Numéro d'ordre : 01/2012-D/IN

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
Université des Sciences et de la Technologie Houari Boumedienne
Faculté d'Electronique et d'Informatique, Département d'Informatique

THÈSE

Présentée pour l'obtention du diplôme de **Doctorat**
En : **Informatique**
Spécialité : Intelligence Artificielle et Bases de données

Par : Mme ALIANE née AOUIDAD HASSINA

Thème :

DECOUVERTE AUTOMATIQUE ET FORMALISATION DES STRUCTURES DE LA LANGUE ARABE CONTRIBUTION A UNE FORMALISATION DE LA TRADITION GRAMMATICALE ARABE

Soutenue le : 12/01/2012, devant le jury composé de :

Mr Badache Nadjib	Professeur	USTHB	Président
Mme Alimazighi Zaia	Professeur	USTHB	Directrice de thèse
Mme Boufaida Zizette	Professeur	UMC	Examinateur
Mr Azzoun Hamid	Maître de Conférences	USTHB	Examinateur
Mr Guessoum Ahmed	Maître de Conférences	USTHB	Examinateur
Mr Okba Kazar	Maître de Conférences	Univ-Biskra	Examinateur

REMERCIEMENTS

Après toutes ces années de travail avec les heurs et les malheurs qu'ont connus tous ceux qui ont préparé une thèse, je dirais que ça fait du bien de mettre un point final (provisoirement car la recherche n'est jamais finie !) et me voilà arrivée sans doute au meilleur moment, celui où on écrit les remerciements.

Aucun mot de la langue française ni de la langue arabe ni d'aucune autre langue ne pourrait exprimer ma gratitude envers ma directrice de thèse, Pr. Zaia Alimazighi pour son aide et pour la confiance qu'elle m'a toujours témoignée.

Je remercie Pr. Nadjib Badache à un double titre, en tant que professeur pour m'avoir fait l'honneur d'accepter de présider ce jury et en tant que directeur du CERIST pour son soutien et sa confiance que j'espère ne pas décevoir.

Je remercie Dr. Hamid Azzoune et Dr. Ahmed Guessoum d'avoir donné de leur temps pour expertiser mon travail et par la suite accepté de participer au jury.

Un grand merci au Pr. Zizette Boufaida ainsi qu'au Dr. Okba Kazar d'avoir accepté de faire le déplacement pour participer au jury.

Merci à mes amis Souad, Nadia, Omar et Djamel, Karima et Houria pour leur soutien moral et pour toutes les choses que nous avons partagées. A ma grande amie Souad, courage, tu vois, c'est possible !

Je remercie les membres de l'équipe ISI pour l'ambiance des réunions avec une pensée particulière pour Latifa et Kamel.

Je remercie Noureddine Meftouh pour son aide toujours diligente et pour tous les articles qu'il m'a fournis à chaque fois que j'en ai exprimé le besoin sans jamais faillir.

Je remercie Djamel Dib pour sa disponibilité et son aide à chaque fois que je l'ai sollicité aussi pour des articles.

Je remercie Salima du service formation pour son aide et sa complicité à chaque fois que j'avais b de m'isoler.

J'ai certainement oublié des personnes, que ceux que je n'ai pas cités me pardonnent.

Je remercie mes parents, ma sœur Hamida et mon frère Rachid ainsi que mon beau- père Ahmed pour leurs encouragements.

Enfin, à mon mari et à mes enfants, je dis tout simplement : Merci d'être là et pardon pour tous les weekends que j'ai passé à bosser ma thèse au lieu de vous préparer des gâteaux, j'essaierai de me rattraper, promis !

Hassina.

« Beaucoup de gens croient qu'ils pensent alors qu'ils remettent seulement en ordre leurs préjugés. »
William James.

Résumé

Aujourd'hui, avec le développement sans cesse croissant de grands volumes d'information textuelle sur le web, le besoin en outils de traitement du langage naturel devient crucial. Si de tels outils sont disponibles pour les langues Indo-Européennes, il n'en va pas de même pour les autres langues comme l'arabe, le chinois, le coréen, les langues slaves, … ce qui fait du TAL, à l'heure actuelle, un domaine en plein essor. Le traitement automatique de la langue arabe constitue aussi le centre d'intérêt de beaucoup de chercheurs arabes et occidentaux. Nous croyons que tout travail sur la langue arabe doit prendre en considération les spécificités de cette langue et pour éviter de développer des systèmes ad-hoc, doit faire référence à la linguistique arabe. La Tradition Grammaticale Arabe (TGA) nous offre un cadre théorique (linguistique) et méthodologique pour entreprendre un travail fondé en TAL arabe. Cependant, la TGA bien qu'appelée par ses fondateurs 'ilm al 'arabiyya' (علم العربية) n'offre pas un modèle formel au sens contemporain et pouvant directement être utilisé par des informaticiens.

L'objectif que nous nous sommes fixés dans cette recherche est la ***''simulation''*** algorithmique de l'approche de la TGA et ultimement ***d'induire une grammaire formelle de la TGA***. Avant d'arriver à l'induction de la grammaire, la simulation algorithmique de la TGA devrait nous permettre d'offrir quelques outils de base pour le TAL arabe. Notre contribution porte donc sur les deux aspects pratique et théorique.

Dans un premier temps, nous proposons une approche non supervisée fondée sur une simulation algorithmique de la TGA pour la découverte automatique des structures de la langue arabe, en l'occurrence les structures morpho- lexicales et les structures syntaxiques simples. Notre approche n'utilise aucune ressource : ni dictionnaire, ni tables de préfixes ou suffixes prédéfinis.

Dans un second temps, nous proposons la formalisation des structures obtenues dans le cadre de la théorie mathématique des catégories.

Abstract

Today, with the increasing development of large volumes of textual information on the web, the need for tools of natural language processing becomes crucial. If such tools are available for Indo-European languages, this is not the case for other languages such as Arabic, Chinese, Korean, Slavic languages, which make NLP, currently, an important growing field.
Automatic processing of Arabic is also the focus of many Arab and Western scholars. We believe that any work on Arabic language must take into account the specificities of this language and to avoid developing ad-hoc systems, must refer to Arabic linguistics. Arabic Grammatical Tradition (AGT) provides a theoretical framework and a methodology for undertaking a founded work in Arabic NLP. However, AGT doesn't provide a formal model in the contemporary sense.
Our aim in this research is an algorithmic "simulation" of AGT approach and ultimately to induce a formal grammar of AGT. The algorithmic simulation of AGT should allow us to provide some basic tools for Arabic NLP. Our contribution is therefore both practical and theoretical.
First, we propose an unsupervised approach based on an algorithmic simulation of AGT for automatic discovery of Arabic language structures. Our approach doesn't use any resources: no dictionary, no tables of predefined prefixes or suffixes and leads to discovering the morpho- lexical paradigms and syntactic structures.
In a second step, we propose the formalization of the structures obtained in the framework of the modern mathematical theory of categories.

Table des matières

Tableau de translitération de l’alphabet Arabe

coup de glotte	’
أ	a
ب	b
ت	t
ث	th
ج	j
ح	ḥ
خ	kh
د	d
ذ	dh
ر	r
ز	z
س	s
ش	sh
ص	ṣ
ض	ḍ
ط	ṭ
ظ	ẓ
ع	‘
غ	gh
ف	f
ق	q
ك	k
ل	l
م	m
ن	n
ه	h
و	w
ي	y
voyelle longue a	ā
voyelle longue i	ī
voyelle longue u	ū

Chapitre 1. Introduction

1. préambule

La Tradition Grammaticale Arabe (TGA) est à ce jour la ‘’Référence’’ en études linguistiques arabes. En fait, quand on parle de linguistique arabe, on parle de tradition grammaticale arabe, du Kitāb de Sibawayh الكتاب(**Le** Livre) car tout ce qui a été fait depuis la TGA a été fait sur et autour de la TGA qui n’a pas encore fini de livrer tous ses secrets.
Le travail que nous présentons dans cette thèse est une tentative pour saisir quelques uns de ces secrets, d’en capturer la finesse, la subtilité à la lumière des paradigmes scientifiques actuels. Ce travail a été et est pour nous, plus qu’un travail préparé en vue de l’obtention d’un diplôme, une recherche d’absolu, une quête du graal. Le graal des scientifiques et des philosophes est de trouver la théorie ou le modèle (formel) ***abstrait*** qui permet de décrire et d’expliquer la réalité de la façon la plus ***économique*** possible. Le graal des informaticiens est de trouver ou de rendre un tel modèle calculatoire. Les modèles des scientifiques et des informaticiens ne peuvent pas rendre compte néanmoins de toute la réalité du monde qui est complexe et hétérogène. La réalité du monde est alors compartimentée en domaines ou univers de discours (physique, biologie, mathématiques, linguistique, …) à partir desquels ou sur lesquels on induit et on applique la théorie ou le modèle. L’absolu s’il est inatteignable de par notre contingence et de par le fait que l’intellect humain ne peut embrasser l’universel dans sa totalité est néanmoins envisageable au travers des opérations universelles de l’intellect (qu’il soit occidental, arabe, chinois, …) comme l’abstraction, la généralisation, l’induction, la déduction, … . Ce travail a été pour nous l’occasion d’un voyage dans cette universalité, et il est remarquable de pouvoir relater une pensée vielle de quinze siècles aux évolutions scientifiques contemporaines.

> « *Nous pouvons dire aussi : l’histoire n’est d’entrée de jeu rien d’autre que le mouvement vivant de la solidarité et de l’implication mutuelle de la formation du sens et de la sédimentation du sens originaire* »
>
> *Frederic Patras*

En fait, les opérations de l’intellect sont universelles, le reste est question de méthodologie. C’est ce que nous allons voir tout au long de cette thèse. Et nous espérons que les résultats modestes que nous décrivons ici soient un début et non une fin de quelque chose.

2. Problématique et objectifs du travail

Aujourd'hui, avec le développement sans cesse croissant de grands volumes d'information textuelle sur le web, le besoin en outils de traitement du langage naturel devient crucial. Si de tels outils sont disponibles pour les langues Indo-Européennes, il n'en va pas de même pour les autres langues comme l'arabe, le chinois, le coréen, les langues slaves, … ce qui fait du TAL, à l'heure actuelle, un domaine en plein essor. Le traitement automatique de la langue arabe constitue, par ailleurs, le centre d'intérêt de beaucoup de chercheurs arabes et occidentaux. Nous croyons que tout travail sur la langue arabe doit prendre en considération les spécificités de cette langue et pour éviter de développer des systèmes ad-hoc, doit faire référence à la linguistique arabe. La Tradition Grammaticale Arabe (TGA) nous offre un cadre théorique (linguistique) et méthodologique pour entreprendre un travail fondé en TAL arabe. Cependant, la TGA bien qu'appelée par ses fondateurs 'ilm al 'arabiyya (علم العربية) n'offre pas un modèle formel au sens contemporain et pouvant directement être utilisé par des informaticiens. Des descriptions linguistiques limitées et non consensuelles (bien que faisant généralement toutes référence à la TGA) sont utilisées dans les travaux de TAL arabe.

L'objectif que nous nous sommes fixé dans cette recherche est la ***''simulation''*** de l'approche de la TGA et ultimement ***d'induire une grammaire formelle de la TGA***. Avant d'arriver à l'induction de la grammaire, la simulation algorithmique de la TGA devrait nous permettre de contribuer à offrir quelques outils de base pour le TAL arabe. Notre contribution porte donc sur les deux aspects pratique et théorique ; nous relaterons aussi la TGA à des questions linguistiques et scientifiques contemporaines et nous verrons comment la TGA peut contribuer à y répondre.

3. Contexte linguistique

Comme l'objet d'étude et d'investigation dans ce travail est une langue naturelle, nous devons le relater à la linguistique moderne qui, depuis Saussure, est la ***science*** qui étudie le langage. Nous parlerons notamment du distributionalisme et du débat empirisme vs formalisme (linguistique de corpus vs linguistique théorique). La TGA a été constituée à partir du corpus de l'arabe fuṣḥa (الفصحى) c'est-à-dire l'arabe parlé par les arabes natifs, autrement dit, la constitution de la TGA s'est effectuée sur la langue en usage ou la

performance pour utiliser une terminologie Chomskyenne. De nombreux linguistes contemporains ont reconnu la nature distributionnelle de la démarche de la TGA. En linguistique contemporaine, le distributionalisme prend origine dans l'école structuraliste américaine dont la figure de proue est Zellig Harris qui a développé la méthode distributionnelle. Le distributionalisme a très tôt été critiqué par Chomsky, lui-même élève de Harris et depuis, le programme de la grammaire générative qui s'articule autour de la notion de compétence (au lieu de performance), d'universaux linguistiques et d'autonomie de la syntaxe a occupé le devant de la scène pendant des décennies, en linguistique mais aussi en grammaire formelle et en informatique (linguistique). Des questions d'ordre calculatoire et épistémologique ont contribué à reposer les problématiques et le paysage aujourd'hui est plus varié, plus riche.

Sur un autre plan, l'avènement de l'Internet et des grands volumes de données textuelles, a remis à l'honneur les corpus et l'analyse distributionnelle a inspiré notamment beaucoup de travaux en TAL non supervisé. En effet, le corpus tant décrié par Chomsky revient aujourd'hui sur le devant de la scène. Les approches non supervisées en TAL trouvent tout leur intérêt pour les langues pour lesquelles il n'existe que peu de ressources et d'outils linguistiques. Les langues latines restant encore les langues les mieux nanties en la matière. L'arabe fait partie des langues pauvres en termes de ressources et d'outils linguistiques disponibles pour le TAL.

4. Méthodologie de travail

Dans un premier temps, à travers l'étude de la TGA et sa comparaison avec le distributionalisme Harrissien, nous montrons l'originalité de la méthode de TGA et nous proposons une méthode complètement inspirée de TGA pour la ***découverte automatique*** des structures et des niveaux de structures de la langue arabe à travers l'analyse d'un corpus électronique brut de textes écrits en arabe standard moderne. L'approche que nous proposons, contrairement aux autres approches n'utilise ni lexique ni tables d'affixes prédéfinies. Ces algorithmes de découverte ou le corpus segmenté obtenu, peuvent servir de base pour différentes applications telles que l'étiquetage en parties du discours, l'extraction de connaissances, le résumé automatique, …

Dans un second temps, partant d'un rapprochement que nous faisons entre l'approche de l'analyse distributionnelle du 'qiyās'(قياس) de la TGA pour rendre compte des structures linguistiques et de l'approche de la théorie des catégories pour les structures

mathématiques, nous proposons la théorie mathématique des catégories (TC) comme cadre pour une lecture moderne et formelle de la TGA, nous montrons que les structures que décrit la TGA sont des catégories au sens de la TC et nous présentons une esquisse d'un modèle catégoriel de la TGA.

5. Contributions

Notre contribution s'articule principalement autour de deux points :

- une contribution à l'état de l'art du TAL arabe : peu de travaux existent encore dans le cadre d'une approche empirique du TAL arabe et les travaux existant utilisent des connaissances sur la langue inspirées principalement des modèles occidentaux. L'originalité de notre approche est qu'elle est entièrement fondée sur l'approche empirique de la TGA : notre approche est une simulation algorithmique de la TGA dont le but est de redécouvrir les structures de la langue à la manière de cette dernière.
- une contribution aux questions linguistiques et scientifiques actuelles : à travers notre formalisation des résultats obtenus dans le cadre de la théorie mathématique des catégories, nous contribuons à la compréhension des mécanismes opératoires mis en œuvre dans la TGA et dont jusqu'à aujourd'hui, les chercheurs n'ont donné que des questionnements et des bribes de réponse en essayant de faire coïncider tel ou tel phénomène linguistique arabe à telle ou telle théorie linguistique moderne. Pour pouvoir arriver à une vision de l'approche de la TGA dans son ensemble, nous avons choisi de nous situer dans le cadre de l'approche scientifique au sens large et non pas dans le cadre d'une quelconque théorie linguistique moderne construite pour les langues indo-européennes. Notre approche peut être aussi envisagée pour les autres langues, car si la théorie des catégories connait de plus en plus d'applications en informatique théorique, notamment en sémantique des langages de programmation, elle n'a encore jamais été utilisée pour la modélisation des structures langagières.

Sur un autre plan, les résultats de ce travail montrent la position de la TGA par rapport au débat formalisme vs empirisme et en particulier à l'hypothèse de l'innéisme linguistique dont le principal défenseur est Chomsky.

6. Organisation de la Thèse

Cette thèse est organisée comme suit : le chapitre 2 décrit les concepts linguistiques contemporains qui serviront de cadre initial pour ce travail à savoir le distributionalisme, la linguistique de corpus, … le chapitre 3 présente un tour d'horizon des approches empiriques en TAL et en particulier les approches non supervisées, le chapitre 4 présente la TGA. Le chapitre 5 présente notre approche algorithmique inspirée de la TGA pour la découverte des structures et les niveaux de structures de la langue arabe, notre approche permet de découvrir tous les morphèmes ainsi que toutes les lexies du corpus et aussi quelques structures syntaxiques de base. Dans le chapitre 6, nous présentons la théorie mathématique des catégories, un retour sur la TGA et le qiyās pour présenter une esquisse catégorielle du qiyās, en particulier nous proposons une codification des structures morpho- lexicales en catégories. Enfin, le dernier chapitre consiste en une synthèse et quelques conclusions.

Chapitre 2. Pré- requis linguistiques : Empirisme Vs Formalisme

1. Introduction

Si les travaux de Chomsky ont marqué le déclin de l'empirisme dans l'étude linguistique vers le milieu du siècle dernier, la masse, sans cesse grandissante d'information textuelle aujourd'hui disponible, rend cruciale la nécessité de disposer d'outils automatiques pour le traitement de cette information d'une part et remet ainsi à l'honneur les études linguistiques sur corpus, d'autre part. Aujourd'hui, la linguistique de corpus est indéniablement liée aux corpus électroniques.

Puisque notre objet d'étude est une langue naturelle, nous allons commencer par présenter les concepts linguistiques qui servent de cadre pour ce travail. Nous présentons donc dans ce chapitre un aperçu sur les grandes idées qui ont marqué notamment le 20ème siècle et qui nous intéressent directement dans ce travail. Nous nous intéresserons particulièrement à la linguistique de corpus et à la notion de distribution qui tout en étant une notion de la linguistique contemporaine, est aussi au cœur de la TGA et donc au cœur de ce travail. Nous commencerons par le structuralisme qui constitue le point de départ de l'institution de la linguistique comme *science* et qui constitue l'école de pensée servant de cadre à la recherche moderne sur le langage et même dans les autres sciences.

2. Le Structuralisme

Le structuralisme prend origine dans la *linguistique structurale* du suisse Ferdinand de Saussure et par la suite des écoles linguistiques de Prague et de Moscou. Vers le milieu du vingtième siècle, le structuralisme est devenu un mouvement intellectuel et se développa pour devenir l'approche scientifique privilégiée dans les milieux académiques et fut appliqué à différents domaines autres que l'analyse du langage, tels que l'anthropologie, la sociologie, la psychologie, la critique littéraire et l'architecture.

La linguistique structurale débute après la publication posthume du *Course in General Linguistics* en 1916, compilé par les étudiants de Saussure à partir de ses conférences (Saussure, 1966). Le livre s'est avéré être très important et les idées de Saussure en fournissant les fondements de la linguistique et de la sémiotique moderne firent de Saussure le père de ces deux disciplines. La linguistique structurale stipule que la langue

doit être étudiée comme *un système doté d'une structure décomposable et dans lequel les éléments ne sont pas étudiés en eux-mêmes mais à travers les relations les reliant les uns aux autres dans le système* (Saussure, 1966). Cette approche de l'étude de la langue est différente des approches traditionnelles qui considéraient la relation entre les mots (signifiants) et les *choses qu'ils désignent* (les signifiés).

Ferdinand de Saussure introduit donc la définition négative du sens : les signes de la langue sont seulement déterminés par leurs relations aux autres signes et non pas par la donnée d'une énumération (positive) de leurs caractéristiques. Autrement dit, les signes de la langue s'arrangent en espaces de sens et leur valeur est seulement déterminée de façon différentielle par la valeur des autres signes eux-mêmes déterminés de manière différentielle (Saussure, 1966).

De Saussure distingue deux sortes de relations entre les signes. Les relations *syntagmatiques* (mots voisins dans une phrase) et les relations *associatives* ou *paradigmatiques.* (Saussure, 1966). Alors que les relations syntagmatiques peuvent être observées à partir de la langue (qui correspond au corpus de textes dans ce travail), toutes les autres relations subsumées par des « associations » ne sont pas directement observables et peuvent être individuellement différentes, d'où la nécessité d'un modèle théorique de la langue.

La linguistique structurale travaille en collectant un **corpus** d'énoncés et en tentant de **classer** les éléments du corpus aux différents niveaux linguistiques : les phonèmes, les morphèmes, les catégories lexicales, les phrases nominales, les syntagmes nominaux, les syntagmes verbaux et les types de phrases. Les méthodes clés de Saussure sont l'analyse syntagmatique et l'analyse paradigmatique qui déterminent les unités lexicales et syntaxiques selon leur contraste aux autres unités du système

Après Saussure, la linguistique structurale suit deux directions. En Amérique, de 1930 jusqu'au milieu des années 1950, Léonard Bloomfield, après lecture du *cours de linguistique générale* jeta les bases du distributionalisme en mettant de côté la question du sens et en préconisant une *approche empirique* et mécaniste de la langue (Bloomfield, 1934). En 1957, après la publication de *Syntactic Structures* par Chomsky (Chomsky, 1957), le programme chomskyen partant de la *Grammaire Générative (approche théorique)* occupa pour longtemps le devant de la scène.

3. Le Distributionalisme

Une contribution importante au structuralisme est attribuée à Zellig Harris qui a tenté de découvrir quelques unes des relations associatives ou paradigmatiques. *Son hypothèse distributionnelle stipule que les mots ayant des sens similaires peuvent être observés dans des contextes similaires* ou comme l'a popularisé J.R. Firth « you shall know a word by the company it keeps» (Firth, 1957). Cette expression peut être considérée comme l'idée maitresse du distributionalisme qui détermine la similarité des mots en comparant leurs contextes. Le distributionalisme ne s'intéresse pas à des occurrences isolées mais plutôt à la distribution d'un mot donné, c'est-à-dire la totalité des contextes dans lesquels il peut apparaitre (Harris, 1946, 1951). La notion de contexte est simplement définie par les éléments de la langue en relation avec le mot. La taille et la structure de ce contexte est arbitraire et différentes notions de contexte produisent différents types de similarité entre les mots partageant ces contextes.

Les travaux de Miller et Charles (Miller & Charles, 1991) ont permis l'opérationnalisation de cette hypothèse et la définition de la similarité de deux signes comme une fonction sur leurs contextes globaux : plus deux mots ont des contextes en commun, plus ils peuvent souvent être inter-changés et le plus ils sont similaires. Ce qui peut donner lieu à des ***procédures de découverte*** qui comparent les éléments linguistiques selon leurs distributions.

3. 1 La méthode distributionnelle

Léonard Bloomfield (Bloomfield, 1934) fut le premier à proposer d'étudier la langue sur la base de la *seule forme* indépendamment de tout autre phénomène linguistique en particulier, sans faire appel au sens. Son élève, Zellig Harris formalisa cette idée à travers ce qui est connu comme la méthode distributionnelle (Harris, 1951).

L'école structurale distributionaliste doit son nom à l'utilisation de la notion de ***distribution.*** (Harris, 1951) présente l'ensemble des méthodes de recherche utilisées en linguistique descriptive ou plus exactement structurale. La méthode consiste à construire un échantillon d'une langue appelé *corpus* afin de décrire la *structure* de cette langue à travers la recherche de *régularités*. L'étude des régularités se base sur la notion de distribution. La distribution d'un élément (phonème, morphème, séquence de morphèmes) est la somme des environnements de cet élément (Harris, 1946, 1951, 1954). Ce seul

critère est utilisé pour catégoriser les éléments. Le sens n'intervient pas dans la démarche. La recherche des régularités se fait en *segmentant* les séquences du corpus pour mettre à jour les régularités entre éléments ainsi segmentés. Nous comparerons par la suite les procédures de Harris aux procédures de la TGA.

La méthode distributionnelle s'articule donc essentiellement sur :

1. L'utilisation d'un corpus,
2. La notion de distribution,
3. L'utilisation de la ***forme*** seule, sans recours au sens.

On trouve dans (Harris, 1954) une présentation générale de la méthode distributionnelle et dans (Harris, 1951) un exposé détaillé des procédures utilisées. L'introduction de (Harris, 1951) restitue bien quel est l'intérêt d'un tel travail pour Harris qui est beaucoup plus méthodologique que pratique (Déjean, 1998). En effet, le travail de Harris est ici à considérer beaucoup plus sur le plan méthodologique qu'opérationnel, d'ailleurs n'écrit-il pas dans cette introduction :

> "Les méthodes particulières décrites dans ce travail ne sont pas essentielles. Elles sont proposées comme des procédures générales d'analyse distributionnelle applicables à des données linguistiques" (Harris, 1951).

Si l'on en croit (Nevin, 1993), Harris n'a jamais prétendu que la méthode qu'il propose permettait de générer une grammaire à partir de textes (Déjean, 1998).

La méthode distributionnelle repose donc, sur une notion centrale : la distribution d'un élément. L'observation de Harris sur la distribution d'un élément est simple :

> "Les parties d'une langue n'apparaissent pas arbitrairement relativement les unes aux autres : chaque élément se rencontre dans certaines positions par rapport aux autres" (Harris, 1954).

De cette notion de distribution découle tout le processus de découverte des structures. Voici la définition que Harris en donne :

> "La distribution d'un élément se définit comme la somme de tous les environnements de cet élément. L'environnement d'un élément A est la disposition effective de ses co-occurrents. C'est-à-dire des autres éléments, chacun dans une position déterminée, avec lesquels figure A pour produire un énoncé" (Harris, 1954).

Ce critère est utilisé pour catégoriser les éléments d'un corpus. Deux éléments ayant une même distribution (critère de similarité) sont considérés comme appartenant à une même *classe distributionnelle* (nous verrons que le regroupement par similarité est aussi le principe qui fonde la recherche des naẓā´ir (نظائر) et la constitution des « bāb »-s dans TGA).

L'analyse distributionnelle est donc basée sur la détermination des rapports syntagmatiques et paradigmatiques d'une unité lexicale. Les rapports syntagmatiques réfèrent à tous les contextes dans lesquels une unité s'emploie. Les rapports paradigmatiques réfèrent aux relations d'équivalence entre une unité et toute autre unité qui pourrait lui être substituée dans un même contexte. Les unités lexicales qui sont en relations paradigmatiques avec une unité donnée ont les mêmes relations syntagmatiques qu'elle dans l'énoncé. Les unités qui n'ont pas les mêmes relations ne peuvent pas appartenir à la même classe paradigmatique.

3.2 Les limites de la méthode distributionnelle

Néanmoins, si la notion de distribution est centrale, elle est aussi problématique et ce point est important pour notre comparaison ultérieure avec TGA. En effet, quiconque doit vouloir effectuer une analyse distributionnelle doit apporter une réponse aux questions suivantes: comment construire les contextes distributionnels, sélectionner les bons contextes et classer les mots ?

3.2.1 Comment construire les contextes ?

Le premier problème rencontré dans l'analyse distributionnelle est celui de la définition du contexte. Nous avons vu que les mots sont regroupés en classes distributionnelles, c'est-à-dire que les mots partageant une même distribution sont regroupés dans une même classe. Quelle est la distribution d'un mot ? Les phrases dans lesquelles, il apparait ? Dans ce cas, aucun mot n'a de distribution semblable et aucun regroupement ne peut se faire. Il faut donc réduire la taille de la distribution. Celle utilisée dans les algorithmes de catégorisation est de quelques mots avant et/ou après (Déjean, 1998).

3.2.2 Comment classer les mots ?

Le deuxième écueil de la méthode concerne la variété de contextes dans lesquels un mot peut apparaitre. Le problème peut être contourné en regroupant les mots qui partagent un contexte assez "proche". La difficulté consiste alors à définir la distance de ressemblance entre deux mots. Certains mots se ressemblent plus que d'autres, ce qui permet d'établir *une hiérarchie sur les classes* obtenues. Mais pour régler ce problème des contextes il faut une *théorie formelle de la langue*. La mise au point d'une analyse distributionnelle a pour tâche principale, la construction de ces contextes distributionnels.

3.3 Critiques adressées à la méthode distributionnelle

De nombreuses critiques ont été adressées à la méthode distributionnelle. Certaines sont d'ordre méthodologique comme celle de Noam Chomsky et d'autres d'ordre pratique synthétisées entre autres, par (Déjean, 1998).

3.3.1 La critique de Chomsky

Le linguiste Noam Chomsky, élève de Harris a très fortement contesté l'intérêt de la méthode. Il condamne fortement le travail basé sur la notion de procédure de découverte et sur l'étude de corpus. Il écrit :

> "Nous pensons qu'il est déraisonnable d'attendre d'une théorie linguistique qu'elle fournisse plus qu'une procédure pratique d'évaluation des grammaires. (…) Autrement dit, elles(les propositions) essaient de formuler des méthodes d'analyses dont un chercheur pourrait réellement se servir, *s'il en avait le temps*, pour construire une grammaire d'une langue directement à partir de données brutes. Il me parait douteux que cet objectif puisse être atteint d'une manière intéressante, et je crains que toute tentative de cet ordre ne conduise à un dédale de procédures analytiques de plus en plus complexes et raffinées, qui laisseront sans solution beaucoup de problèmes importants concernant la nature de la structure linguistique". (Chomsky, 1969)

> "Les allusions à des 'procédures de découverte' ou 'méthodes objectives' présumées bien connues ne font que masquer les conditions effectives où le travail linguistique doit se poursuivre pour le moment". (Chomsky, 1965)

S'il est vrai qu'une génération automatique de grammaire à partir d'un corpus semble un défi assez difficile, les résultats obtenus en essayant de le relever peuvent être très intéressants (Déjean, 1998). C'est ce que nous pensons, et nous montrerons que la méthode des anciens grammairiens arabes est une méthode distributionnelle qui a permis d'induire la TGA en appliquant quelques autres critères méthodologiques formels.

Pour Chomsky, le travail à partir d'un corpus ne peut servir de base à un travail linguistique :

> « Il y a tout d'abord, la question de la manière dont on peut obtenir des informations sur la compétence du locuteur-auditeur, sur sa connaissance de la langue. Comme la plupart des faits intéressants et importants, celle-ci n'est pas accessible à l'observation directe et ne saurait être extraite des données par des procédures inductives d'aucune espèce bien connue. (…) en bref, *il se trouve malheureusement qu'on ne connait aucune technique formalisable adéquate pour obtenir une information solide touchant les faits de la structure linguistique (et cela n'a rien de spécialement surprenant*) (Chomsky, 1965).

Pour Chomsky, l'objet d'étude de la linguistique doit être principalement la *compétence* du locuteur-auditeur, c'est-à-dire la connaissance que ce dernier a de sa langue. Le distributionalisme en travaillant sur le corpus s'appuie sur la *performance* ou l'*usage* de la langue.

3.3.2 Le Problème du Sens

La deuxième critique adressée à la méthode distributionnelle concerne le 'rejet' du sens. Le principe de la méthode distributionnelle est justement de remplacer l'utilisation du sens par la notion de distribution.

> « La description du signifié est (…) le point faible de l'étude du langage » (Bloomfield, 1934).

Mais en réalité, la condamnation du sens chez Harris est beaucoup moins forte (Harris, 1954). En fait, le rejet du sens dans tous les domaines de la linguistique est absurde. Le problème est de définir le champ d'étude des travaux, c'est ce que fait Harris : son objectif est de proposer des méthodes en linguistique descriptive, et pour lui la linguistique descriptive « *ne concerne pas l'ensemble des activités de la parole, mais les régularités dans certaines caractéristiques de la parole »* (Harris, 1951).

3.3.3 L'impossibilité pratique de la méthode

Cette critique est d'ordre pratique :

> « On constate qu'une analyse distributionnelle au sens strict du terme n'a jamais été effectuée, pour une langue. Les applications que l'on connait sont des descriptions où, guidé par l'*intuition sémantique,* le linguiste opère des segmentations et des classements ; mais les arguments qu'il avance en faveur de ces opérations sont de nature distributionnelle. Or les phénomènes distributionnels sont nombreux d'une part, et d'autre part ils ne sont pas tous pris en compte de façon systématique. Il s'en suit que dans l'ensemble des faits de distribution, il y en a qui étaieraient une description, mais on en trouve aussi qui iraient à l'encontre de cette même description. L'analyse distributionnelle dans l'acception stricte du terme (c'est-à-dire sans critère sémantique) est une utopie. (Mahmoudian, 1981).

La critique est simple mais pertinente. La réponse aussi. Devant la complexité de la tâche qui peut s'étonner de ce résultat. ? Et personne ne contredit ces remarques, même Harris y souscrit, l'introduction de (Harris, 1951) va dans ce sens :

> "These procedures also do not constitute a necessary laboratory schedule in the sense that each procedure should be completed before the next is entired upon. In practice, linguists take unnumbered shortcuts and heuristic guesses, and keep many problems about a particular language before them at the same time (…)".

Force est de constater que peu nombreux ont été les chercheurs qui ont poursuivi les traces de Harris. Nous pensons que cela est dû principalement à la place qu'a prise Chomsky au 20ème siècle et le succès de ses grammaires formelles en informatique. Nous verrons qu'aujourd'hui, la méthode de Harris a inspiré beaucoup de travaux en TAL (non supervisé). Il est néanmoins vrai que le manque de formalisme de la méthode présentée par Harris, rend celle-ci inopérante à l'état où Harris l'a laissée (Déjean, 1998).
Nous comparerons plus loin, la méthode distributionnelle de Harris avec celle de la TGA et nous verrons que la méthode de la TGA a permis d'induire les « ḥudūd » (حدود) ou règles de la langue arabe.

4. Empirisme vs Formalisme

Bien que le mot linguistique de corpus n'ait été utilisé ni par Saussure ni par les distributionalistes, nous avons vu que le structuralisme a introduit l'utilisation du corpus dans l'étude de la langue et que cette idée est plus concrète à travers le distributionalisme. Cependant avec les critiques sévères de Chomsky vers la moitié du 20ème siècle, on peut dire que les méthodes empiriques ont été mises de côté pour les approches formalistes théoriques. L'apparition en linguistique anglophone d'un nouveau courant appelé "corpus linguistics" se trouve en quelque sorte officialisé au début des années 90 par une floraison d'articles visant explicitement à faire le point sur ce « nouveau » domaine. Nous citerons en particulier les articles de J. Aarts (Aarts, 1990), qui propose une "évaluation" de la linguistique sur corpus, et de G. Leech (Leech, 1991), qui entreprend d'en établir "l'état de l'art". Ces publications attestent l'existence d'un groupe de linguistes qui se réclament de l'utilisation de corpus, et même se définissent par cette utilisation.

Le courant "corpus" continue actuellement à prendre de l'ampleur boosté par le développement de la puissance des ordinateurs qui permet le développement de corpus électroniques de plus en plus grands.

Du côté arabe, on parle aussi de linguistique arabe de corpus, des workshops sont mêmes récemment organisés autour de cette thématique. En vérité, même si la tradition grammaticale arabe n'est pas largement connue, les auteurs ayant étudié cette langue ont déjà relevé que la TGA est une tradition de corpus (Carter, 1973) (Versteegh, 1997) car elle est avant tout fondée sur l'expérience et l'observation des données, autrement dit sur la langue en usage.

4.1 Approche empirique d'analyse de la langue

Une approche empirique (du grec expérience) est une approche qui est fondée sur *l'observation* des données du monde réel, par opposition à l'utilisation d'exemples artificiellement construits ou l'utilisation de l'intuition. Une approche empirique est aussi une méthode de construction ou d'infirmation d'hypothèses en utilisant l'observation et l'expérience ; c'est une approche inductive (par opposition à déductive) de raisonnement ou de formation d'hypothèses basées sur ces observations (Popper, 1950).

L'empirisme est aussi fréquemment associé au rejet des "idées innées" qui représentent la connaissance présente dans l'intellect humain préalablement à une quelconque expérience sensorielle. L'empirisme radical comme le behaviorisme a été fortement critiqué par (Chomsky, 1959) et (Kuhn, 1962). Pour Kuhn, la création de nouvelles solutions pour des problèmes existants arrive souvent en dehors des frameworks existants et non dans le cadre d'un développement graduel de théories à travers l'observation et l'expérimentation. Pour lui l'expérimentation scientifique n'est pas aussi non biaisée et neutre qu'on le prétend parce que les scientifiques ont déjà des préjugés ainsi qu'une connaissance préalable qui influencent la conduite de l'expérimentation ainsi que l'interprétation des résultats. Voici une description concise d'une approche empirique qui va nous servir de cadre pour notre travail, nous essaierons notamment à travers l'étude de l'approche de la TGA de relater un empirisme 'modéré' à une objectivité mathématique.

Approche empirique

1. observer dans le but de comprendre ou de structurer quelque chose (une matière donnée, dans notre cas linguistique), nous avons besoin d'observations et d'expérimentations sur la matière en question. Ces observations devraient capturer autant que possible d'éléments impliqués mais en même temps ne devraient pas (et ne peuvent pas) être complètes (exhaustives).
2. apprendre sur la base des résultats de l'exploration, il est possible de détecter des *régularités* et de formuler des *hypothèses* à propos des mécanismes responsables des phénomènes observés.
3. vérifier, les résultats obtenus doivent être vérifiés par plus d'observation et d'expérimentation.

Le résultat de ces trois étapes est une description généralisée des mécanismes sous-jacents aux régularités observées sur les données.

4.2 L'hypothèse de l'innéité

En réponse aux tenants de l'empirisme, le nativisme linguistique ou hypothèse d'innéité est le postulat défendu par (Chomsky, 1986) et (Pinker, 1994) parmi d'autres et qui stipule que la faculté du langage est innée chez les êtres humains. Chomsky fut l'instigateur de

cette idée dans la lignée de sa critique de l'empirisme qui en se basant sur l'étude des données observables (le corpus de la langue utilisée) ne peut selon lui, rendre compte des aspects liées à l'acquisition et l'apprentissage du langage qui est ancrée dans les facultés psychologiques des individus. Les facultés innées pour Chomsky et les nativistes sont compartimentées en domaines de connaissances. La faculté du langage est centrée autour d'un certain ensemble de principes innés pour chaque individu et qui constituent ce qu'il appelle la Grammaire Universelle. L'argument avancé en faveur de cette thèse est l'argument de la pauvreté du stimulus (Argument from the Poverty of the Stimulus).

> "…the linguistic input or evidence available for the infant child is so impoverished and degenerate that no general domain- independent learning algorithm could possibly learn a plausible grammar without assistance".

Une réfutation de cet argument serait de démontrer qu'il existe un algorithme capable d'apprendre une grammaire de couverture raisonnable à partir de données langagières (Bieman, 2007). *Les approches empiriques en TAL peuvent être vues comme une telle réponse.*

4.3 Linguistique Théorique vs Linguistique de Corpus

Les approches empiriques du langage naturel utilisent directement ou indirectement la notion de distribution, de contexte et de similarité. La méthode distributionnelle a en particulier inspiré les travaux en TAL non supervisé et la majorité des travaux font référence à Harris. Nous aborderons les travaux en TAL au chapitre 3 mais avant, nous pensons qu'il est intéressant de mieux situer le lecteur dans le débat linguistique théorique vs linguistique de corpus (formalisme vs empirisme), de voir comment ce débat a évolué pour nous permettre à la fin de situer notre travail par rapport à ces deux paradigmes et aux questions scientifiques qui y sont posées. Pour cela, nous allons emprunter à (Lager, 1995) quelques citations très intéressantes pour la compréhension de ce débat.

> "Armchair linguistics does not have a good name in some linguistics circles. A caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.)

> Corpus linguistics does not have a good name in some linguistics circles. A caricature of the corpus linguist is something like this. He has all of the primary facts that he needs, in the form of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations.)
> These two don't speak to each other very often, but when they do, the corpus linguist says to the armchair linguist. "Why should I think that what you tell me is true?", and the armchair linguist says to the corpus linguist, "Why should I think that what you tell me is interesting?"
>
> *Charles Fillmore (1992)*

Fillmore (Fillmore, 1992) est arrivé à l'absurdité de la querelle ayant généré la dichotomie entre les deux approches et conclut (Lager,1995):

> "My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body".
>
> *Charles Fillmore (ibid.)*

De toute façon, à en juger par la littérature, la majorité des linguistes agréent plutôt avec Fillmore sur cette question. En effet, la majorité pense que aussi bien la théorie que les données sont les ingrédients nécessaires de la science. Des théories scientifiques pertinentes ne peuvent émerger qu'à travers l'interaction de la théorie et des données empiriques. Concernant l'objet d'étude de la linguistique, beaucoup de linguistes ne croient pas à la 'réalité' du fossé entre langue et parole, compétence et performance, système et processus ou quelque soit le nom qu'on peut donner à cette dichotomie.

> "It can (...) be argued that the putative gulf between competence and performance has been over emphasized, and that the affinity between (say) the grammar of a language as a mental construct and the grammar of a language as manifested in performance by the native speaker must be close, since the latter is a product of the former."
>
> *Geoffrey Leech (1992)*

Quant à notre manière d'accéder aux données, l'observation, l'intuition et l'introspection sont habituellement pensées se compléter l'une l'autre (sauf peut être par quelques extrémistes de chacun des camps).

> "Grammars do not grow magically from corpora. They require the intelligent analytical mind of a grammarian who draws on knowledge of previous studies, on his or her own intuition as well as on observations of text".
>
> *Stig Johansson (1992)*

Concernant la complémentarité entre linguistique descriptive et linguistique théorique (Leech, 1992) écrit :

> "Both types of linguistics are valid in their own terms, and should be regarded as mutually contributory. Descriptive linguistics can be just as answerable to theory as the "theoretical linguistics" of language universals. In fact, descriptive linguistics is more amenable to theory construction and testing in accordance with the tenets of scientific method, because the nature of its data (e.g. utterances in a particular language) is less abstract and more directly observable."
>
> *Geoffrey Leech (1992)*

Pour résumer, nous pouvons dire que bien que dans la réalité, la dichotomie entre les deux approches est toujours pratiquée, la majorité des linguistes s'accordent à penser qu'une telle dichotomie doit être levée dans l'étude linguistique.
Les linguistes théoriciens travaillent habituellement avec des exemples artificiels plutôt qu'avec de vrais énoncés. Ils sont toujours capables d'avancer des choses intéressantes à propos de ces exemples. Les linguistes de corpus sont habituellement préoccupés par la classification, la description et la synthèse des données sans se soucier si le travail est intéressant du point de vue théorique. Nous allons essayer dans ce travail de proposer une approche à partir de la TGA qui allie théorie et corpus.

5. Dynamique de développement d'une théorie en linguistique

Popper (1972) propose le schéma suivant capturant la 'dynamique' du développement d'une théorie scientifique.

$$P_1 \Rightarrow TT \Rightarrow EE \Rightarrow P_2$$

Un problème P_1 est formulé ; une tentative de théorie est proposée TT ; l'élimination d'erreurs générées par cette théorie EE est effectuée, si des erreurs sont trouvées, elles vont donner lieu à plus de problèmes P_2 et ainsi de suite.

Le développement d'une théorie descriptive en partant du corpus en linguistique être vue comme une instance du schéma proposé par Popper :

- Les problèmes sont détectés, soit comme des faits ne pouvant être pris en compte par la théorie courante, soit sous la forme de contradictions ou autres types d'erreurs dans la théorie courante.
- Des mises à jour (ajouts, suppressions, modifications d'éléments) à la théorie courante, prenant en compte des faits nouveaux, ou corrigeant les erreurs
- La partie concernée de la théorie actuelle est testée à nouveau
- Souvent, de nouveaux problèmes surgissent concernant la manière de corriger les erreurs détectées.

6. Conclusion

Comme nous l'avons déjà dit en introduction, nous n'allons pas étudier la TGA dans le cadre d'une quelconque théorie linguistique moderne pour chercher à y trouver des recoupements, approche que nous jugeons artificielle. Nous essaierons plutôt de la situer dans le cadre plus général et ouvert du développement de théories scientifiques.

Chapitre 3. Approches Empiriques en TAL

1. Introduction

Le développement de grands corpus électroniques a marqué le renouveau des approches empiriques pour l'étude des langues naturelles et en TAL. En effet, si la disponibilité de corpus électroniques met à la disposition du linguiste les données langagières dont il a besoin pour son étude, ce dernier va avoir besoin d'outils pour manipuler ces données et c'est le rôle du TAL que de lui fournir ces outils. Les méthodes formelles à la Chomsky ayant montré déjà depuis longtemps leur inadéquation pour le développement d'applications réelles, les travaux en TAL se tournent de plus en plus vers les approches empiriques basées sur des corpus. Nous allons exposer dans ce chapitre les différentes approches de TAL ainsi qu'une synthèse des travaux en TAL non supervisé.

2. Approches en TAL

La discipline de traitement automatique du langage naturel s'articule autour de deux courants bien établis dont les objectifs diffèrent à la nuance près. D'un côté, nous avons la 'computational linguistics' (CL) ou linguistique informatique, principalement influencée par la pensée linguistique et qui vise à implémenter des théories linguistiques selon une approche rationaliste. L'accent est ici mis sur les problèmes linguistiques plutôt que sur un traitement robuste et efficace. D'un autre côté, nous avons le 'natural language processing' (NLP) qui ne se veut, lié à aucune théorie linguistique. Le but est ici, non pas de comprendre la structure du langage comme une fin en soi, mais d'utiliser la connaissance sur la structure du langage pour développer des méthodes aptes à nous aider dans le traitement automatique des données langagières. Le critère pour le NLP est la performance du système et non l'adéquation à une représentation théorique humaine du langage ; cette approche part donc d'une vision pragmatique mais pauvrement fondée sur le plan théorique. Néanmoins, au vu de la recherche actuelle, les deux courants sont difficilement séparables et s'influencent mutuellement. Les approches en TAL peuvent être généralement différentiés selon les points de vue suivants :

2.1 Approches à base de règles vs statistiques

Techniquement, l'histoire du TAL a été dominée par deux paradigmes majeurs : les approches à base de règles et les approches statistiques. Les approches à base de règles (théoriques) se situent dans la lignée de la tradition CL alors que les approches statistiques (empiriques) sont plutôt du côté du NLP. Le principe des approches à base de règles est de développer des règles de plus en plus sophistiquées pour éventuellement encoder tous les phénomènes du langage. Dans le but d'opérationnaliser l'application des formalismes grammaticaux, de grands systèmes ont été développés avec des milliers de règles de grammaire. Néanmoins, ces règles interagissent les unes avec les autres, par conséquent, l'ajout de nouvelles règles rend très lente l'application des règles déjà existantes ce qui fait de la construction des systèmes à base de règles un processus très couteux. Par ailleurs, ce qui est inhérent aux approches descendantes, guidées par l'introspection des systèmes à base de règles est qu'ils ne peuvent travailler qu'avec le sous-ensemble de la langue couvert par les règles. L'expérience a montré que ce sous ensemble est loin d'être proche d'une bonne couverture de la langue (Bordag, 2007) (Biemann, 2007). Le résultat est que ces systèmes ne peuvent traiter qu'une partie des phrases de la langue.

Avec l'avènement des grands corpus textuels électroniques ainsi que l'évolution de la puissance de traitement des ordinateurs, les approches statistiques de traitement du langage naturel devinrent incontournables. En utilisant les modèles probabilistes au lieu de règles, l'approche *empiriste* a très tôt montré sa capacité d'atteindre plus de précision dans les tâches de traitement du langage naturel que l'approche à base de règles. La partie de travail manuel laborieux qui consistait dans les systèmes à base de règles à instruire la machine directement par des règles décrivant comment traiter les données, consiste dans l'approche statistique à étiqueter des exemples de textes (corpus d'entrainement pour les algorithmes) qui montrent ce que devrait être la sortie du système.

Pour comprendre le langage naturel d'un point de vue linguistique et tester les théories grammaticales, il semblerait que l'approche à base de règles s'impose. Cependant pour développer des applications, les méthodes statistiques ont fait leurs preuves en étant plus robustes et plus rapides à mettre au point.

2.2 Rôle de la connaissance

Une autre échelle sur laquelle on peut classer les méthodes de TAL est la distinction entre approches utilisant des connaissances sur le langage (knowledge-intensive) et les approches n'utilisant pas de connaissances sur le langage (knowledge-free) (Biemann, 2007) (Bordag, 2007). Les approches basées sur la connaissance font un usage excessif de ressources langagières telles que dictionnaires, listes de phrases, ressources terminologiques, réseaux sémantiques lexicaux (comme wordnet), thésaurus, ontologies, … . Comme ces ressources sont nécessairement incomplètes, la couverture du système ne sera pas parfaite et nécessitera toujours plus d'efforts manuels pour améliorer sans cesse les ressources.

Les approches libres de connaissances quant à elles ont pour objectif d'éliminer l'effort et l'intervention humains. L'effort humain consistera ici plutôt à mettre au point des procédures qui permettraient de *'découvrir'* cette connaissance par *induction* et non à la donner. Les méthodes qui utilisent peu d'intervention humaine sont aussi appelées méthodes faiblement fondées sur la connaissance (knowledge-weak) (Biemann, 2007) et essaient de combiner l'avantage de ne pas préparer trop de connaissances et d'obtenir de bons résultats en utilisant des ressources déjà disponibles. Dans le cadre d'une approche empirique, la nature et la quantité de connaissances utilisées, déterminent le degré de supervision du système.

3. L'apprentissage en TAL

Dans le cadre d'une approche empirique, il est naturel que les techniques d'apprentissage automatique soient sollicitées. Ces techniques utilisent l'apprentissage supervisé, semi-supervisé ou non- supervisée avec une préférence de plus en plus accrue pour la méthode non supervisée. En réalité, il y a deux bonnes raisons d'essayer d'appliquer les techniques d'apprentissage non supervisées au langage naturel. La première est d'ordre pratique : L'effort impliqué par la construction manuelle de grammaires, dictionnaires et autres ressources est prohibitif et il est donc naturel d'essayer de les extraire automatiquement à partir des données appropriées (Church & Mercer, 1993), (Greferstette, 1994). La seconde raison est théorique et s'intéresse à la modélisation de l'acquisition du langage humain considéré comme une activité cognitive (Brent, 1993) (Brent, 1997) (Brent & Cartwright, 1997).

Nous pouvons diviser les méthodes d'apprentissage en deux grandes classes, les méthodes symboliques et les méthodes statistiques. Certains chercheurs considèrent les modèles connexionnistes comme une troisième classe, mais nous partageons le point de vue de (Biemann, 2007)) et considérons qu'ils peuvent plutôt être vus comme un type spécifique du modèle statistique. Les méthodes symboliques actuellement utilisées incluent les arbres de décision (Qinlan, 1993), la programmation logique inductive (Mugleton, 1997, 1999), (Mugleton & brain, 1999) et l'apprentissage basé sur la transformation (Brill, 1992, 1993). Les approches statistiques sont aussi très utilisées car elles présentent notamment l'avantage d'être résistantes au bruit et applicables dans un grand nombre de situations (Bishop, 1995). Les techniques les plus couramment utilisées sont les Modèles de Markov Cachés, le clustering distributionnel et les grammaires stochastiques à contexte libre.

L'apprentissage supervisé utilise la majorité des techniques connues d'apprentissage automatique et il existe une grande littérature sur le sujet. Néanmoins, il existe des motivations très pertinentes pour les méthodes non supervisées. Premièrement, les données étiquetées sont difficiles et coûteuses à élaborer, en quantité, elles restent toujours limitées et peuvent contenir des erreurs. Deuxièmement, les données étiquetées imposent un schéma particulier d'analyse qui peut s'avérer inadéquat aux objectifs de l'application visée. Troisièmement et c'est la motivation la plus importante : avec les données en langage naturel, il n'existe pas tout à fait d'agrément sur ce qu'une étiquette devrait être (Bieman, 2007).

L'apprentissage supervisé nécessite des données d'apprentissage associant les données d'entrée aux résultats désirés. Le corpus d'apprentissage restreint de fait le champ des possibilités du système : le système ne sait faire que ce qu'il a appris à faire. Les capacités du système ne dépendent donc pas uniquement de ses propriétés intrinsèques mais avant tout du contenu, de la qualité et de la taille du corpus d'apprentissage.
L'apprentissage supervisé consiste à découvrir les régularités présentes dans un ensemble d'exemples pour ensuite les appliquer à des données différentes. Les techniques d'apprentissage gardent pour ainsi dire « en mémoire » les instances rencontrées lors de l'apprentissage pour les réutiliser au moment de l'analyse. Les méthodes d'apprentissage supervisé nécessitent des corpus d'apprentissage, constitués à partir de ressources existantes ou construites manuellement. Ces corpus déterminent ce que le système est capable de faire à l'utilisation. Or ces données ne sont pas toujours disponibles et leur

construction est une tâche longue et fastidieuse, car il est nécessaire de disposer de corpus d'apprentissage de taille suffisante pour obtenir de bons résultats.
Contrairement aux méthodes supervisées, les systèmes d'apprentissage non supervisé ne nécessitent pas de disposer d'exemples type des résultats souhaités. Les seules données nécessaires sont une liste de mots, éventuellement complétée par un corpus ou des ressources facilement disponibles comme des dictionnaires ou des thésaurus. On pourrait croire – à tort – que l'ensemble des informations utilisables dans une liste de mots pour effectuer par exemple une analyse morphologique complète est très limité. Pourtant la diversité des stratégies adoptées semble au contraire montrer que la plus grande difficulté réside dans la capacité à combiner l'ensemble des indices disponibles pour obtenir les meilleurs résultats possibles (Biemann, 2007). Pour synthétiser, les approches d'apprentissage en TAL sont classées comme suit :

Les approches supervisées fournissent au système des données complètement étiquetées ce qui signifie que tous les exemples sont présentés et classés de façon à ce que la machine puisse les reproduire. Le processus d'assigner des étiquettes à de nouvelles instances est appelé classification. Un classifieur est appris à partir des exemples d'apprentissage en entrée.

Les approches semi- supervisées permettent à la machine de prendre en considération des données non étiquetées. Pour un état de l'art sur les approches semi- supervisées voir (Sarkar & Haffari, 2006)

Les approches faiblement supervisées ou bootsrapping sont basées sur une forme d'apprentissage qui utilise encore moins d'exemples d'apprentissage.

Les approches non supervisées n'utilisent pas du tout d'exemples d'apprentissage et construisent généralement des clusters. Les résultats des algorithmes de clustering sont dirigés par les données donc plus 'naturels' et plus adéquats avec la structure sous-jacente aux données (Biemann, 2007). Cet avantage constitue aussi l'inconvénient de la méthode. En effet, sans la possibilité de dire à la machine quoi faire (comme en classification) il est difficile de juger de la qualité des résultats de façon concluante (Redington & al, 1993) (Redington & al, 1995). Cependant, l'absence d'exemples d'apprentissage rend l'approche non- supervisée très attractive. Nous nous intéressons dans ce travail aux approches non supervisées.

4. Travaux en TAL non supervisé

La méthode distributionnelle de Harris a très tôt inspiré les chercheurs en TAL, et la majorité des travaux y font référence. Nous allons présenter ici quelques travaux en TAL basé sur corpus. Cette approche est surtout utilisée pour l'acquisition ou l'apprentissage des structures du langage soit au niveau morpho lexical soit au niveau syntaxique. L'hypothèse distributionnelle de Harris stipulant que les items de la langue apparaissant régulièrement dans des contextes similaires sont sémantiquement voisins, la méthode est aussi utilisée en analyse sémantique. Nous nous intéressons ici aux travaux portant sur les structures de la langue et nous allons essayer de présenter les travaux les plus importants par ordre chronologique.

4.1 Brill : acquisition de structures de phrases

(Brill & Marcus, 1992a) ont montré que l'acquisition automatique de structures de phrases simples d'une langue naturelle est possible sans supervision, en utilisant seulement une très petite grammaire initiale. La méthode d'apprentissage utilisée est basée sur l'extraction d'informations distributionnelles à partir d'un corpus étiqueté en parties du discours. Le système d'apprentissage est capable d'utiliser les informations extraites avec précision pour analyser syntaxiquement des phrases courtes de l'anglais. L'analyse utilise des procédures distributionnelles similaires à celles proposées par Harris ; elle est fondée sur l'hypothèse que si deux tags (étiquettes) de parties de discours adjacents ont des distributions similaires à celles d'un quelconque tag unique, alors il est probable que ces deux tags forment un constituant de phrase.

Le système apprend une grammaire de règles hors- contexte pondérées. Deux sources d'information distributionnelle sont utilisées. La pondération d'une règle : $tag_x \rightarrow tag_y tag_z$ est fonction de :

- La similarité distributionnelle des parties de discours tag_x ainsi que de la paire $tag_y tag_z$
- Une comparaison de l'entropie de l'environnement tag_x et $-tag_y tag_{z-}$; l'entropie d'un environnement $-tag_y-$ est une mesure de la distribution aléatoire des tags apparaissant immédiatement après tag_y dans le corpus.

A la suite de ce travail qui utilise un corpus déjà annoté en parties de discours pour l'anglais, dans (Brill & Marcus, 1992b), les auteurs présentent une approche pour l'étiquetage automatique en parties du discours (i.e assigner les catégories aux mots : nom, verbe, …) de textes bruts dans une langue inconnue avec un minimum d'assistance de la

part d'un informateur (utilisateur connaissant la langue). Le système d'étiquetage utilise une analyse distributionnelle et utilise les techniques de clustering pour classer les mots.

4.2 Lager

(Lager, 1995) propose un environnement de développement pour les théories de corpus autrement dit qui permet de mettre en œuvre une théorie de corpus : c'est un ensemble d'outils destinés aux linguistes (informaticiens) de corpus. Le système est fondé sur un formalisme logique du premier ordre ; des formules logiques sont utilisées pour exprimer des assertions sur les textes et l'inférence logique est utilisée pour manipuler ces formules afin d'analyser les textes. Le système implémente un ensemble de fonctionnalités nécessaires pour la manipulation et l'utilisation de corpus : recherche, concordances, analyse, étiquetage en parties du discours, lemmatisation, génération d'arbres, apprentissage, … . Le formalisme logique du premier ordre permet par ailleurs les généralisations nécessaires à la production des mêmes descriptions obtenues sur d'autres données textuelles similaires.

Outre son intérêt pratique, ce travail soulève un ensemble de questions intéressantes qui sont d'ordre philosophique et méthodologique : qu'est ce qu'un corpus de textes ? Qu'est ce qu'une théorie de corpus ? Que signifie développer une théorie de corpus ? Que signifie pour une théorie de corpus d'être vraie pour un corpus donné de textes ? Quelle est la relation entre la vérité d'une telle théorie et son utilité pour les objectifs du TAL ?

4.3. L'approche de Déjean : découverte des structures de langues inconnues

Déjean dans sa thèse (Déjean, 1998) présente une approche qui s'inspire de la méthode distributionnelle de Harris pour la découverte automatique des régularités les plus générales contenues dans un corpus, afin d'être applicable à un très grand nombre de langues, et les plus informatives possibles afin qu'elles fournissent des renseignements utiles et pertinents permettant la découverte de la structure syntaxique de la langue. Le travail de Déjan s'articule autour de trois points : la découverte des morphèmes de la langue, puis la construction d'unités supérieures : les chunks (suites de mots supposées correspondre à une structure syntagmatique) et enfin la découverte des relations entre chunks.

La découverte des morphèmes est obtenue par segmentation des mots du corpus en s'inspirant de l'algorithme proposé par Harris et qui se base sur le comptage du nombre d'éléments afin de repérer les frontières de morphèmes dans un mot. Une liste des mots du corpus est d'abord construite, ensuite l'algorithme recherche les débuts et fins de mots les plus fréquents obtenus par découpage selon l'algorithme de Harris. Une fois la dizaine de débuts et fins les plus fréquents de la langue trouvés, ces affixes sont utilisés pour segmenter les mots et découvrir les autres affixes de la langue : pour une séquence de lettres donnée, on regarde la séquence de lettres suivantes ; si la moitié de ces séquences correspond à des morphèmes trouvés, les autres séquences sont aussi considérées comme étant des morphèmes. La segmentation en morphèmes est très importante car les affixes qui jouent un rôle fonctionnel dans la langue jouent un rôle essentiel dans la mise en relation des éléments de la langue. Le principe de l'algorithme est de segmenter le mot avec le morphème le plus long possible.

Ensuite les énoncés sont considérés comme une suite d'éléments : les 'chunks' dont le début et/ou la fin peuvent être caractérisés par une marque morphologique. les signes de ponctuation servent à segmenter le corpus en énoncés. Les chunks correspondent aux syntagmes simples.

4.4 Clark, 2001

Dans sa thèse (Clark, 2001) présente une variété d'algorithmes pour l'apprentissage non supervisé de différents aspects d'une langue naturelle en utilisant des modèles statistiques. L'objectif scientifique du travail décrit est d'apporter une contribution à la réfutation de l'argument de Chomsky pour le nativisme linguistique. L'auteur a mis au point :

- des algorithmes pour l'induction des catégories syntaxiques à partir de textes non annotés en utilisant des informations distributionnelles.

- des algorithmes pour l'apprentissage des processus morphologiques dans plusieurs langues y compris une langue comme l'arabe à morphologie non concaténative.

- des algorithmes pour l'induction d'une grammaire simple, à contexte- libre à partir de textes non annotés.

La réalisation de ces algorithmes permet à l'auteur de conclure que la réalité contredit l'argument de Chomsky.

4.5Creutz, 2006

(Creutz, 2006) décrit un système nommé Morfessor pour l'induction de la morphologie d'une langue naturelle de manière non supervisée et sans utiliser de connaissances préalables sur la langue, ce qui en fait un système indépendant d'une langue particulière. L'induction de la morphologie à partir d'un corpus de textes passe par la segmentation des mots du corpus. Morfessor utilise comme modèle formel les statistiques bayesiennes dans le cadre du MDL (Minimum Description Lenght) (Rissanen, 1978). Morfessor a été utilisé aussi pour différentes langues

4.6 Goldsmith

(Goldsmith, 2006) décrit un algorithme non supervisé pour l'apprentissage de la morphologie d'une langue naturelle. La tâche principale de toute analyse morphologique est la segmentation des mots en ses composants qui interviennent dans sa formation par concaténation. L'algorithme proposé prend en entrée des textes (entre 5000 et 1000 000 de mots) et produit une analyse morphologique partielle de tous les mots de ce corpus. Le but étant de se rapprocher le plus d'une analyse qui serait produite par un humain. L'algorithme n'utilise ni dictionnaire ni règles morphologiques particulières ; il est basé sur les principes du MDL (Minimum Description Lenght) (Rissanen, 1978) et utilise des signatures sur les structures morphologiques analysées. Il permet de construire une grammaire morphologique à partir de n'importe quel texte écrit dans une langue indo-européenne.

4.7Biemann

(Biemann, 2007) définit formellement le paradigme de découverte de structures en arguant pour une approche indépendante d'une langue donnée et libre de connaissance, autrement dit n'utilisant aucune connaissance spécifique sur les langues. Ce paradigme constitue un cadre pour apprendre les régularités structurelles d'une langue à partir de corpus textuels bruts et rendre ensuite ces régularités explicites en les annotant et en les réintroduisant dans les données. Dans ce paradigme, nous n'avons pas besoin de connaissances (règles) sur la langue, ni de supervision ; c'est une approche indépendante d'une langue particulière.

Il s'agit de mettre au point des procédures de découverte qui opèrent sur des données langagières brutes et enrichissent ces données de manière itérative en utilisant les

annotations fournies par les procédures de découverte appliquées précédemment. La figure 1 suivante décrit le fonctionnement de l'approche qui est donc une approche itérative.

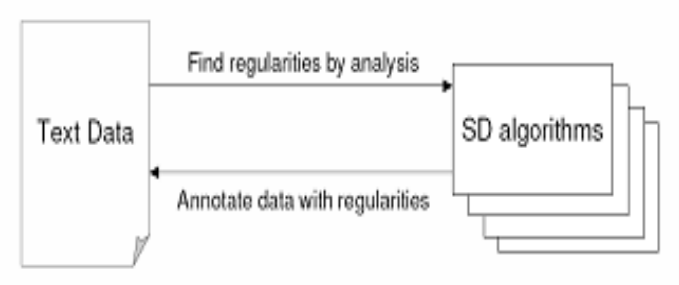


Figure 1. *Processus itératif de découverte de structures*

Pour découvrir les structures de la langue en classant les différentes unités de façon non supervisée, les items de la langue doivent être relatés par des mesures de similarité. Les méthodes de clustering servent à grouper les items (similaires) dégagés en clusters, ce qui permet d'effectuer des abstractions et des généralisations.

Biemann utilise le modèle des graphes pour représenter l'information langagière. Les graphes constituent un moyen d'encoder l'information en nœuds et arcs. Dans un contexte de TAL, les nœuds dénoteraient les unités langagières alors que les arcs représenteraient les relations entre ces dernières. De cette façon, les unités ainsi que leurs similarités sont naturellement et intuitivement traduits en une représentation graphique. Examiner et contraster les effets des processus générant les structures de graphes similaires à celles observées sur les données langagières jette la lumière sur les structures de la langue et leur évolution. Après avoir représenté l'information par des graphes, Biemann utilise un algorithme de clustering de graphes pour classer les unités. Les méthodes de clustering sont le modèle le plus utilisé actuellement pour permettre les abstractions et généralisations nécessaires pour assigner une structure à des textes qui n'ont pas déjà été analysés.

4.8 Bordag

Dans sa thèse (Bordag, 2007) développe un modèle pour l'acquisition lexicale basé sur un ensemble minimal de principes structuralistes, en particulier, la détermination des relations syntagmatiques et paradigmatiques entre les unités linguistiques. Deux hypothèses

essentielles fondent le travail de Bordag : la première est que ces relations opèrent de manière équivalente à tous les niveaux de la langue. La seconde stipule que toute connaissance sur les unités de la langue peut être dérivée de l'observation de l'*usage* de la langue. Tout algorithme basé sur ces deux hypothèses peut être conçu indépendamment d'une langue particulière. Bordag utilise un algorithme « letter successor variety » pour la détection des frontières de morphèmes. Il s'inspire de (Harris, 1954) et de (Déjean, 1998).

4.9 Zeman

(Zeman, 2008) propose un algorithme non supervisé pour l'acquisition des paradigmes morphologiques à partir d'un texte tokenisé d'une langue inconnue. Il utilise aussi une méthode statistique basée sur MDL.

4.10 Chan

(Chan, 2008) décrit de nouveaux algorithmes fondés aussi sur l'analyse distributionnelle pour implémenter des algorithmes non supervisés pour l'induction de la morphologie et des catégories lexicales à partir de corpus textuels.

5. Conclusion

Nous avons vu à travers cette revue des travaux en TAL basé sur corpus que depuis les années 90, la tendance va de plus en plus vers les approches ne nécessitant pas de connaissances à priori sur la langue, d'une part ainsi que celles ne nécessitant pas de ressources d'autre part. Lorsque ces deux critères sont réunis, l'approche est généralement utilisée pour différentes langues et on parle de *découverte* au lieu d'*apprentissage*. Sur un autre plan, en utilisant la méthode distributionnelle, ces travaux sont une réfutation à l'hypothèse d'innéité de par l'existence d'algorithmes permettant l'acquisition de structures simples de la langue, ils sont aussi en faveur de *l'universalité de certains mécanismes distributionnels à l'œuvre dans l'acquisition de structures linguistiques*. Il existe peu de travaux utilisant cette approche pour la langue arabe, nous y reviendrons au chapitre 5.

Chapitre 4. La Tradition Grammaticale Arabe

« Loin d'être l'expression de principes logiques, le langage est pour Sibawayh exactement son contraire : une forme de comportement humain » (Carter, 1973)

1. Introduction

Les études modernes sur la Tradition Grammaticale Arabe partagent l'objectif de reconstruire les fondements méthodologiques, théoriques et épistémologiques de l'approche des anciens grammairiens arabes (Abu-Almakarim, 1975 ; Hassan, 1979 ; Baalabki, 1979 et 1983 ; Bohas & al, 1990 ; Carter, 1973 ; Owen, 1988 ; Suleiman, 1999). Dans sa conception, la TGA est une approche guidée par les données dont l'intérêt principal est une description exhaustive des données plutôt que la construction d'une théorie bien qu'elle aboutit aux « ḥudūd » ou règles de la langue. La TGA ne s'est jamais soucié de rendre explicite ses prémisses théoriques sous-jacentes, ce qui néanmoins ne permet pas de mettre en doute l'existence de ces prémisses (Suleiman, 1999).

2. TGA et la linguistique moderne

Dans l'entreprise d'expliciter ses fondements, la tendance des chercheurs est d'appliquer les modèles des théories linguistiques modernes à la TGA et ce deux manières :
La première et qui semble être l'approche favorite cherche à faire concorder le composant théorique de la TGA ou du moins quelques aspects de ce composant, avec telle ou telle théorie linguistique occidentale. Comme ces modèles théoriques modernes ont été aussi utilisés par les informaticiens en particulier dans l'approche CL (voir chapitre 2), il n'est pas étonnant que cette même approche ait été appliquée du côté du TAL arabe.
Du côté des linguistes, des connexions intéressantes ont été établies entre la TGA et la tradition structuraliste américaine Boomfieldienne (Hassan, 1979), l'analyse en constituants immédiats, version bloomfieldienne (Carter, 1973), les grammaires de dépendance (Owen, 1988), les grammaires transformationnelles (Ayoub & Bohas, 1981 ; Itkonen, 1991). D'autres linguistes ont proposé des connexions entre la TGA et l'hypothèse de la 'Split Morphology' qui stipule la division entre la dérivation et la flexion

comme composants séparés de la grammaire. La dérivation est gouvernée par des règles lexicales alors que dans la flexion des règles syntaxiques seraient mises en œuvre (Ryding, 1993).

La deuxième approche utilise les théories linguistiques modernes de manière plus éclectique comme un instrument d'interprétation au moyen duquel on peut interroger la TGA (Bohas & al, 1990).

Du côté du TAL version CL, les formalismes des théories linguistiques modernes ont aussi fait l'objet d'un certain nombre de travaux. Notre objectif n'étant pas d'utiliser l'un de ces frameworks ni d'entrer dans le débat pour comparer lequel est plus adapté à la langue arabe, nous citerons juste quelques travaux à titre d'exemple. Dans le framework du modèle de la Grammaire Générative (PSG), nous citerons en particulier (Ditters, 2001) qui utilise le formalisme AGFL pour la description de la phrase arabe. Dans le framework des grammaires de dépendance que beaucoup considèrent plus adaptées à la grammaire arabe, (Kassas, 2005) présente un système de génération automatique bilingue arabe-français. Enfin dans le cadre du formalisme des grammaires d'unification, nous citons les travaux de l'équipe de Benhammadou (Bahou & al, 2006) qui adaptent le formalisme des grammaires HPSG pour l'analyse de textes arabes non voyellés et (Aliane & Nehal, 2009) ainsi que les travaux du CRSDLA.

Pour clore cette revue certes non exhaustive, nous dirons avec Yasir Suleiman, que la somme de travail théorique effectué autour de la TGA, témoigne non seulement de l'ingéniosité des chercheurs modernes mais aussi de la richesse et peut-être de manière plus significative de l'élasticité de la TGA (Suleiman, 1999).

En ce qui nous concerne, partant de la nature empirique de l'approche de la TGA et du fait qu'elle semble jusqu'à ce jour résister à rentrer complètement dans un quelconque moule théorique de la linguistique moderne (Suleiman, 1999) (Carter,1973)(Hadjsalah, 1979) (Vesteegh, 1997), nous avons choisi d'adopter aussi une approche empirique qui ne s'oblige pas nécessairement et à priori à retrouver la TGA dans une quelconque théorie linguistique moderne. Comme dans l'approche de (Boas & al, 1990), nous voulons plutôt interroger la TGA mais en restant encore plus ouvert c'est-à-dire interroger pas seulement au moyen des théories linguistiques mais de la méthode scientifique en général. De plus, partant de nos objectifs de TAL, notre approche empirique va consister à essayer de « simuler » la TGA. Tâche ambitieuse, certes, présomptueuse, peut-être mais qui vaut certainement la peine d'essayer. Mais avant d'exposer notre travail, nous allons d'abord

présenter l'approche empirique de la TGA et les éléments nécessaires à la compréhension de ce travail et de ses motivations.

3. La Tradition Grammaticale Arabe comme linguistique de Corpus

« **Le** Livre » الكتاب de Sibawayh est le texte fondateur de la TGA. Il constitue la première analyse complète des structures de la langue arabe et qui continue à être à travers les siècles aussi bien un modèle qu'une source pour les études sur la langue arabe, pour les principes généraux de la grammaire et aussi pour les aspects méthodologiques de son approche du langage.

Le terme arabe qui désigne la grammaire est le « naḥw ». Ce terme est, étymologiquement, un nom verbal associé au verbe naḥāqui signifie « suivre un chemin, cheminer, aller dans une certaine direction ». Le naḥw est donc proprement un « cheminement » ou, dans une acception plus nominalisée, un « chemin », une « voie ». La tradition nous apprend que les premiers grammairiens concevaient leur discipline comme consistant à « suivre la voie » des mystiques anciens, les Arabes bédouins, détenteurs de « l'arabe pur » {al- ʻrabiyya l-fuṣḥa) et que leur science est ainsi devenue « science de la voie » (ʻilm al-naḥw) (Kouloughli, 1999).

À l'intérieur même de la grammaire, le terme « naḥw » a une seconde acception, plus spécifique : il s'y oppose à « ṣarf » (ou taṣrif) qui désigne la composante morpho-phonologique du modèle, et renvoie alors à la partie de la grammaire qui étudie les relations des mots entre eux dans le discours. La préoccupation fondamentale qui domine le naḥw ainsi entendu est le iʻrāb, c'est-à-dire l'assignation des flexions casuelles. Le terme iʻrāb est lui aussi, étymologiquement, un nom verbal. Il est associé au verbe aʻraba qui signifie proprement « rendre à la manière des Arabes », exprimer en « bon arabe »[1], et en est venu à désigner, dans la terminologie technique des grammairiens, « la variation de la finale des mots déterminée par la variation de leurs régissants » (Kouloughli, 1999).

L'élaboration de la TGA s'est donc fondée sur des enquêtes orales (pour bien déterminer les structures, les flexions, et par là permettre une bonne transcription du texte coranique ainsi que des hadiths) suivies de :

[1] Ceci dénote déjà une conception dynamique du rôle du locuteur dans la construction de son discours

- description/inventaire des réalisations
- classification grâce aux outils d'analyse distributionnelle du qiyās que nous décrivons plus loin.
- explication des formes déviantes par rapport au qiyās

3.1 Le corpus de la TGA

Il est de notoriété que la Tradition Grammaticale Arabe est née d'un besoin théologique, en l'occurrence, le besoin impératif de sauvegarder le Coran. Il n'est donc pas étonnant que le corpus ayant servi à son élaboration ait utilisé en grande partie le texte coranique et les hadiths du prophète desquels Sibawayh cite beaucoup d'exemples dans son Kitāb. Ces deux exemples constituent en réalité des exemples de la langue arabe en ***usage*** destinés à être compris par ceux à qui ils s'adressaient c'est-à-dire les utilisateurs ou les locuteurs natifs de la langue arabe qui pour Sibawayh est **la** référence. C'est pourquoi Sibawayh dans son établissement des règles (ḥudūd) en revient toujours au locuteur faṣiḥ bédouin de préférence[2] (corpus de la langue parlée). En termes de textes écrits, les poèmes ont fait aussi parti du corpus considéré. Les anciens grammairiens arabes ne se sont pas assis dans leurs fauteuils comme Chomsky et les tenants de la linguistique théorique à essayer d'imaginer un modèle théorique et l'adapter aux structures de la langue par introspection.

Formellement, l'approche de Sibawayh est appelée « qiyās ». Comme processus de classification et d'explication de réalisations linguistiques basé sur les résultats des descriptions/inventaires, le qiyās est un processus d'analyse exploratoire de données (linguistiques) dont l'objectif est de découvrir les règles et les schémas qui gouvernent le système de la langue.
Le qiyās est fondée sur la théorie du "mawḍi‘ " et sur les concepts de « ´aṣl, far‘, bāb » que nous définissons dans ce qui suit.

3.1.1 Les concepts de ´aṣl, far‘ , bāb

- ***bāb*** (باب)***:*** Le sens lexical du mot "bāb" est entrée. Les grammairiens arabes appellent "bāb" un ensemble d'entités linguistiques qui ont non seulement une propriété commune mais aussi une structure commune" (Sibawayh, 1980) (Hadjsalah, 1979).

[2] N'ayant pas été en contacts avec les nouveaux arrivants à l'islam et dont l'arabe n'est pas fasih.

Le comportement des éléments appartenant à un "bāb" (est décrit par les "mithāl-s" qui sont des modèles de comportement des éléments linguistiques.

En première approximation, le "bāb" peut être vu comme une classe ou plus précisément comme un outil classificatoire. Une caractéristique importante du "bāb" est sa généralité fondée sur le caractère indéterminé et quelconque de ses éléments qui sont uniquement reliés les uns aux autres par des relations spécifiques modélisées par les "mithāl-s"/modèles.

Le "mithāl" est donc *une description dynamique du comportement (et non) des éléments* linguistiques considérés sous le "bāb" dont ils relèvent.

Les éléments d'un "bāb" sont appelés "naẓā´ir" (c'est à dire homologues et constituent des classes d'équivalence sur ce "bāb" représentées par les "mithāl-s".

- ***´aṣl/far'***(الأصل و الفرع) ***:*** Les grammairiens arabes appellent ´aṣl (pluriel ´uṣūl) un élément constant que l'on rencontre dans d'autres éléments (objets ou processus) qui sont considérés des expansions ou des dérivés de ce ´aṣl et qui sont appelés furū' (pluriel de far') (Ibngini, 80)(Sibawayh, 1980)(Hajsalah, 1979)(Gacem, 1997). ´aṣl et furū' sont reliés via le bāb auquel ils sont rattachés et qui est une abstraction des deux.

La transition du ´aṣl au far' est réglée par les mithāl-s ou modèles des comportements (au sens dynamique) possibles des éléments considérés. Ce qui caractérise le ´aṣl est son caractère invariant puisqu'il se retrouve dans tous ses furū'.

Du point de vue logique, les ´uṣūl sont des *prémisses*, des principes admis, des lois qui sont déjà intégrées au système alors que les furū' sont des *observations* problématiques que nous voulons expliquer en les confrontant aux ´uṣūl autrement dit des éléments dont il s'agit de trouver les ´uṣūl.

3.1.2 La théorie du "mawḍi'" "(الموضع)

Un mawḍi' est une position dans un ensemble structuré**,** non pas un système d'oppositions mais un ensemble structuré où les éléments sont mis en correspondance (Hajsalah, 1979). Un mawḍi' est une position abstraite dans un mithāl, c'est le lieu où occure un élément dans le discours (Carter, 1973)(Mosel,1995)(Versteegh, 1997). Ce n'est pas toujours une occurrence matérielle, l'antéposition du complément ne change pas son mawḍi', par exemple.

La position n'est pas définie sur la base de la seule distribution ou des fonctions (grammaticales) des éléments: elle correspond à une position contenue dans un schème

dynamique (le mithāl) qui est abstrait à partir des deux axes syntagmatique et paradigmatique en même temps.

La définition du mawḍi‘ nous permet de définir formellement le mithāl" comme un ensemble de positions (mawḍi‘-s) abstraites, dans chaque position peuvent occurer les éléments d'une classe linguistique spécifique et seulement ces éléments".

Il existe des mithāl-s spécifiques pour chaque niveau de la langue. *C'est le rôle du qiyās que de déterminer pour chaque entité à analyser son mawḍi‘.*

3.1.3 Le Qiyās (القياس)

Le qiyās est une approche de description, analyse et explication des *formes* linguistiques, fondée sur la théorie du mawḍi‘ et les concepts de ´aṣl, far‘ et bāb. Le sens lexical de qiyās est mesure ou unité de mesure. Le qiyās utilise aussi d’autres critères **méthodologiques** comme le critère d’*économie* et le critère de *récurrence (iṭirad).*

Le processus de qiyās commence par une opération classificatoire qui consiste essentiellement à trouver pour chaque entité à analyser le bāb qui représente son ´aṣl et essayer alors de la *reconstruire* en utilisant les schèmes/mithāl-s disponibles pour ce bāb. Nous allons décrire dans le paragraphe suivant la nature opératoire de cette démarche que les orientalistes eux mêmes ont déclaré ne correspondre à aucun processus de logique ou de raisonnement classique.

Afin de bien cerner le contenu du mot qiyās, nous allons examiner tout d’abord un exemple que donne IbnGini pour illustrer l’idée du qiyās (IbnGini, 1980)(Hadjsalah, 1979) et qui consiste en l’observation des constructions suivantes :

- "qāma Zayd**u**n" (Zayd s’est levé)
- "ḍarabtu ‘Amr**a**n " (j’ai frappé ‘Amr)
- "marartu bi-said**i**n " (je suis passé près de Said)

Nous retrouvons là les trois grandes fonctions syntaxiques de la langue et les flexions qui leur correspondent : ce sont là des formes qui sont ***effectivement réalisées*** dans la réalité du langage.

D’autre part, chacune de ces réalisations qui appartiennent aux trois ensembles : nom en fonction de sujet, de complément d’objet direct et indirect est caractérisé par une flexion qui lui est propre.

IbnGinni considère cette uniformité flexionnelle comme un qiyās : la nécessité pour un tel

nom d'être conforme aux autres noms qui ont la même fonction et qui constituent donc son bāb" relève du qiyās.

Notre objectif ici n'étant pas de rentrer dans des considérations purement linguistiques, nous noterons, néanmoins que la combinatoire du qiyās accepte certaines formes qui ne sont pas réalisées ou qui n'existent pas dans l'usage. On parle alors de l'opposition qiyās/usage (isti'mal استعمال). Les bābs qui correspondent à un qiyās non réalisé sont des bāb-s vides.

La mise en qiyās consiste d'abord, en un constat de systèmes de correspondances et dans une seconde étape, par une implication logique, en une extension et une systématisation des rapports existants (Hajsalah, 1979). En quoi consistent ces rapports et ces correspondances ? (Hadjsalah, 1979) nous montre les différents niveaux où interveiennent les mises en correspondances impliquées par le qiyās d'après le kitāb de Sibawayh :

Si nous nous intéressons au niveau lexical, les variations de substance sur un même schème et les variations de forme sur une même substance suggèrent qu'il doit exister sur leur produit E=X x Y une double relation d'équivalence :

R= "avoir le même schème" et
R' = "avoir la même substance".

Ces deux relations établissent sur ce produit une double partition dans E : P et P' de puissance égales au nombre de racines concaves et des schèmes verbaux du trilitère disponibles dans la langue. Les éléments appartenant à un sous ensemble A quelconque obtenu par P ou P' sont équivalents entre eux :

Exemple : A={qāma, , ḍaraba, dhahaba, ..., fa'ala}

Ce sont les naẓā´ir que nous avons vu ci dessus. Fa'ala est ici le symbole du schème/mithāl. Les éléments de l'ensemble précédent sont les naẓā´ir du mithāl fa'ala". L'ensemble des naẓā´ir peut donc être vu comme une classe d'équivalence.

De là, nous pouvons dire le qiyās désigne ici :

- l'équivalence qui existe nécessairement entre les éléments d'un même "bāb" en vertu de leur appartenance à ce même bāb , et par extension,
- le modèle (mithāl) qui symbolise cette équivalence,
- la classe d'équivalence elle-même symbolisée par ce modèle.

Lorsque l'équivalence qu'implique un qiyās a comme référentiel le domaine morpho-lexical, l'équivalence se ramène à une *correspondance de structures*.

En effet, la relation avoir le même schème désigne une correspondance d'arrangement qui s'appliquent à des ensembles finis et ordonnés d'éléments et qui résulte d'une correspondance biunivoque propre à la structure lexicale et compatible avec la composition interne des séquences considérées.

Cependant, le qiyās ne se résume pas au qiyās entre objets que nous venons de voir pour les structures morpho- lexicales.

Une fois en possession d'un certain nombre de constatations et après avoir établi un certain nombre de systématisations, ces dernières constituent un système de ´uṣūl établi qui vont servir pour l'analyse de nouvelles observations.

Il s'agit donc, du passage de certains éléments simples et fondamentaux (´uṣūl) ou prémisses dont le caractère invariant a été constaté par le linguiste à d'autres éléments plus complexes qui les contiennent selon certains processus. Nous nous intéressons alors aux équivalences entre transformations "´uṣūl ↔ furū‘ ".

En fait, l'équivalence entre deux ensembles E et E' de transformations est établie si et seulement si :

- les opérations de E et E' se correspondent terme à terme, en d'autres mots ont un même schème transformationnel,
- leur ordre de succession est le même dans E et E'
- les domaines ou contextes des séquences où elles se produisent sont équivalents.

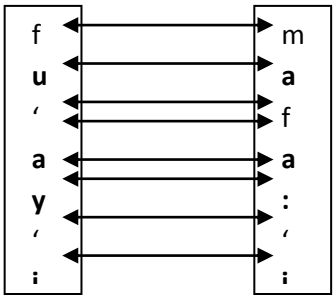
Exemple :

Nous citerons comme exemple, l'assimilation du "taṣghīr" (schème transformationnel du passage à la forme diminutive) au "taksīr" (schème transformationnel du passage à la forme du pluriel interne) (Hajsalah, 1979) car d'après Sibawayh, les opérations qu'ils impliquent sont les mêmes : il y a changement du premier segment dans les deux cas, adjonction d'un "harf mad" en troisième position ainsi que l'ajout de la haraka " i " .

Schème de la forme diminutive : fu‘ayil

Schème du pluriel interne : mafā‘il

Sur la base de la correspondance (muwafaqa, munasaba) (Hajsalah, 1979) de ces opérations de transformations, les linguistes arabes ont admis qu'ils appartiennent à une même classe.



En représentant par C_1 ,C_2 , C_3 , C_4 les variables consonnantiques de ces schèmes et par v_1, v_2, les variables correspondant aux harakat, on a d'après les auteurs arabes:

C_1**u** C_2 **ay** C_3 **i** C_4 = passage à la forme diminutive ex: zuwayrik

$C_1\, v_1\, C_2\, C_3\, v_2\, C_4$

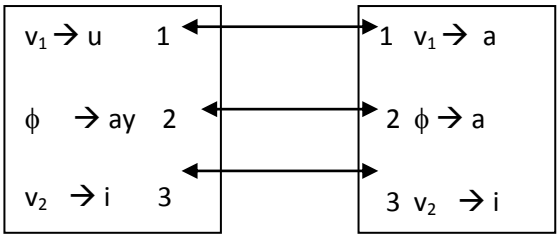
C_1 **a** C_2 **a:** C_3 **i** C_4 = passage à la forme pluriel interne: zawarik

u~a=X, **ay~a**=Y, **i~i** =Z

le *(super) schème* qui peut être abstrait à partir des deux schémas de transformations (diminutif, pluriel interne) précédents est :

$$C_1\, X\, C_2\, Y\, C_3\, Z\, C_4$$

On peut alors établir les correspondances suivantes entre les opérations ayant abouti à ces résultats :



Si nous appelons t_1 , t_2 , t_3 les trois transformations communes dans les deux ensembles précédents. La composition de t_1 et t_2 donne :

$C_1\, X\, C_2\, Y\, C_3\, v_2\, C_4$

Soit $T_1 = t_1 \circ t_2$,

La composition de T_1 et t_3 donne le schème final : $C_1\, X\, C_2\, Y\, C_3\, Z\, C_4$.

(Hadjsalah, 1979) met l'accent qu'à tout les niveaux de mise en correspondance les transformations linguistiques sont réversibles et que ces transformations peuvent être donc représentées par une structure de groupe mathématique.

Pour revenir à l'exemple précédent, il est possible de faire correspondre à tout couple ordonné dans les deux ensembles de transformations un élément unique qui est leur composé autrement dit, il existe dans les deux ensembles de transformations une loi de composition interne.

D'autre part les relations d'ordre qui sont définies dans les deux ensembles (ordre de succession des transformations) se correspondent également.

Il résulte de toute l'analyse précédente que les groupes de transformations qui aboutissent à la forme du diminutif et à celle du pluriel interne sont ***isomorphes :*** cet isomorphisme est basé sur la correspondance biunivoque qui existe entre ces opérations ainsi que sur la correspondance des lois de composition et des relations d'ordre définies sur ces ensembles.

D'autre part, on retrouve dans l'analyse des linguistes arabes, la notion d'opération vide ou élément neutre par rapport aux transformations : il s'agit simplement de l'absence de transformations.

Il existe deux types de transformations : les transformations positives et les transformations négatives. Une opération est positive si c'est une construction qui va d'un élément donné qui n'a subi aucune transformation (´aṣl) à l'élément construit qui en dérive selon un schème/mithāl (far'). La transformation négative est la transformation inverse de la précédente qui consiste à faire reprendre à l'élément construit la forme de son ´aṣl.

3.1.4 Discussion

Le type d'abstraction effectué par le qiyās et que nous avons vu à l'œuvre dans le paragraphe précédent, est appelé par les grammairiens arabes ikhtibar ce qui veut dire tester comment relier un élément à un autre via un bāb.

De là, l'objectif final et fondamental du qiyās est d'*assigner à chaque élément une position dans un mithāl*. Chaque entité sous un bāb est soit un ´aṣl soit un far', un far' pouvant être ´aṣl pour d'autres furū'.

Une caractéristique importante de l'ikhtibar" est son caractère extensif. Par ailleurs, en conférant à chaque élément une position dans une matrice générative et dynamique (le mithāl), le "qiyās" réalise la *synthèse de l'ordre et de la structure* et ceci constitue une des

plus grandes acquisitions de cette approche.

D'autre part, l'inférence du qiyās est fondée sur la relation d'équivalence opératoire définie entre le ´aṣl et le far' et toute mise en qiyās implique une recherche approfondie sur la validité de l'équivalence (entre ´aṣl et far') posée comme hypothèse.

De plus, la mise en équivalence des schémas de réalisation aboutit toujours à un schéma plus large (que les arabes appellent jāmi' ou élément intégrant) qui ne les contient pas mais les intègre à la manière d'un opérateur mathématique qui appliqué à une certaine entité, permet de transformer celle-ci en une autre et qui peut être par là même, commun à plusieurs entités tout à fait hétérogènes. Baalabki dans (Suleiman, 1999) parle de cette intégration ou fusion en ces termes :

> "Coalescence or fusion in the sense of merging two linguistic elements, particularly two morphemes, features in several disparate parts of Sibawayh's Kitāb. Although Sibawayh does not devote a special heading to the description of the rules governing coalescence or the nature of the resulting blend , he does not fail to see how the very concept of coalescence can be used as an extremely helpful tool of grammatical analysis"

L'inférence du qiyās aussi bien dans son agencement que dans la nature du rapport qui relie l'ensemble {´aṣl, far'} avec le jāmi' et celui qui relie ´aṣl et far' entre eux ne correspond exactement à aucun processus logique classique.

En effet, dans la littérature occidentale souvent, le terme qiyās est traduit par syllogisme ou par analogie. Néanmoins certains linguistes ont attiré l'attention sur la différence entre les deux systèmes de logique occidentale et arabe et sur l'inanité d'essayer de comprendre ce dernier à la lumière du premier.

> « …the usual rendering of the term qiyās as 'deduction' by 'analogy' is rather infelicitous, if not altogether misleading for a number of reasons" (Bohas & al, 1990)

Le qiyās procède du spécial au général {´aṣl, far'}↔(jāmi') par composition et c'est en cela qu'il est créatif. Sur le plan de la rigueur du raisonnement, dans le syllogisme, c'est le *rapport d'inclusion* qui assure cette rigueur, alors que dans le processus de qiyās, c'est le *rapport d'équivalence* impliquée par la notion de jāmi' (élément ou schème intégrant).

D'autre part, en assignant à chaque élément une position (mawḍi') dans une structure, le qiyās arrive à établir les règles (ḥudūd) qui règlent le comportement langagier.

Après avoir discuté des concepts qui fondent la TGA, nous allons donner un aperçu du modèle syntaxique dans celle-ci.

4. Le modèle syntaxique

4.1 Le point de départ de l'analyse

En linguistique, comme en logique, la proposition constitue généralement le point de départ de l'analyse. Dans la TGA, la notion de phrase n'existe pas, d'ailleurs le mot « jumla » n'apparait jamais dans le kitāb. Le point de départ de l'analyse est défini formellement sur la base de l'autonomie des séquences dans le discours. Toute séquence verbale autonome peut être considérée comme un message minimal du point de vue de cette autonomie (Sibawayh, 1980)(Hajsalah,79).

Une séquence verbale est autonome lorsqu'elle peut être actualisée dans le discours sans avoir besoin d'autres séquences. Exemple « kitāb » , « muhammad », « ḍaraba » sont des séquences actualisables minimales ; « man », « li » ne sont pas des séquences minimales car elles ont besoin d'autres éléments pour être actualisées. Une unité autonome ainsi définie sur la base de son autonomie ou « isolabilité » dans le discours est une unité formelle puisque définie sur la base du seul signifiant (la forme objective) du discours sans aucun recours au sens (donc la subjectivité). Les unités minimales ainsi définies, définissent un niveau linguistique appelé niveau intra- lexical ou niveau central, ces unités sont appelées lexies ou " lafẓa-s".

Les transformations sur la lexie consistent à générer des séquences dérivées à partir de séquences primitives par ajout ou suppression d'éléments à droite ou à gauche de la séquence de départ.

Ceci nous mène à la définition formelle de la lexie ou "lafẓa" : " toute séquence isolable et indivisible qui peut accepter des ajouts à droite ou à gauche sans perdre la propriété d'être isolable et indivisible est appelée lafẓa" (Hajsalah, 1979)

La lexie constitue le point de départ de l'analyse grammaticale. A partir de ce niveau central, l'analyse du linguiste peut être orientée :

- vers le bas : pour déterminer les segments signifiants qui sont contenus dans les lexies, déterminer les "kalima-s" et les analyser en racine et schème ;
- vers le haut, pour déterminer comment les lexies sont intégrées dans les structures du niveau supérieur c'est à dire les structures syntaxiques.

La possibilité de génération (dérivation) de séquences dérivées appelées furū‘ à partir d’une séquence primitive appelée ´aṣl est appelée par les grammairiens arabes " tamakkun " (capacité) et " tassaruf " (variabilité).

Ces notions permettent notamment d’établir des distinctions formelles entre séquences isolables. En effet certaines séquences possèdent un "tamakkun " parfait car elles peuvent recevoir n’importe quel type d’ajouts. C’est le cas notamment du nom commun (voir figure2). D’autres séquences possèdent une moindre capacité : le nom propre ne peut recevoir l’article ni le complément adnominal. La variabilité de la séquence primitive est soumise à des règles qui sont représentés dans le schème générateur de la lexie (figure 2).

L’isolabilité et la capacité permettent de délimiter sur la base *de la seule forme* de la langue (*du lafẓ*) une première unité dans le discours qui est le nom commun.

A partir de là, on peut définir formellement (à partir de la seule forme) les autres unités contenues dans les ajouts. Cela est possible car elles occurrent en des positions (mawḍi‘) spécifiques qui sont inférées à partir de l’étude de *toutes les occurrences* possibles d’un élément dans le discours. Ces positions déterminent alors les fonctions grammaticales des éléments qui y sont contenus.

L’axe paradigmatique est structuré car il est le siège des transformations et ces transformations sont ordonnées par la relation ´aṣl ↔ far‘ (Hadjsalah, 1979).

Les transformations déterminent les distinctions et les rapports paradigmatiques. Dans l'approche de la TGA, ces rapports ne sont pas considérés en se référant à l’axe syntagmatique à l’intérieur d’une seule classe morpho- syntaxique. Ils sont considérés dans la structure qui résulte de la combinaison des deux axes en même temps.

La figure 2. présente le schème générateur de la lexie nominale (Hajsalah,79). Nous pouvons voir à travers cet exemple les transformations (´aṣl ↔ far‘) à partir de la séquence primitive (morphème minimal sans ajouts).

Retenons que pour chaque niveau de la langue il existe un ou plusieurs schèmes semblables qui modélisent les transformations (´aṣl↔far‘’) régissant le système de la langue.

4.2 Le niveau syntaxique

C’est le niveau supérieur à celui de la lexie. Les unités qui constituent ce niveau ne sont pas le résultat d’une simple combinatoire de lexies. La lexie (ni la kalima) ne constitue pas

l'unité minimale de ce niveau (Hajsalah, 1979)(Sibawayh, 1980). De plus les relations qu'entretiennent les éléments de ce niveau sont de toute autre nature.

Considérons les séquences suivantes :

(1) # abdullahi qaimun #
(2) # qaimun abdullahi #

Sibawayhi affirme que la relation entre les deux lexies n'est pas une simple concaténation mais une relation de construction (ou bina´) puisque la suppression de l'une des deux lexies fait disparaître l'unité. D'autre part ces séquences peuvent se retrouver dans des séquences plus larges.

(3) # inna abdallahi qaimun #
(4) # kāna Abdullahi qaiman #

Nous voyons que ces séquences dérivent de (1) par adjonction de " inna " et "kana". Etant donné qu'il y a une même relation de bina´ entre ces lexies et que (3) et (4) dérivent de (1), il est possible de les faire correspondre terme à terme:

(1) # ϕ abdullahi qaimun #
(2) # kāna abdullahi qaiman #
(3) # inna abdallahi qaimun #

On remarquera que les éléments occupant la colonne de gauche à l'initiale des séquences semblent avoir un rapport avec les désinences contenues dans les lexies.
Ce rapport est justement considéré par les grammairiens arabes comme une rection ; les éléments régissants déterminent en effet la marque désinentielle des éléments régis. Ceci nous permet alors de rapprocher de cet ensemble cette autre séquence de lexies qui comporte une lexie verbale :

(5) # ḍaraba abdullahi ʻAmran #

où "ḍaraba" est un verbe, "abdullahi" est sujet du verbe "ḍaraba". On remarquera d'autre part que (1) comporte vis à vis des autres séquences l'expression zéro du régissant : c'est cette expression que les grammairiens arabes appellent " ibtidā´ ".

D'autre part, il existe un élément parmi ceux que le régissant gouverne qui ne peut jamais être antéposé à son régissant : il s'agit de l'élément régi au "naṣb" (marque 'a') par les verbes de la classe de "inna" et au "raf'" (marque 'u') par ceux de la classe de " kāna " et ceux de la classe de " ḍaraba " autrement dit tous les items qui figurent dans cet ensemble en seconde position. En ce qui concerne la construction qui comporte l'expression zéro du régissant, c'est le "mubtada´ " c'est à dire l'item régi par zéro qui correspond au terme non antéposable.

Dans la terminologie de la TGA, le terme occupant la position du régissant s'appelle « العامل » et les termes occupant une position de terme régi sont désignés par « المعمول » ; le kitāb explicite de façon exhaustive tous les cas de figure de la rection « العمل ».

Pour illustrer les différentes positions qui constituent ces mithāl syntaxiques, (Hadjsalah, 1979) désigne le terme régissant par R, et les termes régis par T_i.

L'item régi obligatoirement en seconde position est celui que Sibawayhi appelle « awalu-ma- yushghalu-bihi-al-'amil » c'est à dire le terme qui absorbe en premier le régissant. Cette subordination (ordre + mise en dépendance avec ce qui précède), Sibawayhi la simule par les séquences sus- citées en y visant l'ordre abstrait :

R (régissant syntaxique) → T_1 (terme régi en premier) T_2 (terme régi en second) où T_1 ne doit pas être antéposé à R. On peut avoir dans le discours (R, T_1, T_2), (R, T_2 , T_1), (T_2 , R, T_1). Il est à remarquer que la construction ou intégration structurelle n'est pas entre R et T_1 mais entre le couple ordonné (R, T_1) et T_2.

Ce couple ordonné peut se trouver seul dans le discours, sans T_2 (ex : # qāma abdullahi #, #qumtu #. Enfin, le contenu de ces entités s'interprète sur le plan casuel ainsi : dans R, on peut avoir soit zéro, soit un verbe exponentiel tel que "kāna " (exposant temporel) ou bien un exposant non verbal de la classe de " inna " (particule de corroboration), soit aussi un verbe non exponentiel, tel que " ḍaraba".

Le contenu de R détermine en fait, le contenu casuel des termes régis. Ainsi, si R= zéro, T_1 a nécessairement comme contenu un " mubtada´ " qui est l'appellation formelle de T_1 mais qui peut s'interpréter sur le plan casuel comme sujet d'un "khabar", ce dernier qui est le contenu de T_2 dans ce type de structure pouvant s'interpréter comme l'item véhiculant une information à propos du terme posé qui est T_1.

Si R= exposant verbal ou non, le noyau de la séquence ne change pas puisque ces

exposants lui sont affectés en tant que tels. On parle de " ´ism et khabar de kāna ou de inna " c'est à dire " ´ism" ou "khabar" régis par ces exposants.

Enfin, si R= verbe non exponentiel, on obtient alors une séquence qui bien qu'identique à la précédente n'en a pas moins ses propriétés propres. T_1 doit avoir alors comme contenu un sujet et T_2 un complément d'objet, T_2 étant alors susceptible d'omission.

La formule (R→ T_1) T_2 constitue en fait, un mithāl/schème générateur capable de caractériser tous les types de noyaux syntaxiques (Hadjsalah, 1979). Ainsi, il existe au niveau supérieur à la lexie, un schème générateur d'items où toutes les *constantes* des niveaux inférieurs sont transformées en *variables* : abstraction du contenu des éléments et abstraction de l'ordre au niveau central à l'exception de l'ordre (R, T_1) sans lequel l'indétermination serait totale sur le plan formel. Aussi, la formule que l'on vient d'examiner permet de limiter considérablement la combinatoire syntaxique.

Les entités syntaxiques sont aussi susceptibles de recevoir comme contenu du niveau inférieur, non seulement des lexies mais aussi des segments signifiants et même des unités syntaxiques de leur propre niveau à savoir la formule (R→ T_1) T_2 elle même.

D'autre part, une telle formule qui relève d'un niveau d'abstraction supérieur à celui de la lexie et du segment signifiant n'est pas nécessairement liée à un palier qui serait matériellement supérieur à celui des autres unités.

En fait, il y a de la syntaxe même à l'intérieur des lexies et jusque dans le noyau de la lexie, ainsi, # ḍarabtuhu# est une lexie (verbale) analysable en (R→ T_1)= ḍarabtu, T_2 = hu, et constitue à ce niveau d'abstraction une structure purement syntaxique.

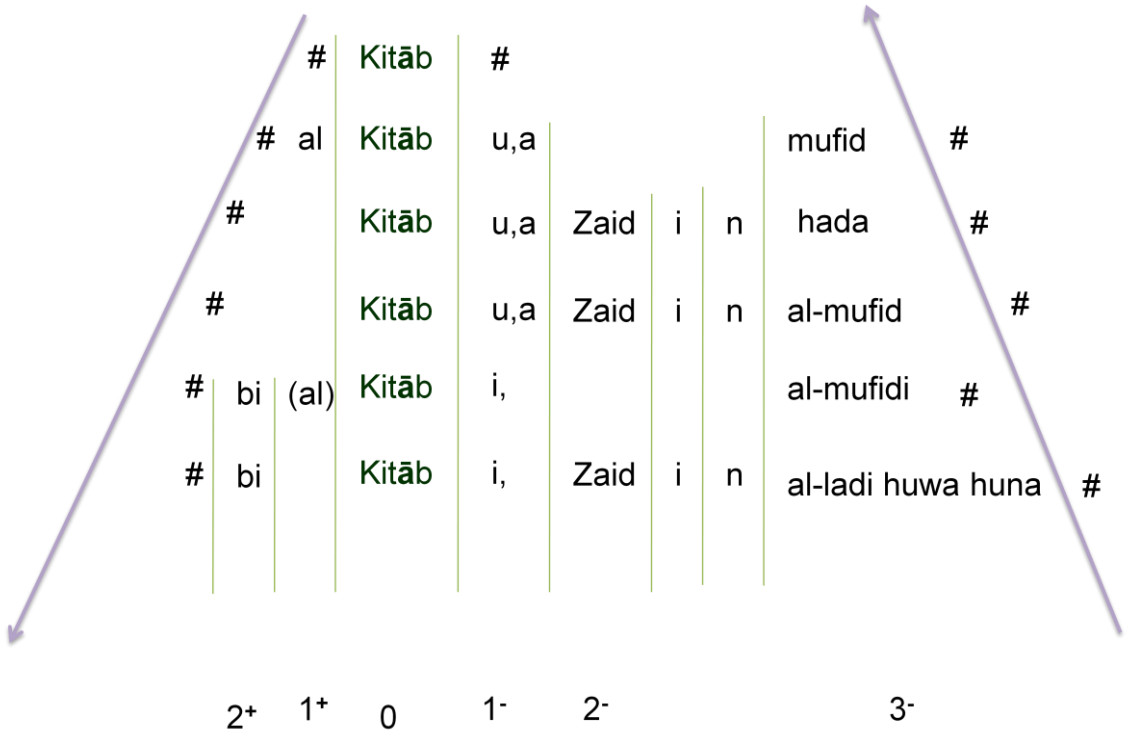


Figure2 - schème générateur de la lexie nominale-

légende

0 = noyau de la "lafda" : généralement un nom commun, c'est la plus petite unité isolable dans le discours.
1^+ = déterminant;
2^+ = préposition;
1^- = désinences casuelles;
2^- = déterminant par annexion directe (mudaf ilayh).
3^- = adjectifs, relatives

4. Conclusion

Bien qu'il soit considéré comme le principal texte fondateur de la tradition grammaticale arabe, et qu'il ait effectivement exercé une influence considérable sur son développement, le Kitāb de Sibawayh n'en reflète pas moins une démarche radicalement différente de celle des grammairiens « classiques ». Cette démarche donne un statut déterminant aux *opérations effectuées par le locuteur en vue de construire une séquence linguistique appropriée à son « vouloir-dire » spécifique* ; ces opérations, organisées en stratégies, mettent enjeu des choix indissociablement formels et sémantiques (Guillaume, 1986).

Nous avons présenté dans ce chapitre les concepts fondamentaux de la TGA, nous avons aussi décrit les niveaux structurels les plus importants et qui nous intéressent dans ce travail (niveau de la lexie et niveau syntaxique simple). Nous reviendrons dans les deux prochains chapitres sur le qiyās comme analyse distributionnelle, d'abord pour proposer une approche empirique non supervisée et n'utilisant aucune ressource pour la découverte automatique des structures de la langue arabe à partirs de texts bruts, ensuite pour interpréter ces structures ainsi que les concepts présentés dans ce chapitre dans le cadre d'un modèle mathématique formel.

Chapitre 5. Découverte des Structures de la Langue Arabe : Que peut nous apprendre une Analyse Formelle ?

1. Introduction

A l'heure actuelle où les besoins pour les outils de TAL se font de plus en plus ressentir, la langue Arabe dispose de très peu de chose en termes d'outils et de ressources. La langue Arabe, de part sa structure différente des langues latines présente des défis intéressants pour les chercheurs. Nous croyons que tout travail dans ce domaine pour ne pas être ad hoc, doit tenir compte des spécificités de la langue et des descriptions qui en ont été faites par ses linguistes. Il n'existe pas de grammaire formelle de la langue arabe au sens contemporain et la seule référence reconnue demeure la Tradition Grammaticale Arabe (TGA) qui bien que ne constituant pas une grammaire formelle au sens du TAL est néanmoins fondée sur une démarche scientifique rigoureuse comme nous l'avons vu au chapitre 4. Nous nous intéressons ici à la simulation algorithmique de l'analyse distributionnelle du qiyās pour la découverte automatique des structures morpho- lexicales et syntaxiques de la langue arabe. La motivation pratique pour une telle approche empirique est la non- disponibilité de ressources pour cette langue. Nous travaillons sans lexique ni tables d'affixes prédéfinis. Allant dans le sens de la TGA de la langue en *usage,* nous travaillons sur un corpus écrit en Arabe Standard Moderne (ASM) qui est la version moderne de l'arabe classique, l'arabe étant une langue restée stable dans ses structures (Massignon, 1954).

2. Travaux en TAL arabe

Bien qu'il existe peu de travaux comparativement aux autres langues pour l'arabe, les approches non supervisées ont aussi été utilisées dans quelques travaux (qui restent peu nombreux) pour le traitement de la langue arabe.

(Snider *et al.,* 2006) proposent une approche non supervisée pour l'induction des classes de verbes de l'ASM en utilisant les ressources de l'Arabic Tree Bank ainsi que l'Arabic Gigaword. (Lee *et al.,* 2003) présentent un algorithme pour la segmentation des morphèmes mené sur un corpus de 110,000 mots segmenté manuellement. Nous avons nous- mêmes utilisé une approche non supervisée mais utilisant des tables de préfixes et de suffixes et un lexique dans le cadre d'un projet de recherche d'information multilingue (Aliane & Alimazighi, 2006a) ainsi que pour l'extraction de connaissances pour peupler une ontologie à partir de textes dans le cadre d'un projet pour la constitution d'une ontologie pour la linguistique arabe (Aliane & al, 2010a). Nous avons utilisé aussi une

méthode d'analyse tabulaire, mais en utilisant les mêmes ressources, lexique et tables de préfixes et suffixes, pour l'extraction des marqueurs du temps et de l'aspect dans les textes en arabe (Aliane& Alimazighi, 2006b). D'autres travaux utilisant une approche non supervisée pour l'analyse morphologique de l'arabe sont présentés dans (Soudi *et al.,* 2007)) mais s'intéressent plutôt à la morphologie non concaténative (prenant en considération les schèmes vocaliques) de l'arabe et s'intéressent au mot en tant que constitué d'une racine consonantique et d'un schème vocalique. Nous ne nous attarderons pas sur ces travaux car l'ASM aujourd'hui est rarement voyellé, donc nous nous intéressons à la morphologie flexionnelle et nous mettons de côté la morphologie dérivationnelle. En effet la prise en compte des procédés dérivationnels de la morphologie n'est pas nécessaire pour la découverte des structures et leur catégorisation. Les schèmes vocaliques apportent plus des effets de sens aux différents mots, et ce que nous recherchons c'est d'abord la structure/forme.

De façon générale, les travaux qui utilisent des approches non supervisées pour le traitement de la langue arabe n'utilisent pas de règles ou de corpus d'apprentissage comme dans l'apprentissage supervisé mais utilisent des ressources comme la prague tree bank (donc textes annotés en parties du discours), et pour la segmentation morphologique des tables de préfixes et suffixes préalablement définis (Khodja & al, 2001) (Diab & al, 2004) (Kassas, 2005).

Notre approche s'inscrit plutôt dans l'esprit de l'approche de (Déjean, 1998) et (Biemann, 2007) donc une approche libre de connaissance (knowledge-free) allant plus dans le sens de procédures de *découverte* que d'apprentissage. L'originalité de notre approche par rapport aux autres travaux c'est qu'en plus d'être libre de connaissance, elle est entièrement fondée sur l'analyse distributionnelle du qiyās de la TGA. Comme nous l'avons déjà mentionné, nous ne nous intéressons pas à la morphologie non concaténative de l'arabe. Nous partons du fait que les textes écrits en ASM ne sont pas voyellés (ou rarement). Nous ne faisons pas une analyse morphologique mais nous essayons de *découvrir* les morphèmes consonantiques de l'arabe à partir de ce corpus de textes, sans dictionnaire ni de tables de préfixes et de suffixes prédéfinis.

Nous avons dit au début de la thèse que notre objectif est de simuler automatiquement la démarche des anciens grammairiens arabes avec deux objectifs :

1. offrir au fur et à mesure des outils de TAL tel que des segmenteurs, étiqueteurs en parties du discours, … la particularité de ces outils sera qu'ils utiliseront la terminologie de la TGA car le travail est fondé sur la TGA.
2. Induire un modèle formel de la TGA

Sur le plan pratique, l'absence de ressources disponibles pour la langue arabe justifie une approche ascendante partant d'un corpus de la langue. Sur le plan théorique, une telle démarche coïncide avec la démarche des anciens grammairiens arabes qui partent du corpus observable de la langue pour élaborer les règles (ḥudūd) régissant le système de la langue. Donc l'idée du travail que nous présentons ici et qui sur le plan pratique rejoint les travaux que nous avons présenté au chapitre 3 est : ***sans utiliser de ressources, que peut nous apprendre une analyse formelle d'un corpus de textes écrits en langue arabe ?*** L'objectif n'est pas de faire une analyse morphologique ou syntaxique mais nous allons plutôt essayer de ***découvrir formellement*** les structures et niveaux de structures de la langue arabe sans l'utilisation d'aucune ressource et avec le minimum de connaissances. Nous allons simuler pour cela l'analyse du qiyās grammatical. Nous nous intéresserons aux structures de base qui sont les morphèmes, les lexies et les structures syntaxiques.

3. Les données du corpus d'étude

> Un corpus se compose par définition de discours, de langue "concrète", et c'est immanquablement sous la forme de textes – écrits ou parlés – que la langue se réalise en discours. (Vergne, 1995)

La langue arabe appartient à la famille des langues sémitiques. Elle s'écrit de droite à gauche, ne dispose pas de lettres majuscules et fait peu usage des signes de ponctuation. Le système d'écriture de l'arabe comprend vingt neuf consonnes prenant différentes formes selon leur position dans le mot. Les voyelles (dites aussi voyelles courtes) n'appartiennent pas à l'alphabet et sont plutôt utilisées comme des diacritiques au dessus ou au dessous d'une consonne pour apporter à cette dernière le son désiré et apporter ainsi à l'ensemble du mot le sens désiré. Quelques auteurs considèrent que l'alphabet arabe est constitué de vingt cinq consonnes et que la hamza "ء" et les lettres dites faibles alif "ا", le wa "و" et le ya "ي" ont le statut de voyelles longues (comme les signes diacritiques sont des voyelles courtes) et ne font donc pas partie de l'alphabet. Nous ne partageons pas cette vision. Selon Sibawayh qui est le fondateur de la TGA, l'alphabet arabe est constitué de 29 lettres. Il les met dans l'ordre suivant, commençant par les laryngales et finissant par les labiales, autrement dit selon leur place d'articulation le long du conduit vocal (Jiyad, 2005).

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

En effet, ces lettres (ḥuruf) considérées comme faibles (glides) quand elles apparaissent dans les verbes sont aussi sujettes à la voyellation selon le son qu'on veut leur apporter, nous préférons par conséquent, garder l'alphabet tel qu'il a été défini en premier par Sibawayh. Les lettres « alif, ya, et la hemza » prennent différentes formes (transcriptions) dans le corpus, nous les prenons toutes en considération dans ce travail.

La langue arabe ne dispose pas de grands corpus disponibles gratuitement pour le TAL. Il existe un certain nombre de corpus payants. Le corpus que nous utilisons a pour source des corpus de presse disponibles sur le net ainsi que des textes tout genre (roman, histoire pour enfants, cours d'arabe). Un travail préliminaire de nettoyage est effectué sur les textes pour enlever les signes diacritiques s'il y en a ainsi que les caractères numériques ou tout autre signe en dehors des signes de ponctuation que nous gardons car ils ont un rôle à jouer dans le travail.

4. Retour sur le qiyās comme méthode distributionnelle **(Aliane & Alimazighi, 2011a)**

Nous avons vu que la Tradition grammaticale Arabe appelée 'ilm al 'arabiyya est définie par ses auteurs comme la science de l'induction des maqāyis ou règles du parler des arabes. Une telle définition nous apprend donc que d'un point de vue méthodologique, la TGA est une linguistique de corpus dans un framework contemporain. Nous avons présenté au chapitre 4, le corpus de la langue arabe sur lequel s'est fondée la constitution de la TGA. Nous allons revenir ici sur le qiyās grammatical en tant que méthode d'analyse de corpus et le relater à la méthode distributionnelle de Harris qui comme nous l'avons vu constitue la référence en linguistique descriptive contemporaine et a inspiré les travaux en TAL non supervisé.

4.1 La recherche des Naẓā´ir : le rasoir d'Occam

Nous rappelons brièvement les concepts clés de la TGA : "bāb", ""´aṣl", "far' ", " naẓīr", "mawḍi'"".

- Un bāb correspond (approximativement) à la notion de classe,

-´aṣl (pluriel ´uṣūl) est un mot polyvalent dont le sens premier est celui de "premier" au sens de prémisse, hypothèse, quelque chose d'où d'autres choses peuvent être inférées ou dérivées; la chose qui est "première" au regard d'autres choses; le mot signifie aussi prototype, origine.

- Far‘ (pluriel furū‘) réfère à ce qui est dérivé ou inféré à partir d’un ´aṣl, signifie aussi branche, instance d’un prototype.
- naẓīr, (pluriel naẓā´ir") signifie similaire et mawḍi'' signifie position, place, distribution

Formellement, le qiyās grammatical est une approche d’analyse exploratoire de données langagières qui cherche à classifier les éléments de la chaine parlée en observant leurs mawḍi‘-s c'est-à-dire position/distribution dans le corpus (Aliane, 2001). Pour relier les ´uṣūl et les furū‘, le qiyās passe par la recherche des naẓā´ir ou éléments similaires à l’élément en cours d’étude en observant donc leur « mawḍi‘’-s ». Le linguiste contemporain M. Carter a été le premier à donner le sens de distribution au sens Harrissien à la notion de mawḍi‘ (Carter, 1973). Nous verrons plus loin que cette notion dépasse en réalité la notion Harrissienne, mais à cette étape du travail, la recherche des ‘naẓā´ir’ dans le corpus tout comme dans la méthode Harrissienne a pour principe la recherche des régularités à travers l’observation des distributions des éléments linguistiques dans le corpus.

Les anciens grammairiens arabes dans le qiyās, procèdent toujours en commençant par la constitution des ´uṣūl à partir de l’observation du corpus. Il s’agit de déterminer ***l’hypothèse minimale ou la prémisse*** qui décrit le mieux la réalité par rapport au faits linguistiques constituant l’objet d’étude en cours ; cette prémisse sera prise comme un ´aṣl. L’observation du comportement des ´uṣūl dans le corpus, c'est-à-dire leurs distributions va déterminer les furū‘ et les bāb-s (classes distributionnelles). Pour cela, le qiyās utilise d’autres critères méthodologiques : la notion d’*autonomie de réalisation et la récurrence*. De ce point de vue, nous pouvons dire que la méthode du qiyās est une méthode minimaliste qui partage les mêmes soucis que les programmes minimalistes contemporains (Chomsky, 1995) par exemple, à savoir l’économie et la couverture linguistique autrement dit arriver à *décrire toutes les structures de la langue en usant du moyen le plus économique possible*. En réalité c’est là le souci de toute méthode scientifique que Guillaume d’Occam a exprimé par ce qui est devenu célèbre par le principe du *rasoir d’Occam*.

Les unités de la langue sont d’abord déterminées sur la base de leur autonomie de réalisation ; ensuite les règles relatant les comportements de ces unités (à travers l’observation de leurs distributions) sont établies sur la base de la récurrence de ces distributions dans le corpus. Le principe d’autonomie de réalisation est utilisé à tous les

niveaux de la description linguistique. La récurrence désignée par le mot « iṭirad » dans le kitāb de Sibawayh, est le critère utilisé pour établir la validité d'un qiyās.

4.2 Découverte Automatique des morphèmes (Aliane & alimazighi, 2011a) (Aliane & Alimazighi, 2011b)

L'analyse distributionnelle Harrissienne a inspiré les approches de TAL non supervisé. Elle est au cœur de beaucoup de travaux qui utilisent une approche ascendante pour induire des connaissances sur la structure de la langue étudiée.

La TGA utilise le principe d'autonomie de réalisation à chaque niveau de la description linguistique. Dans la TGA, on ne trouve pas la notion de phrase comme point de départ de l'analyse linguistique. Le concept de « jumla » (phrase) n'apparait jamais dans le kitāb de Sibawayh. L'unité minimale significative est celle qui peut être naturellement actualisée par le locuteur entre deux pauses : cette unité est appelée "lafẓa" ou lexie. Elle peut être une lexie isolée par exemple « kitāb » qui correspond au lexème dans les théories occidentales ou bien une unité plus complexe (voir chapitre 4). Cette unité définit un niveau central qui constitue le point de départ formel pour l'analyse linguistique qui va vers le bas pour découvrir les morphèmes et vers le haut pour découvrir les unités syntaxiques. Au niveau syntaxique, la TGA cherche à découvrir l'énoncé minimal qui est une combinaison (au sens construction) de lexies qui peut être actualisée par le locuteur et qui se retrouve dans des énoncés plus larges. Enfin, le principe de récurrence appelé (iṭirad) est la condition pour établir les faits de langue comme étant des règles du système de cette dernière (maqāyis, hudūd) (Sibawayh, 1980), (Hadjsalah, 1979) (Bohas & al, 1990). Malheureusement, les auteurs de la TGA ne s'étant jamais soucié de théoriser leur méthode, nous ne disposons pas de modèle formel de la TGA.

En essayant de comprendre la TGA à la lumière de la linguistique contemporaine, la méthodologie de la TGA synthétisée ci-dessus nous a donc conduit aux travaux de l'école américaine, structurale distributionnelle de Bloomfield et Harris (Carter, 1973; Carter, 1980)(Versteegh, 1997).

"Sibawayh seems to be aware of the notion of distribution and environment for he often says that a word, a phrase, … , is an ism (noun) because it occurs in the same position (manzila) as a common noun or any other parts of speech that have already been classified as ism (noun)" (Mosel, 1980)

Comme la TGA avec ses concepts de naẓīr (similaire) et mawḍi' (position/distribution), Harris a présenté des procédures que les linguistes peuvent utiliser pour détecter des entités

distributionellement similaires. Les procédures de Harris sont de nature algorithmique et ont donné lieu à des réalisations informatiques comme nous l'avons vu au chapitre 3.

Nous présentons ici un algorithme général de segmentation qui prend en charge la morphologie flexionnelle de la langue arabe (sans toutefois les voyelles courtes comme marques casuelles car le corpus en ASM n'utilise pas les voyelles courtes.) L'algorithme est capable de segmenter les mots du corpus en séquences de la forme : préfixes + racines + suffixes. En fait, la terminologie de la TGA à ce niveau n'utilise pas les mots préfixes et suffixes. Dans l'analyse distributionnelle du qiyās, on cherche les *ajouts* qui sont appelées « zawā ´id » qui veut dire tout simplement ajouts. Dans les recherches contemporaines on trouve « sawābiq » pour préfixes et « lawāhiq » pour suffixes. Les ajouts comprennent les préfixes, suffixes ainsi que les proclitiques et enclitiques. Mais ce qui nous intéresse dans ce travail, ce n'est pas le statut de ces ajouts mais juste d'identifier pour le moment avec le minimum de connaissance les morphèmes du corpus sans utiliser de lexique, ni de tables de préfixes et suffixes ou de règles. Notre approche permet de découvrir tous les morphèmes du corpus aussi bien les racines que les ajouts ou affixes. Nous ne nous intéressons pas aux infixes où plusieurs racines peuvent correspondre au même radical avec différentes variations inflexionnelles ; nous traitons toutes les racines d'un commun radical comme des unités atomiques séparées. Cette démarche correspond à l'approche de la TGA dont nous nous inspirons. Par ailleurs, dans une visée applicative, l'utilisation de la racine comme morphème plutôt que du radical est plus appropriée pour des applications comme la RI et la TA (Lee, 2003).

4.2.1 *La recherche des morphèmes*

Nous nous intéressons donc, uniquement à ce qu'on peut apprendre de la structure d'un texte en étudiant la forme (lafẓ dans la terminologie de Sibawayh) du texte sans faire appel au sens. Nous n'utilisons pas de lexique mais appliquant le principe du rasoir d'Occam tel que nous l'avons compris de la TGA, nous utilisons un ensemble minimal de prémisses jusqu'à ce que nous ayons expressément besoin de les augmenter. La première prémisse importante dans la TGA est la suivante:

- *la grande masse des mots de la langue arabe est formée des mots à trois consonnes radicales, quelques mots sont constitués de quatre consonnes et rares sont les mots constitués de cinq consonnes. Tous les autres mots sont formés par des procédés de dérivation ou d'affixation à partir de ceux là.*

En réalité cette prémisse constitue un observable du corpus de la langue arabe, observé par les anciens grammairiens à partir du parler des arabes. La forme tri- consonantique du mot arabe au regard de sa majorité dans le corpus, constitue un ´aṣl (prémisse) dans la TGA. En effet, c'est la forme qui est utilisée en premier pour étudier le comportement des éléments de la langue afin d'induire des lois à propos de son système. Le schème فعل est une abstraction de tous les mots tri- consonantiques de l'arabe. Les trois lettres 'ف', 'ع', 'ل' ont été choisies conventionnellement par la TGA de sorte que 'ف' désigne la première consonne du mot, 'ع' désigne la deuxième consonne et 'ل' la troisième. Tout autre kalam (parler) des arabes peut être inféré ou dérivé comme furū' à partir de ce ´aṣl.

Cette prémisse constitue donc notre hypothèse minimale pour l'analyse du corpus et constitue le point de départ de notre algorithme. Nous utilisons aussi le principe d'autonomie de réalisation ainsi que le principe de récurrence. Notre idée est la suivante. A partir de cette seule prémisse, que peut-on apprendre de la structure de la langue arabe en observant le comportement de formes tri- consonantiques dans le corpus ?

Nous prenons arbitrairement un ensemble M constitué de dix mots tri- consonantiques apparaissant dans le corpus et nous cherchons leurs distributions dans ce dernier. Soit X une de ces formes choisies. Dans une première étape, par une décomposition de tous les mots du corpus constitués de plus de trois lettres, nous essayons de retrouver notre X, et ce faisant, nous trouvons de nouveaux éléments concaténés à droite et/ou à gauche de X, ce sont les *ajouts*. X représente le morphème minimal (racine) pouvant apparaître seul dans le corpus et pour utiliser une terminologie Harrissienne, nous pouvons appeler environnements immédiats ces nouveaux éléments concaténés. Nous avons appelé mot ici, toute séquence entre deux blancs, ou entre un blanc et un signe de ponctuation.

Exemple : pour X= كتب nous avons trouvé الكتب بكتب بالكتب مكتب

مكتبة المكتب المكتبة بالمكتبة فكتب كتبت يكتب يكتبونه Ce sont donc les distributions du mot كتب.

Dans une seconde étape, nous considérons ces environnements immédiats obtenus comme des entités formelles significatives pour notre analyse et cette fois- ci, nous utilisons ces environnements (affixes) pour retrouver les autres morphèmes du corpus qui peuvent être tri- consonantiques ou non. En effet, nous n'effectuons pas une analyse morphologique au sens propre mais il est évident que le corpus contient tous les mots permis par le système dérivationnel de l'arabe. Donc, dans cette deuxième étape, nous ne trouvons pas

uniquement les mots tri- consonantiques mais aussi tous les autres mots contenant des infixes. Il est important d'atteindre les morphèmes car nous avons besoin de segmenter le corpus ; les morphèmes contiennent autant d'information sur la structure que les morphèmes grammaticaux et sont essentiels à la découverte des structures syntaxiques par la suite. En sortie de ces deux étapes de segmentation : d'abord utiliser un ensemble initial de morphèmes pour trouver les affixes et ensuite utiliser les affixes découverts pour trouver les morphèmes du corpus est une segmentation de tout le corpus sous la forme b+ X +a, "X" étant le morphème- racine, "a" étant l'ajout à droite et "b" l'ajout à gauche. Le reste des mots n'appartenant pas à M et n'ayant pas été trouvé par la deuxième étape sont considérés comme des morphèmes libres c'est-à-dire ne contenant pas d'affixes.

REMARQUE. — Nous avons choisi de mettre arbitrairement dix mots consonantiques dans notre ensemble M, mais en réalité un seul mot tri- consonantique aurait suffi vu la régularité de la langue arabe et que le comportement d'un seul mot tient lieu de modèle pour tous les autres. Nous avons choisi dix mots pour augmenter les chances de l'algorithme de trouver tous les affixes (ajouts) du corpus.

Comme nous utilisons un corpus électronique écrit en arabe standard moderne, nous considérons que le langage du corpus est celui effectivement utilisé par les arabes (représente la performance) et que le langage du corpus est correct, s'il existe des erreurs d'orthographe ou de transcription, cela n'est pas pertinent pour notre algorithme.

En cherchant les environnements immédiats des morphèmes consonantiques, nous avons implicitement fait usage du principe d'autonomie de réalisation (à l'écrit) : en identifiant, les entités entre deux pauses (la pause étant ici un blanc ou un signe de ponctuation) ; ceci nous emmène à déterminer un deuxième niveau de structure ayant le morphème pour noyau. Aussi longtemps que l'analyse n'apporte pas de contradiction avec notre hypothèse, nous considérons que cette structure correspond à la structure de lexie dans la TGA. Nous appelons L cette structure. Considérons L comme une structure abstraite, les occurrences trouvées dans la recherche des distributions des morphèmes représentent en même temps les différents comportements de L. par exemple, chaque occurrence de كتب trouvée précédemment est une lexie. Comme nous l'avons déjà mentionné, la TGA considère que la lexie constitue le niveau central dans l'analyse linguistique ; nous allons vers le bas pour trouver les morphèmes et vers le haut pour identifier les constructions syntaxiques. La lexie étant justement définie par son autonomie de réalisation.

4.2.2 *Formalisation des structures obtenues*

X pouvant être n'importe quel morphème libre tri- consonantique autre que كتب, nous appelons $P_i \geq 0$ chacun des paradigmes (lexicaux) identifiés ci-dessus. Chacun de ce que nous avons appelé environnement immédiat détermine un paradigme. Le premier critère que nous avons utilisé pour déterminer ce niveau de structure est le principe d'autonomie de réalisation ainsi que la récurrence dans le corpus. Ceci nous permet de définir une structure dont les éléments possèdent un comportement général commun (eu égard à ce principe). A présent, nous voulons juste décrire formellement chaque paradigme. Donc le résultat de cette étape est la re- transcription (étiquetage) du corpus sous la forme:

"$b_{\text{left affixe}} + X_{\text{morpheme}} + a_{\text{right affixe}}$", "a" and "b" étant les éléments concaténés ou ajouts et "X" la racine/morphème. Le second résultat est l'identification de la liste des paradigmes constituant ce second niveau de structure qui correspond au niveau central de la TGA formellement comme suit: $L=\Sigma P_i \geq 0 / P = a + X + b$, X étant n'importe quelle racine tri-consonantique ou n'importe quel autre morphème de base identifié par l'algorithme de segmentation et dont les consonnes appartiennent à l'alphabet arabe.

a ϵ { ال،سي، ,ف، ب، م خ،ي،م ... }, b ϵ { ذ،ج ون، ،وا، ... }.

Comment avons-nous utilisé ici le principe de récurrence ?

1) Comme nous avons considéré que le corpus est correct, l'occurrence d'un élément une seule fois est suffisant pour considérer qu'il appartient en effet à la structure que nous lui avons associé étant donnée notre hypothèse. Par exemple, chaque mot tri- consonantique est aussi une lexie étant donné le principe d'autonomie d'actualisation même s'il apparait une seule fois dans le corpus.

2) le fait que les affixes découverts ne s'appliquent pas seulement à un morphème donné mais à la masse des morphèmes libres du corpus est considéré comme une récurrence qui nous permet d'induire les affixes ainsi que les paradigmes associés comme des faits de langue. Par exemple pour l'exemple كتب ci-dessus, nous avons identifié 13 paradigmes y compris P_0 (morphème sans affixes) et probablement, il existe d'autres comportements ou distributions qui ne sont pas apparues dans le corpus actuel et qui n'ont donc pas été découverts par notre algorithme.

5. Vers la Découverte des Structures Syntaxiques (Aliane & Alimazighi, 2011a)

A ce stade, sans utiliser un lexique ni de tables d'affixes prédéfinies et avec un minimum de connaissances consistant en une seule prémisse, nous avons pu :

1. Découvrir tous les affixes du corpus,
2. Atteindre tous les morphèmes du corpus
3. Définir formellement les paradigmes lexicaux et donc les lexies du corpus.

Nous allons maintenant appliquer encore le principe d'autonomie de réalisation mais cette fois-ci pour découvrir comment se combinent les lexies pour former le niveau de structure suivant. Nous allons donc essayer d'identifier l'énoncé autonome minimal pour ce qui est supposé être le niveau syntaxique.

Nous avons vu l'importance du principe d'autonomie de réalisation dans la TGA. Nous avons aussi vu que nous devons d'abord identifier les structures minimales qui seront prises comme ´uṣūl.

Nous allons donc observer le comportement des P_i dans le corpus autrement dit leurs distributions dans des contextes plus larges. En fait, reconnaitre les limites de phrases dans un texte en cours est une tâche encore plus difficile pour une langue comme l'arabe et ce à cause de l'absence de règles de ponctuation strictes. En effet, il est commun en arabe, d'écrire un discours entier sans un seul point jusqu'à la fin du paragraphe. L'arabe fait plutôt un usage excessif des conjonctions de coordination (و) *wa* and (ف) *fa*. De plus comme pour le chinois, le japonais et le corréen, il n'y a pas de lettres capitales en arabe et les lettres graphiques arabes changent de forme suivant leur position dans le mot, la subordination et la coordination. Par conséquent, nous avons à ce niveau besoin d'augmenter notre ensemble de prémisses. Nous ajoutons donc la prémisse : les mots comportant une seule lettre comme "و" qui apparaissent seuls de façon récurrente dans le corpus ne sont pas partie de constructions basiques de la langue mais jouent un autre rôle dans le système de la langue et nous prenons les réalisations autonomes suivantes:

- les énoncés entre un début de paragraphe et la lettre "و"
- les énoncés entre un début de paragraphe et un signe de ponctuation (s'il y en a)
- les énoncés entre deux signes de ponctuation (s'il y en a)

Comme nous nous voulons découvrir des structures de base qui serviront de ´uṣūl et non des structures complexes et donc pour augmenter les chances que notre algorithme ne

découvre que de telles structures, nous considérons seulement les énoncés comportant deux ou trois P_i. Nous n'allons pas avoir toutes les structures syntaxiques du corpus mais nous espérons bien avoir au moins quelques modèles de structures de base (´uṣūl) du niveau syntaxique.

Nous avons obtenu par exemple pour for P_0 = X; P_1= X+ال, P_2=X+ي les récurrences suivantes:

P_0P_1 خرج الرجل (est sortit l'homme)

$P_0P_1P_1$ كسر الريح الشجرة (brisa le vent l'arbre)

P_2P_0 يذهب الفريق (partira l'équipe)

P_1P_0 البحر هادئ (la mer est calme)

$P_1P_1P_0$ الولد الصغير نائم (le petit garçon dormant)

De la même façon que nous avons procédé pour les paradigmes lexicaux et comme nous n'avons pas plus d'information concernant les structures obtenues, nous les étiquetons simplement comme *S* pour dire que nous avons affaire à des structures syntaxiques.

En fait, les trois premiers énoncés sont des modèles de phrases verbales arabes simples. Les deux derniers sont des modèles de phrases nominales simples comportant un mubtada´ et un khabar. Mais à ce stade nos algorithmes ne donnent pas cette information. Nous ajoutons juste des crochets à notre corpus segmenté pour identifier les structures syntaxiques découvertes en les étiquetant par le label *S*. Néanmoins, l'observation de successions simples des P_i nous a conduit à remarquer que certaines successions apparaissent souvent (récurrence) ensemble dans les mêmes distributions à travers le corpus. Par exemple, nous avons pu observer que les suites suivantes sont substituables dans leurs environnements respectifs ce qui signifie qu'elles ont les mêmes distributions:

يذهبان يذهبون يذهبن سيذهب سيذهبون سيذهبان سيذهبن ذهبت يذهب تذهب

Par conséquent, X+ي , ون+X+ي , ان+X+ي , ن+X+ي , X+ت , ت+X, X+ي+س, ان+X+ي+س , ن+X+ي+س , ن+X+ي

sont substituables et nous sommes autorisés à formuler l'hypothèse que ces P_i peuvent être représentés par un seul paradigme. Néanmoins, nous pensons que pour confirmer cette hypothèse nous allons avoir besoin de plus de connaissances.

Par ailleurs, en considérant les modèles de S, nous avons trouvé que le paradigme $P_0P_1P_1$ comprend par exemple les énoncés suivants :

- ‘ .ذهب الرجل الطويل‘

- ‘"اكل الولد الموز

Dans le premier énoncé nous avons un verbe intransitif, un sujet et un adjectif; dans le deuxième énoncé, nous avons un modèle VSO. Nous avons donc besoin d’autres connaissances pour discriminer des structures apparemment similaires. Ceci pour dire que nos algorithmes déterminent les structures syntaxiques mais ne donnent pas le type de ces structures avec les hypothèses minimales que nous avons choisis. La Figure 3 donne la description générale de notre approche.

Pour synthétiser, les deux premières étapes de notre algorithme découvrent tous les affixes et tous les morphèmes du corpus; la troisième étape n’identifie pas toutes les structures syntaxiques mais découvre les structures syntaxiques de base sans en donner encore le type. Nous les étiquetons *S* juste pour dire que nous avons découvert le niveau syntaxique et que telle est une structure syntaxique. Le résultat est un corpus étiqueté de la manière suivante:

((رجل$_m$ + ال$_a$)$_L$(خرج$_m$)$_L$)$_S$

((ة$_b$ + شجر$_m$+ال$_a$)$_L$ (ريح$_m$ + ال$_a$)$_L$ (كسر$_m$)$_L$)$_S$

((فريق$_m$ + ال$_a$)$_L$ (ذهب$_m$ + ي$_a$)$_L$)$_S$

((هادئ$_m$)$_L$ (بحر$_m$+ ال$_a$)$_L$)$_S$

((نائم$_m$)$_L$ (صغير$_m$ + ال$_a$)$_L$ (ولد$_m$ + ال$_a$)$_L$)$_S$

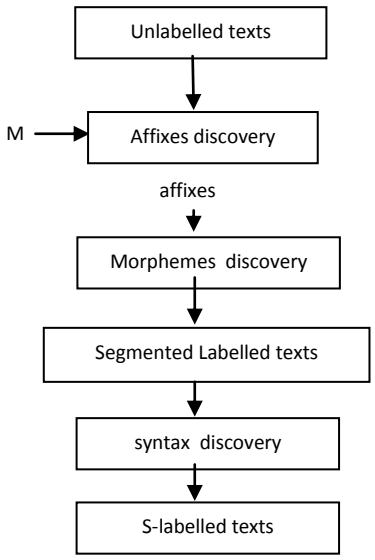


Figure3. ***General description of the system***

Nous générons aussi pour l’utilisateur les affixes à droite et à gauche, les morphèmes du corpus ainsi que les paradigmes lexicaux. Néanmoins si notre algorithme à ce stade découvre tous les morphèmes et tous les paradigmes lexicaux du corpus, il ne détermine pas encore avec le minimum de connaissance fixé toutes les structures syntaxiques mais seulement quelques structures de base.

Ce travail peut déjà être utilisé dans des applications de TAL comme l'étiquetage en parties du discours, le text mining, le résumé de textes ou toute autre application ayant besoin de discriminer les structures linguistiques du corpus ; les morphèmes peuvent servir dans un système de génération automatique d'un dictionnaire électronique de l'arabe.

6. Evaluation et Conclusion

Notre approche permet de découvrir ***tous*** les affixes du corpus et par là tous les morphèmes ainsi que les paradigmes lexicaux apparaissant dans le corpus. Toutefois, il existe des erreurs. Il s'agit en particulier des caractères consonantiques qui appartiennent à la liste des affixes et peuvent aussi faire partie du morphème /racine d'un mot par exemple : une forme comme 'فصوص ' sera segmentée comme ف+ صوص ce qui est incorrect dans la réalité car c'est un mot existant de la langue arabe et le 'ف' ici n'est pas un affixe.

Nous avons présenté dans ce chapitre une approche nouvelle et économique pour la découverte des structures de la langue arabe fondée sur l'analyse distributionnelle de la TGA qui en l'occurrence est proche de la méthode Harrissienne. Le résultat de ce travail est une segmentation du corpus ainsi qu'un étiquetage des unités découvertes. La première étape, en plus de découvrir les affixes, nous a permis de découvrir tous les morphèmes du corpus ainsi que le niveau central de la lexie et ses différent paradigmes. La deuxième étape nous a permis de découvrir le niveau syntaxique et les structures de base de ce niveau. Nous avons utilisé comme fondement de ce travail le principe de l'hypothèse minimale (ʹuṣūl) ainsi que les deux principes d'autonomie de réalisation et de récurrence utilisés par la TGA qui ne sont pas des connaissances linguistiques mais des principes méthodologiques. En fait, l'étiquetage résultant n'est pas un étiquetage au sens classique puisque nous n'assignons pas encore les parties du discours aux unités découvertes. Allant dans l'idée du paradigme de découverte de structures, nous utilisons plutôt comme labels les symboles que nous avons choisis dans ce travail *m, a, b, L, S.* Nous avons essayé d'être cohérents avec l'intuition linguistique dans TGA de sorte à éviter des incohérences dans les développements futurs de ce travail. Le corpus ainsi étiqueté peut déjà servir comme base pour un travail sur l'étiquetage en parties de discours par exemple ou bien pour l'analyse dans d'autres travaux de TAL pour des applications comme le text mining ou le résumé automatique. Concernant le travail qui a été accompli, étant donné les limites d'une

analyse basée uniquement sur la forme des textes, nous croyons que ce que nous avons pu faire avec aussi peu d'information est déjà intéressant en nous faisant l'économie d'un lexique (l'une des suites de ce travail est de générer un lexique à partir du corpus) et de règles sur la langue. Néanmoins, contrairement à un ordinateur, le linguiste qui décrit une langue possède l'avantage de quelques intuitions et hypothèses dans sa tête par son expérience des données linguistiques. En l'absence d'intuition et d'expérience linguistique (ou autre), nous nous arrêtons donc là pour le moment en termes de structures découvertes ; nous allons essayer dans le chapitre suivant d'interpréter ces structures ainsi que l'analyse distributionnelle du qiyās dans un cadre mathématique formel, en l'occurrence la théorie des catégories.

Chapitre 6. Vers une Formalisation de l'Analyse Distributionnelle en Termes de Catégories

1. Introduction

Nous avons présenté dans les chapitres précédents l'approche empirique de la langue et le rôle de l'analyse distributionnelle dans le cadre d'une telle approche ; nous avons vu que l'approche de Harris, somme toute, intuitive et fondée sur le bon sens, a inspiré les travaux en TAL avec le renouveau de l'approche empirique dû à l'avènement des grands corpus électroniques textuels. Nous nous sommes particulièrement intéressé aux techniques non supervisées. L'objet d'investigation étant pour nous la tradition grammaticale arabe, nous avons vu que la méthode d'analyse de la TGA qui est le qiyās est proche dans sa démarche de la méthode distributionnelle Harrissienne. Cependant, si de l'aveu de Harris lui-même la méthode distributionnelle est juste une méthode pour la linguistique descriptive et ne permet pas d'induire une grammaire de la langue, nous avons vu au chapitre 4 que le qiyās justement a permis aux anciens grammairiens arabes d'induire les maqāyis ou règles ou ḥudūd du système de leur langue et qui sert à ce jour de référence pour la grammaire arabe.

Nous allons revenir dans ce chapitre sur les limites du distributionalisme Harrissien et nous allons voir comment le qiyās grammatical est plus que la méthode distributionnelle en montrant qu'on peut en interpréter les concepts dans le cadre de la théorie mathématique moderne des catégories.

2. Retour sur la méthode distributionnelle

Les mots clés d'une analyse distributionnelle sont 'régularités' et similarités', 'classement'. En effet, l'analyse distributionnelle que ce soit chez Harris ou dans la TGA a pour but la recherche des régularités c'est-à-dire des similarités structurelles pour pouvoir classer les éléments similaires en classes distributionnelles chez Harris et en bāb-s dans TGA. Quand on trouve des régularités ou des éléments similaires sous quelque point de vue que ce soit, cela veut dire qu'on trouve des invariants. Notre idée ici est de considérer ces invariants et les contextes distributionnels dans lesquels ils apparaissent sous un point de vue opératoire c'est-à-dire au lieu de nous intéresser aux invariants et aux contextes en eux-mêmes, considérer les changements de contextes comme des opérations. La théorie mathématique moderne des catégories nous semble être le meilleur cadre pour cela. En effet dans la théorie des catégories, aujourd'hui la théorie de structures par excellence, les

objets ne sont pas considérés en eux-mêmes mais seulement à travers leurs interactions avec les autres objets. Ces interactions sont désignées par morphismes ou flèches. Nous allons commencer par présenter la théorie des catégories.

3. La Théorie Mathématique des Catégories

Née au début des années quarante des réflexions de Eilenberg et MacLane sur les relations entre algèbre et topologie (Eilenberg & Maclane, 1945), la théorie des catégories est l'une des premières théories mathématiques à poser le problème de la « *naturalité* » de certaines constructions formelles, où le terme désigne les *propriétés qui sont indépendantes des caractéristiques et des représentations particulières des objets.*

Le concept privilégié dans cette démarche d'abstraction sera celui de transformation, tout d'abord entre des objets (*morphismes*) mais aussi entre des catégories *(foncteurs*). Le processus d'abstraction peut ainsi continuer en considérant des transformations entre foncteurs (*transformations naturelles)* et la possibilité de récupérer, à partir de ces notions, des propriétés caractéristiques des objets de départ. Cette théorie a très vite rencontré un grand succès auprès des mathématiciens, qui l'ont intégré à leur vocabulaire. En effet, il s'est avéré que les catégories fournissent un cadre adéquat pour généraliser des opérations très courantes en mathématiques. La théorie des catégories se présente comme une théorie *unificatrice* en mathématiques : plutôt que de définir de nombreuses notions similaires pour les ensembles, les groupes, les anneaux, les espaces vectoriels, … on définit *une* notion pour les catégories qui comporte toutes les autres comme cas particuliers. Un peu plus tard, Lawvere a même proposé d'utiliser cette théorie comme fondement des mathématiques au lieu de la théorie des ensembles (Lawvere, 1966).

Pour un non mathématicien, il n'est pas facile de saisir les concepts que la théorie des catégories généralise. Nous présenterons ci-dessous les concepts qui fondent cette théorie. Depuis quelques années, la TC connait beaucoup d'applications en dehors des mathématiques. En effet, elle présente un intérêt non négligeable pour certaines branches : elle permet de parler d'objets non pas en termes de leur structure interne, mais en termes de leurs interactions. Les objets forment une sorte de boite noire : seuls leurs échanges sont accessibles dans la démarche catégoricienne (Amiguet, 1998).

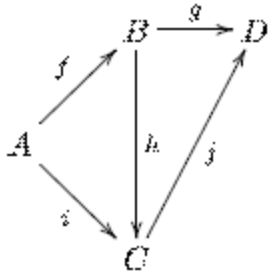
Considérons par exemple le produit cartésien de deux ensembles. L'approche « classique » ensembliste définit le produit $A \times B$ des ensembles A et B comme l'ensemble dont les éléments sont de la forme (a, b) avec $a \in A$ et $b \in B$. Nous verrons que la théorie des

catégories permet de caractériser *A x B* uniquement à travers ses interactions avec les autres ensembles. Mais comme beaucoup l'ont dit, ce qui fait la puissance de cette théorie en fait aussi la difficulté : la théorie des catégories est TRES abstraite. Certains l'ont même appelée « abstract non sense » !

3.1 Les Catégories

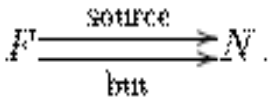
3.1.1 Graphes

La notion de catégorie est basée sur celle de graphe. (Reydeheard & Burstall, 1988) (Awodey, 2003)(Amiguet, 1998). Intuitivement, un graphe est une structure de données permettant de représenter *des liens entre des objets*. On représente généralement un graphe sous la forme de « nœuds » et de « flèches » :



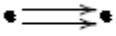
Nous allons adopter une définition abstraite qui se prête bien à l'utilisation en théorie des catégories. Du dessin ci-dessus, nous ne retiendrons que l'ensemble des « nœuds » (ici {A, B, C, D}), l'ensemble des « flèches » (ici {*f, g, h, i, j*} et une manière de trouver, étant donnée une flèche, d'où elle part et où elle arrive. On obtient la :

Définition : Un graphe G est donné par deux ensembles N et F et deux fonctions



Où N est l'ensemble des nœuds et F l'ensemble des flèches de G.

Cette définition admet les « paires parallèles », c'est-à-dire les configurations de la forme



Et les oreilles c'est-à-dire les flèches qui partent et arrivent au même point:

On note $f : A \rightarrow B$ pour signifier que *f* est une flèche de source A et de but B.

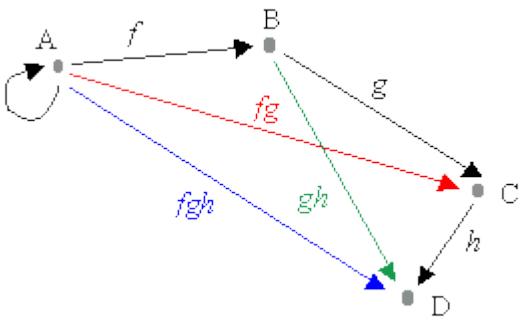
3.1.2 Définition d'une catégorie (Reydeheard & Burstall, 1988) (Awodey, 2003)(Amiguet, 1998).

Définition 1 : Une catégorie est un graphe orienté sur lequel on s'est donné une loi pour composer des flèches consécutives, vérifiant certains axiomes.

Deux flèches *f, g* sont dites *consécutives* si le but de la première est en même temps la source de la seconde: *f:* A ® B, *g*: B ® C. On dit alors qu'elles forment un chemin de longueur 2 de A vers C. Plus généralement, un *chemin* (de longueur *n*) de A vers A' est une suite (f_1, f_2, ...,f_n) de *n* flèches consécutives

f_1: A ® A_1, f_2: A_1 ® A_2, ... , f_n: A_{n-1} ® B.

Une *catégorie* est un graphe dans lequel on définit une composition de flèches, associant à tout chemin (*f, g*) de longueur 2 de A vers C une flèche du graphe de A vers C, dite *composée* du chemin, et notée *fg*. Cette composition vérifie les conditions suivantes:



- *Associativité.* Si (*f, g, h*) est un chemin de longueur 3, les deux composés *f* (*gh*) et (*fg*)*h* qu'on en déduit sont égaux (on les note *fgh*). Il s'ensuit qu'à tout chemin de longueur *n* est aussi associé un seul composé. (Invariance de l'itinéraire)

- *Identités*: A tout sommet A est associée une flèche fermée de A vers A, dite *identité de* A et notée ld_A, dont le composé avec une flèche de source ou de but A est égal à cette autre flèche.

Les sommets du graphe sont aussi appelés *objets* de la catégorie et ses flèches *morphismes* (ou simplement *liens*). Une flèche *f* est un *isomorphisme* s'il existe une flèche *g* (dite son *inverse*) telle que les composés *fg* et *gf* soient des identités (cet inverse est alors unique).

Ainsi une catégorie est formée par des *objets* (les sommets du graphe) et des *liens* entre eux (les flèches ou *morphismes*), mais l'idée essentielle est ***de privilégier les liens sur les objets***. En fait, le succès des catégories dans les domaines les plus variés est dû à la richesse des informations sur les objets qui peuvent être déduites de la seule considération des liens et des opérations sur ceux-ci, quelle que soit la nature et l'anatomie de ces objets.

Définition 2 : une **catégorie C** est

- une collection $\mathbf{Ob_C}$ d'objets, dénotés par a, b, …A, B, …
- une collection $\mathbf{Mor_C}$ de morphismes (flèches), dénotés par f, g, …,
- deux opérations **dom, cod** qui assignent à chaque flèche f deux objets appelés respectivement **domaine** (source**)** et **codomaine** (cible) de **f.**
- une opération **id** assignant à chaque objet b un morphisme $\mathbf{id_b}$ (l'identité de b) tel que dom($\mathbf{id_b}$)=cod($\mathbf{id_b}$)=**b**
- une opération "o" de composition assignant à chaque paire f, g de flèches ayant dom(f)=cod(g) une flèche f o g tel que dom(f o g)=dom(g), cod(f o g) =cod(f)
- de plus, identité et composition doivent satisfaire les conditions suivantes:

la loi d'identité: pour toutes flèches f,g tel que cod(f)= b = dom(g)

$\mathbf{id_b}$ o f =f

g o $\mathbf{id_b}$ **=g**

la loi de l'associativité: pour toutes flèches f, g, h tel que dom(f)=cod(g) et dom(g)=cod(h)

(f o g) o h =f o (g o h).

on écrit f: a → b pour dénoter un morphisme dont la source et la cible sont respectivement a et b. étant donnés, deux objets a et b, la collection de tous les morphismes f tel que f: a→b est dénotée par **C**[a,b]; écrire f ∈ **C**[a,b] est alors une troisième manière d'exprimer dom(f)=a et cod(f)=b. La table suivante liste quelques catégories communes précisant leurs objets et morphismes:

Catégorie	**Objets**	**Morphismes**
Ensemble	ensembles	fonctions
Topologie	espaces topologiques	fonctions continues
Vecteur	espaces vectoriels	transformations linéaires
Grp	groupes	homomorphismes de groupes
PO	ensembles partiellement ordonnés	fonctions monotones

Ce qu'il faut retenir c'est : l'intuition derrière la notion de "catégorie" est de considérer les objets comme une collection "structurée" d'ensembles et les morphismes comme des fonctions "associées" ou "acceptables" relativement à la structure. ***Ce qui correspond exactement à l'intuition derrière la notion de « bāb » dans TGA qui constitue une collection structurée à l'intérieur de laquelle les comportements « acceptables » des objets (linguistiques) sont décrits par les mithāls.***

3.1.3 Quelques définitions catégorielles (Asperti & Longo, 1991) (Amiguet, 1998)(Awody, 2003) (Maclane, 1997)

Comme nous l'avons vu, les axiomes d'une catégorie ne font intervenir que les flèches et leur composition. Une approche purement catégorielle n'utilisera donc que ces notions là. On pourrait penser que l'expressivité d'une telle approche est très limitée (Amiguet, 1998) ; en fait, il n'en est rien. Nous allons présenter quelques constructions simples (mais déjà assez puissantes) des flèches et des objets d'une catégorie.

Objet terminal : Un objet *T* d'une catégorie *C* est appelé objet terminal si pour tout objet *O* de *C* il existe exactement une flèche $O \rightarrow T$.

Exemple : Dans la catégorie d'un ordre partiel cette catégorie coïncide avec celle de *maximum*.

Objet initial : Un objet *I* d'une catégorie *C* est appelé objet initial si pour tout objet *O* de *C*, il existe exactement une flèche $I \rightarrow O$.
Exemple : dans la catégorie d'un ordre partiel, un objet initial est un minimum.

Cette définition est obtenue en renversant les flèches dans la définition précédente. Cela se dit « la notion d'objet initiale est la notion duale de celle d'objet terminal ». En CT, la

dualité est un concept très important ; en effet, chaque fois qu'une construction est définie, la construction duale est aussi définie.

3.2 Les foncteurs

Un foncteur est une correspondance (morphisme) entre catégories.

Définition : soient C et C' deux catégories un foncteur F de C vers C' est une application qui à tout objet A de C associe un objet $F(A)$ de C' et à tout morphisme $f : A \rightarrow B$ de C associe un morphisme $F(f) : F(A) \rightarrow F(B)$ de C' de sorte que :

1. F préserve les compositions : $F(f \circ g) = F(f) \circ F(g)$ dès que $f \circ g$ est défini
2. F préserve les morphismes identités : $F(1_A) = 1_{F(A)}$ pour tout objet $A \in Ob(C)$.

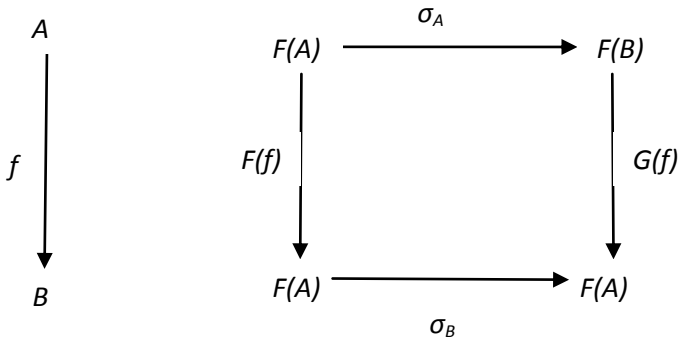
Pour simplifier la notation, on note souvent Ff au lieu de $F(f)$.

3.3 Les transformations naturelles

Une transformation naturelle est une correspondance entre les foncteurs.

Définition : Soit deux catégories C et D, soit deux foncteurs $F, G : C \rightarrow D$. une transformation naturelle σ de F à G noté $\sigma : F \rightarrow G$ de D telle que pour toute flèche $f : A \rightarrow B$ de C

$G(f) \circ \sigma_A = \sigma_B \circ F(f)$.



3.4 Propriétés universelles

La définition par universalité est une technique de définition qui en théorie des catégories, constitue le moyen principal pour caractériser une structure. Nous allons illustrer cette notion de définition universelle par un exemple de la théorie des ensembles (Reydehard & Burstall, 1988) : l'union de deux ensembles *A* et *B* est *le plus petit* ensemble contenant les deux ensembles *A* et *B* ; définir cette union comporte :

- La définition de la classe des objets en question, dans ce cas, la classe de tous les ensembles contenant aussi bien *A* que *B,*
- La définition d'un critère pour distinguer un élément particulier de cette classe par sa relation aux autres éléments. Dans cet exemple, le critère est « le plus petit » c'est-à-dire contenu dans tout élément de la classe.

La théorie des catégories constitue un formalisme adéquat pour exprimer de telles définitions universelles. Aussi bien la classe d'éléments que le critère pour distinguer un élément particulier seront exprimés dans un même langage de « flèches ». La définition par universalité constitue en CT, le format standard pour de telles définitions dans lequel le critère pour distinguer un élément particulier est l'existence d'une flèche unique satisfaisant certaines conditions. S'il se trouve que deux objets satisfont les conditions de la définition, ils sont isomorphes. On dit que les objets sont définis et uniques à *isomorphisme près* (up to isomorphism). Donc, on peut dire que les objets sont vus de manière abstraite indépendamment d'une quelconque représentation particulière. Nous allons présenter quelques exemples de propriétés universelles mais avant, nous allons présenter la notion de diagramme commutatif qui est la base des preuves en théorie des catégories.

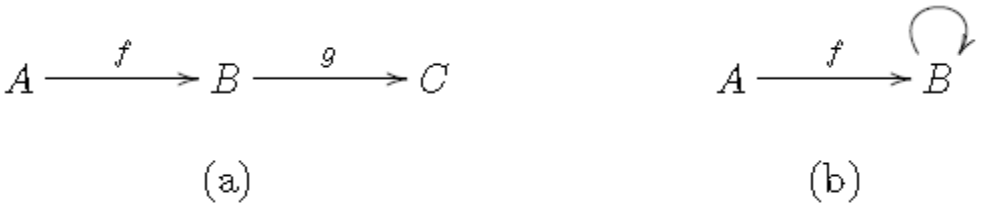
3.4.1 Diagramme (Amiguet, 1998)

Soient *I* et *G* deux graphes. Un diagramme de forme *I* dans *G* est un morphisme de graphe $D : I \rightarrow G$. Le graphe *I* est appelé le graphe de forme du diagramme *D*.

Exemple (Une subtilité dans les diagrammes)

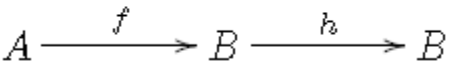
Voici un exemple illustrant quelques subtilités dans le concept de diagramme. Soit *G* un graphe avec les nœuds *A*, *B* et *C* (et peut-être d'autres nœuds) et les arcs

$f : A \rightarrow B$, $g : B \rightarrow C$ et $h : B \rightarrow B$. Considérons ces deux graphes :

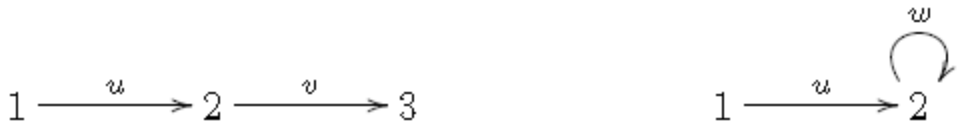


(a) (b)

Ces deux graphes sont de formes différentes (le mot forme est utilisé de manière informelle). Mais le graphe suivant a la même forme que le graphe (a) bien que ce soit le même que le graphe (b):



De manière à saisir la différence illustrée ici entre un graphe et un diagramme, on considère ces deux graphes de forme:



$\mathcal{I}$ $\mathcal{J}$

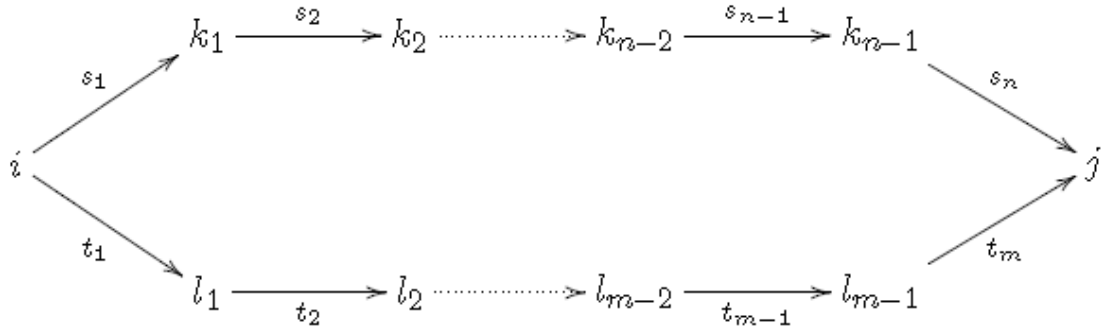
(Par convention, les nombres représentent les nœuds des graphes de formes). Ainsi, le graphe (a) est vu comme le diagramme $D : I \rightarrow$ G avec

D(1) = A, D(2) = B,D(3) = C,D(u) = f et D(v) = g; tandis que le graphe (b) correspond au diagramme *E : J → G* avec *E(1) = A,E(2) = B,E(u) = f* et E(w) = h. De plus, le dernier graphe peut être vu comme le diagramme *D* hormis *v* qui est envoyé sur h et 3 qui est envoyé sur *B*.

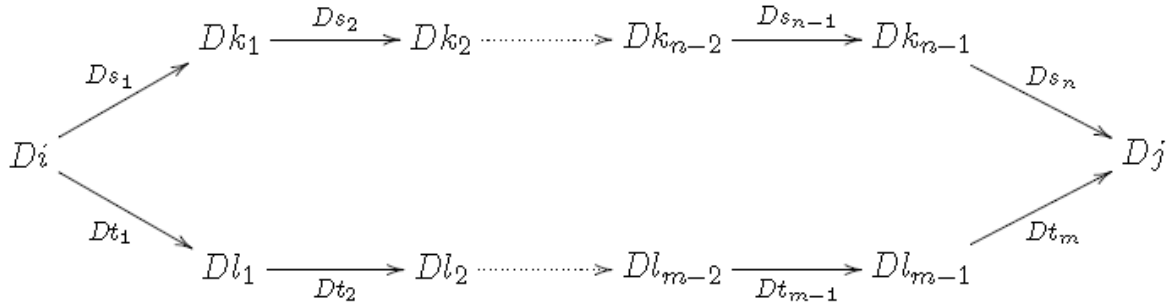
Quand le graphe de destination d'un diagramme est le graphe sous-jacent d'une catégorie, il en résulte plusieurs possibilités. En particulier, le concept de diagramme commutatif qui est le moyen d'expression des équations dans les catégories.

Diagramme commutatif

Soit une catégorie *C* et un diagramme $D : I \rightarrow C$. Le diagramme *D* est commutatif (ou commute) si pour tout nœud *i* et *j* de *I* et tous les chemins



de *i* à *j* dans *I*, les deux chemins :

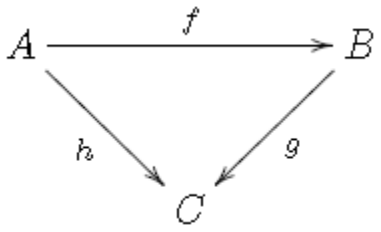


sont égaux dans la catégorie *C*. Cela veut dire que :

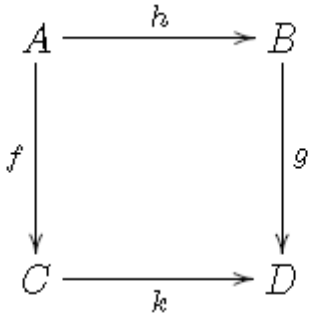
$Ds_n \circ Ds_{n-1} \circ \ldots \circ Ds_1 = Dt_m \circ Dt_{m-1} \circ \ldots \circ Dt_1$

Le triangle : la base de la commutativité

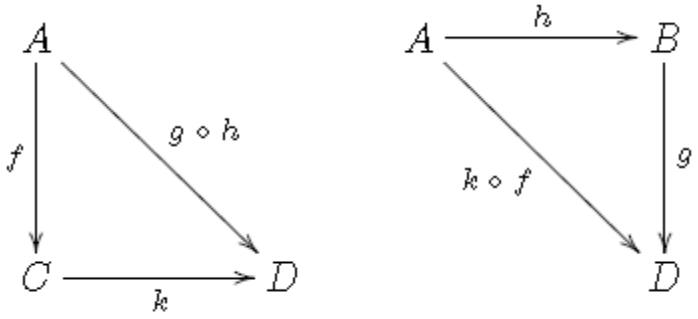
On dit que le triangle suivant commute si et seulement si $h = g \circ f$.



On dit que le triangle est à la base de tout diagramme commutatif (sauf s'il implique un chemin vide) car tout diagramme commutatif peut être remplacé par un ensemble de triangles commutatifs. Un exemple illustre cette proposition. Le diagramme



commute si et seulement si l'un des deux diagrammes suivants commute :



3.4.2 Produit de deux objets dans une catégorie

En commençant par la définition du produit cartésien, il est aisé de comprendre le produit dans les catégories. En effet, le produit dans la catégorie des ensembles est le produit cartésien que nous avons déjà mentionné.

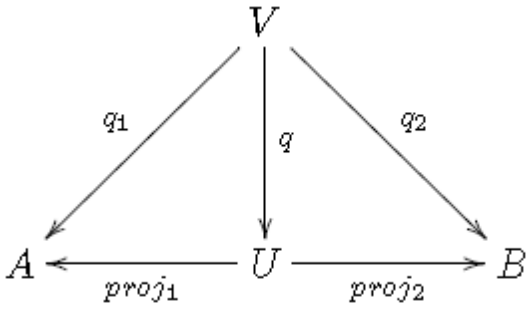
Définition du Produit cartésien

Si *S* et *T* sont deux ensembles, le produit cartésien *S* x *T* est l'ensemble des couples ordonnés avec une première coordonnée dans *S* et une seconde coordonnée dans *T*, autrement dit : $S \times T = \{(s, t) | s \in S \text{ et } t \in T\}$. Pour extraire un élément d'un produit, on utilise des projections $proj1 : S \times T \rightarrow S$ et $proj2 : S \times T \rightarrow T$.

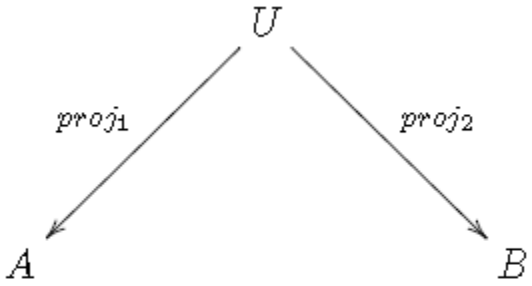
A présent, voici une définition du produit de deux objets dans une catégorie.

Soient *A* et *B* deux objets d'une catégorie *C*. Le produit de *A* et *B* dans *C* est un objet *U* muni de deux morphismes *proj1 : U →A* et *proj2 : U → B* qui satisfait la condition suivante : pour tout objet *V* et les morphismes *q1 : V → A* et *q2 : V → B,* il existe un unique morphisme

q : V → U tel que ce diagramme commute :



Cette spécification crée un objet *U* accompagné de deux morphismes *proj1* et *proj2* dans la catégorie *C*. Le diagramme correspondant



est appelé le diagramme de produit ou cône de produit et les morphismes *proj1* et *proj2* des projections. La paire ordonné *(A, B)* est appelée la base du cône.

3.4.3 La somme de deux objets dans une catégorie

Dans la théorie des catégories, la somme correspond au produit dans la catégorie duale. Voyons la définition d'une telle catégorie, cela permet de mieux comprendre les similarités entre produit et somme.

Définition (Catégorie duale)

Soit *C* une catégorie, on peut construire une autre catégorie, notée C^{op}, en inversant le sens des morphismes. La catégorie duale (ou opposée) C^{op} de *C* est définie ainsi :

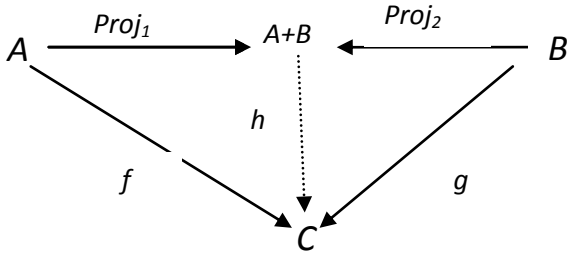
D-1 Les objets et les morphismes de C^{op} sont les objets et les morphismes de *C*.

D-2 Si $f : A \rightarrow B$ est dans *C* alors $f : B \rightarrow A$ est dans C^{op}.

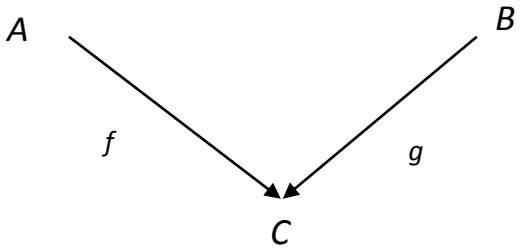
D-3 Si $h = g \circ f$ dans *C* alors $h = f \circ g$ dans C^{op}.

Définition de la somme

Dans une catégorie *C*, la somme (appelée aussi coproduit) de deux objets *A* et *B* est un objet *A+B* ensemble avec deux projections $Proj1 : A \rightarrow A+B$ et $proj2 : B \rightarrow A+B$ tel que pour tout objet *c* de *C* et tout morphisme $f : A \rightarrow C$, $g : B \rightarrow C$, il existe un *unique* morphisme $h : A+B \rightarrow C$ tel que le diagramme suivant commute :

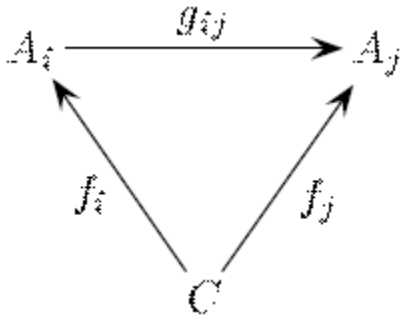


Le diagramme suivant est appelé co-cône de la somme :

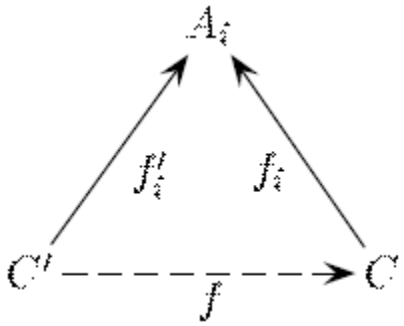


3.4.4 Limite et colimite, catégories complètes et co-complètes

Soit un diagramme constitué par un ensemble d'objets A_1, A_2, ..., A_m et de flèches $g_{ij} : A_i \rightarrow A_j$ entre deux objets de ce diagramme. Un *cône* pour un tel diagramme est un objet *C* de la catégorie avec une flèche f_i pour chaque objet A_i du diagramme de telle sorte que :



commute. Une *limite* pour ce diagramme, c'est un cône de sommet *C* avec la propriété que pour tout autre cône de sommet *C'* il y a exactement une flèche (ou morphisme) $f : C' \rightarrow C$ telle que le diagramme



Commute pour toute flèche f_i vers chaque élément A_i du diagramme.

La notion de colimite se définit de façon duale à celle de limite. Il suffit de reprendre les diagrammes précédents en considérant maintenant que tous les objets du diagramme ont tous une flèche ayant pour codomaine le sommet du diagramme appelé un co-cône.

Une catégorie *C* est complète si tout diagramme de *C* admet une limite. Elle est co-complète si tout diagramme de *C* admet une co-limite. Une catégorie est bi-complète si elle est à la fois complète et co-complète. Un diagramme est fini s'il comporte un nombre fini d'objets et un nombre fini de morphismes entre eux. Une catégorie qui admet une limite pour tous ses diagrammes finis est dite finiment complète. De la même façon on définit une catégorie qui est finiment co-complète.

4. La théorie des catégories en dehors des mathématiques

L'un des premiers apports de la théorie des catégories est de permettre de formuler les structures à partir desquelles sont effectués des calculs. En informatique, ces structures sont habituellement définies à partir de langage de programmation ou de spécification. Par exemple, les *spécifications algebriques* (Ehrig & Mahr, 1985; Bert & *al.*, 1995) peuvent

être généralisées par les catégories qui apportent une nouvelle façon de concevoir les langages de spécification (Goguen, 1991).

D'autres langages informatiques sont liés à des catégories particulières qui ont le même pouvoir d'expression. Par exemple, le λ-calcul est généralisable par une catégorie dite *cartésienne et fermée* (Barr et Wells, 1990). Les bases de données ont aussi été formalisées avec les catégories (Dampney *et al.*, 1992; Rosebrugh et Wood, 1992). Dans d'autres domaines, comme le diagnostic, les catégories apportent une formulation élégante (Li et Pereira, 1995) du problème. Le diagnostic consiste à déterminer les causes d'un comportement anormal d'un système. Les états d'un système sont considérés comme les *objets* d'une catégorie et les changements d'état sont considérés comme des *morphismes* entre objets. La théorie des catégories apporte aussi une solution au problème de l'intégration de l'information à partir de sources de données hétérogènes et distribuées (Healy & Caudell, 2006) (Geldart & Song, 2009) (Ling Hu, 2010).

La Théorie des catégories a aussi été utilisée en biologie pour l'étude des cellules vivantes (Rosen, 1959) et comme outil de modélisation du développement cognitif (Halford G.S. &Wilson W. H., 1980). Les travaux de Piaget en psychologie ont aussi été repris dans le cadre de la théorie des catégories (Henriques, 1990) (Boldini, 1994).

5. Qu'en est-il du langage

Le langage est structure: officiellement cette idée dans la linguistique moderne prend origine, comme nous l'avons vu chez le linguiste Ferdinand de Saussure et constitue donc l'idée fondatrice du structuralisme linguistique. C'est aussi l'idée qui transparait dans la terminologie et l'approche de la tradition grammaticale arabe. En effet, les termes, mithāl, bāb, bina´, mawdi', suggèrent une vision structuraliste de la langue et que le langage est un système :

> "For the Arabic grammarians speech is a system in equilibrium, whether it is the result of a revelation from Allah, or of an agreement between men (istilah). Each and every letter, word, category, has its own place and its own rights. Every phenomenon can and must be explained, and every deviation from the original form (´aṣl) is the result of a well-defined cause ('illa), and occurs according to well-defined rules. It is the task of the grammarian to determine those rules, and thus to codify the inner system of speech, in other words, to unravel the secrets of the Arabic language (asrar al-'arabiyya)" (Versteegh, 1978)

Le qiyās ne fait que découvrir, déterminer et spécifier les différentes structures de la langue ainsi que la position de chaque élément dans le système. Cependant, comme nous l'avons vu au chapitre 4, l'approche du qiyās est une approche opératoire qui comparée à l'approche occidentale met plus l'accent sur les opérations/transformations des éléments de la langue plutôt que sur la structure interne de ces objets. Dans la recherche des naẓā´ir (éléments similaires) pour la classification des objets linguistiques, les objets et les classes ne sont pas pensés sur la base des relations d'appartenance et d'inclusion comme c'est le cas dans l'approche Harrissienne. Les objets linguistiques ne sont jamais regardés en eux même mais seulement à travers la relation ´aṣl ↔ far' ***autrement dit à travers les transformations admissibles permettant d'aller de l'un à l'autre.***

Le structuralisme en posant la langue comme un système de structures voulait certes étudier les objets de la langue dans leurs inter- relations, mais il a échoué car il est resté dans la logique de l'essence et de la relation d'inclusion, ce qui revient à considérer la langue comme un système de structures figées. Nous pensons que l'approche de la TGA a réussi là où le structuralisme linguistique du 20ème siècle a échoué car cette approche a appréhendé le langage, pour emprunter une expression de (Benoist, 2007) pour la phénoménologie, comme un système de structures en mouvement. (Patras, 2005) et (Benoist, 2007) discutent de l'échec du structuralisme mathématique de tradition Bourbakiste (ensembliste), nous pensons que le structuralisme linguistique a échoué dans le même ordre idée en appliquant les mêmes arguments mathématiques au structuralisme linguistique. En fait, Piaget a été le premier à utiliser les structures mathématiques en sciences humaines : il a développé sa théorie du développement des stades de l'intelligence chez l'enfant en la formalisant dans le cadre de la théorie des groupes où l'accent est mis sur les opérations et la réversibilité de ces dernières dans les processus d'apprentissage (Piaget, 1938). L'idée de s'aider des mathématiques pour étudier le langage apparait aussi très tôt dans l'œuvre de Zellig Harris (Harris, 1946). Néanmoins, Harris n'a pas pu appréhender dans un même formalisme les relations syntagmatiques et paradigmatiques des unités de la langue, son analyse distributionnelle permet uniquement de déterminer les relations paradigmatiques en classes d'équivalence quant à la syntaxe, il en aborde les transformations séparément dans la grammaire « opérateur opérandes » où certaines unités de la langue agissent à la manière d'opérateurs (comme en mathématiques) sur d'autres unités (Harris, 1990). Ceci aussi rappelle la formule RT_i de la TGA. Cette forme de grammaire applicative a évolué comme par exemple dans les Grammaires applicatives et cognitives de (Desclés, 1990) où l'auteur met en œuvre des génotypes et des phénotypes

des structures linguistiques sur lesquels ont lieu les opérations (applicatives) et qui rappellent de loin, les notions de ´aṣl et far' de la TGA. Dans une autre approche et qui prend en considération la réversibilité des transformations linguistiques, (Dymetman, 1998) a proposé de modéliser la grammaire d'une langue naturelle par un groupe mathématique en se basant sur la réversibilité des transformations au niveau syntaxique : ces travaux n'ont pas eu de suite.

Pour la langue arabe, Louis Massignon dans (Massignon, 1954) a été le premier à faire l'analogie entre les opérations réversibles que subit un objet linguistique invariant et un objet mathématique ou l'objet invariant et les opérations constituent une structure de groupe.

> A travers toutes ses variétés, la langue arabe maintient l'usage de types de structures, de « groupes » au sens où le « groupe » signifie: une famille de modifications des termes, telles qu'elles maintiennent intacte, dans tous les cas, l'exactitude de leur disposition initiale. (Massignon, 1954)

Des linguistes contemporains comme (Carter, 1973) (Verteegh, 1978), (Guillaume, 1986) (Hadjsalah, 1979) ont mis l'accent sur la nature opératoire de la TGA. S'inspirant des travaux de Piaget, le linguiste Hadjsalah (Hadjsalah, 1979) a aussi fait l'analogie avec la structure mathématique de groupe en faisant remarquer qu'à tous les niveaux de la description linguistique, nous obtenons des structures fermées vu la réversibilité des transformations linguistiques.

Quand on parle de la théorie des groupes, on parle de structuralisme de tradition Bourbakiste. Mais aujourd'hui le structuralisme est 'passé de mode' (Patras, 2003). L'évolution de la pensée scientifique et de la pensée mathématique en particulier avec le développement de la théorie des catégories a adressé de sévères critiques au structuralisme 'figé' de style Bourbakiste. Nous résumons ces critiques dans ce qui suit pour mieux comprendre l'intérêt et le rôle de la théorie des catégories pour notre propos.

On peut considérer un *groupe* comme un ensemble d'éléments muni d'une opération qui permet de combiner deux quelconques de ses éléments sans sortir de l'ensemble, le résultat de l'opération étant donc un élément de l'ensemble. Cette opération, qui prend aussi le nom de « loi de composition interne » et qui, doit respecter certaines propriétés formelles, peut être considérée à juste titre comme « *l'idée mère de la notion de structure* » (Vuillemin, 1962).

Néanmoins, la pensée structurale en mathématiques commence véritablement dès le moment où l'on constate que *« ce qui joue le rôle primordial dans une théorie, ce sont les relations entre les objets mathématiques qui y figurent, plutôt que la nature de ces objets, et que dans deux théories très différentes, il se peut que les relations s'expriment de la même manière; le système de ces relations et de leurs conséquences est une même structure* ***"sous-jacente"*** *aux deux théories »* (Dieudonné, 1987).

Si on applique la théorie des groupes aux structures linguistiques, le problème se réduit à choisir une structure de groupe qui soit pertinente linguistiquement et à en étudier les *invariants* au sens de Klein : « *une fois choisi un groupe de transformations sur un ensemble donné, des propriétés « ne sont pas altérées par les transformations du groupe »* (Klein, 1974)

Les groupes opèrent directement sur les objets et le processus d'abstraction s'arrête au niveau d'une équivalence (i.e. *isomorphisme*) entre objets par rapport à une opération de groupe. Dans un deuxième degré d'abstraction, les opérations ne sont pas considérées en tant qu'objets mais en tant que « substrat » d'une collection de transformations possibles de ces objets. C'est donc en comparant les transformations (et les transformations entre transformations) qu'on peut arriver à établir des critères d'équivalence entre les objets de départ. C'est ce qui ressort des exemples que nous avons présenté au chapitre 4 et qu'on ne peut tout à fait décrire dans le cadre de la théorie des groupes (pour la raison citée plus haut). Il ne suffit pas de mettre l'accent sur la réversibilité des transformations.

En adaptant à la linguistique une réflexion que Patras fait par rapport aux mathématiques, «*les objets existent rarement isolés. Ils prennent tout leur sens lorsqu'ils sont insérés dans un contexte précis (...) et, réciproquement, c'est cette insertion qui leur confère un statut mathématique* » (Patras, 2005). C'est ce qui ressort précisément de la théorie du mawḍi' dans la TGA.

Dans la genèse de la théorie des catégories, il y a eu une « collision » entre l'approche topologique d'Eilenberg et l'approche algébrique de Maclane à partir du concept de «naturalité » des transformations entre des espaces mathématiques. En particulier, l'étude de la propriété de « naturalité » d'un *isomorphisme* en théorie des groupes (Eilenberg et Maclane, 1942) a été le point de départ pour établir une théorie générale de l'équivalence «naturelle » entre structures algébriques, ce qui a conduit à la définition de *foncteur* et à la formalisation du concept fondamental de *catégorie* (Eilenberg et Maclane, 1945). La «naturalité » des transformations entre structures algébriques, en particulier des isomorphismes, concerne la possibilité de définir de telles relations **indépendamment** de

la *présentation* particulière d'une structure. Il s'agit donc d'une notion technique, de même qu'un autre concept à partir duquel la théorie des catégories s'est constitué : le concept d'«universalité ».

Nous avons déjà introduit la notion de propriété universelle en TC. La définition du «produit cartésien » par exemple, indépendamment de la nature des ensembles *A* et *B*, afin de mettre en évidence son caractère d'« universalité », a été historiquement l'un des problèmes qui ont mis en évidence l'utilité d'une approche catégorielle. La notion de produit peut, en effet, être établie *sans aucune référence explicite aux éléments appartenant aux ensembles mais simplement en imposant une propriété de cohérence formelle dans un diagramme de flèches.* Par ailleurs, les concepts de *limite* et de *colimite* d'un diagramme de flèches mettent en évidence le rôle « structural » des flèches dans le processus de description d'un concept mathématique. Comme René Lavendhomme l'affirme, la théorie des catégories permet d'obtenir une « *définition* ***extensionnelle*** *et opératoire d'un concept mathématique qui est bien rendu d'une part par l'ensemble de tous les objets mathématiques répondant à ce concept et, d'autre part, par l'ensemble des transformations compatibles avec ce concept* » (Lavendhomme 2001). Nous pensons que cette affirmation s'applique aux structures linguistiques et qu'elle correspond à la démarche de la TGA.

En ce qui concerne l'application de la théorie des catégories dans le domaine de l'étude des langues naturelles, bien que Lambek ait fait le lien entre la logique et la théorie des catégories depuis plus d'une quarantaine d'années, peu de travaux ont vu le jour dans cette direction. Des travaux récents étudient la CT comme sémantique des langues naturelles, en particulier, la CT est proposée comme sémantique des grammaires catégorielles (Preller, 2005). L'originalité de notre approche consiste en ce que nous faisons l'analogie entre l'approche catégoriale et l'analyse distributionnelle de la TGA et nous proposons la CT comme formalisme de description et de construction des structures linguistiques « from scratch », en ce sens que la CT (ou un modèle de celle-ci) serait **la** grammaire de la langue. Nous allons essayer dans ce qui suit d'interpréter les résultats que nous avons obtenus au chapitre 5 en termes catégoriels.

6. Les structures de la langue arabe comme catégories

Goldblatt (Goldblatt, 1984) donne une définition du processus de construction des catégories dont l'intuition est très proche de ce que fait le qiyās, d'ailleurs, il utilise même

le terme « measure » ; l'extrait suivant dans lequel l'auteur désigne le processus comme 'a pathology of abstraction' aidera sans doute le lecteur à mieux saisir l'analogie entre les deux processus qiyās et élaboration de catégories (sur un univers du discours) :

> "The process of identifying the notion of a category is one of the basic modi operandi of (pure mathematics) it is called *abstraction*. It begins with the recognition through experience and examination of specific situation that certain phenomena *occur repeatedly*, that there are a number of *formal analogies* in the behavior of different entities. Then comes the actual process of abstraction, wherein these common features are presented in isolation …
>
> Having obtained our abstract concept, we then develop its general theory and seek for further instances of it. These instances are called *examples* of the concept or *models* of the axioms that define the concept. Any statement that belongs to the general theory of the concept (i.e is derivable from the axioms) will hold true in all models. The search for new models is a process of *specialization*, the reverse of *abstraction*. Progress in understanding comes as much from the recognition that a particular new structure is an instance of a more general phenomenon as from the recognition that several different structures have a common core. Our knowledge of (mathematical) reality advances through movement from the particular to the general and back again.
>
> An important aspect of specialization concerns the so-called *representation* theorem. These are propositions to the effect that any model of the axioms for a certain abstract structure must be (equivalent to) one of a particular list of concrete models. They ***"measure"*** the extent to which the original motivating examples encompass the possible models of the general notion"

Dans une vision catégoricienne, les éléments appartenant à un même paradigme (classe distributionnelle) comme كتب، ذهب، أكل، ضرب، حضر، sont considérés du point de vue de leur similarité structurelle et non en tant qu'éléments per se. Chaque paradigme correspond à un mithāl, la mise en correspondance de ces paradigmes fait émerger des relations qui sont aussi récurrentes entre les différents paradigmes, ce qui nous fait dire qu'ils partagent encore la même structure à un autre niveau d'abstraction, cette structure générale appelée bāb représente « la classe générale » ou la **catégorie** dans laquelle les éléments peuvent se comporter (morphismes) sans sortir de la structure considérée.

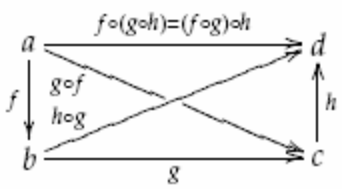
Le rapprochement entre la démarche distributionnelle et la théorie mathématique des catégories (TC) peut sembler assez naturel si l'on considère que comme la méthode distributionnelle chez Harris et TGA, la théorie des catégories part de l'observation des

régularités observées dans le comportement des objets pour construire les catégories. L'originalité et la puissance de la théorie des catégories résident dans le fait que l'objet d'étude n'est pas considéré en lui-même comme objet, mais à travers les opérations (transformations) susceptibles d'agir sur cet objet. La structure des objets est abstraite à travers la notion de morphisme ou flèche. L'idée que nous avançons donc dans ce travail et que de même que pour les objets mathématiques, il serait peut être intéressant d'appréhender les objets linguistiques au travers des invariants et des opérations qu'ils subissent au sens de la théorie des catégories. En réalité André Lentin dans (Lentin, 1999) avait déjà fait allusion à la possibilité d'utilisation de la théorie des catégories en description linguistique mais un tel travail n'a encore jamais été entrepris. En ce qui nous concerne, nous pensons qu'il serait en effet, intéressant d'utiliser le langage de la CT pour la description et la formalisation des structures linguistiques. Nous commençons par appliquer cette idée à la morphologie flexionnelle de la langue arabe où nous pensons que l'analyse distributionnelle que nous avons présentée montre naturellement et intuitivement le passage à une description catégorique. Pour cela, nous allons considérer les morphèmes comme des objets et les opérations d'ajouts d'affixes à droite et à gauche des morphèmes comme des morphismes. Considérons maintenant la structure L que nous avons obtenue au chapitre 5, L est une catégorie ayant pour objets les morphèmes et pour flèches les opérations d'ajouts d'affixes à droite et à gauche de *m*. chaque flèche *f* peut être définie par un domaine et un codomaine. Exemples : كتب→الكتب. Par ailleurs, nous pouvons aussi définir une opération de composition de flèches dans cette catégorie. *f :* كتب → يكتب,

g :يكتب → يكتبون, *gof* : كتب → يكتبون. Cette opération possède un élément zéro (l'identité) qui consiste à laisser l'élément tel quel sans modification. Cette opération est aussi associative : Si on a

f : كتب → يكتب, *g* :يكتب → يكتبون, *h :* يكتبون → يكتبونه alors *h o (gof) = (hof) og* que nous pouvons illustrer par le diagramme commutatif suivant où *a*=كتب , *b*=يكتب , *c*=يكتبون ,

d=يكتبونه .



Formellement, une catégorie est définie comme la donnée d'un 5-tuple ordonné de la forme <O, M, dom, cod, id>, O sont les objets, M, sont les morphismes, dom correspond à la fonction domaine qui assigne à chaque morphisme un objet source ; cod correspond à la fonction codomaine qui assigne à chaque morphisme un objet cible, id pour identité.

Nous proposons donc de modéliser la lexie comme une catégorie au sens de la CT sur laquelle sont vérifiés les axiomes sur l'identité et l'associativité (exemple précédent):

- O : l'ensemble des paradigmes lexicaux autrement dit l'ensemble des mots sur l'alphabet de la langue arabe
- M : l'ensemble des « mithāl » relatifs aux opérations d'ajouts et de suppression d'affixes et permettant le passage d'une lexie à l'autre.
- dom : correspond au ´aṣl c'est-à-dire l'élément originel avant une transformation
- cod : correspond au far‘ c'est-à-dire au résultat d'une transformation (par ajout ou suppression d'affixes).
- id : correspond à l'opération zéro c'est-à-dire absence d'ajout ou de suppression d'affixes.

Les structures obtenues, étiquetées et transcrites au chapitre 5 seront codifiées dans un langage catégorique. Le travail dans son ensemble peut âtre alors vu comme un « compilateur » ayant les textes en langue arabe en entrée et en sortie des structures catégorielles (description en termes d'objets et morphismes). Ces catégories ainsi codifiées seront utilisées dans les développements futurs de ce travail.

Nous pouvons interpréter de la même manière *S*, la structure qui représente le niveau syntaxique. Les lexies L_i composent pour former des structures *S*. La relation entre L_i et *S* est celle d'un foncteur qui prend chaque objet de *L* et lui assigne une position dans *S* (et par là une position dans le discours). Ceci n'est rien d'autre qu'un transport de structure avec préservation des morphismes, compositions et identités.

Nous reprenons ici Goguen: « *to any natural construction on structures of one species yielding structures of other species, there corresponds a functor from the category of the first specie to the category of the second"* (Goguen, 1991)

Il s'agit de trouver le foncteur qui associe à une suite de deux ou trois lexies une structure syntaxique.

Nous avons donc les deux catégories *L* et *S* modélisant les lexies simples et les structures syntaxiques simples découvertes aux chapitre 5. La relation reliant les deux structures est celle de foncteur transportant toute structure *L* dans *S* en en préservant les propriétés (identité et associativité). L'idée est que le foncteur associe à chaque structure de départ une position dans la nouvelle structure, cette position déterminera le statut de la structure de départ *L* dans la nouvelle structure *S*. En linguistique, il s'agira d'un « rôle » et de la flexion casuelle associée.

Exemple : (1) # ϕ abdullahi qaimun # qui est un objet de *S* intégrant deux objets de *L* : a= abdullahi, b= qaimun

Un foncteur F de *L* à *S* est défini comme suit :

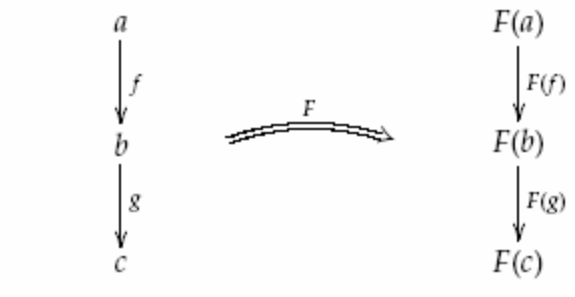
$F : L \rightarrow S$

$(a,f) \rightarrow (b,g)$

tel que pour tout objet *a* et pour toutes flèches composables *f* et *g* dans la catégorie *L:*

$F(\mathrm{id}_a) = \mathrm{id}_{F(a)}$

$F(gof) = F(g)\ o\ F(f)$.



L'idée à retenir est que tout ce qui a été dit pour *L* est préservé par le foncteur.

Une lecture catégoricienne du qiyās

En CT, les structures munies d'une notion de morphisme donnent naissance à une catégorie dont les objets sont toutes les instances (ou modèles) de cette structure et dont les flèches sont tous les morphismes entre ces instances. La structure de catégorie elle-même

n'échappe pas à cette règle et il ya donc une catégorie dont les objets sont les catégories et les flèches sont les foncteurs, bien entendu les collections intervenant dans ces diverses catégories ont des niveaux appropriés pour éviter les paradoxes.

En TGA, les structures munies d'une notion de transformation que nous appellerons, dorénavant morphisme donne lieu à une structure catégorielle qui est le « bāb » dont nous avons dit qu'il correspond à la notion de classe sans se confondre tout à fait avec cette notion car la classe issue de la logique classique et de la théorie des ensembles est fondée avant tout sur la notion d'appartenance et la relation d'inclusion et nous avons vu au chapitre 4 comment le qiyās dépasse cette notion, nous pensons que la notion de « bāb » dans la TGA trouve tout à fait son équivalent dans la notion de catégorie au sens de la CT. Les objets de notre bāb/catégorie sont toutes les instances (ou modèles) de cette structure : ce sont les mithāls et dont les flèches sont les opérations permettant de passer d'un mithāl à un autre. Par exemple pour la catégorie lexie, les mithāls déterminent les différents paradigmes lexicaux et correspondent aux opérations de passage d'un paradigme à l'autre par ajout ou suppression d'affixes. Les exemples que nous avons présentés pour le niveau des structures syntaxiques de base que nous avons découvert au chapitre 5 constituent aussi des paradigmes (modèles/mithāls) pour S.

« … the key lies, not in the particular nature of objects or arrows but in the way the arrows behave" (Goldblatt, 1984).

L'idée fondatrice du qiyās est comme nous l'avons vu d'assigner à chaque objet linguistique une position (mawḍi‘’) dans un (mithāl) en examinant ces objets non pas à travers leurs propriétés internes mais en les appréhendant dans la dynamique des transformations ´aṣl (أصل)↔ far‘(فرع). Nous pouvons donc dire que cette idée coïncide avec l'intuition derrière la CT et qui consiste à abstraire les propriétés des objets à travers les propriétés des morphismes. D'autre part comme dans la théorie des catégories, les transformations sont hiérarchisées dans le qiyās. En effet, comme nous l'avons vu au chapitre 4, on s'intéresse d'abord aux correspondances structurelles (morphismes) comme pour la lexie, ensuite au bina (foncteurs) comme pour le passage du niveau de la lexie au niveau de la syntaxe et enfin, aux correspondances entres foncteurs modélisées en CT par la notion de transformation naturelle comme pour l'exemple du diminutif et pluriel interne.

En termes de construction universelle, la notion de colimite en CT, permet notamment l'intégration de structures d'un niveau donné dans une structure d'un niveau supérieur (aljami‘). Toute structure S peut alors être aussi vue comme une colimite des structures qu'elle intègre. Néanmoins dans ce travail, nous n'allons pas aborder les constructions universelles qu'on pourrait construire en langage catégorique dans TGA, car il nous faudra inclure plus de sémantique grammaticale et les phénomènes linguistiques sont nombreux dont chacun pourrait faire l'objet d'une thèse.

7. Conclusion

La classification des éléments linguistiques dans la démarche de Sibawayh est basée sur deux critères fondamentaux le´aṣl et le far‘. Sur la base de ces deux critères, les linguistes arabes ont établi les notions de mithāl (schème, pattern) et de "bāb" *non pas pour examiner les objets linguistiques en eux mêmes mais pour appréhender et expliquer leur structure générale.* On ne saurait confondre cette classification avec la notion de classification classique fondée sur l'inclusion ensembliste, une telle vision serait réductrice. Nous défendons l'idée que cette démarche correspond à la détermination par universalité (on dit aussi participation universelle) de la théorie des catégories : les exemples : *كتب، ذهب، أكل، ضرب، حضر* **n'appartiennent pas** à la classe du verbe *فعل* mais chacun de ces éléments participe à une structure par les opérations dont il est susceptible d'être l'objet.

Dans l'inférence du qiyās, c'est le rapport d'équivalence comme relation logique fondamentale qui assure la rigueur du raisonnement. Cette équivalence n'est pas une équivalence de co-inclusion mais une équivalence généralisée, opératoire et *extensive*, et qui permet le passage d'un élément relation (implication) à un autre élément sans référence à la classe logique qui pourrait éventuellement les contenir. Aussi, la totalité structurée qui résulte de cette construction ne peut-elle se confondre avec une telle classe puisque les éléments qui la composent n'y sont pas emboîtés mais reliés directement entre eux, ce qui permet de passer de l'un à l'autre par l'intermédiaire de cette relation sans le détour du tout à la partie. Chacun d'eux ne constituant ainsi qu'un cas particulier de la structure qui l'intègre. Ceci est précisément comme nous l'avons vu à la base de la théorie mathématique des catégories.

Dans la théorie des catégories, toute structure est équipée d'une notion de construction ou transformation "acceptable», à savoir un morphisme qui préserve la structure à travers la notion d'invariant.

Ce qui est intéressant aussi dans la théorie des catégories c'est qu'une fois qu'un type de

structure a été défini, il devient rapidement impératif de déterminer comment de nouvelles structures peuvent être construites et comment une structure donnée peut être décomposée. C'est ce que nous avons vu à l'œuvre dans le qiyās, la construction de structures plus abstraites et intégrantes, le jamiʿ à travers des ensembles de transformation qui préservent le ʿaṣl qui représente l'invariant. Nous avons donc proposé dans ce chapitre une lecture de la TGA dans le cadre de la théorie des catégories, nous avons présenté les motivations pour une telle lecture et nous avons proposé une interprétation des structures obtenues (L et S) par notre approche distributionnelle de découverte en termes de catégories.

Conclusion générale

Nous avons présenté dans ce travail une approche non supervisée pour la découverte automatique des structures de la langue arabe à partir d'un corpus électronique brut. Notre approche est basée sur la simulation algorithmique de l'analyse distributionnelle dans la Tradition Grammaticale Arabe et travaille uniquement sur la forme du texte sans référence au sens. Les résultats que nous avons obtenus confirment la réalité linguistique en ce qui concerne les structures de base de la langue.

Par ailleurs, en observant le comportement des paradigmes découverts dans le corpus dans des énoncés plus larges, nous pourrions regrouper certains de ces paradigmes en ajoutant plus de connaissances, ceci constitue une suite de ce travail.

Nous avons alors proposé une interprétation des structures obtenues ainsi que des concepts de la TGA dans le cadre de la théorie mathématique des catégories dans le but d'arriver à un modèle formel de la langue à partir d'une approche ascendante et à partir du corpus observable de la langue. Nous pensons que ce qui manquait au distributionalisme contemporain pour être une théorie rendant compte du système de la langue est justement de considérer dans les distributions des objets, les opérations aboutissant aux changements de forme (les différents environnements correspondants aux différentes formes) au lieu de se concentrer sur les objets dans leurs différents états. Nous pensons que modéliser les structures linguistiques en termes d'objets et de morphismes peut apporter du nouveau dans le domaine comme ça été le cas en biologie par exemple. En effet étudier les objets linguistiques par les propriétés universelles est réellement intéressant. L'idée est de *déterminer* les objets linguistiques par leur ***participation*** à une structure linguistique et non par leur appartenance à cette structure.

En fait, en ce sens, le problème des mathématiciens et des linguistes est quels sont les objets d'analyse à appréhender: les objets isolés ou plutôt les objets à l'intérieur d'une structure. Ceci est exprimé par Resnik cité dans (Stefanik, 1996) et il est remarquable de constater que c'est exactement l'intuition derrière la théorie du mawḍi‘:

"In mathematics, I claim, we do not have objects with an "internal» composition arranged in structures, we have only structures. The objects of mathematics, that is, the entities which our mathematical constants and quantifiers denote are structureless points or positions in structures. As positions in structures, they have no identity outside of a structure."

Pour Resnik, comme pour Sibawayh, les objets (mathématiques), qu'ils soient des nombres, des ensembles, fonctions, … sont des entités qui occurent dans des structures mathématiques. Ces objets occupent des positions à l'intérieur de ces structures, leur identité étant déterminée seulement par leur relation avec les autres positions de la structure, c'est cela la participation par universalité.

Le structuralisme en mathématique comme en linguistique, s'il propose d'organiser le corpus, ne dit rien sur les mécanismes d'apprentissage et de création des savoirs. La science dont il rend compte est une science figée (Patras, 2005). Les idées de Claude Chevally, membre du groupe Bourbaki (décrites par sa fille) malgré qu'elles ne prétendent pas refléter une vision partagée par les autres membres du groupe décrivent bien le type de mathématiques auxquelles Bourbaki visait à aboutir :

> « Le manque de rigueur donnait à mon père l'impression d'une démonstration où l'on marcherait dans la boue, où l'on soulèverait des sortes d'immondices pour avancer. Quand on les avait écartés, on pouvait accéder à l'objet mathématique, une sorte de corps cristallisé dont l'essentiel est la structure. Quand cette structure était construite, il disait être intéressé par cet objet pour le regarder, l'admirer, presque le faire tourner mais certainement ne pas le transformer… Si l'on observe la façon dont mon père travaillait, il semble que c'est davantage cela qui comptait, cette production d'un objet qui alors devenait inerte, mort, en somme. »

Nous reprenons l'argument de Patras, la théorie des catégories codifie et organise au sein d'un symbolisme, ces deux notions fondamentales pour la pensée (mathématique) linguistique que sont celles d'objet et de relation.

Beaucoup ont aussi comparé le qiyās dans la TGA au raisonnement par analogie. Nous avons vu que dans le qiyās ce sont les opérations de transformations qui permettent de comparer les bābs et de passer de l'un à l'autre tout comme les foncteurs permettent de comparer les catégories et de passer de l'une à l'autre. Ce qui différencie la fonctorialité de l'analogie est que la fonctorialité considère en plus des objets eux-mêmes, les applications et relations entre ces objets. Les morphismes exercent un contrôle sur la structure de la catégorie et limitent le champ opératoire. Parce qu'il opère sur les morphismes, un foncteur se doit de mettre en correspondance l'information liée au fonctionnement machinique de la catégorie A vers la catégorie B. Le transport d'une catégorie à l'autre s'effectue non seulement entre des objets, mais aussi et surtout entre les morphismes. Par la considération

non seulement des objets, mais aussi des morphismes, la fonctorialité contraint à impliquer davantage les relations causales entre objets, ce que l'analogie de proportion délaisse.

Enfin, notre approche de découverte des structures et plus, notre proposition d'un « compilateur » fondé sur les catégories pour l'analyse d'un texte brut en langue naturelle, ici, l'arabe, est une contribution à la réfutation de l'hypothèse de l'innéité linguistique, ce que nous apportons de plus par rapport aux autres travaux, c'est la mise en lumière de la possibilité d'arriver à l'acquisition de toutes les structures de la langue dans un même cadre unifié qui est la CT.

Sur un plan pratique, les résultats obtenus peuvent être utilisés dans des applications de TAL nécessitant l'analyse de la langue au niveau de la forme comme la fouille de textes, l'extraction d'information, la RI, le résumé de textes, l'étiquetage, ….

Dans une perspective théorique (avec des retombées pratiques), nous comptons poursuivre le développement du modèle catégoriel que nous avons présenté pour la morphologie flexionnelle. Pour cela, il faut déterminer au fur et à mesure le minimum de connaissances grammaticales à apporter pour pouvoir utiliser toute la puissance du modèle catégoriel, notamment les constructions universelles, comme la notion de colimite par exemple pour l'analyse syntaxique. Dans cette perspective, nous garderons toujours à l'esprit le principe fondateur de la TGA à savoir le fait le plus simple (minimal et économique) qui décrit bien une réalité est à prendre comme prémisse ´aṣl pour l'élaboration de faits plus complexes.

Bibliographie

Aarts, J. (1990), "Corpus linguistics: An appraisal", *Computers in literary and linguistic research*, J. Hammesse & A. Zampolli (eds.), Paris/Genève, Champion-Slatkine, 13-28.

Abu- alMakarim, A. (1975), *Taqwim l-fikr n-nahwi*, Beyrouth: Dar al-taqafat.

Aliane H. (2001), "Abduction et Qiyās : vers une caractérisation de la grammaire de Sibawayh du point de vue de l'intelligence Artificielle et de la philosophie de la science". Thèse de magistère, USTHB, 2001.

Aliane H. & Aliamzighi Z. (2006a), "Une approche pour l'identification automatique des valeurs du temps et de l'aspect dans la langue arabe". *Jetala'06* Oujda, Maroc.

Aliane H & Alimazighi Z. (2006b), "Une Ontologie pour l'Indexation et la Recherche d'Information Multilingue" Conférence *Ai'06*, Maroc.

Aliane H. & Alimazighi Z. (2006c) "Towards a characterization of Sibawayhi's approach from the viewpoint of Artificial Intelligence and Philosophy of Science" *The Second International Conference on Language and Linguistics*. Cairo, Egypt.

Aliane H. & Alimazighi Z. (2010a), "Alkhalil project: Towards an ontology for Arabic linguistics" *Language Evaluation And Resources Conferences*, 2010, Malte.

Aliane H. & Alimazighi Z. (2010b), "A categorical interpretation of Sibawayhi Grammar", accepted at the conference on *"The foundations of Arab Linguistics: Sibawayh and the earliest Arabic grammatical theory"*, Cambridge university.

Aliane H. & Alimazighi Z. (2011a), "Discovering Arabic Morphemes: A Simple Arabic Grammar Tradition Based Approach" in proceedings of the *International Conference on Knowledge Discovery*, ICKD, Chengdu, China.

Aliane H. & Alimazighi Z. (2011b), **"**Discovering Arabic Structure: what can a formal analysis tell us" *Computer Applications in Intelligent Natural Language Processing* special issue of *IJCAT*, inderscience publisher. Vol40n4. Août 2011.

Aliane H. & Nehal D. (2009), "Utilisation des Grammaires d'arbres adjoints pour la modélisation des interrogatives de la langue Arabe", rapport de recherche.

Amiguet M. (1998), " Introduction à la théorie des catégories". Travail de diplôme de mathématicien, université de Neuchâtel.

Asperti A., Longo G. (1991), *Categories, Types and Structures. Category Theory for the working computer scientist*. M.I.T. Press, 1991.

Awodey S. (2003), "Structure in Mathematics an Logic: a categorical perspective". *Philosophia Mathematica* (3) vol. 4(1996) pp 209-237.

Ayoub G. & Bohas G. (1981), "Les Grammairiens Arabes, La Phrase Nominale Et Le Bon Sens", *Historiographia Linguistica*, Volume 8, Numbers 2-3, 1981 , pp. 267-284. J. Benjamin Editions.

Baalabki R. (1979), "Some aspects of harmony and hierarchy in sibawayh's grammatical analysis", *Zeitschrift fur arabish linguistic*, 2, 7-22.

Baalabki R. (1983), "The Relation between Nahw and Balaagha: a Comparative study of the methods of sibawayh an Jurjani" *Zeitschrift fur arabish linguistik* 11, 7-23.

Bahou Y., Belguith L., Aloulou C. & BenHamadou A. (2006), "Adaptation et implémentation des grammaires HPSG pour l'analyse de textes arabes non voyellés", *15ème Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle*, 25-27 janvier 2006 à Tours - France.

Barr M. & Wells C. (1990), *Category Theory for Computing Science*. Prentice- Hall.

Benoist J. (2007), "Mettre les structures en mouvements: La phénoménologie et la dynamique de l'intuition conceptuelle, sur la pertinence phénoménologique de la théorie des catégories»*L. Boi et al. (eds.), Rediscovering Phenomenology*, 339–355.© 2007 *Springer.*

Bert D., Echahed R., Jacquet P., Potet M.-L. et Reynaud J.-C. (1995), "Specification, genericite, prototypage: Aspects du langage LPG". *Technique et Science Informatiques*, 14:1097–1129.

Biemann C. (2007), "Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm", PHD Thesis, Leipzig university.

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*. Oxford University Press.

Bloomfiel L. (1934), *Language and Linguistics*, London, Georges allen and Unwin LTD.

Boldini P. (1994), "Morphismes et catégories: Une lecture formelle de Piaget", *Intellectica* 1994/2.

Bordag S.(2007), "Unsupervised and knowledge Free Morpheme Segmentation and Analysis". PHD dissertation.

Brent M.R. (1993), "from grammar to lexicon: unsupervised learning of lexical syntax" *computational linguistics journal*, vol19 issue 2, 1993.

Brent M. R. (1997), "Syntactic categorisation in early language acquisition: formalizing the role of distributional analysis". *Cognition*, *63*(2), 121–170.

Brent M. R., & Cartwright, T. A. (1997), "Distributional regularity and phonotactic constraints are useful for segmentation." In Brent (Brent, 1997), pp. 93–125.

Brill E. (1992), "A simple rule-based part-of-speech tagger". In *Third Conference on Applied Natural Language Processing*.

Brill E. (1993), "Automatic grammar induction and parsing free text: A transformation-based approach". In *Proceedings of the 31st Annual Meeting of the ACL*, pp. 259–265.

Brill E., & Marcus M. (1992a), "Automatically acquiring phrase structure using distributional analysis". In *Proceedings of DARPA workshop on speech and natural language*.

Brill E. & Marcus M. (1992b), "tagging an unfamiliar text with minimal human supervision" AAI technical report FS-92-04.

Carter M.G. (1973), "An Arab Grammarian of the Eight Century: a Contribution to the History of Linguistics". *Journal of the American Oriental Society,* Vol. 93, No. 2 (Apr. - Jun., 1973).

Carter M.G. (1980), "Sibawayh and Modern Linguistics ", in *HEL*, tome2, Fascicule 1, éléments d'histoire de la tradition grammaticale arabe.

Choukri K. (2009), "MEDAR: Mediterranean Arabic Language and Speech Technology Inventory of the HLT Products, Players, Projects, and Language Resources", In *Proceedings of the 2th International Conference on Arabic Language Resources & Tools.*

Chan E. (2008), "Structures and distributions in morphology learning", PHD thesis, university.

Charles W. & Miller G. (1989), " Context of antonymous adjectives" . *Applied psycholinguistics*, 10.

Chomsky N. (1957), *Syntactic Structures*. Mouton, The Hague.

Chomsky N. (1959), "A review of B. F. Skinner's Verbal Behavior". *Language*, *35*(1), 26–58.

Chomsky N. (1965), *Aspects of the theory of syntax*, MIT Press, Cambridge.

Chomsky N. (1969), *la linguistique cartésienne*, éditions du seuil, Paris.

Chomsky N. (1986), *Knowledge of Language : Its Nature, Origin, and Use*. Praeger.

Chomsky N. (1995), "The Minimalist Program". *Current studies in linguistics* 28. Cambridge, MA: MIT Press.

Church, K. W. & Mercer, R. L. (1993), "Introduction to the special issue on computational

linguistics using large corpora". *Computational Linguistics*, *19* (1), 1-24.

Clark A.S. (2001), "Unsupervised language acquisition : theory and practice". PHD thesis university of Sussex, 2001.

Creutz M. (2006), "Induction of the morphology of natural language: unsupervised morpheme segmentation with application to automatic speech recognition", PHD thesis, Helsinki university of technology.

Dampney C. N. G., Johnson M. S. J. et Monro G. P. (1992), " An Illustrated Mathematical Foundation for ERA". *The Unified Computation Laboratory*, pages 77–83.

Desclés J.P. (1990), *Langages applicatifs, langues naturelles et cognition*, Paris : Hermès

Déjean H. (1998), " Concepts et algorithmes pour la découverte des structures formelles des langues ", thèse de doctorat, université de Caen, 1998.

Diab M., Hacioglu K. and Jurafsky D. (2004), "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks", in *Human Language Technology Conference Proceedings of HLT-NAACL 2004*.

Ditters E. (2001), "A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic" in *proceeding of the EACL 2001 workshop on Arabic Natural Language Processing: Status and Prospects*.

Dieudonné J. (1987), *Pour l'honneur de l'esprit humain. Les mathématiques aujourd'hui,* Hachette éditions.

Dymetman M. (1998), "Group Theory and Linguistic Processing". *COLING-ACL* 1998: 348-352.

Ehrig H. et Mahr B.(1985), *Fundamentals of Algebraic Specification I.* Springer- Verlag.

Eilenberg S. & Maclane S. (1945), "general theory of natural equivalences" *Transactions of the American Society.* 58: 231-294.

Eilenberg S. & Maclane S. (1942), "Natural Isomorphisms in group theory , *Proceedings of the National Academy of Sciences*", USA, pp. 537-543, 1942.

Farghaly A. & Shaalan K. (2009), "Arabic Natural Language Processing: Challenges and Solutions", in *ACM Transactions on Asian Language Information Processing*, Vol. 8, No. 4, Article 14, Pub. date: December 2009.

Fillmore Charles J . (1992), "Corpus linguistics' or 'Computer-aided armchair Linguistics». In : *Directions in Corpus Linguistics . Proceedings of Nobel Symposium*, 4-8 August 1991. Ed . by Jan Svartvik. Berlin, New York: Mouton de Gruyter. 35 -60.

Firth J.C (1957), "A synopsysis of linguistic theory". Palmer, F.R. (eds) Selected papers of J.C. Firth Harlow: Longman.

Geldart J. & Song W. (2009), "Category-based equational reasoning : an approach to ontology integration.", *Journal of logic and computation.*, 19 (5). pp. 791-806.

Goguen J. (1991), *A categorical Manifesto.* Mathematical Structures in Computer Science, Volume 1, No 1.

Goldblatt R. (1984), *Topoï: The Categorial Analysis of Logic*. Elsevier Science Publishers B.V, 1984.

Goldsmith J. (2001), "Unsupervised learning of the morphology of a natural language" *computational linguistics* vol27 n2 2001.

Goldsmith J. (2006), "An algorithm for the unsupervised learning of morphology". *Natural Language Engineering, Vol 12, issue 4, December 2006.* Cambridge University Press.

Grefenstette, G. (1992), "Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis". In *Proceedings of the 30th Conference of the Association for Computational Linguistics, ACL'*92 (pp. 324–326).

Grishman R. (1994), "Computational Linguistics: An Introduction". In *Studies in natural language processing.* Cambridge university Press, 1994.

Guillaume J.P. (1986), "Sibawayhi et l'énonciation : une proposition de lecture" In: *Histoire Épistémologie Langage*. Tome 8, fascicule 2, 1986. pp. 53-62.

Hadjsalah A. (1979), " Linguistique Arabe et Linguistique générale", thèse de doctorat, université de la Sorbonne.

Halford G.S.&Wilson W. H. (1980), "A category theory approach to cognitive developpement*", Cognitive Psychology*, vol 12.

Harris Z. (1951), *Structural Linguistics*, the University of Chicago Press.

Harris Z. (1946), "From Morpheme to Utterance" *Language*, Vol. 22, No. 3.

Harris Z. (1954), "Distributional structure", *word*, 10(2-3).

Harris Z. (1990), « La genèse de l'analyse des transformations et de la métalangue » *Langages*, Année 1990, Volume 25, Numéro 99.

Hassan T. (1979), " *Al-Lughat l-arabiyat mabnaha wa maanaha*" , Le Caire: Al-Hay'at lmisriyyat l- ammat li l-kitÁb. 2e éd.

Healy M & Caudell T.P (2006), "Ontologies and worlds in category theory: implications for neural systems", *Axiomaths* 16(1-2).

Henriques G. & al (1990), *Morphismes et Catégories. Comparer et transformer*, Neuchâtel, Delachaux & Niestlé, 1990.

Itkonen E. (1978), *Grammatical Theory and Metascience*. Amsterdam: Benjamin.

Jackson P. & Moulinier I. (2002), *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. John Benjamins Editions.

Jiyad M. (2005), *A Hundred and One Rules A Short Reference for Arabic Syntactic, Morphological & Phonological Rules for Novice & Intermediate Levels of Proficiency*, Al jalees editions.

Johansson S. (1992), "Comments on a paper by Staffan Hellberg". In J. Svartvik (Ed.). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter.

Kassas D. (2005), « Une Etude Contrastive de l'Arabe et du Français dans une Perspective de Génération Multilingue ». Thèse de doctorat, Paris7.

Khodja S. & al (2001), "An Arabic Tagset for the Morphosyntactic Tagging of Arabic" In *Corpus Linguistics 2001*.

Klein F. (1974), *Le Programme d'Erlangen. Considérations comparatives sur les recherches géométriques modernes*, Paris, Gauthier-Villars, 1974.

Kornai A. (2008), *Mathematical Linguistics*, Springer Verlag, 2008.

Kouloughli D.E. (1999), " Y a-t-il une Syntaxe dans la Tradition Arabe ? " in *HEL* tome21, fascicule 2, 1999.

Kuhn T. (1962), *The Structure of Scientific Revolutions*. University of Chicago Press.

Lager T. (1995), "A logical approach to computational corpus linguistics". Doctoral thesis, Göteberg university, Sweden, 1995.

Lavendhomme R. (2001), *Lieux du sujet. Psychanalyse et mathématique, eds* Seuil, 2001.

Lawvere F. W. (1966), « The Category of Categories as a Foundation of Mathematics », *Proc. Conference Categorical Algebra*, Springer Verlag, 1966.

Leech G. (1991), "The State of the Art of Corpus Linguistics". In K. Aijmer & B. Altenberg (Eds.). *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London and New York: Longman.

Leech G. (1992), "Corpora and Theories of Linguistic Performance". In J. Svartvik (Ed.). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter.

Lentin A. (1999), "Quelques réflexions sur les références Mathématiques dans l'œuvre de Zellig Harris''. *Langages*, volume 25, No 99.

Li R. & Pereira L. M., (1995), "Application of category theory in model-based diagnostic reasoning". *Dans FLAIRS'95 : Florida Artificial Intelligence Research Society*, pages 123–127.

Ling Hu & Wang J. (2010), "Geo-ontology integration based on category theory" *Computer Design and Applications (ICCDA)*, 2010 International Conference.

MacLane S. (1997), *Categories for the Working Mathematician*, Second Edition, Springer-Verlag, 1998.

Mahmoudian M. (1981), *La linguistique*, Paris, Seghers.

Massignon L. (1954), "Réflexions sur la structure primitive de l'analyse grammaticale en arabe *" Arabica*, T1, Fasc 1.

Miller G. & Charles W. (1991), "Contextual correlates of semantic similarity". *Language and Cognitive Processes*, 6 (1), 1–28.

Mosel U. (1980), "Syntactic Categories in Sibawayh's "Kitāb" , in *HEL*, Tome 2, Fascicule 1, *élements d'histoire de la tradition grammaticale arabe*.

Muggleton S. (Ed.). (1997), *Inductive Logic Programming*. Springer-Verlag.

Muggleton S. (1999), "Inductive Logic Programming: issues, results and the LLL challenge". *Artificial Intelligence*, *114*(1-2), 283–296.

Muggleton S. & Bain M. (1999), "Analogical prediction". In *Proceedings of the 9th International Workshop on Inductive Logic Programming (ILP-99)* Berlin. Springer-Verlag.

Nevin B.E. (1993), "A minimalist program for linguist. A perspective on the work of Zellig Harris". *Historiographia Linguistica* 20(2/3).

Owen J. (1988), "The Foundations of Grammar: an Introduction to Medieval Arabic Grammatical Theory". *Studies in the History of the Language Sciences*, 45. Amsterdam: John Benjamins.

Patras F. (2003), "L'horizon sémantique et catégorial de la méthode axiomatique". *Noesis no5 formes et crises de la rationalité au XX^{ème} siècle* ; tome2 'épistémologie'.

Patras F. (2005), "Phénoménologie et Théorie des catégories". *New interactions of mathematics with natural sciences and the humanities*. L.ba ed. Springer Verlag.

Piaget J. (1938), *"* La réversibilité des opérations et l'importance de la notion de «groupe » pour la psychologie de la pensée*"*, *Rapports et compte-rendus du XIe Congrès International de Psychologie*, Paris, 1938.

Pinker S. (1994), *The Language Instinct*. Allen Lane.

Popper K. (1950), "Conjectures and refutations: The Growth of Scientific Knowledge".

Popper K. (1972) "*Objective Knowledge: An Evolutionary Approach".* Oxford: Clarendon Press.

Preller A. (2005), "Category Theoretical Semantics for Pregroup Grammars" *LACL* 2005: 238-254

Quinlan J. R. (1993), *Programs for machine learning*. Morgan Kauffman.

Redington R. & al (1995), "The universality of simple distributional methods: identifying syntactic categories in mandarin Chinese". *Proceedings of the cognitive science of natural language processing*, Dublin.

Redington R. & al (1993), "distributional information and the acquisition of linguistic categories: a statistical approach" *proceedings of the 15th annual meeting of the cognitive science society*. Hillsdale, NJ: LEA.

Rissanen J. (1978), "Modeling by shortest data description". *Automatica*, *14*, 465–471.

Rosebrugh R. & Wood R. (1992), "Relational databases and indexed categories". *Dans CMS'92 : Canadian Mathematical Society Conference*, pages 391–407.

Rosen R. (1959),"A Relational Theory Of Biological Systems II". *Bulletin of Mathematical Biophysics* 21:109-128.

Rydeheard D.E & Burstall D. (1988), *computational Category theory*, Prentice-Hall international series in computer science.

Ryding K. (1993), "Case/Mood Synchretism in Arabic Grammatical Theory: Evidence from the Split Morphology Hypothesis and the Continuum hypothesis" in Investigating Arabic, Raji Rammuni and Dilworth Parkinson editions.

Sarkar A. and Haffari G. (2006), "Inductive semi-supervised learning methods for natural language processing". *Tutorial at HLT-NAACL-06*, New York, USA.

Saussure F. (1966), *Course in General Linguistics*. New York: McGraw-Hill.

Sibawayhi, (1980), *Al-Kitāb*, bulaq editions, le Caire.

Stefanik R. (1996), *Structuralism, Category Theory and Philosophy of Mathematics*. Washington: MSG Press. 1996.

Suleiman Y. (1999), *Arabic Grammar and Linguistics*, Routledge, 1999.

Vergne J. (1999), " Etude et Modélisation de la Syntaxe des Langues Naturelles à l'aide de l'Ordinateur. Analyse Syntaxique non Combinatoire/ Synthèse et Résultats ". HDR, 1999.

Versteegh K. (1997), *The Arabic Language*, New York: Columbia Uinversity Press, 1997.

Versteegh H.M. (1978), "The Arabic Terminology of Syntactic Position", *Arabica*, T. 25, Fasc. 3 (Sep., 1978).

Vuillemin J. (1962), "*La Philosophie de l'algèbre. Recherches sur quelques concepts et méthodes de l'algèbre moderne*", Presses Universitaires de France.

Zeman D. (2008), "Unsupervised acquiring of morphological paradigms from tokenized text". In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, pages 892–899, Budapest.