Introduction générale

« On s'étonnera plus tard (bientôt peut-être) que la description du langage et des langues ait pu si longtemps s'exercer sans faire appel aux données quantitatives »

Charles Muller

Définie comme étant la science qui étudie l'organisation du vocabulaire dans le discours d'un point de vue quantitatif, la lexicométrie renferme un ensemble de méthodes permettant d'opérer « des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes »[2]. Ayant le *texte* au centre de son objet d'étude, la lexicométrie partage de ce fait, des zones plus ou moins vastes avec d'autres disciplines à savoir : la linguistique, l'analyse de discours, la recherche documentaire, l'analyse de contenu, l'informatique, l'intelligence artificielle, la statistique, etc.

Avec l'accroissement fulgurant de l'informatique depuis près de trois décennies d'un côté, et l'essor considérable que connaît la linguistique de corpus dont la manifestation la plus parlante est la multiplicité des banques de données textuelles, de l'autre côté, les études et les recherches lexicométriques accusent des évolutions capitales tant sur le plan des objectifs fixés qu'au niveau des principes méthodologiques adoptés. Parmi ces bouleversements sur le plan méthodologique, le recours inévitable à l'outil informatique pour pouvoir faire face à cette grande envergure des opérations de dépouillement et des calculs statistiques colossaux que nécessitent le traitement et l'analyse lexicométriques ; ce qui a pour conséquence que l'on doive constamment œuvrer à définir « de manière toujours plus précise l'ensemble des règles qui présideront au dépouillement automatique des textes stockés en machine »[3].

---

[2] André Salem, *Pratique des segments répétés. Essai de statistique textuelle*, 1987, p. 315.

[3] Ludovic Lebart et André Salem, S*tatistique textuelle*, 1994, p. 18.

Dans la perspective lexicométrique qui est la nôtre, ce travail qui porte sur un ouvrage célèbre de la pensée arabe médiévale, l'*Imtâʕ wa-l-Muʔânasa* de *Ǧawǧîdî* (IV^{e}/X^{e} siècle) se fixe un triple objectif :

En premier lieu, l'élaboration, pour l'arabe, de ce que dans le domaine de la lexicométrie on appelle une *norme lexicologique*[4], donnant une assise théorique et méthodologique aux travaux lexicométriques futurs sur les textes arabes, *norme* qui a fait défaut, jusque là, aux rarissimes recherches, le plus souvent fragmentaires, dans ce domaine.

En deuxième lieu, la confection du dictionnaire de fréquences (qui peut être hiérarchique ou alphabétique) de notre corpus, *Al-ʔImtâʕ wa-l-Muʔânasa*.

En troisième et dernier lieu, soumettre ce corpus à un certain nombre de méthodes d'analyse et de traitement statistiques propres à la lexicométrie en vue d'en étudier, principalement, la *structure lexicale*[5] mais aussi la *trame radicale*[6].

Il est important de distinguer à cet effet, la structure lexicale, à caractère quantitatif, du contenu lexical, défini par la nature des unités qui forment le vocabulaire. En effet, deux corpus peuvent avoir une forte similitude au niveau de la structure lexicale, mais présenter en même temps une grande disparité au niveau du contenu lexical, et *vice versa*.

---

[4] Notion abordée en détail dans la 2ème partie de cette thèse, *Norme lexicologique*, p. 142-302

[5] Voir la troisième partie de cette thèse, *La trame radicale*, p. 305-442

[6] Voir la quatrième partie de cette thèse, *La structure lexicale*, p. 444-637

1. L'approche quantitative en général et ces différentes applications

L'approche lexicométrique s'inscrit dans la perspective plus générale de l'approche quantitative de l'étude des textes, à la l'intersection de plusieurs disciplines, notamment la linguistique, l'informatique et la statistique.

Cette approche trouve plusieurs applications eu égard aux textes, qu'ils soient pris isolément (préoccupations d'ordre stylistique, didactique, historique, etc.), comparés entre eux (typologies de textes, approche contrastive, etc.), considérés dans leur relation aux auteurs (homogénéité d'auteur, attribution d'auteurs, etc.) ou dans leur relation au temps (séries textuelles chronologiques, spécificité chronologique, etc.).

Dans cette perspective, l'approche des études quantitatives en général et l'approche lexicométrique en particulier permettent non seulement d'étudier le texte d'un auteur mais également d'appréhender la spécificité historico-culturelle de son époque ou de son milieu intellectuel. *Ṭaw Ĵidî*, par exemple, pour avoir voulu défendre son statut et celui des « *udabâ* lésés par des mécènes sans scrupules, et dont le rôle devient d'autant moins défini que l'évolution de la charge de secrétaire et de vizir rend leur fonction quasiment obsolète »[7], cet homme de lettres, cet *adîb*, s'est trouvé pris dans des tiraillements entre trois sortes de pouvoir : un pouvoir politique (princes, vizirs, …), un pouvoir pécuniaire (les mécènes) et un pouvoir politico-intellectuel (vizirs, secrétaires se considérant comme des *udabâ*).

Il ne s'agit en aucun cas, dans cette optique, de considérer l'œuvre d'un auteur comme l'espace de projection d'un réel social ou biographique placé en amont d'elle. Il est question d'appréhender, à travers l'œuvre littéraire, les spécificités de l'auteur et à travers lui, celles de son époque. Il ne faut, ni privilégier la singularité de l'écrivain et minimiser le caractère institutionnel de l'exercice de la littérature, ni dissoudre

[7] Lagrange Frédéric, *La Satire des deux vizirs*, 2004, p. 14

l'existence des créateurs dans le fonctionnement du "champ littéraire" ou du "champ de production culturelle", pour ainsi reprendre la terminologie de P. Bourdieu[8]. Il est vrai que l'aspect institutionnel de la littérature contraint en quelque sorte les comportements des écrivains, mais pour créer, ceux-ci doivent "jouer" de et dans cette contrainte même, quitte à déjouer ses effets comme a tenté de le faire notre auteur *ʾAbû Ḥayyân at-Tawḥîdî pour défendre le "capital symbolique" des udabâʾ*. Nous considérons que le contexte n'est pas et ne doit pas être placé à l'extérieur de l'œuvre ; le texte c'est la gestion même de son contexte. Un fait littéraire doit être conçu comme un acte de langage dans lequel le dit et le dire, le texte et son contexte sont indissociables.

1.1. L'étude de la *structure lexicale*

L'approche quantitative des textes permet entre autres, de décrire la *structure lexicale* d'un corpus en évaluant la variation de ses éléments linguistiques (unités lexicales, mots-formes, lemmes, morphèmes, catégories lexicales, etc.).

Pour que cette démarche soit méthodique et fructueuse, il est nécessaire que les opérations de dépouillement préalables soient opérées selon des règles claires et stables assorties d'une réflexion minutieuse autour des notions de *segmentation*, de *lemmatisation*, de *désambiguïsation*, etc. Et ce n'est qu'après ces étapes de dépouillement et de quantification que les décomptes obtenus seront soumis aux traitements statistiques et à l'interprétation pour pouvoir juger *in fine* des variations des différentes unités linguistiques du corpus dont on veut étudier la structure.

---

[8] Pour les concepts de "champ littéraire" et de "champ de production culturelle", on pourra consulter principalement, par ordre chronologique : "Champ intellectuel et projet créateur", dans *Les Temps Modernes*, n° 246, 1966. *Questions de sociologie*, 1980. *Ce que parler veut dire*, 1982. *Choses dites*, 1987. "Le Champ littéraire", dans *Actes de la recherche en sciences sociales*, n° 89, septembre 1991. *Les règles de l'art. Genèse et structure du champ littéraire*, 1992. *Sur la télévision,* suivi de *L'emprise du jounalisme*, 1996

[9] Voir à cet effet : Lagrange F., *op. cit.*, p. 14-16, et Bergé M., *Essai sur la personnalité morale et intellectuelle d'Abû Ḥayyân at-tawḥîdî*, 1974, p. 404-500

1.2. L’étude contrastive des textes

L’approche quantitative permet également de réaliser des typologies de textes pouvant amener à résoudre par exemple, le problème d’attribution à un auteur. Il est en effet possible, dans cette perspective, d’attribuer un texte dont l’auteur est jusqu’alors inconnu, à tel ou tel auteur dont on dispose des caractéristiques quantitatives de style, recueillies à partir d’autres échantillons de textes.

Cette attribution peut être rendue possible grâce aux comparaisons d’indications quantitatives discriminantes opérées sur les différents textes. Ce qui est mis en exergue dans cette perspective, c’est la recherche de similitude/dissimilitude entre les éléments linguistiques de ces textes sur le plan quantitatif.

1.3. L’étude chronologique/historique des textes

Une autre application de l’approche lexicométrique est de déceler des corrélations/dispersions entre phénomènes au niveau de la structure lexicale pour un auteur, un genre ou une époque donnée. Ce qui permettra par exemple, de résoudre le problème de datation ou d’ordre chronologique de tel ou tel texte d’origine mal établie ou de période d’écriture mal définie.

Outre la datation des textes, les premières apparitions des unités lexicales, leurs actualisations dans le discours ou le passage des unités de ce dernier dans le lexique de la langue, peuvent également être datés ; ce qui fournirait des informations inestimables pour la confection, par exemple, du dictionnaire historique de la langue arabe tant attendu.

1.4. L'établissement de la "langue fondamentale"

La détermination de ce qui est communément appelé la "langue fondamentale", c'est-à-dire les mots les plus fréquents dans une langue donnée en synchronie, représente également une application non moins importante de l'approche quantitative des textes.

Dans cette perspective et pour ce qui est de l'arabe, les travaux pour déterminer les mots les plus fréquents dans le but d'établir l'arabe fondamental, ont commencé vers la fin de la première moitié du siècle dernier avec *Moshe Brill* et son livre *The Basic word list of the Arabic Daily Newspaper* [10] basé sur les décomptes de 136 000 mots de la presse écrite (les quotidiens parus entre 1937 et 1939, principalement le journal égyptien *al-ÞAhrâm* et le journal palestinien *FilasÔîn*).

Vers 1950, c'est au tour de E. M. Bailey de publier, dans *A List of Modern Arabic Words*[11], à partir des décomptes de 200 000 mots de la presse égyptienne, sa propre liste des mots les plus fréquents.

Le premier travail statistique sur des manuels scolaires a été élaboré à Damas en 1953, par *FâÞir Ýâqil* en dépouillant 188 000 mots contenus dans 18 manuels scolaires de lecture pour l'école primaire issus de six pays arabes (Palestine, Liban, Syrie, Égypte, Irak et Arabie Saoudite). Le fruit de ce travail a été publié sous le titre *al-Mufradât al-Þasâsiyya li-l-qirâÞa l-ibtidâÞiyya*[12].

En 1959, sur la base de 136 000 mots de la prose arabe, Jacob M. Landau publie à New York, son livre intitulé *A Word Count of Modern Arabic Prose*[13]. L'ouvrage

---

[10] Moshe Brill, Jérusalem, 1940.

[11] Bailey E. M., Le Caire, sans date.

[12] *FâÞir Ýâqil*, Damas, 1953.

[13] Landau Jacob M., New York, 1959.

inclut une reprise de Brill en ajoutant des romans pour équilibrer le corpus (même nombre de mots).

Par ailleurs, un travail de recoupement et de confrontation de plusieurs listes de mots arabes les plus fréquents a été élaboré par *Rušdî Ïâûir* dans son livre *Qâþimat al-mufradât aš-šâþiÝa fî l-luÈa l-Ýarabiyya*[14], et dans lequel il établit une liste de 1 000 mots les plus fréquents communs à quatre listes, celles de *Moshe Brill*, de Bailey, de *Qadrî Loûfi*[15] et celle de *FâÍir Ýâqil* (seulement, pour ce dernier, les livres égyptiens ont été considérés).

À cela s'ajoutent les efforts de certains ministères de l'éducation dans quelques pays arabes pour la constitution de "l'arabe fondamental".

En Égypte par exemple, le Ministère de l'éducation et de l'enseignement publie en 1965 la liste des mots les plus fréquents que les élèves du primaire doivent connaître[16].

En 1969, le Ministère d'État marocain chargé des affaires culturelles et de l'enseignement publie à son tour, à Rabat, le lexique fondamental pour les élèves du primaire[17].

Un an plus tard, c'est le Ministère de l'éducation et de l'orientation nationale libyen qui publie, lui aussi, une liste des mots les plus fréquents utilisés dans les manuels scolaires de première et deuxième années de l'enseignement primaire[18].

---

[14] *Rušdî Ïâûir*, Égypte, 1955.

[15] Lotfi M. K., *Changes Needed in Egyptian Readers to Increase Their Value*, Chicago, 1948.

[16] *Wizârat at-Tarbiya wa-t-TaÝlîm, ÍaÒîlat al-þalfâÛ l-latî yatalaqqanuhâ t-talâmîÆ fî l-marÎala l-ibtidâÞiyya*, Le Caire, 1965.

[17] *Wizârat ad-Dawla l-mukallafa bi-š-šuÞûn a×-×aqâfiyya wa-t-taÝlîm, Al-MuÝjam al-þasâsî li-talâmîÆ al-madâris al-ibtidâÞiyya bi-l-MaÈrîb*, Rabat, 1969.

[18] *Wizârat at-Tarbiya wa-l-Þirshâd al-Qawmî, Qâþima taštamilu Ýalâ kalimât kutub as-sana l-þûlâ wa-×-×âniya min taÝlîm al-ibtidâÞî*, Tripolie, 1970.

En 1973, Roland Meynet, chargé par le Ministère de l'éducation nationale libanais, au sein du Centre de Recherche et de Développement Pédagogiques, présente son *Projet pour la constitution du lexique fondamental de l'arabe moderne*.

À l'image de l'effort fourni par *Rušdî Ṣâliḥ*, évoqué *supra*, un travail non négligeable de recoupement, de synthèse et d'harmonisation des critères des calculs de fréquence, vit le jour en 1979 à Riyad, il s'agit d'*Al-Mufradât aš-šâḥiŕa fî l-luÈa l-Ŕarabiyya*[19], par *Ḥawâd ŶAṣiyya ŶAbdû*. Cette liste regroupe 3 000 mots les plus fréquents parmi les 700 000 mots que comportent quatre listes, celles de *Fâjir Ŷaqil*, de Jacob Landau et de *Moshe Brill* ainsi qu'une liste, non publiée, de l'auteur lui-même.

Plus récents et apportant une touche de multilinguisme, deux travaux se sont attachés à la confection de lexiques arabes, trilingue (arabe-français-anglais) pour le premier, et bilingue (arabe-français) pour le second.

Le premier, paru en 1991, est celui de Djamel Eddine Kouloughli, *Lexique fondamental de l'arabe standard moderne*[20]. Le second est le *Nouveau lexique bilingue de l'arabe d'aujourd'hui*[21] de Mathieu Guidère et paru en 2004.

Kouloughli a élaboré son *Lexique fondamental* à partir de calculs statistiques sur 200 000 mots de textes usuels tirés, à 90 % d'articles de journaux et à 10 % de pièces de théâtre et de nouvelles et récits courts d'auteurs modernes. Ce lexique comporte 3 000 mots et expressions "*courantes*"[22].

---

[19] Le titre entier est : *Ḥawâd ŶAṣiyya ŶAbdû, Al-Mufradât aš-šâḥiŕa fî l-luÈa l-Ŕarabiyya. Dirâsa fî qawâḥim al-mufradât aš-šâḥiŕa fî l-luÈa l-Ŕarabiyya wa-qâḥima bi-ḥašyaŕ ×alâ×at ḥâlâf kalima fî ḥarbaŕin minhâ*, Riyad, 1979.

[20] Kouloughli Djamel Eddine, Paris, 1991.

[21] Mathieu Guidère, Nantes, 2004.

[22] Kouloughli ne se base pas pour classer hiérarchiquement les mots sur la « simple fréquence d'occurrence, donnée purement statistique » et, selon lui, « trompeuse », mais sur « la notion lexico-statistique d'utilité » (p. 5), ce qui rend difficile l'interprétation de ses résultats du point de vue qui nous attache ici.

Mathieu Guidère, quant à lui, a élaboré son *Nouveau lexique bilingue* à partir d'un corpus journalistique disponible sur Internet couvrant la période 2000-2001. Ce corpus est également composé de 200 000 mots et est tiré de quatre journaux quotidiens : *An-Nahâr*, *Al-Ahrâm*, *AÒ-ÑabâÎ* et *Aš-Šarq al-AwsaÔ*.

Le lexique de Guidère est composé des 3 000 mots et expressions les plus fréquents associés à une première contextualisation des unités du lexique au sein de phrases tirées directement du corpus (un seul contexte est donné pour chaque mot, indépendamment de la polysémie de celui-ci). Cependant les fréquences, aussi bien dans Guidère (2004) que dans Kouloughli (1991), restent limitées à la seule forme des mots indépendamment des différents sens qui peuvent leur être associés. De ce point de vue, ces travaux présentent une régression par rapport à Brill (1940), pour les textes de presse, et à la compilation, intégrant celle de Brill et des textes de romans, de Landau (1959).

Il est frappant de constater que ces efforts non négligeables, du moins de par leur nombre, depuis les années cinquante, dans un cadre didactique de recherche de l'établissement de l'arabe fondamental, n'ont pas été suivis, exception faite de quelques cas que nous mentionnons ci-après, par des travaux ou des études dans d'autres directions de la recherche lexicométrique ou stylométrique.

2. Les travaux lexicométriques ou stylométriques sur des textes arabes

La première de ces applications aux textes arabes de l'approche quantitative, se rapprochant plus de la lexicométrie et de la stylométrie telles que établies aujourd'hui,

fut le travail en 1970, de Jacques Piolle et André Roman[23] consacré à l'établissement, en utilisant l'ordinateur, d'un *Lexique de Concordance du Kitâb Al-Tawahhum*.

En 1985, dans le cadre d'un Doctorat d'État *ès Sciences*, Anis Abi Farah[24] essaya d'établir des règles (mathématiques) et d'écrire un programme informatique pour *la reconnaissance automatique de l'auteur inconnu d'un texte arabe*. Ce travail, qui traite d'une vraie problématique lexicométrique, celle d'attribution d'auteurs, a le mérite, nous semble-t-il, d'avoir fait des propositions adaptables d'un point de vue mathématique ; en revanche, il n'aborde pas les aspects linguistiques.

La thèse de Doctorat, soutenue en 1990, par Katia Zakharia[25] est présentée comme s'inscrivant « dans une double perspective. En premier lieu, une perspective d'ordre littéraire ; en second lieu, une perspective d'ordre général : celle d'une interrogation portant sur les conditions nécessaires à l'utilisation des concepts fondamentaux de la psychanalyse hors de la culture judéo-chrétienne dans laquelle ils sont nés (…). Cette perspective littéraire s'inscrit dans la perspective plus générale du recours à la théorie psychanalytique pour l'examen des textes littéraires »[26]. Bien qu'il renferme une bonne dose de calculs statistiques sur les racines des *Maqâmât Al-Íarîrî* ou sur quelques unités lexicales centrales (*unités instrumentales*, selon la terminologie de l'auteur), ce travail est, dans son projet, une étude littéraire utilisant l'approche psychanalytique plutôt qu'une étude lexicométrique globale. K. Zakharia a eu recours aux calculs statistiques dans un cadre délimité ayant un but bien précis : « Pour démontrer l'effet de l'ouvrage sur l'inconscient de ses lecteurs, nous établirons la présence dans le texte de deux trames, l'une radicale (...), l'autre lexicale (...) dont nous examinerons les principales caractéristiques susceptibles de confirmer notre

---

[23] Piolle J. et Roman A., *Lexique de Concordance du Kitâb Al-Tawahhum, établi sur Ordinateur*, 1970.

[24] Abi Farah Anis, *La reconnaissance automatique de l'auteur inconnu d'un texte arabe*, Paris, 1985.

[25] Zakharia Katia, *Les Maqâmât d'Al-Harîrî. Itinéraire d'un héros imposteur et mystique : Abu Zayd Al-Sarûjî*, Lyon, 1990.

[26] Zakharia K., *op. cit.*, p. 6.

hypothèse »[27]. Cette hypothèse est par ailleurs définie comme suit : « puisque ni le récit des *Maqâmât* d'Al-Ḥarîrî ni les histoires qu'il rapporte n'ont permis de donner un sens cohérent à la conversion finale d'Abû Zayd si, comme nous en avons la conviction cette cohérence existe, elle s'exprime nécessairement dans la structure radicale et lexicale de l'ouvrage. »[28]. Dans le présent travail, nous retiendrons le recours à cette double structure, radicale et lexicale, de l'analyse lexicométrique des textes arabes.

Eu égard aux méthodes utilisés et aux techniques et indices employés, la thèse d'État d'Ayadi Chabir[29], soutenue en 1997 et consacrée à l'application de l'approche lexicométrique aux séances d'*Al-HamaÆânî*, représente le premier travail conséquent s'inscrivant dans une démarche lexicométrique. Son apport principal était, nous semble-t-il, d'appliquer pour la première fois, à un texte arabe, certaines méthodes chères à la lexicométrie à savoir : Les *principales caractéristiques lexicométriques* (l'étendue du corpus et celle de ses parties, …), la notion de *lexicalité et fonctionnalité*, la répartition globale des catégories grammaticales (même si nous ne partageons pas la définition et les frontières qu'il assigne à ces catégories), l'analyse factorielle des correspondances et enfin, la notion de richesse lexicale. Pour ce qui est de la notion, au demeurant importante, de richesse lexicale, Ayadi Chabir propose une formule nouvelle, le *coefficient **R** de richesse lexicale*. Pour des raisons qui apparaîtront au fur et à mesure de ce travail, nous adopterons ici une démarche différente, tout en intégrant les principaux acquis de ce travail novateur.

En outre, SaÝd MaÒlûJ, considéré comme le pionnier, dans le monde arabe, de la stylométrie, publie en 1980 un livre (*MaÒlûJ* 1992 - 3e édition) dans lequel il essaie de jeter les bases théoriques de la stylométrie et de les appliquer aux textes arabes. Il publie également entre 1981 et 1989, dans des revues de langue arabe, quatre études intéressantes (*MaÒlûJ* 1981, 1982, 1987 et 1989) dans lesquelles il applique à des

[27] *Ibidem*, p. 9.

[28] *Ibidem*, p. 192.

[29] Chabir Ayadi, *Approche lexicométrique et Lexique-Index des séances d'Al-HamaÆânî*, Paris, 1997.

textes littéraires arabes (prose et poésie), des méthodes stylométriques. Ces quatre études ont été regroupées et publiées plus tard dans un livre (*MaÒ¡ûÎ* 1990).

Une étude en langue arabe de stylométrie contrastive présentée dans le cadre d'une thèse de Doctorat libanais en Langue et Littérature Arabes, a été soutenue en 1999, par Joseph Chraïm[30]. Cette thèse présente un travail original qui étudie dans une approche contrastive, trois œuvres poétiques, *al-QafaÒ al-mahjûr*, *an-Nây wa-r-rîÎ*, et *Sifr al-Ýawda* de trois poètes libanais, respectivement, *Yûsuf ÇuÒÒûb*, *ËalíÍ Íâwî* et *Salîm Nakad*. Il est clair que ce travail a nécessité un effort considérable. La problématique s'inscrit bien dans le cadre de la stylométrie contrastive et les objectifs sont présentés d'une façon claire. L'auteur a construit sa démarche, (p. 13-14), exclusivement sur "*l'adaptation de la méthode*" de Pierre Guiraud (1954), *Les caractères statistiques du vocabulaire : Essai de méthodologie*.

Un autre travail en stylométrie contrastive est dû à *WafâP Kâmil Fâyid*[31], il s'agit d'un livre paru en 2000 et intitulé *QaÒîdat ar-ri×âÞ bayna šuÝarâÞ al-ittijâh al-muÎâfiÙ wa madrasat ad-dîwân. Dirâsa Þuslûbiyya ÞiÎÒâÞiyya*. Dans cette étude, en langue arabe, l'auteur compare le style de cinq poètes dans des élégies, sur la base d'une figure de rhétorique : la métaphore. Les poètes dont le style est comparé dans cette étude, sont *PAÎmad Šawqî* (11 poèmes élégiaques), *ÍâfiÙ PÎbrâhîm* (11 poèmes élégiaques), *Ýabbâs MaÎmûd Al-ÝAqqâd* (18 poèmes élégiaques), *ÝAbd Al-Çâdir Al-Mâzinî* (6 poèmes élégiaques) et *ÝAbd Ar-RaÎmân Šukrî* (10 poèmes élégiaques).

Par ailleurs, une étude de psycholinguistique quantitative appliquée à l'arabe fait désormais date. Il s'agit d'une étude dans un cadre de psychologie cognitive des langues sémitiques en général et de l'arabe en particuliers, faite par J. Grainger, J. Dichy, M. El-

---

[30] Chraïm Joseph, *Manhajiyyat ad-dirâsa l-Þuslûbiyya l-muÝjamiyya l-muqârana. ("al-QafaÒ al-mahjûr" li-Yûsuf ÇaÒûb wa "an-Nây wa-r-rîÎ" li-Ëalîl Íâwî wa "Sifr al-Ýawda" li-Salîm Nakad)*, Beyrouth, 1999.

[31] *Kâmil Fâyid WafâP*, Le Caire, 2000.

Halfaoui et M. Bamhamed[32]. Cet article expose et analyse les résultats obtenus suite à des expérimentations menées sur l'arabe dans le cadre du Laboratoire de Psychologie Cognitive de l'Université de Provence et l'Ecole Normale Supérieure de Fès, et visant à savoir si la racine joue ou non un rôle dans la reconnaissance des mots arabes. Dans cette étude, deux techniques expérimentales de la psycholinguistique contemporaine ont été appliquées : « la première fait appel à des amorces orthographiques subliminales, la seconde, manipule la fréquence des éléments constitutifs du mot, ici, la fréquence de la racine par rapport à celle de la forme de surface du mot »[33]. Les premiers résultats originaux de ces expérimentations ont démontré « une forte sensibilité à des informations du niveau morpho-lexical, ici, à la racine du mot »[34]. L'un des résultats met en relief les effets d'amorçage par la racine en utilisant le paradigme d'amorçage subliminal. Ce résultat corrobore des recherches menées par Frost et ses collaborateurs (Frost et al. 1997 et 2000) sur une autre langue sémitique, l'hébreu. Un autre résultat non moins intéressant montre que « la fréquence de la racine influence le temps de traitement perceptif des mots arabes (...) la présence d'une racine fréquente dans un mot facilite dans ce cas son traitement perceptif »[35]. Un compte-rendu du travail de réalisation des listes de mots ayant servi à ces expérimentations de psycholinguistique quantitative, ainsi qu'une présentation de plusieurs résultats chiffrés concernant les taux d'ambiguïté, certains phénomènes linguistiques et les réalisations graphiques, sont exposés dans (Abbès & Dichy 2008).

Outre l'importance du cadre psycholinguistique lui-même dans lequel ce travail a été élaboré, en plaçant la racine au centre des préoccupations quant à la structure des mots arabes, ce type d'étude octroie une importance supplémentaire à toute étude quantitative de la structure "radicale" des textes arabes permettant d'en déduire les

---

[32] Grainger, J. et al. (2003), Approche expérimentale de la reconnaissance du mot écrit en arabe, in : Jean-Pierre Jaffré (éd.), *Dynamiques de l'écriture : approches pluridisciplinaires*, n° 22 de la revue *Faits de langue*, p. 77-86

[33] Grainger, J. et al. (2003), *idem*, p. 81.

[34] *Ibid*, *idem*, p. 85

[35] *Ibid*, *idem*, p. 86

principales caractéristiques pouvant nous informer davantage sur le style d'un auteur, d'une époque, d'un genre, etc.

Il est évident que le développement des études quantitatives des textes est étroitement lié aux avancées accomplies dans le domaine du traitement automatique des langues (TAL). En effet, en voulant appliquer des modèles statistiques de plus en plus sophistiqués à des corpus de plus en plus volumineux, il paraît nécessaire que l'on doive développer des outils de traitement efficaces (segmenteurs, analyseurs morphologiques ou morpho-syntaxiques, etc.), répondant à une très large gamme de problèmes, notamment des problèmes de *segmentation*, de *lemmatisation* et, surtout, de *désambiguïsation*.

Bien que le TAL arabe n'ait pas connu les mêmes avancées que celles du traitement automatique des langues indo-européennes, et notamment le français et l'anglais, il n'en est pas moins que des progrès importants ont été accomplis en France depuis deux décennies. Plusieurs travaux ont été effectués dans divers domaines allant de la réalisation de didacticiels à la construction d'analyseurs morphologiques ou syntaxiques ou à la traduction automatique. Il serait trop long de les citer tous ici. Nous nous contenterons donc des travaux dans le prolongement desquels s'inscrit notre propre recherche. Trois étapes importantes marquent ces développements : le coup d'envoi des travaux de recherche sur le traitement automatique de la langue arabe fut le *Rapport Desclés*[36] en 1983. Les recherches se sont poursuivies par la suite dans le cadre du programme de recherche *SAMIA*[37]. Et, enfin, dans le prolongement du programme

[36] Desclés J-P.(dir.), *Conception d'un synthétiseur et d'un analyseur morphologique de l'arabe en vue d'une utilisation en enseignement assisté par ordinateur*, Paris, 1983

[37] SAMIA : **S**ynthèse et **A**nalyse **M**orphologique **I**nformatisées de l'**A**rabe

SAMIA[38], les travaux les plus fructueux en termes d'applications informatiques, sont ceux autour du projet de recherche *DIINAR*[39].

3. Les travaux de recherche en TAL arabe

3.1. Le rapport Desclés

Connu sous le nom de « Rapport Desclés », ce document est, en fait, une étude de « faisabilité » réalisée par une équipe réunie autour de J-P. Desclés à la demande de la sous-direction de la politique linguistique du Ministère des Relations Extérieures. Ce rapport avait pour objet de : « proposer un projet d'ensemble de traitement informatique de la morphologie de l'arabe en synthèse et en analyse, ce projet d'ensemble devant servir de base à la réalisation de programmes d'Enseignement Assisté par Ordinateur (E.A.O.), mais étant également avec diverses autres applications tant en recherche fondamentale (linguistique arabe, traitement formel des langues naturelles) que dans divers domaines d'application (Traduction Assistée par Ordinateur, langages quasi-naturels d'interrogation de Base de Données, etc.) »[40].

Le rapport commence par dresser un tableau de tous les travaux qui existaient en France concernant l'analyse morphologique automatique de l'arabe en commençant par le premier travail dans ce domaine, celui de D. Cohen (1961), et arrivant au dernier en date (à cette époque), celui de M. Hassoun (1982), en passant par les travaux de B. Wakim (1978), de Y. Hlal (1979), de A. H. Moussa (1979), de G. Gheith (1980), et de

---

[38] Pour ce qui est des travaux sur le mot graphique et l'analyse morphologique hors SAMIA, citons notamment : Cohen David (1961/1970), Hlal Yahia (1979), Audebert Claude et Jaccarini André (1986), Jaccarini André (1997), Gaubert Christian (2001), Beesley Ken (1998/2001/2003) et Buckwalter Tim (2002).

[39] DIINAR : **DI**ctionnaire **IN**formatisé de l'**AR**abe

[40] Desclés et al., *op. cit.*, p. 3

Gh. Moghrabi (1980) ; sans oublier les travaux menés à Beyrouth par André Roman et Jacques Piolle entre 1968 et 1972.

L'étude a ensuite défini les choix linguistiques du projet se résumant d'un côté par l'adoption de l'arabe littéraire moderne, appelé aussi arabe standard, et de l'autre côté par le traitement des mots non vocalisés pour les modèles « en analyse » et, selon les besoins, la génération des mots vocalisés ou non-vocalisés pour le modèle « en synthèse ».

Après quelques précisions terminologiques concernant l'opposition entre grammaire de synthèse et grammaire d'analyse, le rapport commence par traiter la question de la synthèse morphologique en présentant trois stratégies différentes puis en détaillant les procédures du traitement informatique de la synthèse morphologique et de la constitution du dictionnaire.

L'objectif visé dans ce rapport concernant l'Enseignement Assisté par Ordinateur est restreint à l'apprentissage de la morphologie arabe par E.A.O. avec simulation. Cinq types d'utilisations possibles d'un programme EAO couplé avec un « synthétiseur morphologique » ont été présentés.

Le rapport conclut à la possibilité de réaliser un synthétiseur morphologique de l'arabe dans un temps qui varie de deux à quatre ans selon les moyens. Parallèlement à l'élaboration de ce synthétiseur, la mise au point d'un analyseur morphologique opérant par analyse directe des mots graphiques et par consultation d'un dictionnaire devient possible.

3.2. Le programme de recherche SAMIA[41]

La recherche commencée dans le cadre du *rapport Desclés* a trouvé son prolongement dans le cadre du programme de recherche SAMIA (***S**ynthèse et **A**nalyse **M**orphologique **I**nformatisées de l'**A**rabe*) au sein de l'UA 1073 du CNRS (Université de Paris 8), dirigée par J.E. Bencheikh. Les travaux étaient menés au Laboratoire d'Informatique documentaire de l'Université Lyon 1, sous la responsabilité scientifique de R. Bouché, alors directeur de ce Laboratoire, et de J. Dichy, alors directeur du Département d'Études arabes de l'Université Lumière-Lyon 2.

Dans le cadre de ce programme, plusieurs travaux ont été réalisés dans une optique d'automatisation du traitement de la langue arabe. La démarche adoptée dans ces travaux était double : une démarche analytique consistant à reconnaître et à décomposer l'unité de traitement (le mot graphique arabe), et une démarche synthétique permettant la génération de mots arabes à partir de ses constituants. Nous présentons sommairement ci-après, la liste des travaux SAMIA les plus importants tant au niveau de la conception linguistique que la conception informatique :

La première application de la notion d'analyse du mot graphique arabe[42] telle que conçue par D. Cohen et améliorée par le *Rapport Desclés* et les études théoriques de l'équipe SAMIA, fut la conception de deux modèles morphologiques arabes de première génération : un modèle de synthèse morphologique (M. S. Ziadah) et un modèle d'analyse morphologique (J. Dichy), ainsi que d'un dictionnaire des bases (M. Hassoun, J. Dichy et G. Awad) accompagnant le synthétiseur et l'analyseur. Ce dictionnaire comportant quatre listes : une liste des mots exceptionnels, une liste des formants du mot, une liste des racines et des pro-racines non-vocalisées et, enfin, une liste des bases et de pro-bases.

---

[41] Pour une vue détaillée de ce programme voir : Dichy J. et Hassoun M. (éd.), *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe, travaux SAMIA I*, 1989

[42] Voir Chapitre 5, Section 1.4.- L'analyse du mot graphique en arabe, p. 207-211

En ce qui concerne la conception informatique, quatre principales réalisations ont vu le jour. Dans son travail intitulé *La constitution automatique du vecteur-mot d'un mot graphique non-vocalisé en arabe*, N. Abu Al-Chay a pu réaliser une segmentation (morphologique) automatique d'un mot graphique non-vocalisé en arabe. Son objectif était « d'associer à chaque mot graphique non-vocalisé placé à l'entrée du système, l'ensemble des segmentations possibles pour ce mot. Le résultat est un ou plusieurs vecteur-mots dont les éléments sont les constituants identifiés du mot considéré »[43].

La deuxième réalisation informatique dans le cadre du programme de recherche SAMIA, fut *Un transducteur phonologique de l'arabe*, élaboré par H. Abaab et constitué d'une grammaire de synthèse et d'un algorithme de génération phonologique. La synthèse phonologique dont il est question dans cette réalisation informatique consistait à « donner une représentation attestable du mot à partir des données présentées dans le vecteur descriptif »[44]. Pour le bon fonctionnement du transducteur, deux types de contrôle sont effectués : « celui de la compatibilité catégorielle, assuré par le générateur phonologique et celui de la compatibilité contextuelle, assuré par la composante de transformations contextuelles »[45].

La troisième réalisation informatique dans la cadre du même programme, fut la *Conception, en PROLOG-Foll, d'un synthétiseur morphologique de l'arabe*, élaboré par F. Mourad. S'appuyant sur la thèse de H. Abaab et sur le modèle linguistique de la grammaire des formants du mot arabe dû à J. Dichy, ce travail a pu mettre au point un programme manipulant un automate qui peut produire les formes nominales ou verbales voulues conformément au modèle sous-jacent.

Quant à la quatrième réalisation informatique dans la cadre du programme de recherche SAMIA, ce fut une *Première Approche d'un système d'information pour*

[43] Najim Abu Al-Chay, *La constitution automatique du vecteur-mot d'un mot graphique non-vocalisé en arabe*, dans : Dichy J. et Hassoun M. (éd.), *op. cit.*, p. 163

[44] Houcine Abaab, *Un transducteur phonologique de l'arabe*, dans : Dichy J. et Hassoun M. (éd.), *idem*, p. 183

[45] *Ibidem*, p. 183

l'E.A.O. de la morphologie de l'arabe, élaborée par L. Bouzidi. L'objectif principal de la conception de ce système d'information était de permettre un E.A.O. de la morphologie de l'arabe. Ce système d'information devrait garantir la mise en place de quatre opérations essentielles dans cette perspective : l'analyse morphologique des mots, la génération et la production des mots, le traitement des formes verbales (en synthèse et en analyse) et, enfin, l'exploitation lexicale ayant trait aux constituants du dictionnaire inclus dans le modèle linguistique. Le système d'information proposé s'inscrit dans la catégorie des « systèmes paramétrés qui n'imposent pas de cursus figés aussi bien au niveau du contenu, qu'au niveau de l'exploitation de système, laissant ainsi l'initiative majeure à l'enseignant avec un minimum de contraintes »[46]

Une cinquième conception informatique dans un cadre didactique, non moins importante que les quatre autres, conçue également dans le cadre des travaux de l'équipe SAMIA par Xavier Lelubre et présentée en 1987 au Concours national de scénarios de logiciels à usage éducatif, organisé par le Ministère de l'Education Nationale. Il s'agit d'un didacticiel de conjugaison baptisé *Dialogaison*, où les verbes sont présentés en contexte au sein de dialogues entres des personnages, au nombre de 1, 2 ou 3, et des deux sexes, ce qui permet de faire conjuguer les verbes aux différents temps et aux différentes personnes. L'utilisateur peut décider du nombre des personnages. Le programme fonctionne en synthèse, où les formes verbales sont générées par un conjugueur mis au point par l'équipe SAMIA, et en analyse, où les réponses sont analysées, également dans le cadre des travaux SAMIA. Ce scénario de logiciel didactique a été primé par le Jury, présidé par le Doyen de l'Inspection Générale de l'Education Nationale[47].

En outre, un didacticiel de conjugaison des verbes arabes anomaux a par la suite été réalisé par C. Ighilaza dans le cadre de son mémoire de D.E.A. en Sciences de l'Information et de la Communication. Les verbes anomaux dont il était question sont

[46] Laid Bouzidi, *Première Approche d'un système d'information pour l'E.A.O. de la morphologie de l'arabe*, dans : Dichy J. et Hassoun M. (éd.), *idem*, p. 228

[47] Lelubre Xavier, *"Dialogaison", scénario de logiciel d'arabe*, 1987.

du type dit *NâqiÒ*. Ce sont des verbes dont la troisième consonne de la racine est un *wâw* « و » ou un *yâ'* « ي ».

Par ailleurs, parmi les travaux antérieurs au programme de recherche SAMIA, il eut la thèse de Doctorat de troisième cycle de Mohamed Hassoun (Hassoun 1982). Ce travail constitua un pilier primordial à toutes les réalisations informatiques qui ont suivi au niveau du programme SAMIA. M. Hassoun fut le premier à avoir élaboré un système opérationnel d'analyse morphologique automatique de la langue arabe au moyen d'une liste de traits morphologiques. Il proposa dans ce travail, un « procédé original d'analyse morphologique adapté à l'arabe, fondé sur le calcul d'une chaîne de variables booléennes représentatives de la présence ou de l'absence de propriétés caractéristiques du mot analysé »[48].

En prolongement ou en approfondissement des travaux présentés ci-avant, toujours dans la cadre du programme de recherche SAMIA, des thèses de Doctorat (de troisième cycle ou d'État) ont été soutenues par les membres du groupe. Nous en citons quelques unes dans un ordre chronologique :

3.2.1. La thèse de Houcine Abaab en 1984

S'intitulant *Contribution au traitement automatique de la langue arabe - Conception d'un synthétiseur morphologique utilisable en E.A.O.*, cette thèse s'inscrit dans une démarche de synthèse visant à produire la représentation phonologique du mot à partir d'un vecteur descriptif. Deux concepts sont introduits : des grammaires à validation et saturation terminales et des automates finis saturés.

Comme indiqué *supra*, ce qui est recherché dans cette réalisation c'est une représentation attestable du mot en se basant sur les données présentées dans le vecteur

---

[48] Mohamed Hassoun, *Utilisation d'une liste de de traits morphologiques pour la réalisation d'un analyseur morphologique automatique de l'arabe* , dans : Dichy J. et Hassoun M. (éd.), *idem*, p. 245

descriptif. Ceci dans le cas où ces données sont cohérentes ; si elles sont défaillantes, le vecteur est rejeté et le processus de synthèse est arrêté.

Les deux types de contrôle mentionnés ci-avant : le contrôle de la compatibilité catégorielle et celui de la compatibilité contextuelle sont bien entendu effectués pour garantir le bon fonctionnement du système. La définition des catégories morphologiques dépend étroitement de l'ensemble des éléments du vecteur descriptif du mot. Et la bonne succession des catégories est contrôlée par cette grammaire grâce à une liste ordonnée de règles de transition.

La réalisation informatique de ce synthétiseur porte donc d'une part sur la définition et la représentation interne de l'automate, et d'autre part sur l'algorithme qui permet de le faire fonctionner.

3.2.2. La thèse d'État de Mohamed Hassoun en 1987

Se basant sur le modèle linguistique conçu au sein l'équipe du programme de recherche SAMIA, cette thèse d'État, intitulée *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'applications*, avait pour objectif principal l'élaboration d'un dictionnaire informatisé pouvant être utilisé dans le traitement automatique de l'arabe. Il a été question dans ce travail de définir une structuration de données permettant la construction d'un dictionnaire informatisé utilisable dans les deux contextes de la synthèse et l'analyse de la morphologie, ainsi que de l'E.A.O. avec simulation d'un modèle linguistique.

La conception du dictionnaire informatisé a été élaborée sous forme d'une base de données relationnelle utilisant le SGBD (Système de Gestion de Bases de Données) Informix et tournant sous le système d'exploitation UNIX.

L'ensemble des tables et l'organisation des relations (au nombre de 14) de cette base de données traduisent les principales règles du modèle linguistique adopté au sein du groupe (schèmes morphologiques attestés, affixes et clitiques, spécificateurs, bases, pré- et post-bases, racines, vecteurs-descripteurs, …).

3.2.3. La thèse de Najim Abu Al-Chay en 1988

Sous le titre *Un système expert pour l'analyse et la synthèse des verbes arabes dans un cadre d'enseignement assisté par ordinateur*, cette thèse s'est fixée comme objectif la conception un modèle adapté au traitement automatique des verbes arabes dans le cadre de l'E.A.O. au sein du programme SAMIA. Najim Abu Al-Chay a conçu un modèle basé sur une structuration des constituants de verbes en préfixe-base-suffixe. Ce système expert est conçu pour permettre l'apprentissage de la conjugaison en tant que composante importante du programme E.A.O de la morphologie arabe.

La structuration et le fonctionnement de ce système opèrent en deux situations différentes : la première étant au niveau de la reconnaissance des formes verbales pour pallier aux difficultés rencontrées par les apprenants au moment de la lecture de l'arabe. La deuxième au niveau de la production langagière en facilitant la construction des verbes arabes au moment de l'expression.

Pour le bon fonctionnement du système, plusieurs listes ont été construites. Elles concernaient le mot graphique, l'ensemble de ses segmentations possibles, leurs combinaisons et leur compatibilité. Cette compatibilité entre les représentations en graphie non vocalisée des morphèmes a été rendue possible grâce à une classification des données linguistiques établies en fonction des relations. Ceci a permis la réduction au maximum du nombre des relations possibles entre les constituants du mot.

Le processus fonctionne en deux étapes : la première étape est l'extraction de la base du mot (identification de la pré-base, identification de la post-base, comparaison des deux en fonction de leur compatibilité), et la deuxième étape est l'identification de

la base du mot et la comparaison de celle-ci à la liste des bases enregistrées dans le dictionnaire provisoire.

3.2.4. La thèse d'État de Joseph Dichy en 1990

Même si la notion d'analyse du mot graphique a été initiée, comme nous l'avons vu plus haut, par D. Cohen et améliorée par d'autres membres du programme de recherche SAMIA, c'est à Joseph Dichy que revient, nous semble-t-il, la paternité du modèle linguistique dans sa dernière formulation (modèle du mot graphique, de ses relations et des spécificateurs morphosyntaxiques qui lui sont associés). C'est en effet, cette version finalisée par lui qui a été utilisée et adoptée par tout le groupe du programme SAMIA.

Dans sa thèse de Doctorat d'État intitulée *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*, surtout le dixième chapitre, ainsi que dans (Dichy 1997-b), Joseph Dichy aboutit non seulement, à finaliser et à faire la synthèse de ce modèle linguistique désormais sous-jacent à tous les travaux de TAL basés sur la notion du mot graphique arabe, mais il l'enrichit par des données primordiales : l'inventaire fini des spécificateurs morphosyntaxiques du domaine du mot. Ce qui caractérise le plus son approche dans ce modèle linguistique, c'est qu'elle met précisément en exergue la relation entre lexique et grammaire.

À ces spécificateurs morphosyntaxiques revient la tâche de gérer les relations qui peuvent exister entre la base nominale ou verbale, d'un côté, et les autres formants qui représentent des extensions à l'une ou l'autre des deux bases, de l'autre côté. D'où la distinction que J. Dichy fait entre formants-noyau (Fn) et formants-extension (Fe)[49]. Les spécificateurs morphosyntaxiques sont de deux types : les spécificateurs relatifs aux bases nominales, au nombre de six, et les spécificateurs relatifs aux bases verbales, au nombre de sept.

---

[49] Voir la section 1.4.2 du chapitre 5, p. 210-211

3.3. Les travaux de recherche autour de DIINAR[50]

Tous les travaux et réalisations informatiques dans le cadre du programme de recherche SAMIA ont abouti à une conclusion importante : pour que les différents outils développés dans la perspective du traitement automatique de l'arabe (analyseurs morphologiques ou morphosyntaxiques, conjugueurs, correcteurs orthographiques, voyelleurs, étiqueteurs, segmenteurs, lemmatiseurs, indexeurs, etc.) soient fonctionnellement efficaces, robustes et performants, il faut parvenir à confectionner un dictionnaire informatisé, une base de données lexicale renfermant toutes les unités de la langue arabe, leurs relations (d'ordre et de collocation dans le vecteur de représentation du mot graphique) et leurs spécificateurs morphosyntaxiques, et ce à la lumière de tout ce qui a déjà été accompli dans les travaux précédents (notamment la conception informatique de M. Hassoun) et sur la base du modèle linguistique finalisé par J. Dichy.

Cette grande base de connaissances lexicale attendue et pour la confection de laquelle les travaux de recherche ont repris en 1991, il a été décidé de la baptiser *DIINAR* (***DI****ctionnaire* ***IN****formatisé de l'****AR****abe*). Et c'est autour de ce projet (considéré comme la deuxième génération du programme SAMIA ou son prolongement) que l'équipe a été reconstruite : presque totalement au niveau de ses membres et partiellement au niveau de sa direction. Géographiquement, c'est désormais Lyon qui accueille cette équipe. Le *pôle lyonnais* de traitement automatique de la langue arabe est né.

Les travaux autour de DIINAR ont été menés, au départ, au CRTT (*Centre de Recherche en Terminologie et Traduction*) de l'Université Lyon 2, puis dans le cadre d'ICAR (*Interactions, Corpus, Apprentissages et Représentations*, UMR 5191, CNRS – Université Lyon 2 et ENS-LSH) d'un côté, et au CERSI (*Centre d'Etudes et de Recherche en Sciences de l'Information*), devenu par la suite SII (*Système*

[50] Voir le site : http://silat.univ-lyon2.fr/ → Rubrique : « Outils pour le TAL » → DIINAR

d'Information et Interface) de l'ENSSIB (Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques) de l'autre côté, et ce en collaboration directe avec l'IRSIT (*Institut Régional des Sciences Informatiques et des Télécommunications*), aujourd'hui *IT.COM*. Les travaux sont dirigés par J. Dichy et X. Lelubre (Université Lyon 2), Mohamed Hassoun (ENSSIB) et S. Ghazeli et A. Braham (IRSIT).

3.3.1. La conception et la réalisation de la Base de données lexicale DIINAR.1 (Dichy 1998-a), (Hassoun & al 1996) et (Braham 1998)

Comme nous le mentionnions plut haut, la base de données lexicale DIINAR est construite sur le modèle linguistique conçu par le groupe SAMIA et finalisé par J. Dichy. La structuration des données choisie permet la construction d'un dictionnaire informatisé utilisable dans les deux contextes de synthèse et d'analyse morphologiques avec simulation des modèles linguistiques. Le dictionnaire, conçu sous forme d'une base de données relationnelle, est géré par un système de gestion des bases de données (SGBD) qui assure une parfaite indépendance entre les données et les traitements, et qui permet surtout l'évolutivité du dictionnaire.

Au départ, les données linguistiques ont été saisies dans une base de données multi-fichiers c'est-à-dire avec un système SGF (Système de Gestion de Fichiers) qui fonctionnait sous le système d'exploitation MS-DOS® (Gader 1996) et (Ghenima 1998). Avec cette ancienne version quelque 20 000 entrées verbales et 39 000 entrées nominales ont été saisies. L'inconvénient majeur de cette base est que les données étaient dépendantes des programmes qui les gèrent.

Riadh Zaafrani a été chargé de la migration de cette base de données vers une base gérée par le SGBD Access® et tournant sous Windows®. Il a conçu la modélisation de la nouvelle version de la base de données DIINAR et en a fait la réalisation informatique (Zaafrani 2002).

Outre la base de données verbale et la base de données nominale existantes dans l'ancienne version, deux autres bases de données ont été ajoutées dans la nouvelle version de DIINAR : une base de données des mots-outils et une base de données des noms propres. La modélisation et la réalisation informatique des deux bases ont été faites par R. Zaafrani.

Quant aux données linguistiques représentant le contenu de ces deux bases de données, c'est nous qui avons fourni, en collaboration avec J. Dichy, modélisé et saisi les 442 mots-outils simples et plus de 11 000 mots-outils composés, pour la base de données des mots-outils d'une part, et c'est Abdelmoneim Jedamy qui a fourni et saisi, également en collaboration avec J. Dichy, les 1 384 noms propres, pour la base de données des noms propres d'autre part.

La base de données lexicale DIINAR comprend actuellement, environ 20 000 verbes, 30 000 noms, 10 000 pluriels brisés, 70 000 dérivés nominaux, 1 300 noms propres et 442 mots-outils simples (11 000 mots-outils composés sont générés à partir des mot-outils simples). Aux entrées de cette base sont associés des spécificateurs morpho-syntaxiques gérant les relations lexique-grammaire du niveau du mot.

Les ressources lexicales de la base de données DIINAR sont disponibles à l'achat chez l'organisme européen ELRA / ELDA[51] (**A**ssociation **E**uropéenne pour les **R**essources **L**inguistiques / **E**valuations and **L**anguage resources **D**istribution **A**gency).

3.3.2. Méthodologie d'élaboration des entrées d'une Base de données lexicale de l'arabe (Ezzahid 1996)

L'objectif principal du travail de Samia Ezzahid était de proposer une méthodologie d'élaboration des entrées d'une base de données lexicale de l'arabe basée sur le modèle Sens-Texte d'Igor Mel'cuk. Ceci s'inscrit dans la perspective de doter la base de données DIINAR d'un ensemble de spécificateurs syntactico-sémantiques.

[51] Le site internet de l'organisme est : http://www.elra.info/fr/ et http://www.elda.org

S. Ezzahid présente une assise théorique concernant la théorie et le modèle Sens-Texte (TST et MST), discute la relation grammaire/lexique en arabe selon E. Ditters et A. Fassi Fehri, expose les fonctions lexicales (FL) de Mel'cuk en les illustrant par des exemples en arabe, et finit par étudier les fonctions lexicales dans le livre *Fiqh al-luÈa* de Õa Ý âhibî (mort en 429/1038).

Après avoir élaboré quelques entrées lexicales de l'arabe selon la méthode lexicographique du DEC (Dictionnaire Explicatif et Combinatoire) de Mel'cuk, S. Ezzahid applique le système de paraphrasage de la TST à la langue arabe et finit par proposer une interface de saisie des entrées lexicales de l'arabe selon la méthode DEC. Cette interface est conçue avec Microsoft Access®.

3.3.3. La correction orthographique des textes arabes : Le système *CorTexAr* (Gader 1992 et 1998)

Dans ce travail, Nabil Gader montre la faisabilité de l'automatisation de la vérification et de la correction des fautes dans les textes arabes non-voyellés. Le système *CorTexAr* (***Cor****rection des* ***Tex****tes* ***Ar****abes*) permet de vérifier les textes sur le plan lexical par décomposition de ses mots graphiques en prébases, bases et postbases, et de corriger un type d'erreur par constituant, ce qui permet donc, de corriger au plus trois erreurs par mot. Il permet également de proposer un mot voisin en cas d'erreur.

La méthode de vérification et de correction adoptée par N. Gader n'exclut pas la possibilité d'avoir plusieurs lexiques, ou plus précisément un lexique général et plusieurs vocabulaires propres à des domaines de spécialité différents.

CorTexAr se présente sous la forme d'un module additif au traitement de texte Word® de Microsoft.

3.3.4. Un système de voyellation des textes arabes : Le système Voyla (Ghénima 1998)

Malek Ghénima présente dans ce travail, un système de voyellation de textes arabes, basé d'une part sur un analyseur morphologique de l'arabe et d'autre part sur une micro-syntaxe de mots-outils.

L'analyseur morphologique utilisé dans le système de voyellation, utilise un lexique d'environ deux millions de mots, généré à partir de la base de données lexicale DIINAR. La micro-syntaxe des mots-outils est un ensemble de règles relatives aux contextes avant et après un mot-outil. Le système *Voyla* fonctionne sous MS-DOS®.

3.3.5. Définition des unités linguistiques intervenant dans l'indexation automatique des textes en arabe (Abbas 1998)

Wijdan Abbas propose dans ce travail, une catégorisation des unités linguistiques nécessaires à l'élaboration d'une grammaire de reconnaissance des syntagmes nominaux en arabe. Elle étudie la question de variables syntaxiques, flexionnelles et lexicales dans le but d'enrichir le lexique de la base de données DIINAR afin de pouvoir surmonter les solutions "parasites" dans l'analyse des textes et arriver ainsi à une analyse linguistique robuste en examinant les relations morphosyntaxiques existant entre ces catégories. Elle propose une typologie des syntagmes nominaux (SN simples, SN annectifs, SN anaphoriques). Elle établie également une liste des critères logico-sémantiques du SN.

En se basant sur la question de la détermination (par l'article défini et par annexion) et de l'indétermination (*tanwîn*), W. Abbas arrive à distinguer entre logique intensionnelle et logique extensionnelle.

Les règles élaborées par W. Abbas peuvent permettre d'extraire un grand nombre de SN contenus dans le corpus étudié. Cependant, certains problèmes d'ambiguïté liés aux contraintes d'accord en genre, en nombre, en personne ou en cas demeurent. Pour résoudre ces problèmes d'ambiguïté, elle propose de compléter le dispositif par des règles EAG (Extend Affix Grammar) dont l'intérêt est de pouvoir intégrer les affixes de ces variables (genre, nombre, personne et cas).

3.3.6. Modélisation des verbes arabes en vue d'un traitement automatique de la conjugaison (Ammar & Dichy 1999)

Ce travail, établi par Sam Ammar et Joseph Dichy et publié dans la célèbre collection *Bescherelle*, présente les 129 modèles de conjugaison de l'arabe. Ces modèles ont permis de conjuguer les 21 000 verbes recensés dans la base de données des verbes contenue dans la base de connaissances lexicale DIINAR.

Ce résultat a été atteint grâce à une modélisation à partir des listes de préfixes, de suffixes et de bases verbales (Dichy 1993). Dans cette modélisation, les bases schématiques sont donc les mêmes pour toutes les formes relevant d'un verbe modèle donné. Il suffit alors au système en synthèse d'avoir deux informations pour conjuguer un verbe : le numéro du modèle qui lui est associé et sa racine.

3.3.7. Modèle probabiliste associé à un analyseur morphologique pour la levée d'ambiguïté (Tout 2001)

Dans cette thèse, Mohamed Tout se fixe comme objectif principal de proposer un modèle probabiliste de levée d'ambiguïté, associé à l'analyseur morphologique développé dans le cadre du groupe de recherche autour de DIINAR.

Après avoir présenté toutes les méthodes proposées pour la levée d'ambiguïté dans d'autres langues, M. Tout considère que le choix le plus adéquat à la langue arabe

est la méthode utilisant des tables des fréquences conditionnelles relatives. Cette méthode utilise une approche probabiliste pour reconnaître la structure syntaxique des chaînes de 2 à n catégories morphologiques consécutives dans l'échantillon.

3.3.8. La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe (Ouersighni 2002)

S'inscrivant dans le cadre de la compréhension automatique en général des textes écrits, le travail de Riadh Ouersighni se fixe pour objectif principal la conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe et d'un système de détection et de diagnostic des erreurs.

Après avoir présenté et discuté les problèmes que l'analyse morpho-syntaxique de l'arabe soulève et les différentes conceptions proposées, R. Ouersighni présente l'architecture générale, les composants et le fonctionnement de son système *AraParse*. Une bonne synthèse des méthodes d'analyse syntaxique et des grammaires formelles est présentée.

Avant de faire une présentation détaillée du système *AraParse* et d'une évaluation des résultats de l'analyse morpho-syntaxique, R. Ouersighni définit une stratégie globale de robustesse pour le traitement des phénomènes extra-linguistiques observés dans les textes et met l'accent sur la réutilisation des modules de l'analyse morpho-syntaxique dans le développement d'un système de détection et de diagnostic des erreurs.

3.3.9. Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère (Zaafrani 2002)

L'objectif de cette thèse est l'élaboration d'un environnement informatique d'aide à l'apprentissage lexical et grammatical de l'arabe langue seconde ou étrangère. Riadh Zaafrani vise dans ce travail : l'élaboration de ressources linguistiques et d'outils informatiques pour le traitement automatique de la langue arabe dans le but de les utiliser dans la construction du système baptisé *Al-MuŶallim* permettant l'individualisation de l'apprentissage de l'arabe notamment par la gestion d'un modèle de l'apprenant.

Le système *Al-MuŶallim*, écrit en Visual Basic et tournant sous Windows®, est un environnement d'apprentissage fonctionnant autour d'un schéma d'apprentissage en trois volets : exposition/compréhension de textes, rétention du lexique et maîtrise de la grammaire.

R. Zaafrani a d'abord construit un générateur automatique dont la tâche est la production paramétrée d'un lexique généré à partir de la base de données lexicale DIINAR, permettant à un analyseur morpho-syntaxique de fonctionner. Il a également développé un certain nombre d'applications à savoir : un outil d'étiquetage morpho-syntaxique de textes bruts, un "concordanceur" qui permet de sélectionner des exemples d'utilisation d'une unité dans un corpus ou de générer des activités d'apprentissage, et un programme pour le calcul des fréquences des unités lexicales.

C'est à R. Zaafrani que revient la conception et la réalisation informatique de l'interface graphique de la base de données des mots-outils de DIINAR qui nous a permis de saisir en 1999 les données linguistiques des 442 mots-outils simples établis en collaboration avec J. Dichy et plus de 11 000 mots-outils composés.

Il est à noter également que la collaboration de R. Zaafrani nous a été précieuse dans le développement du segmenteur semi-automatique (Mouelhi 2008-a) dont nous rendons compte plus loin dans ce présent travail[52].

3.3.10. La conception et la réalisation d'un concordancier électronique pour l'arabe (Abbès 2004)

Le travail de Ramzi Abbès aboutit à la réalisation d'un concordancier électronique de l'arabe. Après avoir montré l'insuffisance pour l'arabe des méthodes de parcours de surfaces (KWIC) utilisées dans la majorité des autres concordanciers, R. Abbès explique ses choix quant à la génération du lexique à partir de la base de données lexicale DIINAR.

La question des ambiguïtés de la langue arabe est posée de façon à permettre à une concordance automatique de répondre à des heuristiques pour réduire la multiplicité des solutions tout en restant interactive, permettant ainsi d'assister l'intervention humaine.

3.3.11. Indexation des documents multilingues d'actualités incluant l'arabe (El Hachani 2005)

Dans cette thèse, Mabrouka El Hachani commence par mettre l'accent sur l'importance et en même temps l'insuffisance des thésaurus comme solution pour l'indexation des documents multilingues. Cette insuffisance au niveau de la recherche d'équivalence de langue à langue est due au fait que la documentation propose des solutions inspirées de la terminologie ; ce qui ne prend pas en considération le fait que les langues et les cultures découpent le réel de manière différente.

La question que pose alors M. El Hachani est de savoir si l'indexeur (humain) s'appuie plus sur ses propres connaissances pour résoudre le problème de recherche

[52] Voir la section 3.1.1 du premier chapitre, p. 68-76.

d'équivalence ou sur des connaissances externes. Et c'est pour répondre à cette question principale qu'elle élabore un protocole d'expérimentation pour le contexte multilingue en se basant sur les travaux réalisés pour l'étude du processus cognitif en indexation monolingue.

M. El Hachani expose les résultats de l'expérimentation qui ont montré que l'indexeur s'appuie sur les outils disponibles mais également sur ses connaissances. C'est pourquoi elle a pu établir un schéma du processus cognitif permettant de voir le cheminement de l'indexeur lors de son activité d'indexation en contexte multilingue. Les stratégies établies par M. El Hachani fournissent des informations utiles pour élaborer des outils automatiques multilingues pertinents pour la gestion des connaissances.

3.4. Le projet DIINAR-MBC (1998-2001)

DIINAR-MBC (***DI****ctionnaire* ***IN****formatisé de l'****AR****abe,* ***M****ultilingue et* ***B****asé sur* ***C****orpus*) est un projet européen financé par la Commission des Communautés Européennes (DG XIII, N. 961 791). Il regroupe trois institutions scientifiques européennes et trois institutions de recherche dans le monde arabe.

Les partenaires européens de ce projet sont : l'Université Lumière-Lyon 2 (sous la direction de J. Dichy et X. Lelubre), l'ENSSIB (sous la direction de M. Hassoun) et l'Université de Nimègue (représentée par E. Ditters). Les partenaires du monde arabe sont : l'IRSIT de Tunis (sous la direction de S. Ghazali et A. Braham), l'ERI du Caire (sous la direction de N. Hegazi) et l'IERA de Rabat (sous la direction de A. Fessi-Fehri). La coordination de DIINAR-MBC est assurée par l'Université Lyon 2 (J. Dichy).

La tâche principale des partenaires dans ce projet était la conception et la réalisation d'une boîte à outils pour le traitement automatique de la langue arabe dans une perspective multilingue.

Un certain nombre de délivrables ont été soumis en fin de projet à la commission européenne (Dichy 2000-b). Ces délivrables sont regroupés en trois secteurs : corpus, lexiques et analyse syntaxique. Dans chacun de ces secteurs se trouvent des ressources linguistiques, des logiciels et outils informatiques, ainsi que des documents théoriques et des modes d'emploi.

On y trouve par exemple, au niveau des corpus : *ARCOLEX OCR* (un corpus scanné de 10 millions de mots), *ARCOLEX textual dB* (une base de données textuelles encodée selon le format TEI) et *ARCOLEX tagged corpus* (corpus étiqueté parenthésé) du côté des ressources linguistiques, et du côté des logiciels et outils informatiques, *ARTINDEX input & consultation interface* et *TEI procedures for Arabic text encoding (using EMACS)*.

Au niveau des lexiques : *Arabic Monolingual PROLEMAA*, *Arabic-English bilingual – PROLEMAA* et *Arabic-French bilingual – PROLEMAA* du côté des ressources linguistiques, et du côté des logiciels et outils informatiques, *PROLEMAA interface*.

Quant aux logiciels et outils informatiques pour l'analyse syntaxique : *LARUSA Analyser* et *CATAD input & consultation interface*.

3.5. Le groupe de Recherche SILAT

Aujourd'hui les recherches dans le domaine du TAL arabe et de l'ingénierie linguistique se poursuivent dans le cadre du groupe de recherche lyonnais SILAT[53]

[53] http://silat.univ-lyon2.fr/

(***S**ystèmes d'information, **I**ngénierie et **L**inguistique **A**rabes, **T**erminologie*) sous l'égide de deux organismes de recherche lyonnais : ICAR (***I**nteractions, **C**orpus, **A**pprentissages et **R**eprésentations*, UMR 5191, CNRS – Université Lyon 2 et ENS-LSH) et ELICO (***É**quipe de **L**yon en **I**nformation et **C**ommunications*, EA 4147, ENSSIB et Université Lyon 3). Le groupe de recherche SILAT est co-dirigé pour ICAR-CNRS, par J. Dichy et X. Lelubre (Université Lyon 2), et pour ELICO, par Mohamed Hassoun (ENSSIB) et Mabrouka El Hachani (Université Lyon 3).

4. De la nécessité d'une norme lexicologique pour la lexicométrie arabe

La statistique lexicale (devenue lexicométrie par la suite) et la stylométrie ont connu, après P. Guiraud, une importante expansion, surtout avec les travaux, les développements, les réaménagements, les critiques et les refontes successives d'un Charles Muller, d'un Etienne Brunet, d'un André Salem, d'un Maurice Tournier, d'un Pierre Lafon, d'un Charles Bernet, d'un Dominique Labbé, d'un Benoît Habert, et de beaucoup d'autres ; et elles ne cessent de se développer.

Ces deux approches du domaine des études statistiques des textes ont donc connu d'énormes progrès notamment au niveau méthodologique. Parmi les impératifs méthodologiques que le développement de la discipline a pu imposer, depuis Charles Muller et l'équipe de Lexicologie Politique de Saint-Cloud autour de Maurice Tournier, à tous les travaux s'inscrivant dans le domaine des études quantitatives, c'est justement la nécessité d'établir et puis de suivre une *norme de dépouillement* au moment du prétraitement des données textuelles du corpus étudié. Loin des différentes connotations que peut avoir, surtout en linguistique, le terme *norme*, la notion de *norme de dépouillement* doit être comprise comme « une exigence de standardisation provisoire des textes contenus dans un corpus. Cette standardisation est destinée avant tout à les

rendre comparables, à les stabiliser le temps d'une expérience »[54]. Mais l'établissement d'une telle *norme de dépouillement* ne va pas sans susciter des difficultés de nature différente. L'une de ces difficultés est de satisfaire aux exigences parfois contradictoires de la linguistique et de la statistique et pour lesquelles « la norme devrait être acceptable à la fois pour le linguiste, pour ses auxiliaires, et pour le statisticien. Mais leurs exigences sont souvent contradictoires. L'analyse linguistique aboutit à des classements nuancés, qui comportent toujours des zones d'indétermination ; la matière sur laquelle elle opère est éminemment continue, et il est rare qu'on puisse y tracer des limites nettes […]. La statistique, dans toutes ses applications, ne va pas sans une certaine simplification des catégories ; elle ne pourra entrer en action que quand le continu du langage a été rendu discontinu »[55].

Dans cette perspective, une question importante s'impose, tant au niveau des principes méthodologiques sous-jacents qu'au niveau des objectifs fixés et qui ne peuvent échapper aux "pressions" parfois contradictoires du "double patronage" de la linguistique et de la statistique : Quels sont, dans chaque étude lexicométrique, les choix arrêtés et les décisions prises au moment du dépouillement lexical du corpus à étudier ?

Outre la rigueur scientifique exigée dans tout travail de recherche quant à la définition des unités utilisés, l'ensemble des choix et décisions formant la *norme lexicologique* vont pouvoir servir de base à toute étude contrastive lexicométrique ou stylométrique. Sans cette harmonisation des critères de dépouillement des textes, il serait difficile sinon impossible de pouvoir comparer deux corpus sur la base de leurs indices lexicométriques ou stylométriques respectifs, des similitudes/dissimilitudes qui peuvent exister entre eux, ni même pouvoir juger des éventuelles corrélations/dispersions décelées entre les éléments de chacun d'entre eux.

Il convient donc désormais de définir pour les études lexicométriques arabes une *norme lexicologique* comportant les règles de saisie et/ou d'harmonisation, la nature ou

[54] Habert B., Nazarenko A. et Salem A., *Les linguistiques de corpus*, 1997, p. 187

[55] Muller Ch., *Initiation aux méthodes de la statistique linguistique*, 1992, p. 113

les frontières des unités sur lesquelles porteront les décomptes, la définition des lemmes ou les limites, plus ou moins claires, des catégories lexicales auxquelles seront rattachées les formes du texte étudié. Le présent travail vise donc, dans l'un de ses trois axes principaux, à traiter cette question qui semble être demeurée dans l'ombre.

5. Les trois axes principaux de cette thèse

5.1. Le premier axe : La norme lexicologique

Alors que les recherches en lexicométrie française ou anglo-saxonne ont fait d'énormes progrès tant au niveau théorique et méthodologique que sur le plan empirique ou encore applicatif (informatique), les études lexicométriques arabes n'ont pas encore pris leur envol. Cette hésitation peut s'expliquer, nous semble-t-il, sur le plan méthodologique, entre autres, par l'absence d'une *norme de dépouillement lexicologique*[56] dans une perspective de traitement automatique de l'arabe en général, et de lexicométrie en particulier.

C'est pour apporter notre modeste contribution, sur le plan méthodologique, à l'essor de la lexicométrie arabe, que nous nous attachons dans le premier des trois axes de cette thèse, à proposer une *norme lexicologique* regroupant une panoplie de règles basées sur des considérations méthodologiques assorties, à chaque fois que cela s'avère nécessaire, de réflexions théoriques faisant appel à des concepts linguistiques propres à l'arabe, ou à des exigences imposées par le(s) formalisme(s) qui caractérise(nt) la recherche actuelle en TAL.

[56] Voir dans la deuxième partie de ce présent travail « Norme lexicologique », le chapitre 5 intitulé « Norme de dépouillement » p. 197-302.

5.2. Le deuxième axe : Le dictionnaire de fréquences

Le deuxième axe de cette thèse est la confection du *dictionnaire de fréquences* de toutes les unités lexicales composant le corpus étudié, fruit naturel de toute étude lexicométrique globale de cette nature, appliquée aux textes.

Le *dictionnaire de fréquences* traduit et synthétise les réorganisations formelles opérées sur la séquence textuelle d'origine, ainsi que le résultat des différentes analyses statistiques qui ont porté sur le vocabulaire du texte, sujet de l'étude lexicométrique (lemme, catégorie lexicale, fréquence globale, fréquence par partie, coefficient de spécificité, formes fléchies du lemme, etc.).

5.3. Le troisième axe : L'application de l'approche lexicométrique

Quant au troisième et dernier axe de cette thèse, il concerne l'application à notre corpus, le premier volume d'*al-ÞImtâÝ wa-l-MuÞânasa*, des méthodes de traitement, d'analyse et d'interprétation propres à l'approche lexicométrique, et ce dans le but d'étudier les caractéristiques lexicométriques de la structure du vocabulaire susceptibles de révéler des particularités ou de spécificités quantitatives du corpus.

La *richesse lexicale*, l'*accroissement du vocabulaire*, la répartition des catégories lexicales, la *connexion lexicale*, etc., représentent tant d'éléments et d'indices pouvant caractériser le style d'un auteur, d'un genre ou d'une époque. Tous ces éléments seront donc étudiés et analysés à la suite, bien entendu, d'une phase de prétraitement dans laquelle nous appliquerons à notre corpus-même la *norme de dépouillement lexicologique* que nous avons établie à cet effet et que nous soumettons à l'approbation des spécialistes de TAL arabe en général et de lexicométrie arabe en particulier.

À la suite de ces traitements/analyses, outre l'interprétation, raisonnée, des résultats basée sur les différents indices obtenus, un certain nombre de fichiers lexicométriques sous-jacents renfermant des données quantitatives importantes seront ainsi récupérés, à savoir : la distribution de fréquences (rapports établis entre les fréquences et les effectifs et entre les effectifs et les classes successives de fréquence), le tableau lexical entier (TLE), le vocabulaire spécifique du corpus entier et de ses parties, les *hapax* et les mots rares, la répartition des catégories lexicales, etc.

6. L'organisation générale de la thèse

Comment se présente la *trame radicale* d'*al-ÞImtâÝ wa-l-MuÞânasa et qu'est-ce qui la caractérise le plus par rapport à celles d'autres corpus ? Comment est organisée la structure lexicale de notre corpus et les différents faits qui la caractérisent : les principales caractéristiques lexicométriques*, la *richesse lexicale*, l'*accroissement du vocabulaire*, la *connexion lexicale* ? Mais avant tout, sur quelle base et selon quelles règles doivent être dépouillés les mots graphiques composant, au départ, le corpus ?

Pour répondre, entre autres, à ces questions et pour mener à bien ce travail selon les trois objectifs définis, nous avons structuré cette thèse en cinq parties :

Volume 1

6.1. La première partie

La première partie portera sur la présentation et la description des différents moments de l'analyse lexicométrique à savoir, la constitution du corpus, le dépouillement lexicologique, le traitement du corpus et enfin, l'analyse et l'interprétation des résultats.

Nous commencerons par la présentation, sous forme d'une chronologie abrégée, de l'auteur de notre corpus, *Abû Ḥayyân at-Tawḥîdî* et de son œuvre. Nous passerons ensuite en revue les étapes de la constitution du corpus et les difficultés qui l'ont accompagné, en présentant successivement le texte-source, l'enregistrement du texte en machine, son apurement des différentes erreurs de saisie, l'harmonisation primaire ainsi que le repérage et la ligature des noms propres et des unités polylexicales. Après cette phase initiale de la constitution du corpus, nous décrirons le moment critique du dépouillement lexical et de ses étapes qui sont la segmentation, la lemmatisation, la catégorisation et l'encodage. Nous n'omettrons pas de définir à chaque fois ces opérations, d'en décrire le déroulement et de présenter les outils utilisés pour les effectuer (**chapitre 1**).

Nous nous attacherons ensuite à présenter quelques données statistiques concernant la différence entre le dépouillement segmenté et le dépouillement en mots graphiques. Le but de ces calculs est, d'un côté de révéler l'impact de la segmentation aussi bien sur l'étendue du corpus que sur son vocabulaire, et de l'autre côté de présenter la distribution des unités lexicales résultantes eu égard aux différents types de mots graphiques. La productivité segmentale des mots graphiques représente un bon indicateur de la restructuration du corpus (**chapitre 2**).

Nous continuerons la présentation des moments de l'analyse lexicométrique par la description de la phase du traitement et de l'analyse du corpus, dans laquelle nous définirons la quantification d'un corpus et présenterons le logiciel de traitement lexicométrique que nous avons utilisé, *Lexico3*[57]. La récupération des données quantitatives dans *Excel* puis la constitution d'une base de données sous *Access*, seront également décrites. Nous terminerons enfin cette partie par la description de la phase de confection du dictionnaire de fréquences ainsi que de l'utilisation d'autres outils pour la présentation des résultats de l'analyse lexicométrique (**chapitre 3**).

---

[57] *Lexico3* est un logiciel développé par André Salem, Serge Fleury, Cédric Lamalle et William Martinez. Le site officiel de *Lexico3* est : http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/lexico3.htm

6.2. La deuxième partie

La deuxième partie nous donnera l'occasion de présenter le premier des trois axes centraux de notre travail : la *norme lexicologique* que nous proposons aux spécialistes ou aux futurs spécialistes de lexicométrie, arabisants ou arabophones, mais aussi à tous ceux qui s'intéressent d'une manière ou d'une autre au Traitement Automatique de la Langue arabe ou à la linguistique de corpus.

Étant donné que le prétraitement des données textuelles s'opère en deux moments : le moment d'enregistrement du texte en machine, de son apurement et de son harmonisation, et le moment consistant à restructurer le corpus en délimitant et regroupant ses unités, à lever les ambiguïtés et à opérer de nouvelles codifications sur les unités ainsi obtenues, cette *norme lexicologique* comportera donc deux volets : le premier est une *norme de saisie et d'harmonisation* comportant un certain nombre de règles à suivre au moment de saisir le texte ainsi que des règles d'harmonisation aussi bien primaire que régulatrice (**chapitre 4**).

Le deuxième volet est une *norme de dépouillement* incluant la définition, les critères et les règles de chacune des quatre grandes opérations qui composent ce que l'on appelle *l'analyse lexicale* : il s'agit du processus qui, analysant le texte, va définir les limites des unités de décompte (*segmentation*), assigner le lemme (*lemmatisation*) et la catégorie lexicale (*catégorisation*) à chaque unité en fonction de ses voisins (*désambiguïsation*) (**chapitre 5**).

Volume 2

6.3. La troisième partie

La troisième partie s'intéressera principalement à l'étude de la *trame radicale* du corpus étudié. Une assise théorique sera posée et des considérations méthodologiques précisées.

Nous nous arrêterons d'abord sur la notion de racine et sur son statut dans la langue arabe, en évoquant les différentes définitions données à cette notion ainsi que les différentes positions quant au problème de dérivation des unités lexicales à partir de la racine. Nous décrirons ensuite les décisions que nous avons prises pour l'extraction des racines de notre corpus. Et nous présenterons enfin les différentes méthodes de calcul des racines théoriquement possibles aussi bien par les grammairiens arabes classiques, que par les deux méthodes des *arrangements avec répétition* et des *arrangements sans répétition* utilisées en analyse combinatoire (**chapitre 6**).

Nous examinerons ensuite, parmi ce nombre gigantesque des racines théoriques, la part exacte des racines attestées dans la langue arabe, le ***mustaÝmal***, et celle des racines possibles mais non attestées, le ***muhmal***. En plus du nombre donné par le lexicographe *Az-Zubaydî (m. 379/989)* concernant les racines attestées, nous examinerons en détail, non seulement le nombre de ces racines, mais aussi la distribution interne des différentes racines groupées selon la première radicale ainsi que l'ordre décroissant de ces groupes de racines selon leur fréquence, et ce à l'intérieur de la base de connaissances DIINAR.1[58], et de deux dictionnaires classiques, *Lisân al-Ýarab* et *AÒ-ÑiÎâÎ* (**chapitre 7**).

Nous entreprendrons dans le dernier chapitre de cette partie, d'étudier la trame radicale d'*al-ÞImtâÝ wa-l-MuÞânasa*. La distribution interne des différents groupes de racines, selon la première radicale mais également selon le nombre des consonnes radicales, sera étudiée. Sera également examinée la productivité des groupes de racines

[58] DIINAR : **DI**ctionnaire **IN**formatisé de l'**AR**abe.

en formes mais aussi en occurrences. Nous calculerons le coefficient général de productivité des racines, celui de productivité des racines triconsonantiques ainsi que celui de productivité des racines quadriconsonantiques. Le calcul de ces coefficients se fera également à l'intérieur des racines groupées selon la première radicale. Une brève comparaison sera ensuite faite entre les racines les plus productives et les racines les plus fréquentes.

Nous nous attacherons ensuite à étudier les racines dans d'autres corpus, à savoir : les *Maqâmât al-HamaÆânî* (de Badiʿ Az-Zamân Al-HamaÆânî 358-398/969-1008), *Al-ÞAdab al-kabîr* (d'Ibn Qutayba 213-276/828-889), et les racines verbales dans *Le Coran* ; ceci pour arriver en fin de parcours à apporter quelques éléments de comparaison d'abord entre les corpus eux-mêmes *Al-ÞImtâʿ wa-l-MuÞânasa* (IV^{e}/X^{e} siècle), *Maqâmât al-HamaÆânî* (IV^{e}/X^{e} siècle) et *Al-ÞAdab al-kabîr* (III^{e}/IX^{e} siècle), et ensuite entre chacun d'entre eux et chacun des deux dictionnaires et la base de ressources lexicales DIINAR étudiés auparavant (**chapitre 8**).

6.4. La quatrième partie

La quatrième partie sera consacrée au troisième axe principal de notre thèse, l'étude de la *structure lexicale* de notre corpus *Al-ÞImtâʿ wa-l-MuÞânasa*. Dans cette partie qui correspond à l'application directe des méthodes et de la démarche lexicométriques, nous commencerons par présenter les principales caractéristiques lexicométriques du corpus ainsi que de ses parties, les *Nuits*[59].

[59] À l'image des *Mille et une nuits*, *Al-ÞImtâʿ wa-l-MuÞânasa* [*Le plaisir offert et la sociabilité partagée* (trad. de F. Lagrange)] est organisé en 40 Nuits (le premier volume comportant 15 Nuits plus le préambule). Dans ce livre, Abû Ḥayyân at-Tawḥîdî (932-1024), compile à la demande de son ami et protecteur Abû l-Wafâʾ al-Muhandis, le compte-rendu des entretiens nocturnes qu'il avait eu avec le vizir, Ibn Saʿdân (m. en 985), du prince bouyide Ṣamṣâm ad-Dawla (964-998). Ces entretiens autour de sujets très variés de nature littéraire, philosophique, scientifique, etc., forment la matière de ce livre.

Par *principales caractéristiques lexicométriques du corpus* (PCLC), il faut entendre : l'étendue du corpus et celle des *Nuits*, l'étendue du vocabulaire du corpus et celle du vocabulaire de chacune des *Nuits*, la fréquence maximale du corpus et celle de chaque partie, les lemmes correspondant à ces fréquences maximales, les *hapax*, la répartition générique, etc.

Nous montrerons l'existence d'une corrélation significative entre le rang de l'étendue des *Nuits* et de celui de l'étendue de leur vocabulaire en utilisant un test statistique, le coefficient de corrélation des rangs de Spearman ; et nous présenterons, dans un tableau, tous les paramètres de la description statistique des données quantitatives relatives aux *Nuits*. Nous décrirons ensuite la répartition générique (prose et poésie) et intertextuelle (citations coraniques, prophétiques et poétiques) des unités lexicales du corpus (**chapitre 9**).

Nous nous attarderons ensuite sur un aspect fondamental de l'étude des textes dans une perspective lexicométrique : l'étude de la *richesse lexicale*. Après avoir défini la notion de richesse lexicale et exposé brièvement les problèmes qu'elle soulève et les solutions proposées, nous présenterons les raisons qui nous amené à choisir cinq des méthodes de mesure de la richesse lexicale existantes.

Nous commencerons ensuite par appliquer à notre corpus, chacune des cinq méthodes retenues : la méthode de comparaison des indices, la formule de Guiraud ($V/\sqrt{N}$), l'indice W de Brunet, la méthode binomiale de Muller et l'indice $\boldsymbol{V_m}$ de Yule-Herdan. Pour chaque méthode, nous présenterons d'abord son bien-fondé théorique, son assise méthodologique et le déroulement de ses étapes, avant de l'appliquer à l'*ImtâÝ wa-l-MuÞânasa*.

En guise de bilan, nous comparerons les résultats obtenus par l'application des quatre méthodes (la méthode des indices étant écartée), en évaluant par le test de Spearman, les corrélations des rangs et les liens qui pourraient exister entre les

différents classements. Nous montrerons que les corrélations les plus fortes sont, par ordre décroissant, entre les méthodes (*Muller, Yule-Herdan*), (*Muller, Brunet*) et (*Brunet, Guiraud*).

Pour affiner cette analyse contrastive, nous utiliserons l'analyse factorielle des variables latentes du classement des 15 *Nuits* par les quatre méthodes. Après l'interprétation des deux axes de la représentation graphique de l'analyse factorielle, l'axe F1 (60 % de l'inertie) et l'axe F2 (40 % de l'inertie), et l'interprétation globale, nous tenterons de dégager quatre sous-ensembles classés selon la richesse lexicale des *Nuits* qu'ils regroupent. En classant les *Nuits* à l'intérieur de chacun des sous-ensembles, nous arriverons de proche en proche à un classement, inféré à partir de la représentation graphique, qui serait la résultante des classements des quatre méthodes. Nous comparerons ensuite chaque classement calculé à ce classement inféré.

À la suite de la mesure de la richesse lexicale des *Nuits*, nous calculerons la richesse lexicale du corpus entier selon trois de ces méthodes (la méthode Muller ne s'appliquant qu'à des parties d'un corpus) afin de la comparer à la richesse lexicale d'un autre corpus, celle des *Maqâmât* d'*al-Hamaδânî* sur la base de calculs faits par nous à partir des données puisées dans la thèse de Chabir Ayadi. Nous montrerons qu'*al-ÞImtâÝ wa-l-MuÞânasa* est plus riche lexicalement que les *Maqâmât* d'*al-Hamaδânî* et ce selon chacune des méthodes de mesure utilisées confirmant ainsi, à la fois, cette tendance et la validité des méthodes choisies. En guise de synthèse, nous conclurons ce chapitre par des suggestions quant à l'utilisation de telle ou telle méthode pour la mesure de la richesse lexicale dans les études lexicométriques arabes (**chapitre 10**).

Nous nous consacrerons ensuite à l'étude d'un autre fait, dynamique cette fois-ci, de la structure lexicale. Il s'agit de l'*accroissement du vocabulaire*. Il sera donc question, avec la notion d'*accroissement du vocabulaire*, de voir la façon dont l'étendue

du vocabulaire (V)[60] se constitue au fur et à mesure que le texte croît jusqu'à son dernier mot. Nous étudierons d'abord l'accroissement réel du vocabulaire (général et par classe de fréquence), nous calculerons ensuite l'accroissement théorique du vocabulaire d'*al-ÞImtâÝ wa-l-MuÞânasa* selon la loi binomiale proposée par Charles Muller, pour enfin comparer les deux courbes d'accroissement du vocabulaire réel et théorique et les deux courbes d'accroissement cumulé du vocabulaire réel et théorique. Le but sera d'observer les fluctuations, à chaque fois, entre les deux courbes afin d'évaluer et d'interpréter les écarts des valeurs observées par rapport aux valeurs théoriques (**chapitre 11**).

Nous nous attarderons par la suite, sur un fait d'une importance capitale qui est la répartition des mots du corpus en catégories lexicales. Cette répartition ajoute au panorama structurel dessiné par les autres faits déjà étudiés, un aspect relatif au contenu lexical.

Nous commencerons par présenter les données relatives à la lexicalité et à la fonctionnalité aussi bien au niveau du corpus qu'au niveau de ses parties. Nous examinerons ensuite la répartition des catégories lexicales de base au niveau du corpus, mais également celle des sous-catégories au sein de chacune des catégories de base. Les mêmes répartitions seront également examinées mais cette fois-ci, au niveau des *Nuits*.

Nous entreprendrons ensuite de calculer les effectifs théoriques des catégories lexicales ainsi que les écarts entre eux et les effectifs réellement observés dans le corpus pour pouvoir évaluer, pour chaque catégorie lexicale, les écarts significatifs (positivement ou négativement). Nous nous attacherons ensuite à déceler les écarts significatifs mais dans une autre perspective qui consistera à examiner chaque *Nuit* en fonction de l'excédent ou du déficit qu'elle peut avoir en telle ou telle catégorie lexicale.

---

[60] À la différence de l'étendue du corpus (N) définie comme étant le nombre de toutes les occurrences des mots d'un corpus, l'étendue du vocabulaire (V) représente, quant à elle, le nombre des mots différents d'un corpus, donc le nombre de ses vocables.

Dans le but d'étudier le système d'opposition/similitude qui règle le jeu des catégories lexicales dans notre corpus, nous appliquerons un certain nombre de tests de corrélation de Pearson, d'un côté aux effectifs réels et aux effectifs théoriques des catégories lexicales, et de l'autre côté aux écarts réduits entre effectifs réels et effectifs théoriques afin de comparer, d'abord les classes lexicales entre elles, et ensuite les *Nuits* entre elles. Nous appliquerons enfin l'analyse factorielle des correspondances des 15 *Nuits* par les 5 catégories lexicales de base (*Verbes*, *Noms primitifs*, *Noms dérivés*, *Adjectifs* et *Mots-outils*), pour interpréter en dernier la représentation graphique issue de cette analyse (**chapitre 12**).

Nous terminerons cette quatrième partie par l'étude d'un fait de structure, qui est par ailleurs relatif au contenu, il s'agit de la connexion lexicale. Nous présenterons la méthode de calcul, les données concernant la connexion lexicale des vocables et celles relatives à la connexion lexicale des occurrences. Nous terminerons par une présentation et une interprétation des dix couples de *Nuits* ayant la connexion lexicale la plus forte et ceux ayant la connexion lexicale la plus faible (**chapitre 13**).

Volume 3

6.5. La cinquième partie

Représentant le cœur de notre travail, la cinquième et dernière partie de cette thèse, quant à elle, nous permettra de présenter le fruit de la démarche lexicométrique, et correspondant au deuxième objectif principal que nous nous sommes fixé. Il s'agit du *dictionnaire de fréquences* hiérarchique total.

Synthétisant la majorité des données quantitatives relatives aux différentes réorganisations formelles opérées sur la séquence textuelle d'origine de notre corpus, le *dictionnaire de fréquences* représente également le résultat d'un certain nombre de traitements statistiques qui ont porté sur le vocabulaire du texte. Y figurent : chaque lemme du corpus, le code de la catégorie lexicale correspondante, sa fréquence globale dans le corpus, la fréquence du lemme dans chacune des parties du corpus, le coefficient de spécificité du lemme dans chacune des parties, les différentes formes fléchies du lemme en question, le numéro de la page dans laquelle la forme fléchie est rencontrée et enfin, la fréquence de la forme dans la page.

Volume d'annexes

Dans ce volume, nous présentons quelques annexes mentionnées ici et là dans la thèse. La première annexe présente quatre listes, celles des proclitiques et des enclitiques, simples et composés (**Annexe A**). La deuxième présente la liste des verbes concaves homographes (**Annexe B**). L'**annexe C** présente la liste des *maÒdar* et pluriels homographes. Nous présentons dans l'**Annexe D**, les 9911 racines avec, pour chaque racine son effectif et sa fréquence dans *Al-ÞImtâÝ wa l-MuÞânasa*. Quant à l'**Annexe E**, elle présente les racines trilitères classées par fréquence décroissante et le nombre de formes produites par chacune d'entre elles. Le même tableau est présenté dans l'**Annexe F** mais pour les racines quadrilitères. Les principales caractéristiques lexicométriques des catégories lexicales sont présentées dans l'**Annexe G**, à l'exception des spécificités du

vocabulaire qui sont, quant à elles, présentées dans l'**Annexe G**[bis]. L'**Annexe H** enfin comporte, *dans la colonne de gauche*, le texte de référence d'*Al-ImtâÝ wa-l-MuÞânasa,* le même texte segmenté dans la colonne de milieu et celui lemmatisé dans la colonne de droite.