# INTRODUCTION

Notre travail s'inscrit dans le cadre des travaux du groupe de recherche LISIA (Linguistique Informatique et Systèmes d'information appliqués à l'arabe), commun à l'université Lyon 2 et à l'ENSSIB (voir présentation en page 7. Dans toutes les langues il existe des suites qu'un étranger (voire certains locuteurs natifs) ne peut comprendre même s'il connaît le sens de chacun des termes la constituant. Pourquoi ? tout simplement parce que le sens de ces suites n'est pas déductible de celui de ses éléments. Ces suites sont appelées expressions figées et doivent être retenues en entier comme un mot simple par tout locuteur, car dans l'acquisition d'une langue, le fait de maîtriser les faits de locutionalité, c'est-à-dire de pouvoir les produire et les interpréter aisément, constitue un seuil déterminant. Mais se limiter à la non compositionalité sémantique d'une suite pour en faire une expression figée serait réducteur car on occulterait l'impossibilité d'y avoir une liberté syntaxique. En définitive, le figement d'une expression se définit par l'ensemble des traits morphosyntaxique, sémantique restreignant l'application d'opérations syntaxiques d'une part et d'autre part par la compositionnalité sémantique de cette suite. Ce n'est qu'à partir de ces deux caractéristiques que nous pouvons conclure que c'est une unité plus ou moins soudée.

Le concept d'expression figée (التعبير المسوك)(*at tᶜbîru l maskûk*), couramment utilisé, n'avait jamais été formellement défini puisque toutes les définitions tournaient autour de la fixité de la forme et la non-compositionnalité de son sens, car même le *dictionnaire de linguistique* (1994), pourtant destiné à faire référence définit le figement comme étant «un processus linguistique qui, d'un syntagme dont les éléments sont libres, fait un syntagme dont les éléments ne peuvent être dissociés.» Cette définition est réductrice puisqu'elle ne tient pas compte des phrases.

Les auteurs classiques ont utilisé plusieurs termes pour désigner ces expressions, de ***synthème*** utilisé par MARTINET (1965) à la **lexie composée** de POTTIER (1987) en passant par le terme **synapsie** de BENVENISTE (1967), cette différence terminologique reflète l'absence de points de vue théorique entre linguistes quant aux expressions figées. Parmi les auteurs contemporains, Igor MEL'CUK a été le premier à parler de l'expression figée en tant que **phrasème** et à lui donner une définition complète tenant compte de ses différents degrés de figement sémantique. Il définit le phrasème comme étant «une locution prise dans une seule acception bien déterminée et munie de tous les renseignements qui spécifient son comportement dans un texte» (1995, p.57) et distingue trois types de phrasèmes majeurs :

-Le phrasème complet, qui est un phrasème au sens opaque car n'incluant le sens d'aucun de ses constituants.

-Le semi-phrasème, est un phrasème dont le sens inclut celui de l'un de ses éléments.

-Le quasi-phrasème, est un phrasème dont le sens inclut ceux de chacun de ses éléments avec un surplus imprévisible.

Ces différents degrés d'opacité sémantico-syntaxique ont souvent conduit à un blocage lors du traitement automatique des phrasèmes. Il nous

reviendra alors, lors de notre travail, de trouver un système permettant une reconnaissance plus aisée des phrasèmes arabes lors du traitement automatique. Nous appelons donc phrasème ce que d'autres appellent clichés, idiotismes, locutions métaphoriques, proverbiales, familières, gallicismes.

Dans notre travail, l'étude des phrasèmes ne se fera pas à partir des catégories grammaticales habituellement utilisées (nominale, verbale, adjectivale et adverbiale) car nous avons décidé de classer chaque phrasème selon le premier terme le composant. Notre classement des phrasèmes ignore donc la fonction que ce phrasème peut occuper dans une phrase pour se limiter au premier terme apparent de cette suite Ce qui donnera les cinq familles suivantes : (1) les phrasèmes à initiale nominale, (2) les phrasèmes à initiale verbale, (3) les phrasèmes à initiale prépositionnelle, (4) les phrasèmes à initiale pronominale et (5) les phrasèmes à initiale adjectivale. Nous pourrons donc avoir, dans une même famille un phrasème à fonction adjectivale et un phrasème à fonction adverbiale exemple le phrasème à fonction adverbiale : صباح مساء (*SabâHa masâ'a*) : *matin et soir* (I.2.2.4 : numéro du phrasème dans l'A.MI.FO.P) et le phrasème à fonction adjectivale : صاحب الطأس و الكأس (*SâHibu T Ta'si wa l ka's*) : *ivrogne* (I.1.2.1.2) appartiennent tous à la famille des phrasèmes à initiale nominale. Le phrasème suivant ضرب في الأرض (*Daraba fi l 'arDi*)(II.3.1) qualifiant *un homme ayant beaucoup voyagé* ne sera pas classé selon sa fonction adjectivale mais bien d'après son premier constituant qui est un verbe, on le retrouvera donc parmi les phrasèmes à initiale verbale.

L'unité de base de la lexicologie est la lexie, celle-ci est soit un mot pris dans une acception bien définie dans ce cas, elle est appelée lexème, soit une locution prise elle aussi dans une acception bien spécifique et dans ce cas, elle est appelée phrasème. Cette seconde forme de lexie qu'est le phrasème va donc être l'unité centrale de notre travail. Quelle est sa définition exacte ? pourquoi avoir choisi cette dénomination ? quelles sont les différentes structures morphosyntaxiques des phrasèmes arabes ? comment procéder à leur modélisation tout en simplifiant leur reconnaissance lors du traitement automatique ? Telles sont les principales questions qui vont guider notre étude.

Ce travail comportera donc trois parties. Dans la première, composée de trois chapitres, nous procéderons à l'étude théorique des phrasèmes. Nous y définirons le phrasème en général, dans le premier chapitre, et le phrasème arabe dans le troisième chapitre. Le second chapitre sera consacré, lui, à la notion de figement. Nous y étudierons notamment les différences existant entre les phrasèmes et les proverbes.

La seconde partie de notre travail, composée, elle, de cinq chapitres, sera consacrée à l'étude des structures morphosyntaxiques des phrasèmes arabes. Nous y étudierons les différents éléments composant le phrasème ainsi que son degré de figement syntactico-sémantique. La seule condition pour qu'un phrasème soit étudié étant l'existence et la présence dans notre classement d'au moins deux exemples. Nous allons donc relever, dans cette partie, toutes les combinaisons syntaxiques possibles d'être produites en

figements. Pour atteindre ce but, tout phrasème devra répondre aux conditions générales suivantes :

1) la non compositionnalité du sens. Le sens d'un phrasème n'est pas le résultat des sens de ses constituants.

2) impossibilité d'y appliquer les transformations syntaxiques telle que la pronominalisation du complément d'objet d'un phrasème :

ضرب في الأرض (*Daraba fî l 'aDi*) : ضرب فيها (*Daraba fîhâ*) : *il a frappé dans elle* , car le sens de la suite change.

3) actualisation des constituants impossible, exemple l'actualisation du même phrasème donne : يضرب في هذه الأرض (*yaDribu fî hâdhihi l 'arDi*) : *il frappe dans cette terre*, signifié éloigné de celui du phrasème.

4) l'insertion d'un élément nouveau y est impossible, exemple l'insertion d'un complément d'objet direct au même phrasème :

ضربه في الأرض (*Darabahu fî l 'arDi*) : *il l'a frappé par terre* change son sens

5) la substitution synonymique y est impossible exemple :

ضرب في التراب (*Daraba fî t turâb*) : *il a frappé dans la terre* .

Ces conditions sont valables pour tout phrasème quelle que soit sa famille, en d'autres termes, nous remarquons une homogénéité quant aux caractéristiques de figement. Toutefois, lors de l'étude de chaque catégorie, nous donnerons les caractéristiques de figement particulières à chacune d'entre elles.

Chaque chapitre traitera d'une famille de phrasèmes. Ainsi, le premier chapitre sera consacré aux phrasèmes à initiale nominale qui ont pour caractéristiques de figement :

1-impossibilité d'y insérer un élément nouveau
2-substitution synonymique impossible
3-détermination globale
4-sens non compositionnel

le second, aux phrasèmes à initiale verbale, avec pour caractéristiques de figement :

1-substitution synonymique ou paradigmatique impossible
2-détermination globale
3-actualisation des éléments impossible
4-blocage des transformations syntaxiques (la pronominalisation, la passivation, la relativation)
5-opacité sémantique

le troisième, aux phrasèmes à initiale prépositionnelle, avec pour caractéristiques de figement :

1-détermination contrainte
2-substitution synonymique impossible
3-adjonction d'un modifieur impossible
4-sens non compositionnel

le quatrième, aux phrasèmes à initiale pronominale qui auront les mêmes caractéristiques que les phrasèmes à initiale nominale (si le terme qui suit le

pronom est un nom) ou verbale (lorsque le terme suivant le pronom est un verbe),
le cinquième aux phrasèmes à initiale adjectivale, dont les caractéristiques de figement sont les suivantes :

1-détermination globale
2-insertion d'un nouveau terme impossible
3-substitution synonymique impossible
4-nominalisation de l'adjectif impossible
5-non compositionnalité du sens

Notre travail devra donc être multilatérale. Ce qui nous oblige à étudier le phrasème selon les critères suivants :

1) critère référentiel

Un phrasème doit toujours avoir un référent unique. Il désigne un objet, une personne, un fait déterminé.

2) critère sémantique

Le phrasème a un signifié non compositionnel, car même le quasi-phrasème a un surplus sémantique ne dépendant pas du sens de ses constituants. Exemple : ساعة رملية (*sâᶜatun ramliyyatun*) : *un sablier* qui n'est pas une montre de sable servant à lire l'heure mais une sorte de chronomètre. Cette suite répond au critère référentiel et au critère sémantique malgré la compositionalité apparente de son sens. Ce critère nous permettra de classer les phrasèmes selon leur degré de figement sémantique. Ainsi, un phrasème au sens opaque sera un phrasème complet, un phrasème au sens semi-opaque sera un semi-phrasème, un phrasème au sens compositionnel avec un surplus imprévisible sera un quasi-phrasème encore appelé par certains idiotisme, terme défini dans le dictionnaire de linguistique (1994) comme étant une «construction qui apparaît en propre à une langue donnée et qui ne possède aucun correspondant syntaxique dans une autre langue». Le critère retenu est donc l'impossibilité de traduire cette suite mot à mot d'une langue à une autre. Les critères de figement sémantique et syntaxique sont ignorés contrairement à la définition que nous lui donnons.

3) critère syntaxique

Il y a blocage des opérations syntaxiques dans les phrasèmes alors qu'elles sont réalisables dans des syntagmes libres.

4) critère de la présence d'au moins deux exemples dans notre étude.

Tout phrasème dont nous n'aurons pas trouvé au moins deux exemples (exemples pris dans les différents dictionnaires et ouvrages consultés), ne sera pas étudié. Toutefois, il figurera dans la partie reservée à la classification de toutes les structures morphosyntaxiques des phrasèmes arabes (voir annexe). Le fait de travailler selon ces critères nous permettra d'atteindre un degré de description tenant compte de toutes les caractéristiques dégagées par ces critères.

Le but de notre étude étant la constitution d'une base de données lexicales, nous allons donc, dans une troisième partie, concevoir l'Arbre de Mise en Format des Phrasèmes (A.MI.FO.P). La conception de cet arbre ne pourra être éffective sans la création d'un système qui nous permettra de l'utiliser. Ces deux thèmes vont être étudiés dans deux chapitres. Le premier

sera consacré à l'Arbre de Mise en Format des Phrasèmes qui se présente sous forme d'arbres élémentaires reliés les uns aux autres. On y distingue des racines (au nombre de cinq), des noeuds (au nombre de trente cinq), des branches (au nombre de cent neuf) et des feuilles (au nombre de quatre cent dix neuf).

Le deuxième chapitre traitera du système d'utilisation de cet Arbre. Ce système a pour base morphosyntaxique l'arbre élémentaire et comme base sémantique l'arbre de dérivation. L'arbre élémentaire est l'unité de base de notre système, il correspond à une entité sémantique, ce qui en fait la plus petite représentation de notre arbre. En d'autres mots, un arbre élémentaire représentera un phrasème complet car il est sémantiquement opaque. Les autres phrasèmes seront représentés par des arbres dérivés, qui sont des arbres issus de l'assemblage de plusieurs arbres élémentaires. Les traits sémantico-syntaxiques des éléments de chaque phrasème (exemple les traits Humain/non-humain, concret/abstrait, le genre, le nombre etc...) se retrouveront aussi dans l'arbre élémentaire (pour le phrasème complet) ou dans l'arbre dérivé (pour les semi et quasi-phrasèmes). L'arbre de dérivation nous renseigne, lui, sur le degré de figement sémantique du phrasème, car dans une représentation arborescente la représentation des dérivations caractérise la combinaison d'arbres élémentaires en nous informant sur quels noeuds ils ont été combinés.

L'arbre de dérivation d'un phrasème tel que ضرب في الأرض (*Daraba fî l 'arDi*) : *il a beaucoup voyagé* sera le suivant


A partir de cet arbre de dérivation nous déduisons que ce phrasème est un semi-phrasème avec un sujet non figé.

Le système procèdera donc en plusieurs étapes. Il commencera par lire les informations données par la structure morphosyntaxique du phrasème, puis il formera l'arbre élémentaire ou dérivé correspondant. En concevant l'arbre de dérivation, le système vérifie, d'une part, que c'est bien un phrasème et, d'autre part, que sa représentation dans l'A.MI.FO.P (Arbre de Mise en Format des Phrasèmes) est un arbre élémentaire (pour le phrasème complet) ou un arbre dérivé (pour le semi ou quasi-phrasème).

**Annexe relative au groupe de recherche inter-établissements (ENSSIB et université Lumière-Lyon 2) "Linguistique Informatique et Systèmes d'Information appliqués à l'Arabe dans une perspective Multilingue" (LISIAM)**

Le groupe de recherche inter-établissements « *Linguistique informatique et systèmes d'information appliqués à l'arabe dans une perspective multilingue* » (LISIAM) accueille depuis le début des années 1990, années 10 à 15 étudiants en DEA et doctorants appartenant, respectivement à l'Ecole Nationale Supérieure de Sciences de l'Information et des Bibliothèques (ENSSIB, Villeurbanne) et à l'université Lumière-Lyon 2. Ce groupe est dirigé par Joseph Dichy, professeur de linguistique arabe à Lyon 2 et Mohamed Hassoun, professeur de sciences de l'information à l'ENSSIB, avec la collaboration de Mathieu Guidère et Xavier Lelubre, maîtres de conférences à Lyon 2. Il donne lieu au séminaire de recherche doctorale LISIAM. Une demie douzaine de thèses ont été soutenues dans le cadre d'une co-direction entre J. Dichy et M. Hassoun, soit en linguistique arabe, soit en sciences de l'information.

Dans le prolongement de cette recherche a été conçue et réalisée la base de connaissances **DIINAR.1** (« Dictionnaire informatisé de l'arabe - version 1 »), en arabe, **Ma'âlî,** abréviation de « Mu'jam al-'Arabiyya l-'âlî ». La base comprend environ 129.000 entrées, soit autour de 20.000 entrées verbales, 79.000 entrées déverbales, 29.000 entrées nominales (près de 10.000 formes de pluriel "brisé" sont en outre associées aux noms correspondants), 1.000 noms propres et 450 mots-outils, ainsi que l'ensemble complet des enclitiques, proclitiques, préfixes et suffixes de cette langue. Il ne s'agit nullement de simples listes : les interfaces de saisie et de consultation des données ont été conçues de manière à permettre l'association aux entrées lexicales de spécificateurs morphosyntaxiques destinés à la génération (en écriture vocalisée) et la reconnaissance (en écriture non-vocalisée) des mots.

DIINAR.1 a été conçue et réalisée en commun à Lyon (Université Lumière-Lyon 2 et ENSSIB) et à l'IRSIT de Tunis, où a eu lieu la saisie des données. Elle constitue l'une des ressources qui ont été à la base du projet européen DIINAR-MBC [1]. La diffusion-valorisation est actuellement en cours de négociation avec ELRA (European Language Ressources Association, Paris – http://www.elda.fr).

---

[1]. **DIINAR-MBC** ("DIctionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus") est un projet soutenu par la Commission européenne (projet n° 961791 du programme de Coopération avec les Pays Tiers et les Organisations Internationales – INCO-DC). La durée de ce projet, qui s'est achevé en décembre 2000, était de 30 mois. La coordination scientifique a été assurée par J. Dichy, *Université Lumière-Lyon 2* (avec la participation de X. Lelubre), assisté par EZUS-Lyon 1 pour les aspects administratifs et de gestion. DIINAR-MBC avait pour autres partenaires l'*École Nationale Supérieure des Sciences de l'Information et des Bibliothèques* (ENSSIB, France - M. Hassoun), l'*Electronics Research Institute* (ERI, Égypte - N. Hegazi), l'*Institut d'Etudes et de Recherche pour l'Arabisation* (IERA, Maroc - A. Fassi-Fehri), l'*Institution Régionale des Sciences Informatiques et des Télécommunications* (IRSIT, Tunisie - A. Braham, S. Ghazali) et l'*Université Catholique de Nimègue* (Pays-Bas - E. Ditters). **Résultats principaux :** un analyseur morphosyntaxique de haut niveau de performance (Ouersighni, 2002) ; un ensemble d'interfaces et de lexiques (dont un prototype de lexique bilingue arabe-français et arabe-anglais) ainsi que des procédures de traitement et d'indexation des données textuelles ou des corpus. Un corpus de 10 millions de mots a été compilé à l'université de Nimègue ainsi qu'à l'IRSIT (Tunis).