Conclusion générale

« Le savant n'est pas l'homme qui fournit les vraies réponses ; c'est celui qui pose les vraies questions »

Claude Levi-Strauss

Notre objectif principal était triple. Élaborer une *norme lexicologique* relative à la saisie, l'harmonisation et le dépouillement des textes arabes dans une perspective de lexicométrie en particulier et de traitement automatique en général. Confectionner, pour notre corpus *Al-ÞImtâÝ wa-l-MuÞânasa*, le dictionnaire de fréquences, condensé de toutes les applications lexicométriques entreprises. Et enfin, appliquer à ce même corpus un certain nombre de méthodes d'analyse lexicométrique dans le but d'en étudier la structure lexicale ainsi que la trame radicale.

Outre l'intérêt de la démarche descriptive et exploratoire d'une partie de ces méthodes, la possibilité de réutilisation des résultats obtenus dans des recherches contrastives n'en est pas moins importante.

Au terme de ce travail que nous avons structuré en cinq parties servant les trois axes principaux, nous espérons avoir atteint nos trois objectifs.

Nous avons pu mettre en place une *norme lexicologique* que nous jugeons stable, composée de deux volets correspondant aux deux moments du prétraitement des données textuelles. Le premier volet est consacré à la saisie et à l'harmonisation des textes arabes en vue d'un traitement automatique. Le second volet, quant à lui, se rapporte au dépouillement *stricto sensu* et concerne les aspects de *segmentation*, de *lemmatisation* et de *catégorisation*. Cette *norme de dépouillement* comprend également des règles et des choix correspondant à la phase de *désambiguïsation* qui est une opération pouvant accompagner, précéder ou succéder à chacune des trois autres.

En nous inspirant des percées ouvertes par les autres langues dans le domaine du TAL – considération évidemment faite des spécificités de chaque langue naturelle – nous avons essayé d'établir la *norme lexicologique* dans une relation dialectique entre

les exigences théoriques de la langue arabe, ses règles et ses mécanismes, d'un côté, et les considérations pratiques, les différents contextes d'application et les difficultés que nous avons rencontrées lors du dépouillement de notre corpus, de l'autre côté.

Les formalismes établis par les linguistiques contemporaines et leurs multiples applications dans le domaine du TAL nous ont inspiré sinon guidé dans cette démarche ; formalismes qui n'entrent bien entendu pas en contradiction avec la nature de la langue arabe. Celle-ci semble se prêter aisément – et bon nombre de linguistes arabisants modernes le prouvent (D. Cohen, J. Dichy, A. Braham, C. Audebert, …) – à ces formalismes notamment au niveau de l'analyse du mot graphique arabe et de ses prolongements dans plus d'un domaine du TAL.

Dans le prolongement de cette *norme lexicologique*, nous avons également proposé une traduction en arabe, des termes *lemme* [أَصْلَم], *lemmes* [أَصَالِم], *lemmatiser* [أَصْلَمَ / يُؤَصْلِمُ], *lemmatisé* [مُؤَصْلَم], *lemmatisation* [أَصْلَمَة] et *lemmatiseur (automatique)* [مُؤَصْلِم (آلِي)]. Ces équivalents arabes pourraient contribuer au développement et à la diffusion dans l'espace arabophone, des études lexicométriques et du TAL arabes.

Nous avons pu observer la distribution interne des racines dans *Al-ÞImtâÝ wa-l-MuÞânasa*. Le diagramme de Pareto de la production des formes nous indique que 21 % des formes différentes sont produites par 61 % des racines. Celui de la production des occurrences révèle que 22 % des occurrences sont produites par 54 % des racines. Le coefficient général de productivité dans notre texte, est de 4,64 pour les formes et de 11,95 pour les occurrences. Notons que ce coefficient n'atteint, chez Al-HamaÆânî, que 2,57 pour les formes et 6,6 pour les occurrences (Chabir 1997). Nous avons pu rendre compte également de la disparité entre les racines les plus fréquentes et les racines les plus productives.

Il nous a donc été possible de montrer comment, à partir de tous ces éléments et indices, se dessinent au fur et à mesure, les caractéristiques qui distinguent la trame radicale d'*al-ÞImtâÝ wa-l-MuÞânasa*. À partir de là, nous avons pu établir différentes comparaisons, sur la base de la distribution des racines, d'une part entre *Al-ÞImtâÝ wa-l-MuÞânasa* , *Maqâmât al-HamaÆânî* et *Al-ÞAdab al-kabîr*, et d'autre part, entre eux et

la base de ressources lexicales DIINAR.1 et deux grands dictionnaires arabes médiévaux : *Lisân al-ÝArab* et *AÒ-ÑiÎâÎ*.

Ces comparaisons pourraient constituer un début à des études contrastives, sur la base de la trame radicale, appliquées à d'autres textes de l'œuvre d'At-TawÎîdî ou à d'autres œuvres de la même époque ou d'époques différentes.

Nous avons longuement examiné la structure lexicale du corpus. Un certain nombre d'analyses, dont quelques unes utilisées pour la première fois sur des textes arabes, ont été opérées pour faire parler les faits de structure qui caractérisent cette œuvre. Le but était d'éprouver l'application à un texte arabe de quelques procédés de lexicométrie et d'en mesurer l'efficacité.

Pour l'étude de la richesse lexicale par exemple, nombreuses sont les méthodes proposées pour apporter une solution objective, à un problème auquel les réponses n'ont été, pendant longtemps, que subjectives, approximatives et impressionnistes. De toutes ces méthodes, nous avons retenu cinq que nous avons jugées les plus adéquates et les plus fiables et nous les avons appliquées.

On pourrait formuler une objection quant au bien-fondé d'utiliser, pour des textes arabes, des méthodes de mesure de la richesse lexicale élaborées principalement à partir d'autres langues. Mais, ce qu'il faut savoir c'est que ces différentes méthodes utilisent des formules statistiques manipulant incontestablement des chiffres et des indices (étendues, fréquences, effectifs, racines carrées, variances, écart-types, etc.). Ces chiffres traduisent la distribution des unités du vocabulaire dans le corpus et non pas la nature des ces unités. La richesse lexicale, rappelons-le, est un fait de structure et non de contenu. Ce qui est donc manipulé dans le calcul des indices de richesse lexicale, ce sont des valeurs chiffrées totalement indépendantes de tout contenu lexical et de toute langue.

Le bilan contrastif des méthodes de mesure, complété, en fin de parcours, par une analyse factorielle, nous a offert la possibilité, non seulement d'inférer un classement final des différentes parties du corpus sur la base de la richesse lexicale, mais aussi de comparer la richesse lexicale d'*al-ÞImtâÝ wa-l-MuÞânasa* à celle des *Maqâmât al-HamaÆânî*. Il a révélé que le premier texte est plus riche lexicalement que

le second et ce selon les trois méthodes de mesure utilisées. Il n'est malheureusement pas possible, avec les méthodes de mesure existantes actuellement (et ce pour toutes les langues), de dire selon quelle composante du vocabulaire, stylistique et/ou thématique, *Al-ÞImtâÝ wa-l-MuÞânasa* est plus riche que *Maqâmât al-HamaÆânî*. Cette tâche pourrait être accomplie en faisant appel à d'autres types analyses stylométriques que nous nous proposons de faire dans un futur proche. En guise de synthèse de ce bilan, nous avons proposé des suggestions quant à l'utilisation des méthodes de mesure de la richesse lexicale des corpus arabes.

Un autre fait nous a permis de décrire la structure du vocabulaire d'*al-ÞImtâÝ wa-l-MuÞânasa* : l'*accroissement lexical*. En effet, après avoir exposé l'accroissement réel du vocabulaire de notre corpus tant au niveau général que par classe de fréquence et après en avoir commenté la distribution des valeurs, nous avons calculé, pour chaque partie du corpus, l'accroissement théorique du vocabulaire selon la méthode de calcul établie par Ch. Muller et basée sur la loi binomiale. Nous nous sommes ensuite attaché à évaluer la distance, pour chaque partie du corpus, entre l'accroissement réel et l'accroissement théorique. Ceci nous a donné la possibilité de classer les *Nuits* d'après les déviations observées par rapport à l'accroissement théorique du vocabulaire.

Ce classement a été enfin comparé à ceux que fournissent d'un côté l'ordre chronologique des parties, et de l'autre côté l'étendue relative du vocabulaire. Pour ce qui est de la comparaison, par exemple, entre le classement selon l'accroissement et celui selon l'ordre chronologique des *Nuits*, il serait légitime de remettre en question l'ordre, chronologique, d'écriture des *Nuits* 10, 15 et 16. Ces *Nuits*, ont-elles vraiment été rédigées dans l'ordre qui correspond à leurs numéros respectifs ? Ceci, nous semble-t-il, est loin d'être acquis.

L'étude de la distribution des mots du corpus en catégories lexicales a représenté un moment d'analyse d'une importance capitale dans ce travail. En plus de son aspect structurel, cette répartition apporte une vision imprégnée de contenu lexical. Au niveau du rapport entre *lexicalité* et *fonctionnalité*, les mots lexicaux représentent 40 % de l'ensemble des occurrences du corpus alors que les mots-outils en représentent 60 %[322].

[322] Il n'est pas sans intérêt de noter ici que ces comptages sont opérés sur des textes segmentés et non sur des mots graphiques.

Ceci confirme, à une échelle réduite, une tendance générale vérifiée dans pratiquement toutes les langues à partir de grands corpus de textes écrits.

La prééminence des *particules* non seulement sur toutes les autres catégories des mots-outils[323] mais aussi sur toutes les catégories lexicales au niveau du corpus n'est pas un fait surprenant. En revanche, la primauté au niveau des catégories des mots lexicaux, de la catégorie des noms primitifs sur les verbes, les adjectifs et les noms dérivés est un indice qui est fortement discriminant du style dans *Al-ÞImtâÝ wa-l-MuÞânasa*. Il est à noter que, dans *Maqâmât al-HamaÆânî*, les noms arrivent en deuxième position après les verbes (Chabir 1997).

Outre les effectifs réellement observés dans le corpus, nous nous sommes employé à construire un modèle théorique de distribution des catégories lexicales par le calcul des effectifs théoriques. Nous avons fait ensuite une comparaison entre les effectifs observés et les effectifs théoriques dans le but d'évaluer les écarts significatifs entre effectifs réellement observés et effectifs calculés car ces écarts représentent des caractéristiques discriminantes du style de l'auteur. Cela nous a permis de voir, par exemple, que les verbes et les noms primitifs sont largement au dessus de la moyenne dans la *Nuit* 10, alors que les adjectifs, les noms dérivés et les mots-outils y sont extrêmement sous-utilisés. À l'inverse, il nous a été également possible de passer en revue, une par une, toutes les *Nuits* afin de juger de leur déficit ou de leur excédent en telle ou telle catégorie lexicale.

L'étude de la structure du vocabulaire d'*al-ÞImtâÝ wa-l-MuÞânasa* nous a permis, somme toute, de dégager ce qui caractérise le plus, quantitativement, le style d'Abû Íayyân at-TawÎîdî. L'on trouve, parmi ces caractéristiques, le vocabulaire spécifique positif et négatif, les indices permettant d'estimer la richesse lexicale de tout le corpus et des *Nuits*, les déviations observées entre l'accroissement réel et l'accroissement théorique du vocabulaire, les rapprochements/éloignements thématiques qui existent entre les *Nuits* composant *Al-ÞImtâÝ wa-l-MuÞânasa* et enfin, la répartition des catégories lexicales au sein du corpus ainsi que le déficit ou l'excédent qui distingue chacune des *Nuits* en telle ou telle catégorie lexicale.

[323] Voir pp. 260-302.

Il convient également de souligner l'importance primordiale qu'a le dictionnaire de fréquences que nous avons pu confectionner au terme de ce travail et qui représente le condensé des informations quantitatives relatives à toutes les unités du corpus, formes ou lemmes. Sa consultation, que nous souhaitons aisée, permet non seulement d'avoir à la portée, les informations recherchées concernant telle ou telle unité du vocabulaire d'*al-ÞImtâÝ wa-l-MuÞânasa* , mais aussi de faciliter le retour aux contextes de l'unité dans le corpus par le biais des indications concernant les parties, les pages et les fréquences du mot dans chaque partie et par page.

Cette étude lexicométrique se doit d'être élargie à toute l'œuvre d'Abû Íayyân at-TawÎîdî, voire même à d'autres auteurs de son époque et à des époques différentes. Multiplier et approfondir les travaux lexicométriques de cette nature pourrait, non seulement contribuer à la constitution d'une banque de données textuelles arabe renfermant des corpus à la fois bruts et annotés, mais aussi accélérer la mise en place de travaux substantiels pour la confection du dictionnaire historique de l'arabe tant attendu.

En outre, bien qu'il renferme quelques comparaisons partielles, notre travail ne s'inscrit pas dans une perspective de lexicométrie contrastive. Comme nous l'avions déclaré d'emblée, notre démarche dans ce travail reste purement et simplement une démarche descriptive. Néanmoins, les analyses entreprises, les résultats obtenus et les indices calculés pourraient bien constituer à l'avenir de bonnes bases pour des comparaisons globales dans le cadre d'une perspective de lexicométrie ou de stylométrie contrastives. Dans ce cadre, les comparaisons les plus tentantes, dans le contexte de l'arabe classique, seraient celles qui mettraient vis-à-vis du vocabulaire d'at-TawÎîdî (dans l'ensemble de son œuvre) celui de certains de ses prédécesseurs (Al-JâÎiÛ [776-869], …), de ses contemporains (Ar-Rummânî [908-994], Miskawayh [932-1030], As-Sijistânî [m. après l'année 1000], …) et de ses successeurs (Al-MaÝarrî [973-1058], Ibn Íazm [993-1064], Al-Çazâlî [1058-1111], ...). Ceci pourrait permettre, par exemple, de retracer la formation du vocabulaire du *Þadab* relatif à ces trois générations, et de mesurer, entre autres, la part d'at-TawÎîdî dans cette élaboration.

La norme lexicologique que nous proposons dans ce travail, est principalement établie sur la langue générale. Une prise en considération future de langues de spécialités (vocabulaire économique, juridique, médical, …) ne peut que l'enrichir en

l'ouvrant au domaine de la terminologie avec les multiples applications que l'on connaît.