions qui posent parfois des difficultés telles que la transitivité ou l'intransitivité. L'apport le plus significatif est celui de COBUILD dont les définitions sous forme de phrase font clairement apparaître les structures syntaxiques sans pour autant contraindre le lecteur à aller chercher la signification de tel ou tel code. Notons que ce procédé permet en outre, surtout en compréhension, de rester dans le même niveau de lecture du texte, l'utilisateur passe aux phrases du dictionnaire et évite ainsi le métalangage des codes grammaticaux.

6.4.5 Les fréquences et les registres

Deux éléments d'informations peuvent assister l'apprenant en production : la fréquence et le registre.

Les dictionnaires basés sur corpus permettent d'associer des indications sur la fréquence de chaque entrée dans le corpus. L'apprenant est ainsi informé de l'utilisation réelle d'une entrée et peut en mesurer, par exemple, son côté démodé ou au contraire dans le vent. Le COBUILD par exemple applique systématiquement une échelle de 0 à 5 diamants (pour les plus fréquents).

Le COBUILD aborde aussi les problèmes de pragmatique en appliquant des registres spécialisés aux entrées du dictionnaire tels que *journalisme*, *légal*, *littéraire*, *démodé*, *parlé*, *écrit*, etc.

6.4.6 Conclusion

Cette section a montré toutes les améliorations apportées aux dictionnaires classiques afin de répondre aux attentes aux apprenants de langue étrangère surtout lorsqu'ils les utilisent pour une tâche de production. Ainsi, de nouveaux dictionnaires pédagogiques sont apparus dont la principale caractéristique est l'utilisation de définitions et d'exemples plus accessibles aux apprenants débutants.

Cependant quelques problèmes persistent encore et rendent parfois ces dictionnaires inefficaces. Le problème d'accès lexical reste particulièrement présent dans les dictionnaires arabes et augmente la durée de consultation du dictionnaire, ce qui nuit au

processus de lecture des textes. Nous verrons dans la section suivante, si les dictionnaires électroniques pourraient pallier ce problème et concilier les apprenants avec le dictionnaire.

6.5 Les dictionnaires électroniques

Depuis le début des années 90, le grand public a pu avoir accès à de nombreux dictionnaires électroniques[56]. Ces dictionnaires ont apporté avec eux un nouveau mode d'usage et quelques améliorations qui ont permis de pallier les problèmes rencontrés avec les dictionnaires sur papier. Il convient donc d'examiner les principaux apports de ces dictionnaires, dont notamment l'amélioration de l'accès lexical et la facilité de navigation dans le dictionnaire.

6.5.1 L'accès lexical

Les possibilités de traitement automatique et de recherche d'un mot dans un index sont un des grands atouts des dictionnaires électroniques par rapport à leur équivalent papier. Ce procédé efficace permet d'accéder au mot très rapidement, ce qui s'avère très important dans un processus qui doit être le plus court possible.

En effet, passer de la forme fléchie dans le texte à la racine qui constitue l'entrée du dictionnaire papier, puis sélectionner la forme canonique appropriée n'est pas évident pour l'apprenant surtout lorsque des clitiques sont attachés au mot et lorsque le texte est non vocalisé.

Pur résoudre ce problème, les dictionnaires électroniques ont recours à deux solutions : une première, non envisageable pour la langue arabe qui consiste à lister l'ensemble des formes fléchies (cf. § 3.1) et une seconde qui fait fonctionner un analyseur morphologique.

Bien évidemment, l'accès au dictionnaire PROLEMAA pourrait être résolu par l'utilisation de l'analyseur morphologique (cf. chapitre 3). Cependant, le problème ne sera pas complètement résolu et deux problèmes au moins subsisteront :

- Le problème de l'homonymie : que faire lorsque deux mots s'écrivent de la même manière ? Lorsque les deux mots n'ont pas la même catégorie grammaticale (*KTB*'verbe' : écrire et *KTB*'nom' : livres), le problème peut être résolu par l'ajout, de la catégorie grammaticale à côté du mot dans la liste des entrées. Par contre, si ce n'est pas le cas, aucun moyen n'est possible de les distinguer. Il faut donc lire la définition de chaque entrée pour rester ensuite sur celle qui nous intéresse, à moins que le texte ne soit étiqueté par des informations de niveau sémantique comme ceux que

[56] Il convient de signaler que parmi le grand nombre d'éditeurs des dictionnaires électroniques, le Cobuild semble, pour l'instant, le seul éditeur à avoir produit des dictionnaires électroniques pour apprenant.

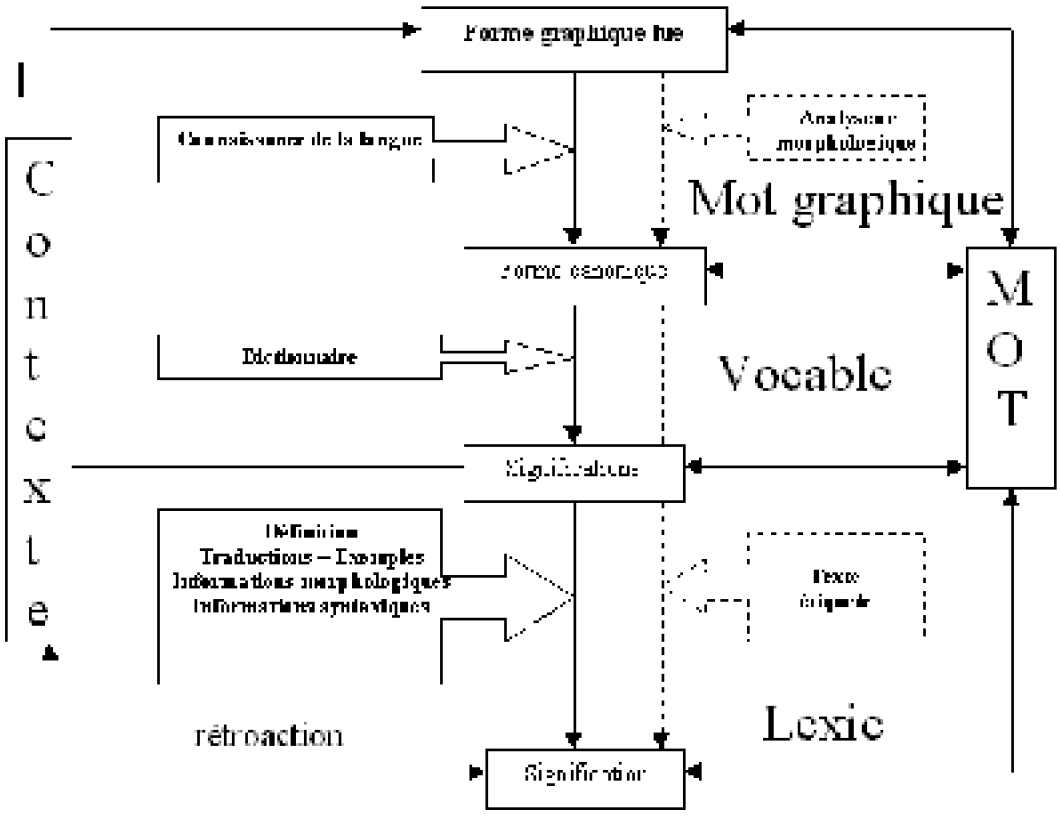
nous proposerons dans le cadre de l'environnement d'apprentissage « *AL-Mu*^{c} *aLLiM* ».

- Le problème des collocations et expressions semi-figées : Là, les solutions proposées sont très différentes. On peut proposer un index contenant principalement des expressions semi-figées mais qui sont listées telles quelles et même avec leur variation. Il faut taper exactement les premières lettres de l'expression pour avoir accès à sa traduction. Inutile de dire que cette fonctionnalité ne sera que de peu de secours pour celui qui précisément ne connaît pas l'expression ou qui ne la devine pas dans un texte. On peut aussi employer un moyen plus souple qui consiste à taper un mot ou une forme canonique de l'expression et obtenir les différentes expressions associées. C'est cette solution qui sera retenue, vu le nombre important de variations de chaque expression en arabe.

La figure (6-17) ci-dessous résume les différentes étapes permettant le passage d'un mot graphique arabe à sa signification dans le dictionnaire PROLEMAA. Le schéma est divisé en deux colonnes : la colonne de gauche correspond à un accès classique alors que celle de la colonne de droite correspond à une méthode d'accès optimisée que nous emploierons dans le cadre de l'environnement « *AL-Mu*^{c} *aLLiM* ».

L'utilisation du dictionnaire PROLEMAA peut être donc envisagée de trois manières différentes :

- L'apprenant cherche la signification d'un mot dont il ignore la forme canonique (i.e. l'entrée du dictionnaire). Dans ce cas, il pourrait avoir recours à l'analyseur morphologique qui lui propose un ensemble de solutions possibles parmi lesquelles figure la bonne forme canonique du mot.
- L'apprenant cherche la signification d'un mot dont il connaît la forme canonique. Dans ce cas, il entre directement l'entrée du dictionnaire et le système lui propose les différentes significations de cette forme parmi lesquelles figure la bonne signification du mot recherché.
- L'apprenant cherche la signification d'un mot à partir d'un texte électronique étiqueté. L'apprenant sélectionne le mot en question et le système le fait directement passer à la signification du mot choisi.



C'est ce dernier mode d'accès qui sera le plus utilisé dans l'environnement d'apprentissage, puisqu'on n'utilisera que des textes complètement étiquetés. Ce mode d'accès répond en effet aux besoins des apprenants débutants, qui n'ont pas assez de connaissances pour arriver à passer de la forme fléchie à la forme canonique. Les apprenants d'un niveau intermédiaire ne perdront non plus le fil de lecture du texte, puisqu'ils ils passeront peu de temps lors de la consultation du dictionnaire.

6.5.2 L'interactivité

Le second avantage des dictionnaires électroniques par rapport aux dictionnaires sur papier est l'interactivité qui facilite la navigation entre les différentes informations du dictionnaire et améliore l'efficacité de la consultation. L'utilisateur peut influer sur la nature et la quantité des informations qui lui sont présentées et s'affranchit des limites du papier. Le support informatique n'est en effet plus tributaire du manque de place des versions papier.

Le dictionnaire électronique ne doit surtout pas reproduire à l'écran l'agencement de la version papier. Il doit par contre permettre d'agir sur la présentation en laissant la possibilité à l'utilisateur d'afficher tel ou tel élément d'information. On peut avoir seulement les définitions, ou les exemples, ou les règles grammaticales, ou les traductions etc. ou bien des combinaisons de plusieurs d'entre eux, comme les définitions et les règles grammaticales (figure 6-15). Cette fonctionnalité facilite le parcours de l'entrée et limite les inconvénients de l'agencement de la version papier.

Les dictionnaires électroniques permettent aussi les renvois dans et entre différentes sources qui constituent le dictionnaire, les liens directs avec les autres applications, les possibilités de recherche complexe à l'intérieur du texte des entrées du dictionnaire et les interactions possibles avec les utilisateurs pour aider à développer le vocabulaire et les facilités de consultation.

6.5 Conclusion et perspectives

Nous avons recensé dans ce chapitre les informations que devrait contenir un dictionnaire électronique adapté aux apprenants de la langue arabe langue seconde et ses modalités d'usage. Le prototype PROLEMAA, nous a permis de réaliser un premier dictionnaire riche en informations qui est très proche des dictionnaires semi-bilingues et qui peut aider l'apprenant débutant dans ses tâches de compréhension et de production grâce à une présentation des informations plus conviviale et plus sélective.

Ce prototype doit par conséquent être enrichi par de nouvelles informations très utiles aux apprenants : les illustrations, les synonymes, les antonymes, les fréquences, etc. La qualité linguistique des définitions doit être revue en utilisant un vocabulaire définitoire contrôlé.

L'enrichissement de PROLEMMA pourrait être envisagé à l'aide des applications informatiques développées : le programme d'injection automatique des listes du définitions, programme de calcul de fréquences, etc.. On pourrait aussi faire participer des apprenants avancés à ce travail d'enrichissement sur des parties du dictionnaire. Dans le cadre de l'environnement « *AL-Mu* ^{c} *aLLiM* », nous avons commencé à explorer cette piste en proposant aux apprenants l'introduction d'illustrations, de synonymes et d'antonymes par le biais su dictionnaire personnel (cf. § 9.5).

Nous avons proposé d'autre part une solution pour résoudre le problème d'accès lexical qui représente l'un des principaux obstacles quant à l'utilisation du dictionnaire par les apprenants. L'utilisation de textes complètement étiquetés par des informations sémantiques, permet en effet, le passage direct du mot à son sens par une simple sélection du premier.

Les difficultés relatives à la compréhension des textes étant minimisées, nous aborderons dans le chapitre suivant le problème du choix des activités lexicales et grammaticales qui permettront la mémorisation du lexique consulté.

Chapitre 7 Les activités lexicales et grammaticales

« Malheur à qui ne se corrige pas, soi et ses œuvres ! Il faut se corriger, eût-on quatre-vingts ans. Je n'aime point les vieillards qui disent : « J'ai pris mon pli. » Ah ! vieux fou, prends en un autre. Rabote tes vers, si tu en as fait, et ton humeur si tu en as. » VOLTAIRE

7.1 Introduction

Les études psycholinguistiques ont montré que la rétention du vocabulaire et la maîtrise de la grammaire sont favorisées par la quantité de travail effectuée par l'apprenant sur la langue cible (cf. chapitre 5). La lecture des textes et l'exposition des sens des mots à elles seules ne suffisent pas à l'acquisition du vocabulaire, la présence d'activités permet d'accélérer l'acquisition lexicale et d'améliorer les compétences grammaticales des apprenants.

Il s'agit dans ce chapitre de spécifier les activités qui feront partie de l'environnement « *AL-Mu*^{c} *aLLiM* », et qui vont permettre à l'apprenant d'acquérir des connaissances lexicales factuelles et des connaissances procédurales simples (i.e. règles grammaticales).

La conception de ces activités se base sur un ensemble de principes fondamentaux pédagogiques et ergonomiques que nous exposerons au début de ce chapitre. Nous présenterons ensuite les principales formes d'activités qui seront utilisées et le processus de génération automatique qui permet de les obtenir à partir des différentes ressources du système (la base de données lexicale, le corpus de textes étiquetés et le modèle de l'apprenant). Nous verrons que contrairement aux exercices conçus sur un support papier, les activités électroniques doivent être définies d'une manière générique et plus précise. L'enseignant rédigeant les épreuves sur support papier fait en effet abstraction de beaucoup de paramètres que nous ne pouvons nous empêcher de les considérer. L'ALAO impose une vue plus générale et plus complète de la conception d'activités.

7.2 Principes de construction des activités[57]

Toute compétence bénéficie d'un exercice systématique. L'apprenant doit, à un moment donné, pouvoir exercer systématiquement les compétences encore incomplètement stabilisées. Une activité doit assurer quatre fonctions essentielles :

- permettre à l'apprenant d'intégrer l'information qui lui est présentée,
- maintenir l'attention de l'apprenant,
- informer l'apprenant sur son niveau de maîtrise des objectifs et permet en cela, pour autant que ceux-ci soient clairement annoncés, d'autoréguler son comportement,
- informer le système sur le niveau de maîtrise des objectifs par l'apprenant, ce qui permet au système de transformer l'activité en interactivité,

Pour assurer la présence de ces fonctions dans les activités, nous devons tenir compte d'un certain nombre de principes *pédagogiques* et *ergonomiques* que nous énumérons dans cette section.

7.2.1 Précision des objectifs

Apprendre c'est réaliser des activités qui se rapprochent des objectifs définis. Ceci semble une affirmation triviale, mais beaucoup d'environnements d'apprentissage contiennent des activités qui ne sont pas liées aux objectifs, parfois parce que les objectifs eux-mêmes n'ont pas été clairement spécifiés.

57
Les principes pédagogiques qui seront présentés dans cette section, sont synthétisés à partir d'informations récupérées du site de l'unité (TECFA), active dans le domaine des technologies éducatives et qui fait partie de la Faculté de Psychologie et des Sciences de l'Education de l'Université de Genève.: http://tecfa.unige.ch/themes (dernière consultation - Octobre 2001).

Le point de départ de la conception d'une activité, c'est de décrire de façon **claire** et **précise** ce que les apprenants doivent savoir faire au terme de leur interaction avec l'activité. Imaginons que l'objectif de l'activité est : "les apprenants devront connaître un élément X (le participe actif en arabe par exemple)". Comment alors l'auteur pourrait-il créer des activités pour vérifier que les apprenants ont atteint cet objectif ? Il peut leur demander :

1. de citer la définition de X,
2. de démêler X dans un ensemble d'éléments plus vaste (reconnaître des participes actifs parmi d'autres formes de déverbaux),
3. d'établir des rapports entre X et des items se rapportant ayant avec lui des liens définis au préalable (dériver des participes actifs à partir de verbes).

L'apprenant peut réussir dans l'une des épreuves tout en échouant dans une autre. Ces trois types d'activités dénotent des compétences différentes et permettent de les mesurer. L'objectif général spécifié ci-dessus ne permet pas de déterminer si une activité a été efficace ou non par rapport à cet objectif. Il est recommandé aux auteurs de définir des objectifs, clairs et opérationnels, c'est-à-dire de donner une description précise de ce que les apprenants devraient être capables de faire en effectuant l'activité prescrite pour garantir que l'objectif cible a été atteint et que la compétence visée a été acquise. Formuler l'objectif dans l'énoncé "démêler X dans un ensemble d'éléments plus vaste" est plus opérationnel que "Reconnaître un élément X" car il décrit de quelle manière on peut vérifier si l'objectif est atteint ou non.

Un objectif opérationnel comprend trois éléments :

- Les actions ou **comportements** que l'apprenant doit réaliser. Les actions définissent des observables sans lesquels il n'y a pas d'évaluation. Les verbes savoir, connaître, comprendre,... ne définissent pas d'action observable. Ces connaissances ou compétences doivent être traduites en actes observables : citer, souligner, encadrer,...
- Les **conditions** dans lesquelles l'apprenant doit réaliser ces comportements. L'objectif "pouvoir transformer des verbes de l'inaccompli vers l'accompli" est très imprécis. Le même objectif couvre des connaissances bien différentes selon que le sujet saisit l'accompli ou le sélectionne parmi un ensemble de réponses possibles. Les conditions dans lesquelles le sujet doit réaliser sa tâche peuvent faire varier du tout au tout la nature de cette tâche et sa complexité. Si c'est le cas, il convient donc de préciser ces conditions.
- Les **critères de qualité ou le niveau de performance**. Dans certains cas, il est nécessaire de préciser un critère permettant de considérer si l'objectif est atteint. Il est par exemple parfois utile d'indiquer la précision acceptée : "pouvoir générer l'inaccompli des verbes simples tout en vocalisant la deuxième lettre radicale". Dans certains cas, le critère renvoie au temps maximal de réponse. Le critère temps doit être utilisé avec précaution car il peut générer un stress important chez l'apprenant, ce qui biaise les résultats de l'évaluation. Ces critères sont particulièrement

importants et doivent être explicitement encodés dans le programme.

La spécification des objectifs a trois grands avantages :

- Elle permet une évaluation plus sérieuse de l'apprenant.
- Elle permet de communiquer clairement à l'apprenant ce qu'on attend de lui. Cette information rend l'enseignement plus efficace, probablement parce que l'apprenant perçoit mieux la pertinence des informations qu'il reçoit.
- Lorsque les objectifs sont clairement définis, la conception des activités d'apprentissage peut reposer sur ces objectifs, et non sur la simple tendance qu'a l'être humain de reproduire des activités passées.

7.2.2 Principes pédagogiques

Dès les premiers travaux en matière d'enseignement assisté par ordinateur, l'efficacité de l'ordinateur est essentiellement attribuée à son potentiel en matière d'**individualisation**, c'est-à-dire à sa capacité d'adapter les interventions aux caractéristiques de chaque apprenant.

Selon les informations dont il dispose, le système pourra modifier son comportement de différentes façons : choisir un feed-back spécifique, augmenter le nombre d'exercices ou la difficulté des exercices, choisir une méthode d'apprentissage (par exemple, inductif versus déductif). Ces trois exemples sont issus de principes pédagogiques très connus dont nous reprenons ici les plus importants d'entre eux :

- Principe de feedback : L'apprenant doit être informé de l'adéquation de son comportement par des feed-back spécifiques et pourra adapter son comportement en conséquence. Le feedback ne se limite pas aux énoncés de type "c'est correct" ou "c'est faux", mais doit surtout prendre d'autres formes comme des liens sur des ressources du système, des nouvelles questions (activités) ou des tableaux de synthèse au sein desquels le sujet peut avoir un récapitulatif du déroulement de l'activité.
- Principe de motivation : L'ALAO a souvent compté sur une motivation extrinsèque à la tâche, soit liée à l'effet de nouveauté du média (...), soit à l'utilisation des procédés multimédias. Il est préférable de chercher une motivation **intrinsèque** à la tâche qui repose sur un énoncé clair et concret précisant les compétences à acquérir et leur **utilité** pour l'apprenant. Sur un autre plan, l'activité doit être **positive**, c'est-à-dire le nombre de succès doit être largement plus important que le nombre d'échecs. Il semble évident qu'un taux élevé d'échec est susceptible de décourager ou démotiver les apprenants, du moins ceux qui n'ont pas confiance en eux ou ceux qui n'ont pas une grande motivation. En outre, il convient d'ajouter que les erreurs ne sont source d'apprentissage que sous certaines conditions, en particulier lorsque l'apprenant dispose de l'information nécessaire pour comprendre en quoi son comportement est erroné et comment le résoudre.

- Principe de progressivité : C'est un des plus vieux principes pédagogiques que de décomposer les apprentissages complexes en apprentissages plus simples. Ce principe est cependant controversé. La décomposition implique souvent une décontextualisation qui prive l'apprenant d'informations précieuses sur la fonction de chaque étape. Quelle que soit la finesse de la progressivité, l'apprenant a souvent besoin d'aide pour réaliser une tâche : indices, suggestions,....
- Principe de participation : Les théories qui privilégient les aspects sociaux de l'apprentissage accordent une grande importance au **guidage**, via le processus de **participation**. Le principe de participation désigne un partage de la tâche entre l'apprenant et le système de telle sorte que, les deux ensemble résolvent la tâche fixée. La partie assumée par le système doit idéalement diminuer jusqu'au moment où l'apprenant assume seul la tâche.
- Principe de la multiplicité : Il n'existe pas de méthode unique, de représentation unique, d'activité unique. Un environnement d'apprentissage sera d'autant plus riche qu'il dispose d'une multiplicité d'activités didactiques, des formes de représentation. L'apprenant aura des connaissances d'autant plus riches et robustes qu'il dispose de représentations multiples. Il est en outre important de souligner que l'apprenant dissocie rarement les connaissances acquises du contexte dans lesquelles elles ont été enseignées. Il est donc essentiel pour le rendre capable de transfert, de l'exercer à appliquer ses compétences dans une variété de contextes

7.2.3 Principes ergonomiques

L'utilisation de l'ordinateur fait que souvent les effets cognitifs de l'activité sont influencés par les aspects moteurs de l'activité. Si par exemple, on demande à l'apprenant de sélectionner le groupe sujet de la phrase suivante (الولد الذي سقط على الأرض عجر من المدرسة : L'enfant qui était tombé par terre est revenu de l'école). Une difficulté consiste à savoir s'il faut fournir uniquement le centre du groupe sujet (الولد : l'enfant) ou tout le groupe sujet (الولد الذي سقط في الأرض : l'enfant qui était tombé par terre). La réponse de l'apprenant sera déterminée par ce que le concepteur aura défini comme objet 'clickable' ou par la longueur du champ d'entrée de la réponse attendue (i.e. si la longueur du champ de réponse varie de question en question). L'importance de ces détails illustre l'importance d'avoir un scénario pédagogique très détaillé : certains détails de l'implantation peuvent modifier totalement la nature des activités cognitives induites.

D'un autre côté, et dans la mesure du possible, le concepteur doit éviter que la communication de la réponse au système, c'est-à-dire l'utilisation de l'interface, constitue en soi une activité complexe. En théorie, l'apprenant ne devra pas perdre du temps à apprendre à utiliser le système lui-même. En pratique, il est difficile de réduire à zéro le temps consacré à cet apprentissage non pertinent. Il n'existe pas d'interface totalement "transparente", quels que soient les progrès en matière de manipulation directe. Par exemple, dans le cadre de l'environnement « *AL-Mu*^{c} *aLLiM* », nous avons évité au maximum de saisir les réponses à partir du clavier après avoir constaté que la manipulation du clavier arabe (certaines graphies sur ce clavier peuvent être facilement

confondues), les déconcentrait de l'activité. Pour les activités nécessitant l'utilisation du clavier, nous avons ajouté à l'interface un clavier écran qui permet de nuancer les caractères et de les saisir par des simples clicks. L'allégement de l'activité motrice des apprenants avait pour conséquence un meilleur rendement de la part des apprenants et la diminution des erreurs de type typographique.

7.3 Conception et réalisation des activités

La conception et l'informatisation d'une activité est une opération complexe et contraignante. La réalisation de chaque activité implique un ensemble de choix de natures diverses : fixer l'objectif de chaque activité, réfléchir sur le contenu et le traduire sur l'interface, prévoir les retours, les corrections et l'aide à fournir aux réponses de l'apprenant, expérimenter l'activité et apporter des correctifs, etc. Dans cette section, nous essayerons de décrire les différentes activités utilisées dans l'environnement d'apprentissage « *AL-Mu*^{c}*aLLiM* » en soulignant leurs principales caractéristiques.

7.3.1 Caractéristiques des activités

Les activités proposées ici sont toutes informatisées et ne peuvent pas être reproduites sur papier. L'intérêt computationnel prévaut et les moyens que l'informatique offre et les ressources développées sont tous utilisés (analyseur morpho-syntaxique, base de données lexicale, textes étiquetés, modèle de l'apprenant, dictionnaire général et dictionnaire personnalisé).

Chaque activité est reproductible, c'est-à-dire que l'apprenant pourra avoir de nouvelles épreuves à volonté. Néanmoins, leur mise au point ne révèle pas une explosion combinatoire au-delà d'un certain nombre d'items considérés. Le temps de réponse du système, tant au niveau de la préparation des épreuves que dans l'aide qu'il apporte, est convenable et acceptable pour que l'activité puisse se dérouler normalement.

Les activités sont pertinentes en termes d'apprentissage, c'est-à-dire que l'apprenant doit fortifier ou valider son acquisition grâce à elles. Elles se prêtent à la mise en place d'un système d'aide qui confortera le rôle pédagogique de l'environnement, c'est-à-dire que le système ne se contentera pas de répondre vrai ou faux, mais aiguillera l'apprenant pour trouver la solution.

Il existe dans la littérature plusieurs types d'activités pour l'apprentissage des langues qui mettent en jeu des approches différentes ou qui traitent des aspects différents. Il y a par exemple les activités en compréhension ou en production, les activités communicatives ou non communicatives, les exercices hors contexte ou en contexte.

Au vu des moyens informatiques dont nous disposons (analyseur des mots graphiques), il n'est pas possible que l'ordinateur intervienne sur une production totalement libre d'une phrase par l'apprenant en vue d'une analyse et d'une correction automatique. Cependant, le problème n'est pas vraiment de l'ordre de la production libre

ou non. La langue est un phénomène riche et complexe et tous ses aspects ne peuvent être abordés en même temps. Il convient donc d'en isoler les mieux appropriés à nos ressources, pour pouvoir les travailler plus efficacement. Ainsi, dans le cadre de l'environnement d'apprentissage « *AL-Mu*^{C} *aLLiM* », nous avons retenu un certain nombre d'activités que nous pouvons diviser en quatre catégories : Les activités à réponses fermées, les activités à réponse ouverte ou construite, les activités de découverte guidée et les activités de type ludique.

7.3.2 Activités à réponses fermées

On parle d'activités à réponses fermées, lorsque le sujet choisit sa réponse dans un ensemble fini de propositions. La gamme des activités possibles de ce type est assez vaste. Elles diffèrent surtout par les modalités de saisie des réponses introduites :

- le sujet choisit sa réponse parmi les N boutons proposés,
- le sujet clique sur une des N zones ou objets sensibles définis (ces zones ou objets peuvent parfois être nombreux),
- le sujet déplace un objet dans une des N zones définies,
- le sujet presse une des N touches considérées,
- le sujet sélectionne un des N items des M menus définis,

La correction de ces activités ne pose pas de problèmes particuliers (i.e. le travail effectué par l'apprenant est comparé aux différentes attentes du système). Par contre, la conception est bien plus délicate mettant en évidence de nombreux facteurs et contraintes pesant sur les 'distracteurs' (proposition correspondant à une erreur classique des sujets) par rapport au stimulus et sur l'établissement des tests.

Nous allons à l'aide de quelques formes d'activités à réponses fermées, montrer leur processus de conception et leur fonctionnement. L'utilisation de plusieurs formes d'activités, permet d'éviter un enseignement monotone et la démotivation de l'apprenant.

1) Question à choix multiple (QCM) : Les activités QCM ont été intensivement utilisées dans l'évaluation pédagogique car elles permettent un traitement rapide, objectif et facilement programmable des réponses. Elles ont cependant souvent été critiquées, car la plupart d'entre elles étaient mal construites et ne fournissaient pas une mesure valide des compétences. La plupart de ces défauts ne sont cependant pas intrinsèques aux QCM. Certains QCM peuvent posséder un pouvoir diagnostic supérieur aux questions ouvertes, par exemple en incluant parmi les propositions un ou plusieurs 'distracteurs'.

Dans le cadre de l'apprentissage des langues, ce type d'activité permet d'explorer différents aspects (i.e. on peut demander une traduction en langue maternelle, un synonyme, un antonyme, une définition, un déverbal, une conjugaison, une déclinaison, etc.). On fait intervenir généralement des mots seuls, isolés de la phrase et du texte. A chaque question correspond un certain nombre de réponses (au moins quatre de préférence[58]), une ou plusieurs étant la ou les bonne(s) réponse(s), les autres étant des distracteurs. Tout l'intérêt de ces activités réside donc dans le choix de ces distracteurs,

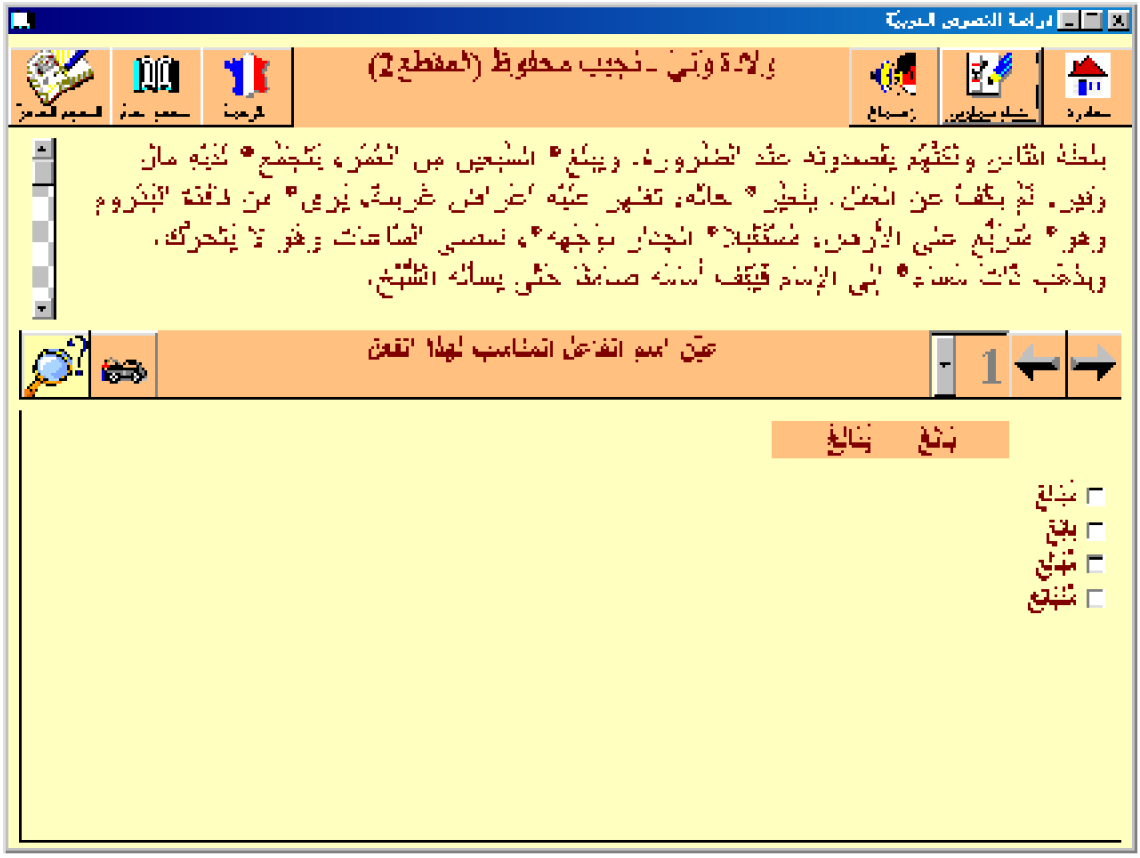
qui doivent être générés à partir de critères préalablement établis par l'auteur.

Par exemple, supposons que nous devons définir les distracteurs d'une activité QCM, qui a pour objectif de permettre à l'apprenant de reconnaître le schème du participe actif des verbes de la forme (IV) comme celui du verbe (بَـتَـاكَ - يُـبِـتَـاكُـي, *KâTaBa – YuKâTiBu*) = (« écrire à quelqu'un ») parmi d'autres déverbaux. La figure (7-1), montre cette même activité générée à partir du verbe (غَـلَّـابَ – يُـغَـلِّـابُـي : *BâLa G a – YuBâLi G u*) = (« exagérer »), qui a été sélectionné à partir du texte à étudier[59].

Etant donné l'objectif de cette activité, les distracteurs doivent avoir des formes très proches de la réponse correcte. Nous avons choisi comme premier distracteur le participe passif du même verbe : (مُـبَـتَـاكَـم, *MuKâTab*) dont la seule différence avec la bonne réponse est la voyelle de la deuxième lettre radicale qui devient (َ : *a*) au lieu de (ِ : *i*). Nous avons choisi comme second distracteur un autre participe actif très proche aussi de la réponse correcte : (مُـبَـتَـاكِـتَـم : MuTaKâTib) qui est généré à partir du verbe obtenu à partir de la même racine à la forme VII (تَ – بَـتَـاكَـتَـي : *TaKâTaBa – YaTaKâTaBu*) = (« s'écrire »). Nous avons enfin choisi le participe actif du verbe de la première forme avec la même racine : (بَـتَـاكَ : *KâTib*) qui induit souvent les apprenants en erreur.

[58] Augmenter le nombre de propositions permet de réduire le rôle du hasard. En outre, compter un score négatif pour les réponses erronées incite le sujet qui ne connaît pas la réponse à s'abstenir de répondre, plutôt que de répondre au hasard. On peut alternativement inclure un bouton 'Je ne sais pas'.

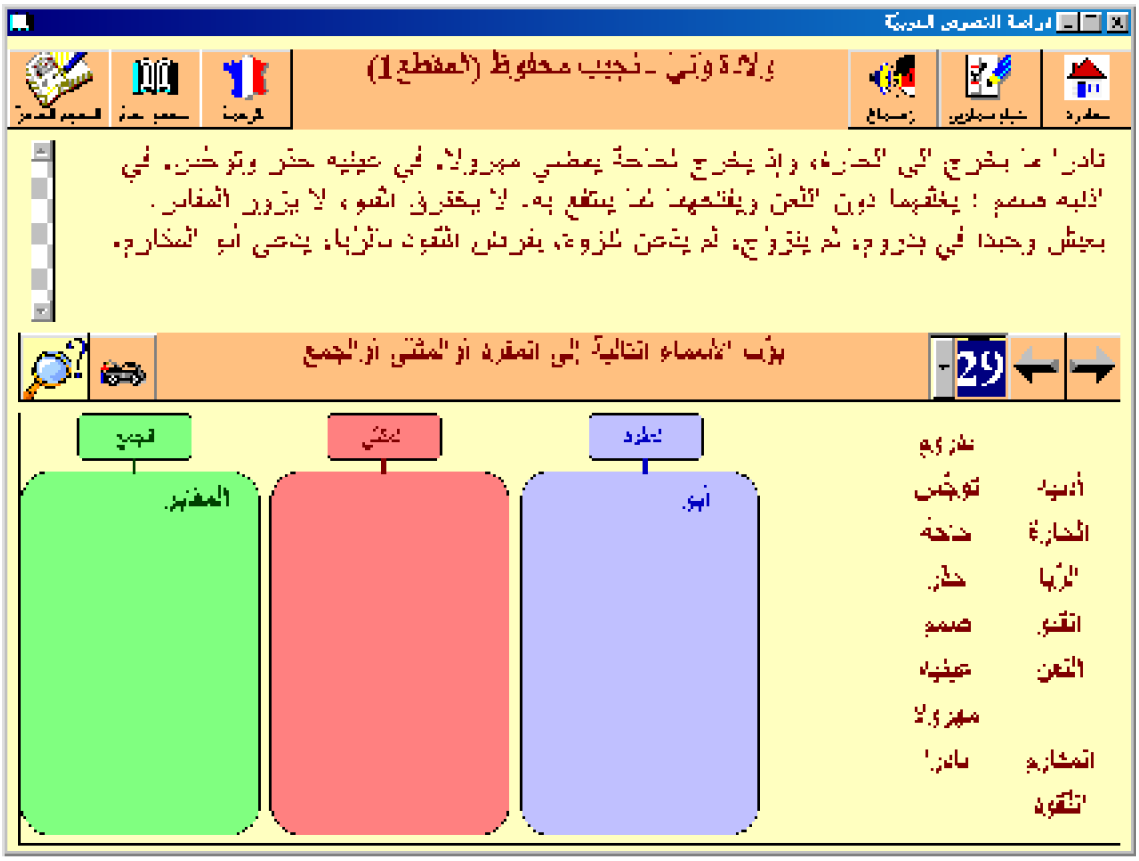
[59] L'interface d'étude de texte de l'environnement « *AL-Mu c aLLiM* » se compose de deux parties : une partie pour l'affichage du texte qui est toujours visible et une seconde pour l'affichage de l'activité en cours ou les propriétés d'un item du texte à partir du dictionnaire général.



Ces distracteurs doivent être, par conséquent, définis avec le plus grand soin pour qu'ils soient **pertinents** : Afin de multiplier le nombre de propositions, le concepteur a parfois tendance à ajouter des propositions fantaisistes que le sujet peut écarter sans aucune difficulté. En plus, le concepteur doit aussi prévoir pour chacun de ces distracteurs, un commentaire adapté susceptible d'avoir un fort impact pédagogique sur l'apprenant, au cas où il serait sélectionné par ce dernier. Il peut aussi prévoir d'autres formes d'aide, comme par exemple proposer des liens hypermédias sur une leçon de grammaire ou sur une partie du dictionnaire ou sur une nouvelle activité relative au distracteur sélectionné.

Par ailleurs, le raisonnement de l'apprenant sera plus complexe si on ne lui indique pas à l'avance le nombre de propositions vraies. La bonne réponse ne doit pas être disposée suivant un système observable (comme toujours en dernière position), le système informatique doit en en effet placer aléatoirement la réponse dans la liste des propositions et ne pas mettre la proposition correcte au même endroit dans deux questions consécutives.

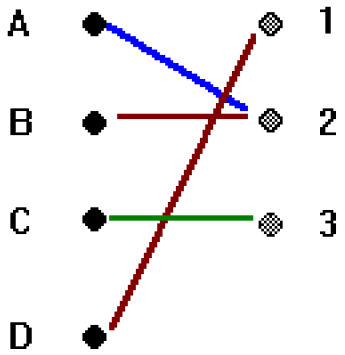
2) Exercices de trous : Dans ce type d'activité, l'apprenant doit placer une liste de mots extraits de leur contexte dans une série de phrases trouées. Ces phrases sont produites automatiquement par le concordanceur à partir du corpus textuel étiqueté. L'auteur doit préalablement spécifier le nombre et les propriétés des mots à extraire. Ces activités permettent à l'apprenant d'asseoir ses connaissances lexicales, en réexaminant l'emploi et le fonctionnement des items lexicaux déjà rencontrés dans de nouveaux contextes. Les différentes propriétés du mot sont illustrées de manière plus naturelle et plus complète que dans des activités hors-contexte. Le choix des propriétés (morpho-syntaxiques et/ou sémantiques) des mots à chercher par le concordanceur, dépend des objectifs de l'activité. La correction de ce type d'activité est aussi simple à effectuer.



3) Exercices de rangement : Ces activités permettent d'opposer deux ou plusieurs valeurs d'une propriété linguistique donnée. L'apprenant doit ranger un ensemble de mots ou de phrases dans les deux ou les contenants correspondants. Les propriétés de chaque contenant doivent être clairement explicitées par l'auteur. Les mots peuvent être

directement sélectionnés à partir du texte à étudier ou proposés dans un ensemble d'éléments plus vaste. Pour la correction de ces activités, le système vérifie simplement si les propriétés des mots correspondent ou pas à avec celles du contenant. L'exemple de la figure (7-2) ci-dessus, montre une activité de ce type où l'apprenant doit placer tous les mots du texte dans les trois contenants selon leur nombre (singulier, duel, pluriel).

4) Exercices d'appariement : Il existe une forme plus complexe d'activités à réponses fermées, les **exercices à appariement,** qui consistent à mettre en relation des propositions fournies dans deux listes distinctes. Dans ce cas, le nombre de réponses possibles est fortement accru, ce qui réduit la part laissée au hasard. La présentation classique des questions par appariement est celle présentée dans la figure (7-3) : le sujet relie par un trait les propositions qu'il désire associer. Dans cette même figure, nous proposons une seconde forme de la même activité[60].



	1	2	3
A		X	
B		X	
C			X
D	X		

L'activité d'appariement est plus complexe si la relation entre les deux listes n'est pas bijective, c'est-à-dire si une proposition de la première série peut-être associée à plusieurs propositions de la seconde série et réciproquement. Lorsque deux listes de quatre items doivent être appariées, le nombre de réponses possibles est de 24 (factorielle de 4) si la relation est bijective et de 256 (44) si la relation n'est pas bijective.

Si les deux listes ont la même longueur, les sujets peuvent induire à tort que cela implique une relation bijective. L'auteur peut soit proposer des listes de longueurs différentes, soit préciser la nature de la relation dans le tronc de la question ('Plusieurs flèches peuvent partir du même point ou arriver au même point'; 'Plusieurs croix peuvent être placées dans la même ligne ou dans la même colonne').

Si les deux listes sont de longueurs différentes, il est préférable de placer la plus longue à droite. Imaginons que les deux listes comprennent respectivement 8 et 3 items. Les sujets apprenant l'arabe ont tendance à lire la première proposition de la liste de droite puis à chercher son correspondant dans la liste de gauche. Si celle-ci ne contient que quelques items, les sujets les retiendront assez rapidement et pourront délibérer sur

[60] Lorsqu'une interaction facile à réaliser sur papier se révèle plus difficile sur écran, il est préférable de lui chercher une substitution que d'obstiner à la transposer fidèlement.

chaque item de droite sans relire à chaque fois toutes les propositions de la liste de gauche.

Dans le cadre de l'environnement « *AL-Mu*^{c} *aLLiM* », nous avons réalisé un certain nombre d'activités dont les plus intéressantes sont celles relatives aux propriétés sémantiques des unités lexicales (i.e. synonymes, antonymes, etc.). La conception de ces activités est simple : L'auteur se contente de spécifier la relation entre les items des deux listes, le nombre d'items de chaque liste et le système se charge de la génération (i.e. personnalisée au profil de l'apprenant) et de la correction des activités.

7.3.3 Activités à réponse ouverte ou construite

On parle de question ouverte ou de question à réponse construite lorsque le sujet construit sa réponse, en particulier lorsqu'il répond par du texte écrit. Du point de vue de l'apprenant, les réponses de type 'texte' lui permettent en effet de construire librement sa réponse. Toutefois, du point de vue de la machine, les réponses de type texte sont analysées par rapport à un ensemble de classes de réponses. Ces questions peuvent donc être considérées comme des questions fermées.

Au vu des moyens dont nous disposons, il n'est pas possible que l'ordinateur intervienne sur des productions totalement libres de phrases. La correction sémantique relève d'un niveau que nous n'avons pas encore abordé et nécessite des travaux plus élaborés. Néanmoins, le fait de disposer d'un analyseur morpho-syntaxique du mot graphique arabe, nous permet d'aborder des productions libres du niveau du mot. Il convient donc d'utiliser ce moyen afin de générer des activités permettant de travailler plus efficacement la production de l'apprenant.

En tenant compte de cette contrainte, nous avons conçu et réalisé plusieurs activités à réponses ouvertes pour l'environnement « *AL-Mu*^{c} *aLLiM* ». La conception de ce type d'activité est analogue à celle des activités à réponses fermées. Par contre, le processus d'analyse des réponses est plus délicat. L'auteur doit définir parfois des classes de réponses considérées comme acceptables et des classes d'erreurs. Ces sous-ensembles sont des espaces de variation autour d'une **réponse-type**. Le processus de comparaison d'une réponse et d'un pattern porte le nom de 'pattern matching'.

Une des formes les plus appropriée pour ce type d'activités est l'exercice à trous où l'apprenant doit remplir librement sa réponse. On utilisera l'analyseur pour déterminer les propriétés de la réponse entrée par l'apprenant. La réponse est ensuite comparée à la réponse correcte et aux classes d'erreurs définies par l'auteur. Dans le cas, où elle figure dans l'une de ces classes, le système propose des aides adaptées préalablement définies par l'auteur.

Dans le cadre de l'environnement « *AL-Mu* ^{c} *aLLiM* », nous avons réalisé plusieurs activités à réponse ouverte. L'exemple de la figure (7-4) ci-dessus, montre un exemple d'une activité mixte où une partie de la réponse est entrée librement par l'apprenant et une seconde partie doit être sélectionnée parmi une liste prédéfinie d'items.

En résumé, pour ce type d'activités, l'auteur doit spécifier la forme de l'activité, définir les propriétés de la réponse correcte et celles des classes d'erreurs. Le système se chargera alors de la génération automatique de l'activité et de la correction des réponses à l'aide de l'analyseur des mots graphiques. Le système peut ainsi expliquer les erreurs

de l'apprenant et l'orienter vers des modules de l'environnement d'apprentissage pour compléter son savoir-faire et ses connaissances.

7.3.4 Activités de découverte guidée

Dans ce type d'activités, le système assiste l'apprenant durant le déroulement de l'activité. Dans le cadre de l'environnement « *AL-Mu*^{c} *aLLiM* », nous nous sommes basés sur le processus d'étiquetage des textes (cf. chapitre 4) pour proposer une activité de ce type. Le système propose un ensemble de phrases ou un texte à l'apprenant, à partir des textes préalablement étiquetés par le système, qui devront être analysés mot par mot par l'apprenant.

Pour chaque mot qui présente des analyses équivoques, le système propose les différentes analyses possibles et l'apprenant devrait choisir l'une des solutions proposées. Seuls les mots qui présentent une solution unique sont directement proposés aux apprenants. L'apprenant a ainsi l'impression de participer au processus d'étiquetage et qu'il assume une partie de la tâche. Cette activité permet ainsi à l'apprenant de suivre pas à pas le fonctionnement du système et de simuler le processus d'étiquetage. Le système contrôle les choix des apprenants et peut expliquer son fonctionnement : Dans le cas où l'apprenant ne choisirait pas la bonne réponse, le système fournit les raisons du choix de la bonne réponse.

7.3.5 Activités de type ludique

Les jeux peuvent jouer un rôle très important dans l'apprentissage des langues, même si les recherches dans ce domaine sont peu nombreuses. Cependant, il faut mettre en grade contre certains jeux qui se focalisent sur la force d'attraction du média et s'éloignent des principes pédagogiques qui doivent guider leur réalisation : Il ne suffit pas que les apprenants s'amusent, mais il faudrait qu'ils apprennent en jouant.

Ainsi, nous avons réalisé quelques activités de type ludique dont nous exposons à titre d'exemple l'activité du pendu (figure 7-5 ci-dessus), qui permet d'aider l'apprenant à mémoriser les unités lexicales découvertes dans le texte ou traitées dans le dictionnaire personnel. Ce jeu consiste à construire des paires à partir d'un ensemble de 16 cases fermées, où se cachent 8 unités lexicales et leurs 8 illustrations. Pour que l'apprenant gagne dans ce jeu (i.e. ne soit pas pendu), il doit minimiser le nombre d'échecs c'est à dire éviter de cliquer successivement sur deux cases qui ne se correspondant pas.

Cette activité ne demande aucune intervention préalable de l'auteur. Les unités lexicales proposées sont en effet générées automatiquement par le système à partir du dictionnaire personnel ou du texte à étudier.

L'association de l'image au texte dans cette activité favorise la rétention des unités proposées. L'image dans le cadre de la didactique des langues, a en effet quatre fonctions :

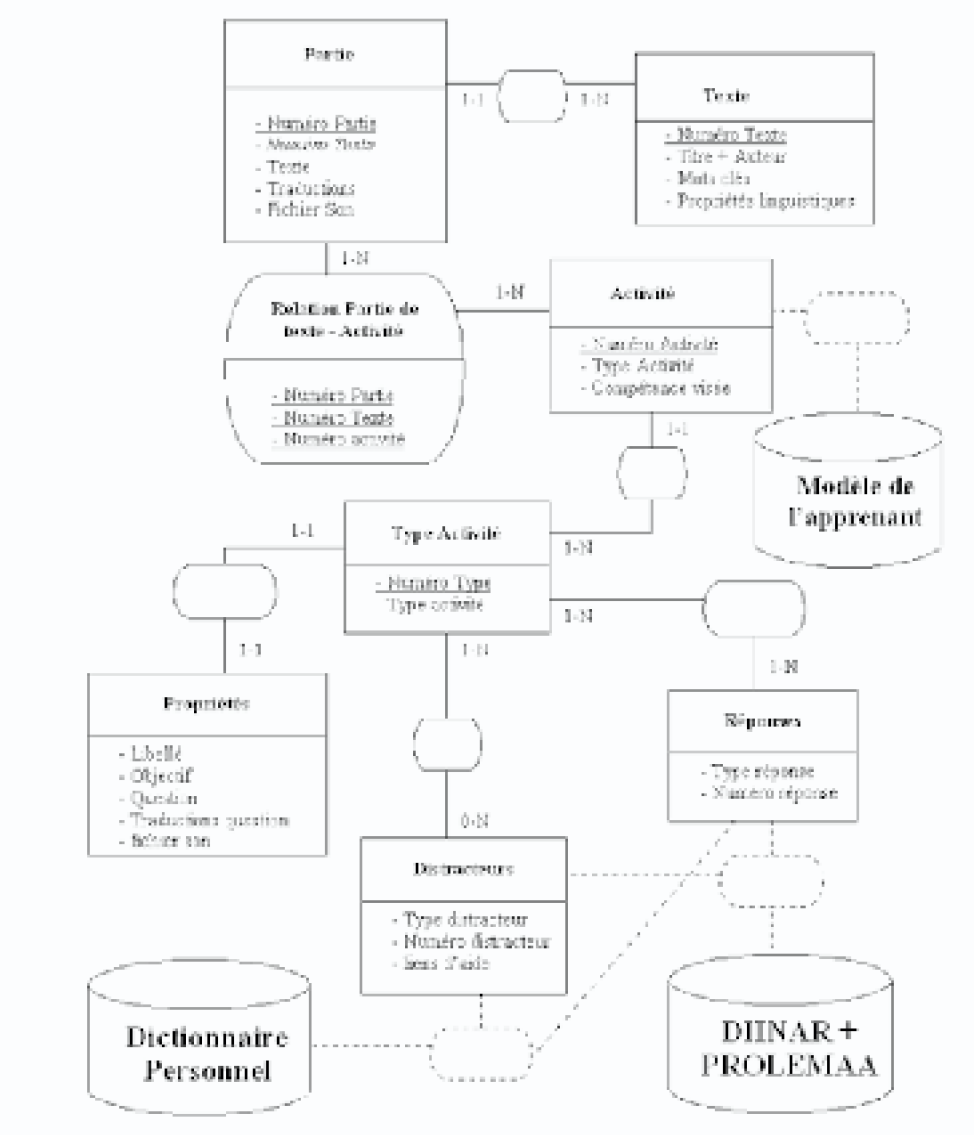
- une fonction psychologique de motivation,
- une fonction d'illustration ou de désignation puisqu'il y a association d'une représentation imagée du terme et de l'objet qu'il désigne,
- une fonction inductrice puisque l'image est assortie d'une invitation à décrire, à raconter,
- une fonction de médiateur intersémiotique, sorte de liaison entre deux systèmes linguistiques, la langue maternelle (L1) et la langue-cible (L2).

7.4 Processus de définition et de génération des activités

Dans cette section, nous allons montrer que cette grande variété d'activités que nous venons de présenter, n'est pas générée aléatoirement. Même si l'utilisation des possibilités de génération aléatoire confère de multiples avantages, il faut se garder de proposer une grande quantité d'exercices choisis au hasard : l'apprenant prendra son temps à traiter de nombreux exercices qui ne lui apporteront rien, parce que trop faciles, trop difficiles ou simplement non pertinents.

Dans le cadre de l'environnement « *AL-Mu* ^{c} *aLLiM* », la génération des activités répond à deux contraintes : d'une part aux spécificités des textes à étudier et d'autre part aux besoins des apprenants en matière d'apprentissage. L'auteur peut, en effet, définir pour chaque partie du texte à étudier un ensemble d'activités statiques pour tester le degré de compréhension du texte par les apprenants. A chacune de ces activités, il doit associer les niveaux des apprenants visés pour que le système n'utilise que les activités les plus appropriées.

D'autre part, l'auteur peut aussi sélectionner des activités génériques que le système adaptera automatiquement aux connaissances des apprenants. Ces activités lexicales ou grammaticales ne peuvent être efficaces, que si elles touchent les compétences précises que l'on veut renforcer chez l'apprenant.



Pour atteindre cet objectif, chaque activité a été définie avec la compétence linguistique visée. Le système se base d'une part sur les lacunes de l'apprenant, à partir des données fournies par le modèle de l'apprenant (cf. chapitre 8), pour choisir l'activité la plus appropriée, et d'autre part, sur les unités lexicales traités dans le dictionnaire personnel (cf. § 9.5), pour choisir les items de l'activité. Par exemple, l'activité de la figure (7-1), a été générée pour un apprenant qui ne maîtrise pas très bien le participe actif des verbes de la forme (IV). Le choix du verbe (غْلِابُي – غَلَّابَ : *BâLa G a – YuBâLi G u*) = (« exagérer »), a été dicté par le fait qu'il était présent dans le texte à étudier et que l'apprenant l'a traité dans son dictionnaire personnel. Ainsi, si la construction précise de ces activités repose sur la fonction aléatoire, celle-ci est paramétrée.

Le module de définition et de génération des activités, se base sur une base de données de type relationnelle, dont nous avons schématisé ses principales entités et relations (figure 7-6). Cette base est en relation avec d'autres bases de données (DIINAR, PROLEMAA, le modèle de l'apprenant, le dictionnaire personnel) que nous avons présenté sur le schéma en pointillés. Nous reviendrons dans le dernier chapitre, sur les

relations entre ces différentes bases de données.

7.5 Conclusion

Nous avons présenté dans ce chapitre, les différents types d'activités utilisés dans le cadre de l'environnement d'apprentissage « *AL-Mu*^{C} *aLLiM* » et leur processus de génération automatique. Nous avons vu que ce processus n'est pas aléatoire, mais tient compte d'un certain nombre de paramètres dont notamment le niveau des connaissances de l'apprenant. Dans le chapitre suivant, nous allons nous intéresser à ces connaissances et voir comment elles sont modélisées et les processus qui les mettent à jour.

Chapitre 8 Modélisation de l'apprenant

« N'aie pas honte de te faire aider ; car tu te proposes de faire ce qui est utile, comme le soldat à l'assaut des murs. Quoi donc ! si tu es boiteux et si tu ne peux monter seul au créneau, mais si c'est possible, grâce à un autre ? »
MARC-AURÈLE

8.1 Introduction

Le besoin de disposer d'une représentation des connaissances d'un apprenant est une condition nécessaire pour la personnalisation des activités de l'environnement d'apprentissage. Cette nécessité nous a conduit à induire un modèle de l'apprenant (MA) en s'appuyant sur son comportement observable. Ce modèle doit d'une part sauvegarder toutes les informations résultantes des agissements de l'apprenant et les exploiter pour adapter adéquatement l'enseignement au profil de l'apprenant d'autre part.

En effet, si beaucoup d'environnements d'apprentissage possèdent un modèle de l'apprenant, peu de réalisations se sont intéressées à l'aspect d'évolution ou de synthèse de ce modèle au fur et à mesure de l'interaction apprenant-système. Dans ces systèmes, la gestion des contradictions dans les réponses de l'apprenant et la prise en compte de la fréquence des réponses correctes par rapport aux réponses erronées pour un même type de questions sont, en règle générale, ignorées.

Ce chapitre se divise en deux grandes parties. Dans la première partie, nous décrirons les différents types et contenus des MA existants et nous choisirons parmi ces modèles celui qui répond au mieux aux connaissances de notre environnement. Dans la seconde partie, nous présenterons le modèle de l'apprenant réalisé, à savoir non seulement son contenu, mais également le diagnostiqueur qui est l'ensemble des processus qui élaborent et mettent à jour ce MA.

8.2 Typologie des modèles de l'apprenant

Le modèle de l'apprenant est une structure de données, au sens informatique, qui caractérise, pour l'environnement d'apprentissage, l'état d'un sous-ensemble des connaissances de l'apprenant du point de vue du système.

Il va se définir par l'écart entre les propres connaissances (supposées) de l'apprenant et les connaissances cibles, enjeu de l'apprentissage, telles qu'elles sont représentées dans le système. La façon de concevoir cet écart conduit à distinguer deux grandes classes de modèles :

- **Les modèles d'expertise partielle ou de superposition :** (overlay model, Goldstein et Carr, 1977), dans lesquels la connaissance de l'apprenant n'est qu'un sous-ensemble de la connaissance cible. L'idée sous-jacente à ce type de modèle est que l'apprenant présente des lacunes ou des connaissances encore mal assurées, en quelque sorte des faiblesses, qu'il s'agit d'identifier pour lui permettre de progresser. L'objectif du système d'apprentissage est alors de compléter les connaissances de l'apprenant pour qu'il acquière l'ensemble des connaissances précisées dans le modèle.
- **Les modèles différentiels :** (Wenger, 1987), qui incorporent des " connaissances fausses ", correspondant à des perturbations des connaissances expertes ou des préconceptions erronées. En effet, des études montrent que de nombreuses erreurs ne sont pas dues à un comportement erratique des apprenants, mais à l'application correcte de procédures fausses. Pour élaborer un modèle des connaissances des apprenants, il faut prendre en compte ces erreurs de type systématique, que les chercheurs vont désigner par le terme " bogue " (*bug*).

Alors qu'un modèle d'expertise partielle invite à des stratégies d'enseignement centrées sur le fait de combler les lacunes de l'apprenant, les modèles différentiels vont induire des stratégies basées sur la remédiation. Dans le domaine de l'ALAO d'une langue seconde, l'analyse des erreurs de l'apprenant s'avère être une tâche très ardue. Plusieurs processus cognitifs peuvent être proposés comme sources possibles d'explications des erreurs, dont l'interférence avec la langue maternelle.

8.3 Contenu d'un modèle de l'apprenant

Deux méthodologies peuvent être suivies pour déterminer le contenu du modèle de l'apprenant (Danna, 1997). L'approche fonctionnelle vise à décrire ce contenu en termes des fonctions que le MA doit assurer pour permettre l'individualisation de l'apprentissage. L'approche extensionnelle, s'appuyant éventuellement sur les résultats obtenus lors des études de la première méthode, tente d'énumérer les informations qui doivent apparaître dans le MA.

Comme les informations sur l'apprenant sont principalement destinées aux modules pédagogiques du système (choix de l'activité, choix de l'aide, choix du texte à étudier, etc.), il semble nécessaire de s'intéresser particulièrement aux fonctions que ces modules requièrent de la part du MA. Self (Self, 1987) énumère six fonctions que le MA doit pouvoir assumer que nous résumons brièvement :

1. Fonction corrective : le MA doit pouvoir servir à corriger les erreurs de l'apprenant. Cette fonction doit permettre :

- une *prise de conscience* des erreurs : une description en langage naturel peut être associée à l'erreur, ou à la connaissance incorrecte (l'erreur profonde) plutôt qu'à l'erreur observée (l'erreur de surface), et être montrée à l'apprenant. Le message peut être instancié en fonction du contexte ;
- une *correction indirecte* : on montre à l'apprenant seulement un résumé ou une partie de la connaissance correcte qu'on veut lui enseigner et pour laquelle le système a des preuves du manque de maîtrise de l'apprenant ;
- une *correction directe* : on donne directement la bonne solution à l'apprenant. Cela peut être fait quand l'erreur est primitive (en d'autres termes, quand l'erreur ne concerne que des connaissances factuelles), ou quand le système ne peut déterminer la connaissance erronée mais qu'il sait quelle est la connaissance qui aurait dû être employée, ou lorsque les techniques plus indirectes ont toutes échoué ;
- la *génération de contre-exemples* : Il existe deux types de contre-exemples : ceux qui montrent explicitement à l'apprenant qu'il se trompe et ceux qui amènent l'apprenant à se rendre compte de ses erreurs ;
- l'*analyse des étapes* : le système devine les étapes de l'apprenant. Si les inférences sont suffisamment fiables, le système laisse l'apprenant réfléchir sur son processus de résolution ;
- la *rétrospection* : elle permet de traiter les erreurs profondes telles que les interférences entre deux domaines (par exemple la langue maternelle sur la langue enseignée), ou bien encore entre deux solutions proposées par l'apprenant à différents moments.
- la *génération d'activités* du même type : on peut utiliser une analyse analogue à celle

servant pour la génération de contre-exemples, ou consulter une banque d'activités ;

- le *rappel* : quand on craint que l'apprenant ait oublié, au moins en partie, certaines connaissances qu'il avait maîtrisées antérieurement, le tuteur peut rappeler à l'apprenant certaines informations.

Fonction élaborative : le MA doit permettre d'augmenter l'ensemble des connaissances correctes de l'apprenant, i.e. de choisir la prochaine information à lui enseigner. Ce choix peut être : 1.

- *basé sur le curriculum* : si la connaissance à enseigner et la connaissance de l'apprenant sont toutes deux représentées sous forme d'un curriculum, une comparaison de ces deux structures permet de choisir le prochain thème à enseigner.
- fait en *comparant les réponses de l'apprenant et du système* : on détermine le thème suivant en comparant directement les réponses du système et les réponses de l'apprenant que le MA permet d'inférer pour le problème en cours, sans référence à aucun curriculum.
- fait en fonction des résultats d'une *analyse interne de la connaissance de l'apprenant* : on détecte des déficiences structurelles, des redondances, etc.
- *laissé à l'apprenant* : parmi une liste d'activités choisies en fonction du modèle de l'apprenant et du curriculum.

Fonction stratégique :le MA est utilisé pour mettre au point la stratégie d'enseignement suivie par le système. On peut par exemple, être amené à changer de plan pédagogique lorsque l'apprenant n'arrive pas à le suivre. 1.

Fonction diagnostique : le MA doit décrire les ambiguïtés qu'il contient. Deux niveaux d'ambiguïtés existent : 2.

- *ambiguïté du MA :* quand le MA est ambigu, les informations qu'il décrit doivent permettre de choisir entre les hypothèses concernant l'état cognitif de l'apprenant.
- *ambiguïté des connaissances de l'apprenant* : quand le MA indique que l'apprenant possède des doutes sur certaines connaissances, le système peut amener l'apprenant à lever cette ambiguïté.

Fonction prédictive : le MA peut être utilisé par le système pour prédire le comportement de l'apprenant face à un problème. La prédiction peut porter sur : 1.

- *la performance de l'apprenant :* on peut alors utiliser les différences entre la prédiction de la réponse de l'apprenant et celle donnée effectivement par l'apprenant afin de focaliser l'analyse de sa réponse sur des informations nouvelles.
- *Sur les effets des actions didactiques :* si le modèle contient les procédures d'apprentissage de l'apprenant, ces procédures peuvent être appliquées pour chaque action didactique envisagée afin de sélectionner la plus bénéfique.

1. Fonction évaluative : les informations contenues dans le MA doivent être représentatives de l'apprenant et du système, afin de pouvoir servir à évaluer l'apprenant avec le système.

Une des remarques qui nous paraît évidente est, qu'il est difficile d'intégrer toutes ces informations afin d'obtenir des MA reflétant exactement et précisément l'apprenant. En effet, la nature des informations touche des domaines de natures diverses, dont celles sur l'état cognitif et celles sur des traits psychologiques de l'apprenant. Ces dernières sont non seulement difficiles à déterminer, mais aussi à faire évoluer en cours d'interaction avec l'apprenant et à utiliser pour améliorer la situation d'enseignement.

Certains chercheurs tentent d'attirer l'attention sur le fait que cette tâche est éventuellement inutile d'un point de vue pratique (Self, 1994). Ces chercheurs mettent l'accent sur la finalité informatique du MA, plutôt que la finalité psychologique. Selon eux, la qualité d'un MA doit être jugée principalement par rapport à son adéquation avec les autres modules du système, et notamment avec le module pédagogique. Les données que contient le MA ne sont pertinentes que si elles sont utilisables par le reste du système.

S'appuyant sur ce raisonnement, nous avons élaboré un MA restreint aux compétences linguistiques de l'environnement d'apprentissage « *AL-Mu*^{c} *aLLiM* » et qui répond surtout aux besoins des modules pédagogiques du système (générateur d'activités, choix du texte).

8.4 Élaboration du modèle de l'apprenant

Pour élaborer le MA, nous avons effectué une première expérimentation, à l'aide d'une série d'activités grammaticales à réponse ouverte du niveau du mot graphique arabe (Zaafrani, 97). Cette expérimentation a montré que dans ce cas précis, le recensement et l'étude des bogues des apprenants ne permettaient pas de déterminer avec certitude les types d'intervention à effectuer. On s'attendait un peu à ce résultat, étant donné le nombre important des corrections possibles pour chaque mot erroné. Il n'y a qu'à voir la diversité des propositions des correcteurs orthographiques pour une erreur donnée, pour se rendre compte que toute interprétation d'une erreur serait hasardeuse.

Et ceci est également valable pour les propositions reconnues par le système, mais qui ne correspondent pas aux réponses correctes. Par exemple, supposons que l'apprenant répond dans une activité à réponse ouverte par la forme « بتكتب : TKTB » = (« tu écris »), au lieu de « بتكتن : NKTB » = (« nous écrivons »). L'analyseur du mot graphique détecterait que l'apprenant aurait utilisé un pronom préfixe erroné et on pourrait déduire que l'apprenant maîtrise mal la conjugaison des verbes simples à l'inaccompli indicatif. Le système produirait un message d'erreur dans ce sens, qui pourrait s'avérer inadapté si l'apprenant aurait simplement fait une faute de frappe. Le risque serait alors de voir se développer chez l'apprenant une méfiance vis à vis du système, à cause d'une suite de commentaires inappropriés aux erreurs commises.

En résumé, cette expérimentation (zaafrani, 1997) a montré qu'étiqueter un comportement dans le cadre de l'apprentissage grammatical de l'arabe, ne donne pas *ipso facto* les informations nécessaires pour expliquer le comportement de l'apprenant et choisir une thérapeutique adaptée. Nous avons par conséquent opté pour un MA de type expertise partielle (cf. § 8.2) et à limiter le diagnostic aux erreurs qui peuvent être analysées avec certitude. Nous avons par ailleurs réalisé des activités à réponses fermées (cf. § 7.3.2), pour à la fois faciliter l'expression de l'apprenant et la contraindre.

Le modèle de l'apprenant proposé ici, vise à ce que l'apprenant possède une connaissance exhaustive des différents aspects linguistiques de la langue arabe. La figure (8-1) présente l'interface principale d'initialisation du MA, qui permet à l'apprenant de déclarer lui-même ses lacunes au système : L'apprenant classifie les principales difficultés qu'il rencontre par ordre décroissant et donne un jugement approximatif sur ses propres compétences.

Les compétences linguistiques traitées par le MA ont été divisées en huit catégories

distinctes : connaissances du vocabulaire = (« معارف في المعجم والمفردات »), compétences de conjugaison («خبرات في التصريف »), compétences grammaticales (« خبرات في النحو »), compétences morphographiques (« خبرات في الخطّ والرسم »), compétences de compréhension écrite (« خبرات في فهم المكتوب »), compétences de compréhension orale (« خبرات في فهم المنطوق »), compétences de production écrite (« خبرات في التعبير الكتابيّ ») et compétences de production orale (« خبرات في التعبير الشفاهيّ ») (figure 8-1).

Chaque catégorie est à son tour divisée en d'autres sous-catégories et ainsi de suite jusqu'à obtenir des catégories élémentaires correspondant à des activités du système. Par exemple, les compétences de conjugaison ont été divisées en deux sous-catégories : les compétences relatives à la conjugaison des verbes et celles relatives à la dérivation des noms et des déverbaux. Les premières sont divisées en d'autres sous catégories correspondant aux propriétés morphologiques des unités lexicales verbales concernées (racines tri-consonantiques et quadri-consonantiques, normales et anomales, verbes simples et/ou augmentés, etc.) et aux aspects/modes de conjugaison des verbes.

Les secondes sont divisées en des sous catégories correspondant au type du déverbal ou au genre et au nombre de l'unité lexicale concernée (Masculin/féminin, singulier/duel/pluriel) (Voir figure 8-2). A Chaque catégorie élémentaire correspond un ensemble d'activités, permettant d'évaluer le comportement de l'apprenant. La figure (8-3) montre un second exemple de cette interface relative à l'initialisation des sous catégories des compétences morphographiques.

Pour chaque catégorie élémentaire, le système attribue une évaluation qui reflète le niveau de l'apprenant à un instant t. L'évaluation de la catégorie principale est obtenue en faisant la moyenne des évaluations des sous-catégories qui la composent.

Lors de l'initialisation du MA, l'apprenant peut donc s'auto évaluer en affectant à chaque catégorie du MA, l'un des cinq attributs suivants : débutant = (« مبتدئ »), faible = (« ضعيف »), moyen = (« متوسّط »), bien = (« جيد »), excellent = (« ممتاز »). Le système se basant sur ces informations, initialise le MA qui sera par la suite modifié selon le comportement de l'apprenant.

Le contenu du MA étant défini, nous examinerons dans les sections suivantes, le système de modélisation de l'apprenant qui permet la mise à jour du MA, au fur et à mesure de l'interaction apprenant-système.

8.5 Caractéristiques d'un Système de Modélisation de l'Apprenant

Indépendamment du domaine enseigné, un système de modélisation de l'apprenant (SMA) doit posséder trois principales caractéristiques pour construire des MA fiables et précis : il doit être capable de suivre l'évolution de l'état cognitif de l'apprenant, de prendre en compte les éventuelles contradictions que celui-ci peut avoir parmi ses connaissances et de gérer l'incertitude relative aux informations acquises sur lui (Danna, 97) :

- Évolution de la connaissance : La prise en compte de l'évolution cognitive de l'apprenant par le SMA est primordiale. De par le fait que l'apprenant est placé en situation d'apprentissage, sa connaissance est censée évoluer dans le temps. L'apprenant peut donner dans un premier une réponse incorrecte puis une réponse correcte et vice versa. Le SMA doit être capable par conséquent, de suivre un raisonnement *révisable*.
- Contradictions : Une deuxième caractéristique de la situation d'apprentissage est que l'apprenant peut avoir des contradictions en tête sur certains concepts. Dans le cas où le SMA devient conscient de ses contradictions, il peut gérer un ensemble d'hypothèses à propos de la connaissance dont il n'est pas sûr et attendre des informations ultérieures avant d'agir. Le SMA doit être capable par conséquent, de représenter et de raisonner sur des connaissances *inconsistantes*.
- Incertitudes du diagnostic : La dernière caractéristique concerne les connaissances du MA qui peuvent être erronées. La cause de ces erreurs est généralement due à un diagnostic incertain. L'apprenant peut utiliser les concepts du système pour aboutir à une forme de surface correcte, comme il peut s'appuyer sur d'autres concepts. Pour des raisons pédagogiques, le diagnostic optimiste peut être initialement préféré. L'apprenant est alors considéré avoir utilisé les concepts maîtrisés par le module expert du système. Si une interaction ultérieure prouve que l'apprenant ne maîtrise pas certains de ces concepts, le premier diagnostic doit être changé. C'est pourquoi le SMA doit être capable de suivre un raisonnement *hypothétique*.

Ainsi, sont définis les principes qu'un bon SMA doit intégrer dans son analyse du comportement de l'apprenant. Faut-il insister encore sur le rôle des activités qui doivent minimiser les incertitudes sur les connaissances de l'apprenant. Les activités à réponses fermées permettent particulièrement une analyse stricte des réponses et facilitent ainsi l'interprétation des raisonnements cognitifs de l'apprenant, ce qui permettra l'obtention d'un MA fiable et précis.

8.6 Formalisation du Système de Modélisation de l'Apprenant

Comme nous l'avons déjà annoncé (cf. § 8.4), nous avons opté pour une modélisation de l'apprenant à partir de ses comportements sans chercher les raisons qui sous-entendent ses comportements. L'apprentissage lexical et grammatical met, en effet, davantage des connaissances factuelles qui s'apprêtent à ce genre de raisonnement.

8.6.1 Représentation des connaissances

Pour représenter la connaissance comportementale de l'apprenant, nous avons choisi la logique probabiliste comme formalisme. Cette technique de représentation formelle, utilisée par Danna (1997), permet comme nous allons le montrer d'implanter un raisonnement à la fois révisable, paraconsistant et hypothétique.

En utilisant la logique probabiliste, le modèle comportemental est implanté sous la forme d'un ensemble d'*entités comportementales*. Chaque élément de cet ensemble dénote une association entre un type de comportement affiché par l'apprenant et un coefficient de certitude relatif à ce comportement. Un *type de comportement* synthétise les associations faites par l'apprenant entre un type de question et un type de réponse. Le *coefficient de certitude* correspond à la probabilité que l'apprenant se comporte de la façon indiquée par le type de comportement associé, plutôt que suivant un autre type de comportement modélisé. Le calcul de ce coefficient est donné par la théorie des probabilités qui requiert que la somme des probabilités associées aux propositions contradictoires du langage soit égale à 1.

Dans notre cas, une proposition dénote un événement qui est l'association par l'apprenant d'un type de comportement à un type de question. Tous les évènements portant sur le même type de question sont mutuellement exclusifs puisque chacun sous-entend que l'apprenant affiche un certain type de comportement, à un instant t donné, en face d'une certaine catégorie de question. Le coefficient est donc calculé en divisant le poids associé au type de comportement considéré (noté α) par la somme des poids associés aux autres types de comportement pour le même type de question : , où n est le nombre de types de comportement représentant la connaissance de l'apprenant et la propriété associée à l'entité comportementale *i* à l'instant *t*.

8.6.2 Acquisition et synthèse des informations comportementales de l'apprenant

Le SMA doit d'abord acquérir des informations en analysant le comportement de l'apprenant, puis synthétiser ces nouvelles informations avec celles acquises

antérieurement. Lorsque l'apprenant initialise le MA, on attribue au comportement correct un coefficient qui varie selon l'évaluation choisie : débutant : 0,1 / faible : 0,3 / moyen : 0,5 / bien : 0,7 / excellent : 0,9. Afin de respecter la contrainte portant sur l'unité des probabilités, on attribue également un second coefficient égal au complément à un du premier coefficient sélectionné (i.e. c'est à dire 0,9 si l'apprenant s'est jugé comme un débutant), correspondant à un comportement d'un apprenant n'ayant aucune connaissance du système expert.

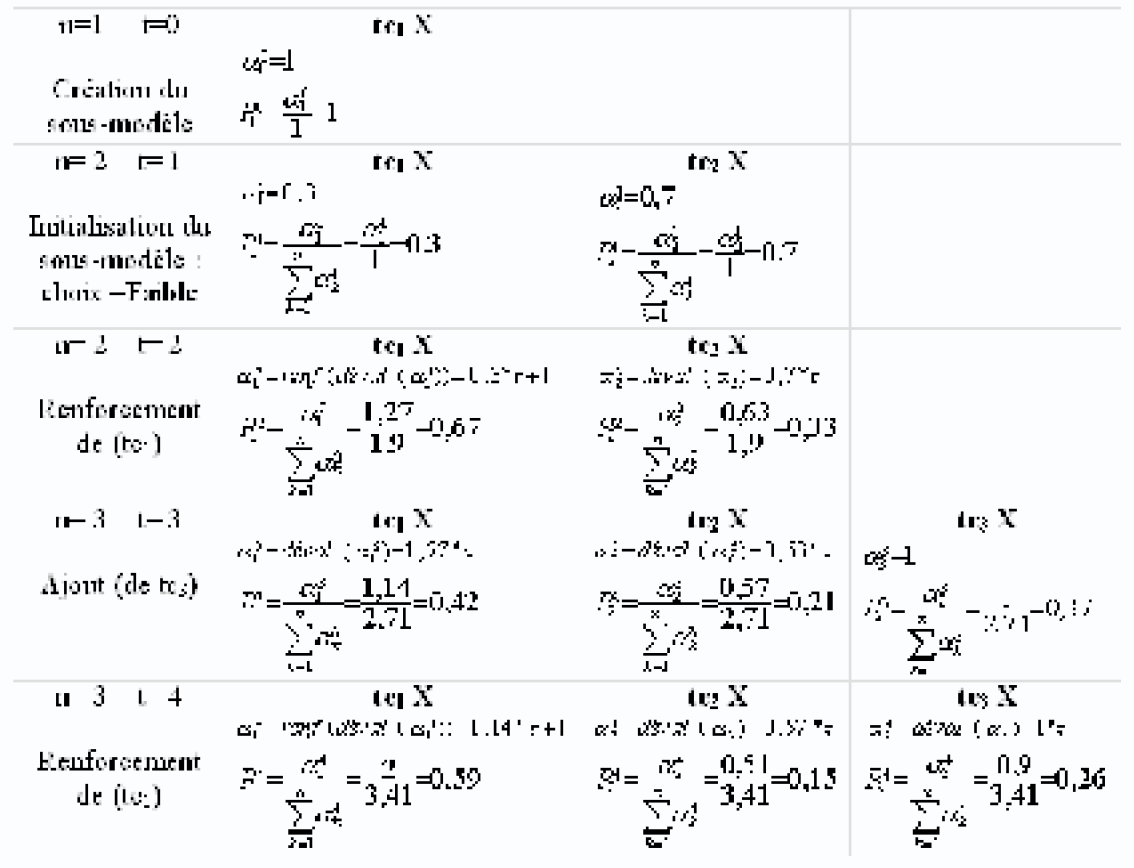
Les modifications sont effectuées après l'analyse de chaque réponse de l'apprenant. Elles consistent à faire évoluer les entités comportementales correspondant à l'activité concernée ainsi que les coefficients de certitude. A chaque nouvelle réponse de l'apprenant, on associe une nouvelle entité comportementale. Les entités comportementales d'une activité sont regroupées sous formes de sous-modèles.

Les différentes modifications à apporter sont effectuées en deux phases. Tout d'abord, les poids associés aux entités comportementales sont modifiés. Les nouvelles probabilités sont ensuite calculées. A la première étape, l'algorithme suivant est appliqué :

- En premier lieu, une dévaluation est effectuée pour tous les poids qui sont associés à chaque entité comportementale, selon le même principe utilisé par Danna (1997). Cette opération est effectuée via une multiplication par une valeur (inférieure à 1). La valeur en est donc donnée par . Cette diminution des coefficients associés aux anciens types de comportements (ceux qui dénotent que l'apprenant a donné un autre type de réponse pour la même catégorie de question que celle considérée), implémente la notion de relativisation temporelle. Nous reprenons la même évaluation que celle de Danna (1997), aura la valeur 0,9.

- Quand la nouvelle entité comportementale (la $n^{ième}$) n'est pas encore modélisée dans le sous-modèle, soit parce que le sous-modèle est encore vide (auquel cas *n* vaut 1), soit parce que l'entité dénote un type de comportement non encore montré par l'apprenant ($n \geq 2$), il faut l'y ajouter. Le poids qui est associé à cette nouvelle entité est 1.
- Si cette entité appartient déjà au sous-modèle (la $i^{ème}$ par exemple), son poids est recalculé. Il doit être augmenté afin de respecter la théorie des probabilités. Nous utilisons pour cela la fonction de renforcement **renf** (renf(x) = x + 1). On obtient donc .

La figure (8-4) ci-dessous montre un exemple d'évolution des coefficients lorsque de nouvelles entités sont ajoutées au modèle.



La seconde étape, visant à mettre à jour les probabilités à partir de nouveaux poids, correspond alors simplement à la répartition de la somme des poids de manière à vérifier la contrainte portant sur l'unité des probabilités.

8.6.3 Exemple de construction de modèle comportemental

Dans cette section, on se basera sur un exemple d'activité à réponses fermées présenté dans la section (§ 7.3.2) figure (7-1), et qui est associé au sous-modèle comportemental relatif au participe actif des verbes de la forme (IV). A travers une série de réponses de l'apprenant, à partir de la même activité moyennant un changement de la racine et de l'ordre des réponses, on va décrire dans ce qui suit l'évolution de ce sous-modèle comportemental.

Lors de l'initialisation de ce sous-modèle, l'apprenant choisit un niveau faible. A l'instant **t=1**, le sous-modèle a par conséquent deux types de comportements : un premier

comportement **tc $_1$** correspondant à un apprenant maîtrisant parfaitement la dérivation de ce type de déverbal pour cette catégorie de verbe avec une probabilité **= 0,3** et un second comportement **tc $_2$** correspondant à un apprenant ignorant les règles dérivation de ce type de participe actif avec une probabilité **= 0,7**.

A l'instant **t=2**, l'apprenant choisit la réponse correcte. Après les calculs du renforcement de la bonne réponse et des dévaluations, on obtiendra pour le comportement **tc $_1$** une probabilité **= 0,67** et pour le comportement **tc $_2$** une probabilité **= 0,33**.

A l'instant **t=3**, l'apprenant choisit un le distracteur du participe passif. On ajoute un nouveau comportement **tc $_2$** correspondant à un comportement confondant les participes actifs et passifs de ce type de verbe. Après les calculs de dévaluations des autres comportements, on obtiendra pour le comportement **tc $_1$** une probabilité **= 0,42**, pour le comportement **tc $_2$** une probabilité **= 0,21** et enfin pour le nouveau comportement **tc $_3$** une probabilité **= 0,37**.

A l'instant **t=4**, l'apprenant choisit une nouvelle fois la réponse correcte. Après les calculs du renforcement de la bonne réponse et des dévaluations, on obtiendra pour le comportement **tc $_1$** une probabilité **= 0,59**, pour le comportement **tc $_2$** une probabilité **= 0,15** et enfin pour le comportement **tc $_3$** une probabilité **= 0,26**.

La figure (8-4), présente avec plus de détails les calculs correspondant à l'évolution de ce sous-modèle comportemental. D'autres activités peuvent être associées à ce sous-modèle. Le deuxième comportement correspondant généralement à un apprenant ignorant les règles du sous-modèle, est renforcé lorsque le système n'est pas en mesure de déterminer la nature de l'erreur de l'apprenant dans une activité à réponse ouverte.

8.7 Conclusion

Les points forts du SMA réalisé résident, d'une part, dans son contenu qui englobe toutes les compétences de la langue arabe que l'apprenant doit maîtriser et, d'autre part, dans l'utilisation de mécanismes permettant de suivre l'évolution de l'état cognitif de l'apprenant, de prendre en compte ses éventuelles contradictions et de gérer l'incertitude relative aux informations acquises sur lui.

Dans la mesure où les informations qu'il gère sont fiables et précises, le SMA permet à l'environnement d'individualiser la situation d'apprentissage. Chaque activité doit être directement liée à un sous-modèle comportemental du MA, pour qu'elle puisse être utilisée afin de corriger les comportements de l'apprenant le cas échant. Dans le dernier chapitre, nous nous décrivons cette relation Modèle Apprenant - Module des activités en la plaçant dans l'architecture générale de l'environnement « *AL-Mu ^{c} aLLiM* ».

Chapitre 9 Architecture de l'environnement « *AL-Mu c aLLiM* »

« Un noble philosophe a dit de l'architecture qu'elle est une musique pétrifiée, et ce mot a dû exciter plus d'un sourire d'incrédulité. Nous ne croyons pouvoir mieux reproduire cette belle pensée qu'en appelant l'architecture une musique muette. » Johann Wolfgang von GOETHE

9.1 Introduction

Les principaux éléments de l'environnement ont été présentés dans les chapitres précédents. Ils ont illustré les réponses que nous apportions aux différents problèmes abordés au cours de ce travail de thèse. Il reste quelques autres modules, non moins importants, que nous décrivons dans ce chapitre.

Afin d'avoir une meilleure vue d'ensemble, nous dressons tout d'abord un récapitulatif de l'environnement. Nous passons en revue ensuite les quatre modules qui n'ont pas été encore abordés : le module *choix du texte*, le module *compréhension de texte,* le module *gestion du dictionnaire personnel* de l'apprenant et le *module de l'enseignant* qui permet d'enrichir et de mettre à jour les ressources de l'environnement. Nous évaluerons enfin le système à partir de l'expérimentation qui a été menée avec les étudiants de l'université

Lyon 2 (cf. § 5.5).

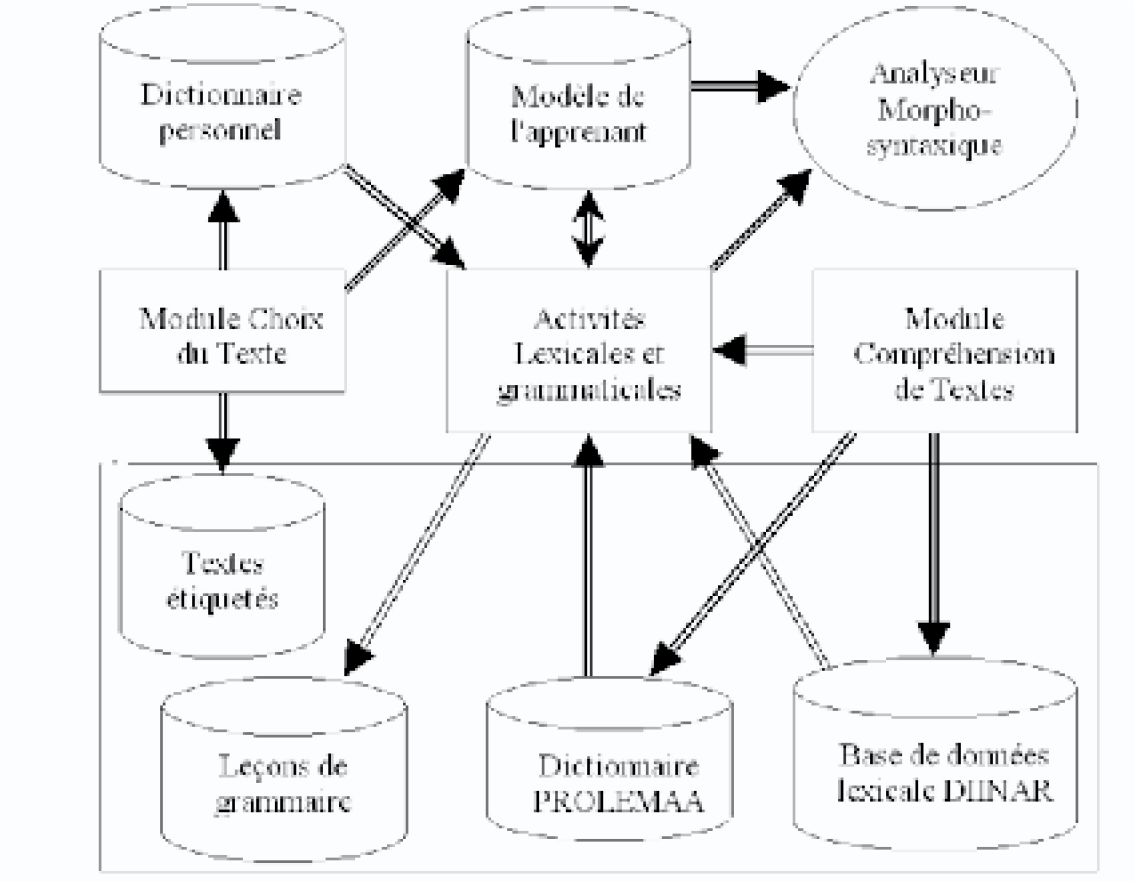
9.2 Présentation de l'environnement

La conception de l'architecture de l'environnement d'apprentissage « *AL-Mu* ^{c} *aLLiM* », a été donc inspirée de l'expérience d'enseignement de l'arabe langue seconde à l'université lumière (Lyon 2). Cet environnement a été conçu afin de compléter le travail de l'enseignant en classe et d'aider les apprenants à enrichir leu vocabulaire et à maîtriser la grammaire d'un façon autonome.

Nous avons ainsi construit un environnement à partir des ressources élaborées :

- Le **corpus de textes étiquetés** qui vont servir de base de travail à l'apprenant.
- La **base de données lexicale DIINAR**, d'où sont extraites les propriétés morphologiques et les structures syntaxiques enseignées aux apprenants.
- Le **dictionnaire électronique PROLEMAA** qui contient les ressources lexicales : propriétés sémantiques, équivalents en langue cible et associations entre mots.

A ces différentes ressources nous avons ajouté une base de données permettant d'accueillir les **leçons de grammaire** définies par l'enseignant et une **aide en ligne** (figure 9-3) accessible à partir de n'importe quel module de l'environnement.



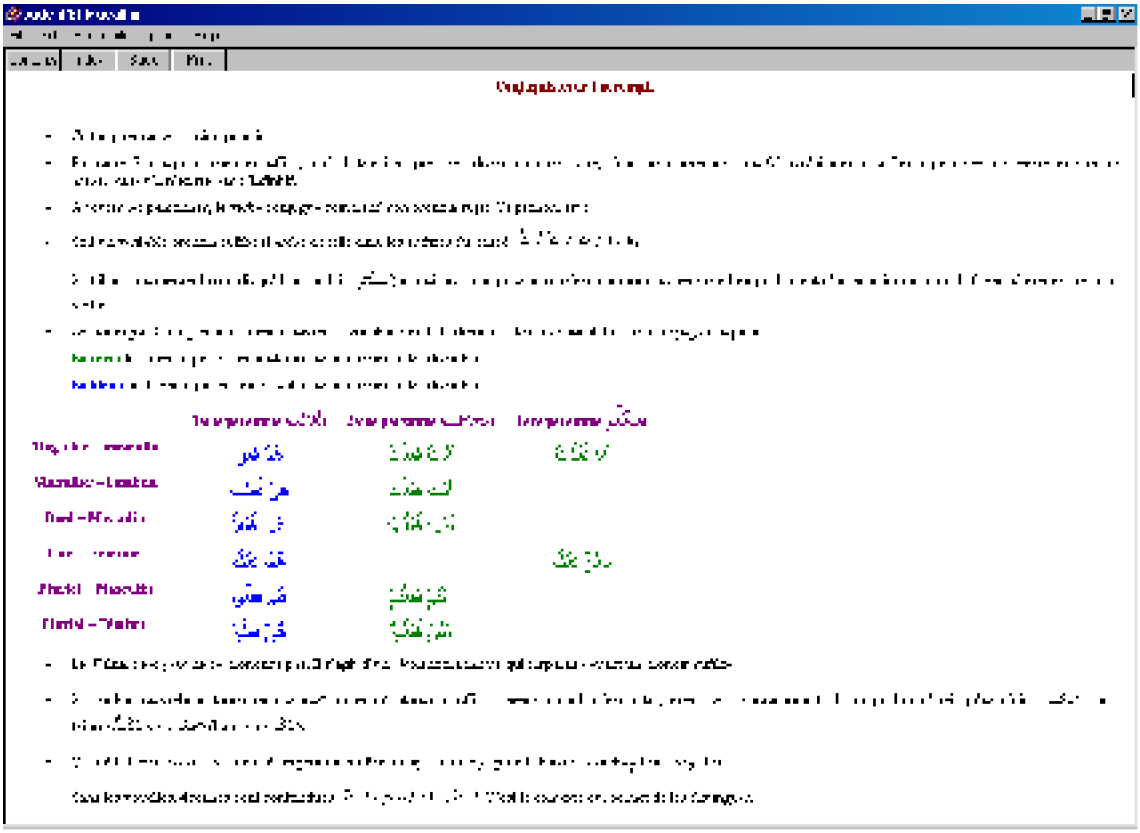
Autour de ces ressources nous avons réalisé cinq principaux modules (figure 9-1):

- Un module **choix du texte** qui donne la possibilité à l'apprenant de choisir un texte ou de rechercher un texte à partir de certains critères qu'il définit.
- Un module **compréhension de texte**, où l'apprenant étudie le texte choisi.
- Un module **activités lexicales et grammaticales** qui lui permet de pratiquer des exercices ne se rapportant pas à un texte particulier.
- Un module de gestion du **modèle de l'apprenant** qui contient des informations sur ses compétences et sur les tâches qu'il a effectuées.
- Un module de gestion du **dictionnaire personnel** qui permet à l'apprenant d'organiser lui-même les informations sur le vocabulaire qu'il a étudié.

Plusieurs scénarios d'utilisation du système sont proposés à l'apprenant. L'écran principal de l'environnement permet d'accéder directement à ces différents scénarios (figure 9-2).

Le principal scénario est celui de la *compréhension de texte*. L'apprenant peut réaliser plusieurs tâches : écouter et/ou lire un texte sélectionné, s'attarder sur les significations et les propriétés morphologiques et syntaxiques des mots, s'exercer sur des activités générées par le système ou enrichir son dictionnaire personnel par les mots qu'il juge intéressants de retenir.

Ces différentes tâches sont directement accessibles à partir de la page d'accueil de l'environnement (figure 9-2). L'apprenant peut par exemple avoir besoin de consulter le *dictionnaire général* ou *une leçon de grammaire*, après son travail en classe ou au cours de la lecture d'un texte sur support papier. Il peut aussi avoir besoin de noter un nouveau mot dans son *dictionnaire personnel*, sans avoir à le reproduire d'un texte de l'environnement.



9.3 Le module de choix du texte

Dans l'environnement d'apprentissage « *AL-Mu c aLLiM* », l'apprenant peut soit demander au système un texte correspondant à son profil, soit choisir lui-même son texte de travail.

Dans le premier cas, le système utilise les informations du modèle de l'apprenant et du dictionnaire personnel pour la sélection du texte. En effet, à chaque texte étiqueté du système, est associé une entité recensant les fréquences des différentes catégories lexicales, propriétés morphologiques et syntaxiques et lemmes du texte. Le système choisit alors le texte qui présente le nombre le plus important de la propriété linguistique que l'apprenant ne maîtrise pas et/ou celui qui présente le nombre le plus important des lemmes travaillés par l'apprenant dans son dictionnaire personnel.

Dans ce second cas, nous proposons à l'apprenant deux interfaces qui lui permettent de définir ses critères de choix et assurer la recherche dans le corpus textuel.

La première interface permet à l'apprenant d'effectuer une recherche documentaire dans le corpus textuel (figure 9-4). Nous avons retenu trois critères de recherche : Auteur, titre et mots clés. La recherche à partir de ce dernier champ est la plus intéressante puisqu'elle permet à l'apprenant d'effectuer une recherche à partir d'un sujet donné. Tous les textes du corpus sont préalablement indexés manuellement par le biais de l'interface de saisie "ARTINDEX", que nous avons développé dans le cadre du projet DIINAR-MBC (Voir Annexe 3). Le programme de recherche, utilise le même algorithme que celui qui a été utilisé dans le cadre du système CATHIE (CATalog Hypertextuel Interactif et Enrichi) et qui supporte des critères de filtrage et de reformulation de la requête (Zaafrani & Ihadjadene & Bouché, 2000).

La seconde interface, la plus utilisée par les apprenants, permet d'effectuer des recherches à partir de la spécification du type de texte à rechercher, i.e. son domaine

lexical. L'apprenant peut demander par exemple, un texte qui lui permet d'examiner un certain type de déverbal ou un certain aspect de conjugaison. En retour, le système propose parmi les textes qui n'ont pas été encore traités par l'apprenant, ceux qui présentent le plus la catégorie ou la propriété demandée.

9.4 Le module de compréhension de texte

Le module *compréhension de texte* constitue le principal module de l'environnement. L'interface correspondante se divise en deux parties distinctes : une partie fixe qui présente le texte à étudier et une partie qui permet d'effectuer des activités (figure 9-6) ou d'accéder au dictionnaire général (figure 9-5). Ce module a un double objectif : la compréhension du texte et la mémorisation de son vocabulaire.

Pour faciliter la compréhension du texte, l'interface de ce module permet d'accéder à deux fonctionnalités : La traduction instantanée et intégrale du texte dans la langue cible, l'écoute du texte et la consultation des significations et des propriétés du vocabulaire du texte. La première et la seconde fonctionnalité sont accessibles par un simple clic sur respectivement les boutons (الترجمة = Traduction) et (الإستماع = Ecoute) alors que pour accéder aux propriétés d'un mot, il suffit de le sélectionner dans le texte (figure 9-5). On résout ainsi le problème de l'accès lexical, qui présente le principal obstacle que l'apprenant rencontre lors du décodage d'un texte.

La mémorisation du lexique se fait par la production d'un certain nombre d'activités lexicales et grammaticales relatives au texte. Ces activités dont une partie est statique et définie par l'auteur ou l'enseignant (cf. § 9.6), permettent à l'apprenant d'étudier le texte et de s'exercer sur le nouveau lexique. Chaque activité est évaluée par le système et une note globale de l'apprenant pour la session en cours est retournée après chaque activité.

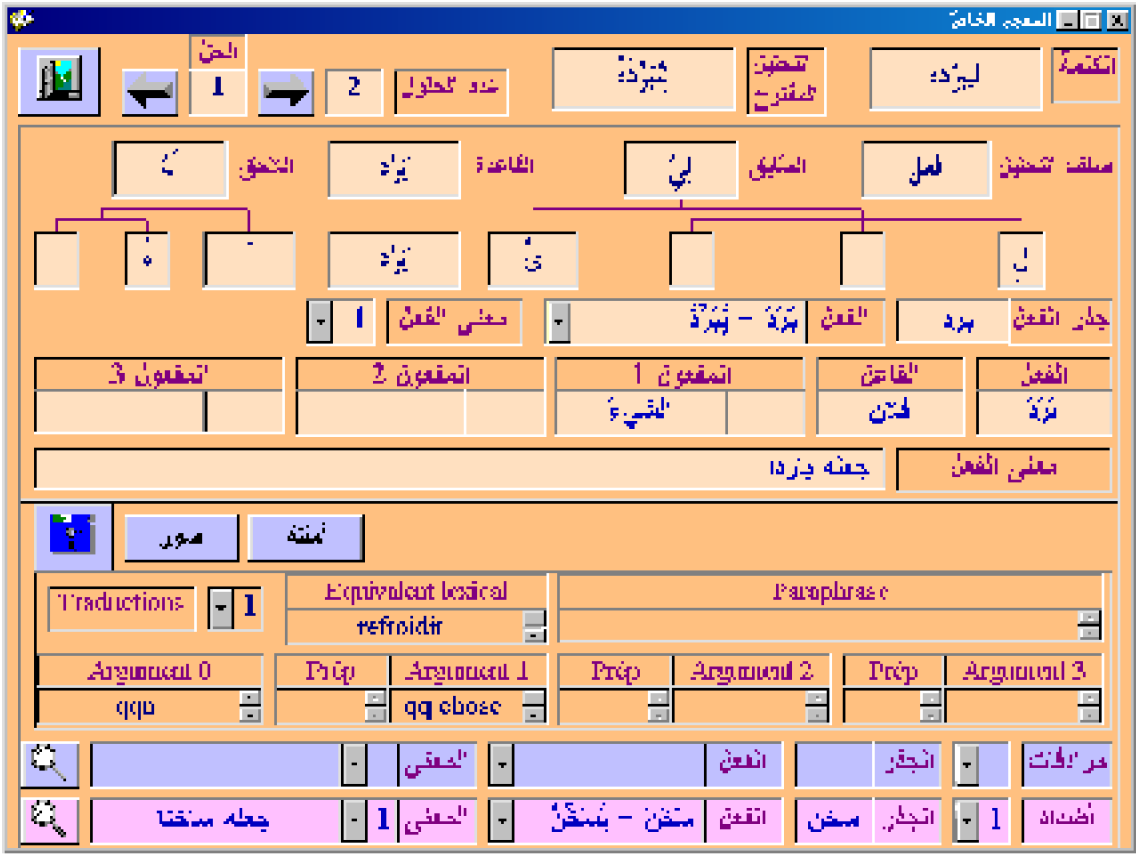
9.5 Le dictionnaire personnel

Nous avons élaboré un module pour permettre à l'apprenant de construire et gérer les informations nouvellement acquises dans un **dictionnaire personnel** (figure 9-7). Les travaux de Goodfellow (1995) ont montré l'importance de ce module visant l'assistance de l'apprenant à noter et à organiser le vocabulaire en partie connu. Il est donc tout à fait pertinent de tenir compte de ces informations pour la spécification des activités lexicales et le choix des textes.

Afin d'éviter que le dictionnaire personnel ne soit une reproduction du dictionnaire général, nous avons ajouté trois nouveaux champs qui permettent de définir le *synonyme* et l'*antonyme* et une illustration de l'unité lexicale acquise. Les informations relatives à ces trois champs par l'apprenant ne peuvent pas être contrôlées par le système. Par contre, l'apprenant est contraint de choisir les informations relatives aux autres champs parmi celles du dictionnaire général. La question reste ouverte de savoir s'il faut contraindre l'apprenant dans ses choix ou le laisser libre avec le risque de mémoriser des informations erronées.

Le processus d'enrichissement du dictionnaire personnel, consiste à spécifier l'unité lexicale en entrée qui est analysée par le système. L'apprenant choisit alors une des solutions proposées par le système. L'apprenant pourrait remplir ensuite l'ensemble des propriétés de l'unité lexicale ajoutée, par les informations récupérées. Il pourrait par exemple, associer un ou plusieurs exemples de phrases du mot, des traductions, des synonymes, des antonymes, des illustrations, etc. (figure 9-7).

L'apprenant pourrait à tout moment accéder au dictionnaire personnel pour le mettre à jour. La méthode d'accès est exactement la même que celle utilisée pour accéder au dictionnaire général. On a prévu aussi des écrans qui synthétisent les unités lexicales sous forme de listes selon leurs catégories ou leurs propriétés. Ce sont en effet ces listes qui vont servir à personnaliser les activités d'apprentissage.

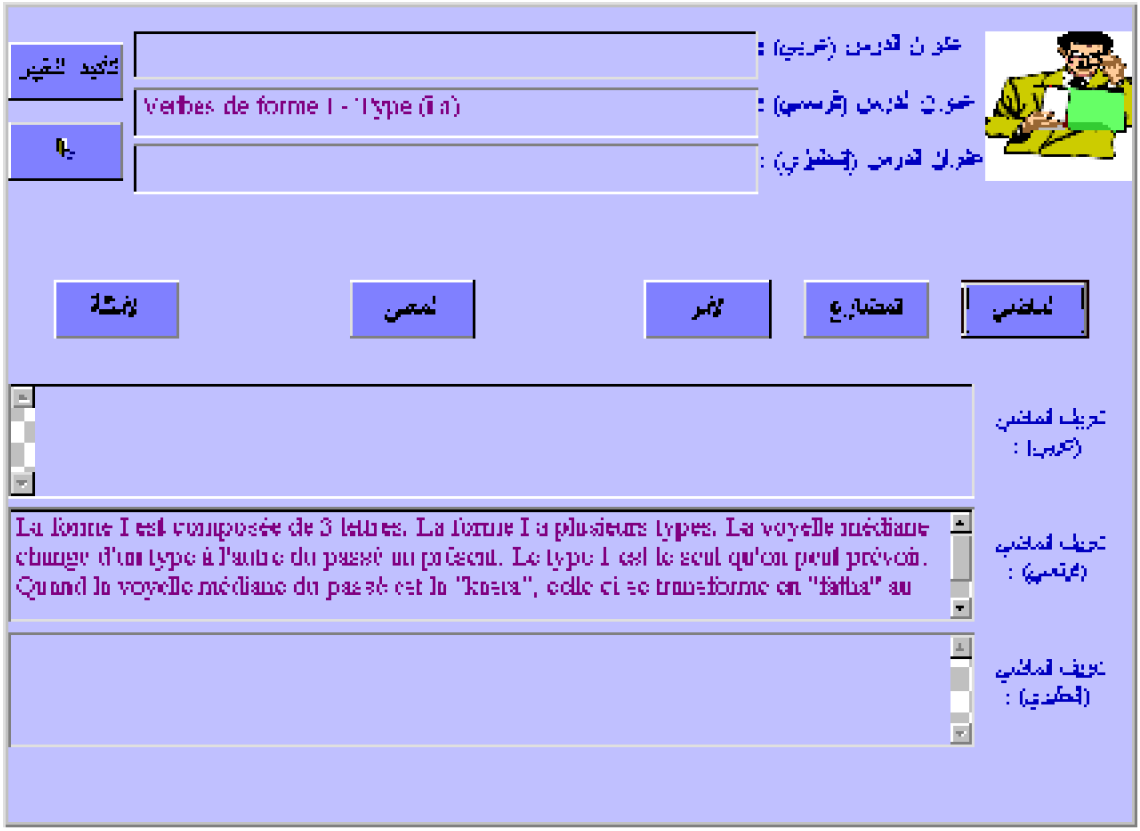


9.6 Le module de l'enseignant

Dans le but de permettre à un non informaticien d'accéder aux différentes ressources de l'environnement et de les enrichir nous avons créé un module de l'enseignant qui est en quelque sorte un système auteur. Ce module se divise en deux parties : une première partie permet à l'enseignant d'enrichir le corpus textuel par un nouveau texte étiqueté et une seconde lui permet d'introduire une leçon de grammaire.

Pour introduire un nouveau texte dans le corpus, nous avons prévu une interface qui permet à l'enseignant, de diviser son texte en une ou plusieurs parties s'il est assez long, de définir sa traduction en langue(s) cible(s), d'assister le système lors du processus d'étiquetage du texte en utilisant l'outil informatique (cf. chapitre 4) et de définir des activités.

Chaque texte a en effet certaines propriétés : style, acteurs, etc., qui peuvent être sujettes à des activités intéressantes qui ne peuvent être générées automatiquement par le système. A l'aide des maquettes des activités, l'enseignant pourrait donc définir pour chaque texte un certain nombre d'activités statiques.

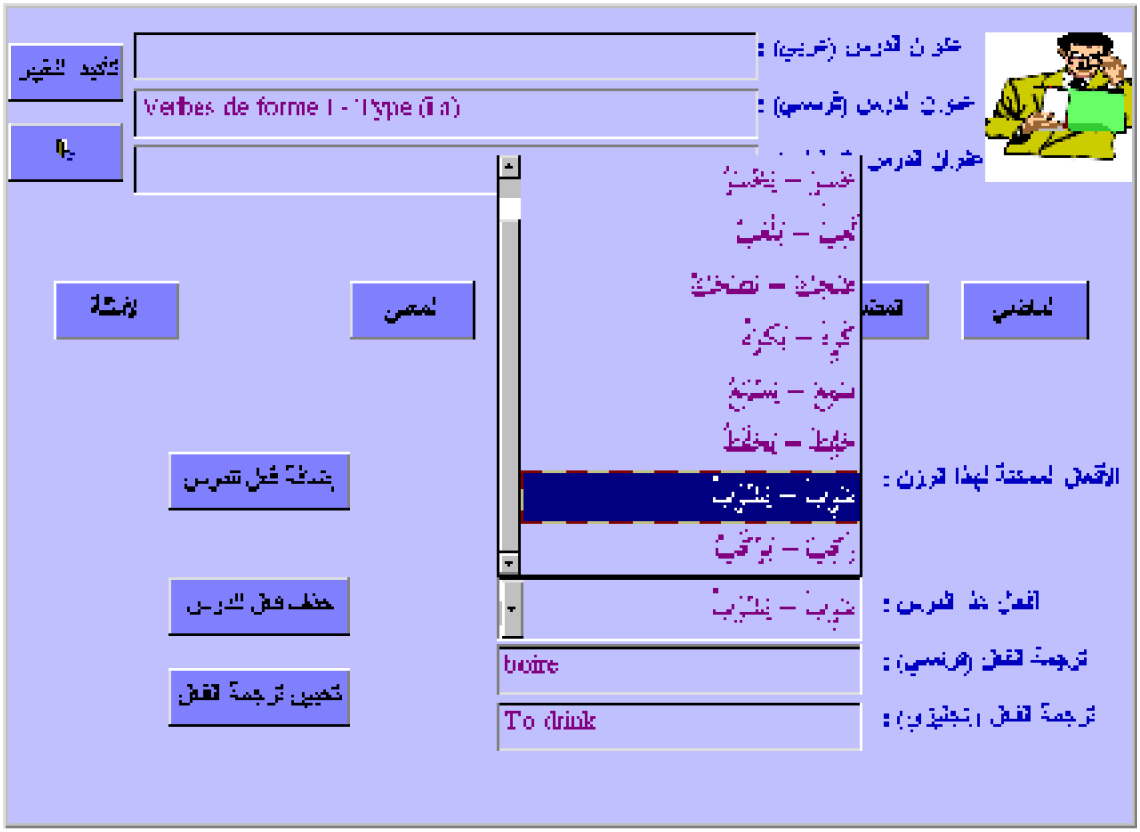


Pour définir des leçons de grammaire, nous avons prévu plusieurs interfaces génériques dont chacune est spécifique à un certain type de leçons morphologiques ou syntaxiques.

L'enseignant doit d'abord choisir le type de la leçon (morphologique ou syntaxique) et fixer le niveau de difficulté de la leçon par rapport aux autres leçons déjà intégrées dans l'environnement. Ensuite, selon le type de la leçon, il doit se conformer aux contraintes du système pour définir ses différents éléments.

Par exemple, pour ajouter une leçon sur les propriétés morphologiques d'un type de verbe (figure 9-8), l'enseignant doit définir le titre de la leçon y compris en langue cible, expliquer les règles de conjugaison du verbe pour chaque aspect de conjugaison, la

signification de ce type de verbes et sélectionner un ensemble de verbes exemple qui seront montrés à l'apprenant (figure 9-9).



L'enseignant pourrait enfin définir des activités pour chaque leçon (définition de la question, des distracteurs, des réponses correctes, explications, etc.) et lui associer un sous-modèle de comportement du modèle de l'apprenant.

9.7 Évaluation du système

L'expérimentation s'est déroulée pendant les quatre dernières séances du cours (cf. § 5.5). La durée de chaque séance était d'environ deux heures. A l'époque de l'expérimentation, l'environnement ne permettait d'étudier que trois textes, à partir desquels les apprenants peuvent s'exercer sur les activités correspondantes pour enrichir

leur vocabulaire.

Afin d'évaluer le système, les étudiants notaient au fur et à mesure les problèmes qu'ils rencontraient sur un papier libre et il leur a été demandé d'évaluer globalement le système à la fin de la dernière séance. La majorité des remarques obtenues étaient d'ordre technique. Nous avons pu néanmoins tirer quelques conclusions de cette première expérimentation :

- La durée de l'expérimentation était malheureusement trop courte pour pouvoir évaluer les différents modules du système (le module « choix de texte » n'était pas utilisé, les sous-modèles comportementaux des apprenants étaient à peine initialisés, les dictionnaires des apprenants étaient très peu utilisés).
- L'effet de cette expérience était très positif sur les apprenants et ils étaient très motivés pour continuer l'expérience chez eux sur de nouveaux textes.
- Les apprenants s'intéressaient beaucoup au système d'évaluation des activités et cherchaient à obtenir la meilleure évaluation.
- Certaines fonctionnalités ne doivent être accessibles qu'à partir d'un certain moment précis de la session. Par exemple, certains apprenants préféraient directement lire la traduction du texte ou à obtenir rapidement les corrections des activités.

9.8 Conclusion

Dans ce chapitre nous avons présenté les différents modules de l'environnement « *AL-Mu c aLLiM* » et les relations qui les lient. Nous avons exposé aussi les résultats de l'expérimentation effectuée et qui devrait nous permettre d'évaluer dans quelle mesure les ressources et les modules disponibles aidaient l'apprenant à la compréhension des textes, à la rétention du vocabulaire et à la maîtrise de la grammaire.

Les résultats obtenus n'ont malheureusement pas apporté de réponse à cette question. L'expérimentation doit en effet être effectuée sur une longue période avec un nombre plus important d'apprenants et avec des ressources beaucoup plus riches.

La constitution de ces ressources n'est pas une tâche aisée. Elle doit être effectuée avec le plus grand soin par un expert du domaine, qui doit être préalablement formé au système. Un travail colloboratif impliquant plusieurs experts pourrait à notre avis être la solution pour accélérer la construction de ces ressources et mener une véritable expérimentation de l'environnement.

Conclusion générale

« Le savant n'est pas l'homme qui fournit les vraies réponses ; c'est celui qui pose les vraies questions ». Claude LEVI-STRAUSS.. Le Cru et le cuit

L'objectif de ce travail de recherches a été de concevoir un prototype d'environnement informatique favorisant l'acquisition lexicale et la maîtrise grammaticale de l'arabe pour des apprenants de langue étrangère. Le prototype doit être en conformité avec les caractéristiques spécifiques à la langue arabe, les ressources et les outils informatiques à disposition et les possibilités de la machine (exploration et individualisation).

Le domaine qui a été pris en compte est le mot graphique. Ce choix s'explique par le fait que la structure particulière de cette unité en arabe rend nécessaire une étude approfondie des phénomènes qui se produisent à son niveau pour l'apprentissage de la langue.

Autour du mot graphique arabe, nous avons défini un ensemble de ressources linguistiques :

1. Une base de données lexicale (DIINAR) que nous avons enrichi par des informations d'ordre syntaxique et sémantique (PROLEMAA) afin de pallier les insuffisances des traits linguistiques. L'environnement d'apprentissage doit, en effet, améliorer les différentes compétences de l'apprenant : compétences linguistiques (morphologiques, syntaxiques et sémantiques) et compétences communicatives.
2. Un corpus textuel étiqueté par des propriétés morphologiques, syntaxiques et sémantiques.

La construction de ces ressources a relevé des outils de TAL arabe développés :

1. L'analyseur des mots graphiques arabes
2. L'étiqueteur semi-automatique de textes bruts.
3. Le concordanceur
4. Le programme de calcul des fréquences.

A partir de ces ressources et de ces outils, il fallait construire un environnement d'apprentissage autonome permettant l'acquisition du vocabulaire et la maîtrise de la grammaire.

Dès lors, nous avons conçu un environnement d'apprentissage qui fonctionne autour d'un schéma type d'apprentissage en trois volets : choix et compréhension d'un texte, rétention du lexique et maîtrise de la grammaire. Ce schéma a décidé des constituants de l'environnement :

1. Un module pour le choix du texte,
2. un corpus de textes étiquetés pour l'exposition à de nouveaux mots,
3. un dictionnaire comme outil d'aide à la compréhension,
4. un module d'activités d'apprentissage permettant à l'apprenant de s'exercer et favorisant ainsi sa maîtrise de la langue,
5. un dictionnaire personnalisé pour organiser le nouveau lexique et faciliter sa rétention,
6. un modèle de l'apprenant permettant l'individualisation de l'apprentissage,
7. un module enseignant permettant de définir les activités.

Le choix du texte tient compte du niveau de l'apprenant et du vocabulaire qu'il maîtrise. L'utilisation de textes complètement étiquetés, permet le passage du mot à sa signification par une simple sélection du premier et résout ainsi le problème d'accès lexical au dictionnaire. Le corpus textuel et le dictionnaire électronique ne doivent pas être considérés uniquement sur le plan de la consultation et de l'aide. Ils servent de matériaux de base pour générer automatiquement des activités lexicales et grammaticales renouvelables.

C'est certainement le point fort de l'environnement puisque les activités permettent de répondre aux deux objectifs que nous nous sommes définis : rétention du lexique et maîtrise de la grammaire. La personnalisation de ces activités est assurée par le modèle de l'apprenant qui est en mesure d'évaluer l'état cognitif de l'apprenant à tout moment.

Les perspectives de recherche concernent dans un premier temps l'amélioration des ressources de l'environnement. Le corpus textuel doit être étendu de manière à définir les fréquences des différentes unités lexicales sur un corpus représentatif de la langue et ainsi proposer aux apprenants débutants des textes avec les mots les plus courants. Concernant le dictionnaire général, les définitions doivent être redéfinies avec le vocabulaire définitoire et certaines parties doivent être ajoutées, comme par exemple la

synonymie et l'antonymie. Le modèle de l'apprenant doit être testé sur un échantillon important d'apprenants et sur une longue période. Les maquettes d'activités doivent être aussi enrichies et associées aux sous-modèles comportementaux des apprenants.

Dans un second temps, l'environnement doit être réévalué. La perspective d'implanter l'environnement sur le WEB doit être envisagée. Ceci permettrait de mener un travail collaboratif entre enseignants et apprenants pour enrichir les ressources, évaluer le système actuel et le faire évoluer.

Références bibliographiques

Abbes R. (1999) : *Conception et réalisation d'un prototype de concordancier électronique de la langue arabe*, Mémoire de DEA en Sciences de L'information et de la Communication, ENSSIB, France.

Abbes R., Hassoun M. (1999) : « conception d'un prototype de concordancier de la langue arabe : des éléments de réflexion », *Colloque Génération Systématique et Traduction Automatique*, Rabat, 15-17 novembre 1999.

Abu Al-chay N. (1988) : *Un système expert pour l'analyse et la synthèse des verbes arabes dans un cadre d'enseignement assisté par ordinateur,* Thèse de doctorat, Université Claude Bernard - Lyon I.

Aitchison J. (1987) : *Words in the mind*, Oxford, Blackwell.

Al-Hakkak G., Neyreneuf, M. (1996) : *Grammaire active de l'arabe*, Editions LM : Les langues modernes, Le livre de poche, France.

Ammar S., Dichy J. (1999) : *Les verbes arabes,* Collection Bescherelle, Paris, Hatier.

Baccouche T. (1992) : "*Attasrif al-arabi min kilal ilm al-asouat al-hadith* ", Editions sociétés Abdelkarim Ben Abdallah, 3ème édition modifiée, Tunisie.

Balacheff N., Baron M., Desmoulins C., Grandbastien M., Vivet M. (1997). « Conception d'environnements Interactifs d'apprentissage avec ordinateur Tendances et perspectives », *PRC-GDR IA'97*, pp. 316 - 337.

Beeslay K. R. (1996) : «Arabic finite-state morphological analysis and generation », *In*

COLING-96 Proceedings, volume 1, pages 89-94, Copenhagen. Center of Sprogteknologi. The 16th International Conference on Computational Linguistics.

Ben Hamadou A. (1991) : *Vérification et correction automatiques par analyse affixale des textes écrits en langage naturel : cas de l'arabe non voyellé,* Thèse de doctorat, Université de Tunis, Avril 1991.

Bogaards P. (1994) : *Le Vocabulaire dans l'Apprentissage des Langues Etrangères*, Langues et Apprentissage des Langues, CREDIF, ENS St-Cloud, Hatier/Didier.

Bogaards P. (1995) : « Dictionnaires et compréhension écrite », *Cahiers de Lexicologie 67*, 1995-2, pp. 37-53.

Bogaards P. (1998) : « Des dictionnaires au service de l'apprentissage du français langue étrangère », *Cahiers de Lexicologie 72*, 1998-1, pp. 127-167.

Braham A., Ghazeli S. (1998) : « ##### ######## ######### ## ##### #### ######## ##### ##### ##### : ##### ##### », ######## ######## ###### – ##### 32 – ##### ######## #### – ##### 1419 ## - ###### 1998 – ALESCO.

Bruillard E. (1997) : *Les machines à enseigner*, Editions Hermès, Paris, 320 pages.

Chanier T (1992) : « Perspectives de l'apport de l'EIAO dans l'apprentissage des langues étrangères : modélisation de l'apprenant et diagnostic d'erreurs » *ICO : Revue de liaison de la recherche en Informatique Cognitive des Organisations.* Montréal. vol 3, n.4. pp 25-34.

Chanier T., Fouqueré C., Issac F. (1995) : « AlexiA : Un environnement d'aide à l'apprentissage lexical du français langue seconde », *Conférence Environnements Interactifs d'Apprentissage avec Ordinateur (EIAO'95)*, pp 79-90, Eyrolles, Paris.

Chanier T., Selva T. (2000) : « Génération automatique d´activités Lexicales dans le système ALEXIA », *Sciences et Techniques Educatives (STE)*, vol 7, 2. Editions Hermès : Paris pp 385-412.

Cohen D. (1961/70) : « Essai d'une analyse automatique de l'arabe », 1961 (T.A. informations), in D. Cohen, *Etudes de linguistique sémitique et arabe*, Mouton, Paris,1970.

Danna F. (1997) : *Modélisation de l'apprenant dans un logiciel d'enseignement intelligement assisté par ordinateur, Application à un tutoriel dédié aux composés anglais*, thèse de l'Université Rennes 1, Janvier 1997.

Demaizière F. (1986). : *Enseignement assisté par ordinateur*. OPHRYS, Collection Autoformation et Enseignement Multimédia, 569 p.

Desclés J.P., Abaab H., Dichy J., Kouloughli D. E., Ziadah M.S. (1983) : « Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement Assisté par Ordinateur », Rapport rédigé sous la direction de J.P. Desclés, à la demande du Ministère français des Affaires étrangères (sous-direction de la Politique linguistique), souvent cité comme le *Rapport Desclés*.

Desclés J.P. (1989) : « La linguistique informatique et le programme SAMIA », in Dichy & Hassoun (1989), pp. 13-25.

Dichy J. (1987) : « The SAMIA Research Program, Year Four, Progress and Prospects », *Processing Arabic Report n°2, T.C.M.O., Université catholique de Nimègue*, pp. 1-26.

Dichy J. (1990) : *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*, Thèse pour le Doctorat d'état (ès Lettres), Université Lyon 2.

Dichy J. (1993) : « Knowledge-systems simulation and the computer-aided learning of Arabic verb-form synthesis and analysis », *Processing Arabic Report n°6/7, T.C.M.O., Université de Nimègue*, pp. 67-84, 92-95.

Dichy J. (1997) : « Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot », *communication aux IVe Journées scientifiques du réseau "Lexicologie, Terminologie et Traduction" de l'AUPELF-UREF*, (Lyon, 28-30 Sepr. 1995), in Méta Vol.42, n°2, juin 1997, Québec, Presses de l'Université de Montréal, pp. 291-306.

Dichy J., Hassoun M. (1989) : *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe, travaux SAMIA I*, Fondation postuniversitaire interculturelle (Conseil International à la Langue Française).

Dichy J., Hassoun M., Mouelhi Z., Zaafrani R. (2002) : « Vers un dictionnaire de l'arabe (I) : les mots outils », à paraître.

Dichy J., Hassoun M., Zaafrani R. (2002, a) : « Vers un dictionnaire de l'arabe (I) : les noms », à paraître.

Dichy J., Hassoun M., Zaafrani R. (2002, b) : « Vers un dictionnaire de l'arabe (I) : les noms propres », à paraître.

Dichy J., Hassoun M., Zaafrani R. (2002, c) : « Vers un dictionnaire de l'arabe (I) : les adjectifs », à paraître.

Dichy J., Hassoun M., Zaafrani R. (2002, d) : « Vers un dictionnaire de l'arabe (I) : les verbes », à paraître.

Ditters E. (1992) : *A Formal Approach to Arabic Syntax. The Noun Phrase and the Verb Phrase*, Thèse de Doctorat, Université Catholique de Nimègue, 1992.

Farrington B. (1984) : « A program to simulate the translation lesson », *In AILA Brussels 84. Proceedings. Vol. 2.* pp. 680-681.

Fuchs C., (1993), *Linguistique et traitements automatiques des langues*, éditions Hachette université, 1993.

Gader N. (1996) : « Vers un analyseur morphologique de l'arabe non vocalisé, partiellement vocalisé ou complètement vocalisé », *Proceedings of 5th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, 11-13 Avril, Cambridge, Angleterre, pp. 3.15.1-3.15.8.

Ghenima M. (1998) : *Un système de voyellation de textes arabes*, Thèse de doctorat d'état, Université Lumière - Lyon II.

Goldstein I., Carr B. (1977). « The computer as a coach: an athletic paradigm for intellectual education ». *Proc. of 1977 ACM annual conference*, Seattle, p. 227-233.

Goodfellow, R. (1994) : *A computer-based strategy for foreign language vocabulary learning*, Unpublished PhD thesis, Institute of Educational Technology, Open University.

Goodfellow, R. (1995) : « A Review of Types of Programs for Vocabulary Instruction », *Computer-Assisted Language Learning 8*, 2-3, pp. 205-226.

Habert B., Nazarenko A., Salem A. (1997) : *Les linguistiques de corpus*, Armand Colin/Masson, Paris.

Haddar K., (2000) : *Caractérisation formelle des ellipses de la langue arabe et processus de recouvrement,* Thèse de doctorat, Université Tunis III, juillet 2000.

Hassoun M. (1987) : *Conception d'un dictionnaire pour le traitement de l'arabe dans différents contextes d'application,* Thèse de doctorat d'état, Université Claude Bernard-Lyon I.

Issac F. (1997) : *Analyse syntaxique et apprentissage des langues*, thèse de doctorat, Université Paris-Nord.

Jaccarini A., Audebert C. (1986) : « à la recherche du HABAR, outils en vue de l'établissement d'un programme d'enseignement assisté par ordinateur », *Annales islamologiques*, Tome XXII, Institut français d'archéologie orientale du Caire.

Johns T. (1991) : « Should you be persuaded - Two examples of data-driven learning », *English Language Research Journal 4*, pp. 27-45.

Lelubre X. (1985) : « Projet pour un didacticiel de conjugaison de verbes arabes », *Ministère de l'éducation nationale*.

Lelubre X. (1993) : « Coursware for the theory and practice of arabic conjugation », *in : Processing Arabic Report, n°6/7, TCMO, Université catholique de Nimègue*, pp. 85-89 et 92-95.

Levy, M. (1997). *Computer-Assisted Language Learning : context and conceptualization. Oxford* : Oxford University Press.

Lyons J. (1978/90) : *Sémantique linguistique*, 1ère édition 1978, traduction française J.DURAND et D. BOULONNAIS, Paris : Larousse, 1990.

Michael L. (1997) : *Computer-Assisted Language Learning Context and Conceptualization*, Oxford : Clarendon Press, 298 pages.

Ouersighni R. (1998) : « An approach for the conception of an arabic parser based on Affix Grammars over Finite Lattice », *Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, April 16-19, 1998, Cambridge, England.

Ouersighni R. (2001) : *La conception et la réalisation d'un analyseur morpho-syntaxique en vue de la vérification grammaticale assistée par ordinateur de textes écrits en arabe*, Thèse de doctorat, Université Lyon 2, Décembre 2001.

Papert S. (1981) : *Jaillissement de l'esprit. Ordinateurs et apprentissage*. Paris : Flammarion (Collection Champs).

Paribakht T. S., Wesche M. (1997) : « Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition », *J. Coady, T. Huckin (Eds), Second Language Vocabulary: a rationale for pedagogy*, Cambridge, Cambridge University Press, pp 174-200.

Rézeau J., (1997) : « Concordances, Cédérom et Internet au service de l'enseignement du français aux adultes », *The Dong-Eui International Journal (Corée)*, Juin 1997.

Roman A. (1990) : *Grammaire arabe*, Paris, Presses universitaires de France - collection "que sais-je".

Sabah G. (1989) : *L'intelligence artificielle et le langage*, Volume 2, Processus de compréhension, Hermès, Paris.

SAMIA (Groupe de recherche), signature collective (1984) : « Enseignement Assisté par Ordinateur de l'arabe : Simulation à l'aide d'un modèle linguistique - La Morphologie », *in : Actes du colloque international "E.A.O. 84"*, Paris, Agence de l'Informatique, pp. 81-96.

Self J. (1987) : « Students Models : what use are they ? » *Actes de IFIP / TC3*, p. 73-85 ,.

Self J. (1994) : « Formal approaches to student Modelling » *Actes des journées student modelling : The key to individualized Knowledge-Based Instruction*, éd. Par Greer (Jim E.) et Mac Calla (Gordon I.). p. 295-354, Springer-Verlag.

Selva T. (1999) : *Ressources et activités pédagogiques dans un environnement informatique d'aide à l'apprentissage lexical du français langue seconde,* Thèse d'Université, Université de Franche-Comté, Besançon, octobre 1999, 210 pages.

Tréville M.-C., Duquette L. (1996) : *Enseigner le vocabulaire en classe de langue*, Paris, Hachette.

Tribble C. & Jones G., (1997) : *Concordances In The Classroom : a resource book for teachers*, Houston : Athelstan, 1997, nouvelle édition, 114 pages.

Wegner E. (1987) : *Artificial Intelligence and Tutoring Systems : computational and cognitive approaches to the communication of knowledge*.Los Altos (CA): Morgan Kaufmann Publishers.

Weidenfeld G., Caillot M., Cochard G-M., Fluhr C., Guérin J-L., Leclet D., Richard D., (1997) : *Techniques de base pour le multimedia*, édition MASSON, février 1997.

Zaafrani R. (1997) : « Morphological analysis for an Arabic Computer-aided learning system », Proceedings of DIALOGUE'97, *International Conference on computational linguistics and its applications*, June 10-15, 1997, Yasnaya Polyana, Russia.

Zaafrani R. (1998, a) : « Al-Mucallim 2 Software : An Arabic Computer Learning System Using Conceptual Sentence Generation », *Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, April 16-19, 1998, Cambridge, England.

Zaafrani R. (1998, b) : « Prototype d'un système d'EIAO de l'écriture de phrases en Arabe à partir d'une Base de données lexicale », *Atelier de travail sur le traitement automatique de la langue arabe. Journées de l'AFEMAM (Association française pour l'étude du monde arabe et musulman)*. 2-4 Juillet, 1998, Lyon, France

Zaafrani R., Ihadjadene M., Bouché R. (2000) : « The dynamic nature of searching and browsing on Web-OPACS : The CATHIE experience », *ISKO Conference*, faculty of Information Studies, University of Toronto, juillet 2000, Toronto, Canada.

Zock M. (1991) : « SWIM or Sink: The problem of communicating thought », in *Bridge to International Communication: Intelligent Tutoring Systems for Second-Language Learning,* Swartz M.& Yazdani M (eds). New York: Springer-Verlag.

ANNEXE 1: TABLE DE TRANSLITTÉRATION

Cette translittération est librement empruntée à A. Roman (1990). Pour des raisons de portabilité informatique, il n'est fait usage d'aucun caractère particulier. Les emphatiques et la constrictive vélaire sont en caractère gras. Le soulignement distingue, lorsqu'il y a lieu, la constrictive de l'occlusive correspondante, ou un graphème d'un graphème "voisin" (il ne s'agit pas d'une transcription phonétique au sens strict !); les voyelles longues sont surmontées d'un accent circonflexe. On a :

Consonnes					
ب	B		ط	T	
ت	_		ظ	D	
ث	_		ع	'	
ج	J		غ	G	
ح	H		ف	F	
خ	X		ق	Q	
د	D		ك	K	
ذ	D		ل	L	
ر	R		م	M	
ز	Z		ن	N	
س	S		ه	H	
ش	$		و	W	
ص	S		ي	Y	

Voyelles brèves					
	a		u		i

Voyelles longues						

Morphogrammes							

ANNEXE 2 : LA BASE DE DONNÉE LEXICALE DIINAR.1

1) Les partenaires de cette réalisation, liés par une convention d'études et de recherche, sont :

- L'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), 17-21, boulevard du 11 nov. 1918, 69623 Villeurbanne Cedex, France
- L'Institution de Recherche en Sciences Informatiques et Télécommunications (IRSIT), Parc Technologique des Communications - Route de Raoued Km 3.5, 2083 Ariana, Tunisie
- L'Université Lumière-Lyon 2, 86, rue Pasteur, 69365 Lyon Cedex 07, France

2) Le produit réalisé :

Un produit de valorisation, intitulé **DIINAR.1**, a été conçu et développé en commun par le consortium **ENSSIB / IRSIT / LYON 2**.

Ce produit est constitué d'un lexique informatisé complet et voyellé de l'arabe. Il prend la forme d'une base de données comportant environ :

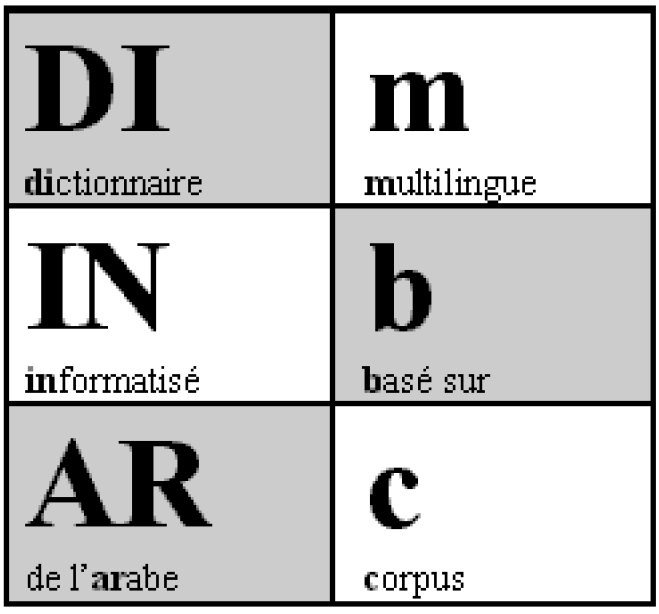
- 20000 entrées verbales,
- 70000 entrées déverbales,

- 29000 entrées nominales, auxquelles sont associés 10000 formes de pluriel interne
- 1000 noms propres
- 450 mots-outils et
- l'ensemble complet des enclitiques, proclitiques, préfixes et suffixes de cette langue.

Les différentes entrées sont renseignées, et précisent en particulier :

- la façon dont le mot se fléchit (conjugaison, flexion, déclinaison, …)
- les relations dérivationnelles (« fléchages » entre formes du singulier et pluriels internes, entre formes verbales et déverbales, …)
- la façon dont il se combine avec des enclitiques et les proclitiques
- la racine ou la pro-racine à laquelle l'entrée est rattachée.

ANNEXE 3 : LE PROJET DIINAR-MBC



DIINAR-MBC

(DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus)

Titre complet de DIINAR-MBC: Short-term achievement of a corpus-based multilingual basic Arabic Lexical dB and related resource-productive tool-box

EC Programme & numéro du contrat : Durée: Période: Project Officier : Coordinateur du projet:	LE[illegible] 36 mois [illegible] décembre 2003 [illegible] [illegible]
Partenaire 1: LYON2 Université Lumière Lyon 2 86 rue Pasteur 69365 Lyon Cedex 07 France	Partenaire 2: ENSSIB [illegible] [illegible] [illegible]
Partenaire 3: KUN Katholieke Universiteit Nijmegen, [illegible] P.O. Box 9103, 6500 HD Nijmegen, The Netherlands	Partenaire 4: INSIT [illegible] [illegible] 1002 Tunis Tunisie
Partenaire 5: ERI Electronics Research Institute [illegible] [illegible] Egypt	Partenaire 6: INRA [illegible] [illegible] Agdal – Rabat Maroc

Site Internet du projet : http://www.univ-lyon2.fr/langues/promodiinar/Accueil.htm

Descriptif du projet

L'objectif général du projet DIINAR-MBC (**DI** ctionnaire **IN** formatisé **AR** abe- **M** ultilingue et **B** asé sur **C** orpus) était de munir la langue arabe d'un ensemble d'outils et de ressources de traitement de la langue destinés aux linguistes, lexicographes et les professionnels des technologies de la langue.

Le projet DIINAR-MBC a délivré deux boîtes à outils (Voir figure ci-dessous : DIINAR-MBC Tools & Resources Diagram) :

La première boîte à outils renferme : 1.

2.

i. Un corpus textuel brut représentatif de l'arabe moderne standard de 10 millions de mots (ARCOLEX : Arabic Raw Corpora for Lexical purposes) collectés à partir de textes représentatifs de la langue arabe par les différents partenaires participants au projet et encodés selon la norme de la TEI (Text Encoding Initiative).

ii. Une petite partie de ce corpus (200.000 mots) a été étiquetée manuellement et automatiquement. L'étiquetage manuel ayant servi à évaluer les résultats obtenus automatiquement à l'aide de l'analyseur morpho-syntaxique (LARUSA : a Lexical-purpose Arabic Unvowelled Sentence Analyser).

iii. Une sélection de dix mille lemmes les plus représentatifs de l'arabe, qui a permis de réaliser un prototype d'une base de données lexicale multilingue arabe-français et arabe-anglais (PROLEMAA : Prototype de Lexique Multilingue à partir de l'Arabe).

Le deuxième paquet se compose d'un ensemble d'outils qui permettent la réalisation du lexique multilingue basé sur l'Arabe. Il comprend notamment : 1.

Des interfaces utilisateur ergonomiques pour la saisie du lexique PROLEMAA. Elles permettent de saisir tous les spécificateurs morpho-syntaxiques et sémantiques qui sont associés aux entrées lexicales. 2.

Un analyseur syntaxique de textes arabes non voyéllés pour le traitement automatique du corpus. 3.

DIINAR-MBC Tools & Resources Diagram

ARCOLEX

ARCOLEX utilise deux niveaux de codage des documents :

Un premier niveau de codage de la structure grossière (divisions) jusqu'au niveau du paragraphe pour l'ensemble du corpus qui est automatisé. 1.

Un deuxième niveau pour une partie du corpus textuel (200.000 mots) doit avoir un codage fin des éléments internes au paragraphe. 2.

J'étais chargé de la réalisation d'une application qui permet de baliser le premier niveau de codage du texte dans le format de la TEI (I1). D'autre part, j'ai réalisé une interface WEB pour la saisie des notices bibliographiques du corpus textuel et la génération automatique des entêtes des documents en utilisant la DTD de la TEI LITE.

PROLEMMA

Dans le cadre de PROLEMAA :

- J'ai élaboré les différentes bases de données lexicales de PROLEMAA (arabe – arabe : LR5), (arabe – anglais : LR6) et (arabe – français : LR7).
- J'ai réalisé les interfaces de saisie et de mise à jour des spécificateurs syntactico-sémantiques du dictionnaire PROLEMAA (arabe – arabe : I5), (arabe – anglais : I6) et (arabe – français : I7).