LES CAHIERS DE L'ANALYSE DES DONNÉES

J.-P. BENZÉCRI

Sur l'étude des textes arabes d'après les occurrences des formes de mots

Les cahiers de l'analyse des données, tome 19, n° 1 (1994), p. 65-84

<http://www.numdam.org/item?id=CAD_1994__19_1_65_0>

Les Cahiers de l'Analyse des Données
Vol. XIX 1994 - n°1 - pp.65-84.

SUR L'ÉTUDE DES TEXTES ARABES D'APRÈS LES OCCURRENCES DES FORMES DE MOTS

[MOTS ARABES]

J.-P. BENZÉCRI

0 Introduction: analyse philologique et analyse automatique

La stylométrie a déjà été appliquée aux textes arabes; et, dans certaines études, des tableaux de correspondance ont été construits, puis analysés. Le travail le plus important, dans ce genre, est la grande thèse soutenue à Paris par un Professeur de Beyrouth, Anis Abi FARAH. Un essai plus modeste, [STYLE ARABE], a paru dans le *Cahier* offert en hommage à Étienne ÉVRARD: *CAD*, Vol.XIII, n°1, (1988).

Cette dernière étude, est fondée sur la partie publiée de relevés étendus, effectués par le Pr. Gérard LECOMTE, en distinguant les catégories syntaxiques et les fonctions des mots: c'est-à-dire, en appliquant à la langue arabe une méthode analogue à celle de recherches poursuivies à Liège pour le Grec et le Latin: e.g., dans les articles [FRÉQ. CAT. LATIN] et [MÉT. ARISTOTE] du *Cahier* déjà cité. Bien qu'il procède par une autre voie, A. Abi FARAH recourt, lui aussi, à une élaboration préalable des textes, tout autre qu'automatique, mettant en jeu la compétence grammaticale du statisticien. Aussi dirons-nous que la stylométrie de l'Arabe, a recouru jusqu'à présent, à des analyses philologiques.

Au contraire, des corpus littéraires en Grec, Latin et Espagnol, ainsi que des articles et documents en Français, ont pu être analysés en traitant les textes bruts par des algorithmes de dénombrement de formes, dont la conception et la mise en œuvre ne requiert du statisticien que la saisie des textes, sans aucune élaboration lexicale ou grammaticale; (cf., e.g., [TEXTES GRECS, 1 & 2], Vol. XVI, n°1, et XIX, n°2; [TEXRES LATINS], Vol.XVI, n°4; [SIÈCLE d'OR], Vol.XVII, n°4; [*CAD* XII-XVII, 1-3], Vol.XVIII, n°1).

Même si le choix du lexique des formes à dénombrer peut être guidé par la connaissance de la langue, le dénombrement lui-même est purement

Les cahiers de l'analyse des données 0339-3097/94/01/65 20/$ 4.00/© Gauthier-Villars

mécanique: une forme étant définie comme un segment minimal de texte délimité par des espaces blancs ou des signes de ponctuation.

Il est facile de critiquer l'usage de telles unités: des segments identiques peuvent représenter des formes différentes: soit qu'il s'agisse, simplement, e.g., d'un même verbe, mais avec des associations {mode-temps-personne} différentes; soit, ce qui est plus grave, de sens et de catégories syntaxiques distincts: comme pour 'part', dans 'il part' et 'la part'.

On défendra les dénombrements de formes en alléguant, d'abord, que, tout en s'étendant facilement à de grands corpus, ils ont montré leur pertinence dans de nombreuses études typologiques. On poursuivra en critiquant tous les dénombrements, plus coûteux, qu'on est tenté de leur préférer: car, par exemple, la lemmatisation, ou dénombrement des formes sous les mots du lexique, réduit à 'être' les trois formes {est, soit, soyons}; dont les deux dernières évoquent certains contextes que la première n'évoque aucunement.

Il faut concéder que le dénombrement des formes accepte une hiérarchisation contingente des éléments du langage: celle qu'a adoptée l'écriture du discours; mais la structure du système que proposent grammaire et lexique n'est pas, elle non plus, dépourvue d'arbitraire. Reconnaissons qu'ici comme là, on s'efforce de saisir l'insaisissable.

Mais, prêt à accepter, avec gratitude, les données, de toute forme, que des linguistes diligents auront recueillies...; nous nous appliquons, particulièrement, à acquérir une vue globale de l'expression verbale fondée sur l'analyse de purs dénombrements de formes.

Reste que la valeur linguistique des formes, mécaniquement définies, varie selon les langues: variation d'ailleurs liée à celle qu'on trouve entre les grammaires et lexiques adoptés pour ces langues. Sans approfondir, en termes de linguistique générale, le problème du mot, on découvre, à première vue, que si les noms, singuliers ou pluriels, les verbes à la conjugaison étendue, les prépositions, les pronoms etc., donnent au français et à l'italien des inventaires de formes analogues, la conjugaison réduite de l'anglais, ou les déclinaisons du russe engendrent de tout autres systèmes.

Sans expérimentation préalable, on ne peut donc affirmer que les succès obtenus en analysant des textes grecs, se répéteront quand on appliquera la même méthode à des textes arabes.

Le présent article rend compte de l'état des recherches effectuées sur un corpus restreint de tels textes.

1 Ecriture de l'arabe et dénombrement des formes

Ce n'est pas le lieu d'enseigner l'écriture de l'arabe: cependant, sans connaître, de cette écriture, certaines règles qui se reflètent dans l'enregistrement d'un fichier de texte arabe, on ne peut saisir la structure du programme qui passe d'un tel fichier à une liste ordonnée de formes de mots; ni, par conséquent, les particularités que présente cette liste, du seul point de vue linguistique.

Dans ce §, les signes arabes seront figurés selon la police naskh du Macintosh, (modifiée pour être acceptée par tous les programmes de traitements de texte); et accompagnés de la valeur des octets qui leur correspondent dans la police courrier.

1.1 L'alphabet arabe et ses ligatures

Bien que fondée sur un alphabet, l'écriture de l'arabe diffère notablement de celle du latin ou du français. À première vue, un texte arabe imprimé se signale par une souplesse de tracé qui évoque la calligraphie manuscrite. Il n'y a pas de distinction entre minuscules et capitales - bas de casse et haut de casse, disent les typographes - mais, en règle générale, chaque lettre admet quatre graphies: isolée, initiale, médiane et finale; certaines lettres, toutefois, ne se liant jamais à gauche (i.e. à la lettre qui les suit).

ففف	فـ	ف	ÿµí	ÿ µ í	m
قققق	قـ	ق	Ÿðì	Ÿ ð ì	n
ككك	كـ	ك	⁄Σî	⁄ Σ î	o
لل	لـ	ل	¤Πï	¤ Π ï	p
مم	مـ	م	‹πñ	‹ π ñ	q
نن	نـ	ن	›∫ó	› ∫ ó	r

Voici, sur un tableau, les six lettres {f q k l m n}. De chaque lettre, on a, sur une ligne, écrit les quatre formes, suivant l'ordre propre à l'arabe, de droite à gauche; puis un mot, formé des trois formes liées. On voit que, sous toutes ses formes, la lettre f (fa) comprend une boucle surmontée d'un point: si la lettre doit être liée à droite ou à gauche - i.e. si elle est respectivement précédée ou suivie d'une autre lettre, au sein d'un mot - la boucle est prolongée d'un trait horizontal, du côté convenable. De plus, si le f n'est suivi d'aucune lettre, on le termine à gauche par un crochet ascendant. Le q (qaf) est très semblable au f: à ceci près que la boucle y est surmontée de deux points; et que, s'il y a un crochet à gauche, celui-ci est fortement cambré. Etc.

On notera, en passant, que, tandis que les lettres isolées sont représentées par les mêmes octets que diverses minuscules, les autres formes correspondent

à des lettres accentuées, ou à des symboles de toute sorte: en fait, les multiples signes arabes s'approprient tous les octets non attribués aux caractères de commande, aux chiffres, ou à la ponctuation.

Une deuxième particularité de la graphie arabe est l'abondance des ligatures. Certes, de celles-ci, les manuscrits anciens - latins et surtout grecs - offrent de nombreux exemples: mais presque rien n'en subsiste - sinon {æ œ fi fl ...} - dans la typographie moderne, issue d'autres traditions, notamment de celle des inscriptions monumentales. En arabe, une diversité de tracé virtuellement infinie défie l'ingéniosité des dessinateurs de polices. Voici, sans exhaustivité, des tracés distingués par la norme typographique du Macintosh et qui tous seront finalement transcrits par nous avec le caractère ÿ:

En bref, la lettre considérée est une dent soulignée de deux points. À ce schéma, les traits de liaison droit et gauche apportent une diversité attendue; et le tracé de la boucle finale se signale par son élégance. Mais, de plus, la précision des liaisons requiert des variantes; et il y a une ligature fÿ, et de multiples ligatures lÿ. Enfin les deux points peuvent disparaître, ou être remplacés par un autre signe: on aborde ici des particularités linguistiques intimement mêlées aux règles de l'écriture.

1.2 ***Scriptio plena*** **et** ***scriptio minima***

En toute rigueur, la notation graphique d'un texte arabe n'est presque jamais complète: et peut-être la *scriptio plena* n'est-elle pas définie jusqu'au point de rallier l'unanimité des grammairiens.

On sait que l'alphabet a été inventé par les Phéniciens pour noter une langue très proche de l'hébreux; et étroitement apparentée à l'arabe. Dans cette écriture, de laquelle procèdent toutes les nôtres, seules étaient notées les consonnes. Il semble qu'on soit redevable aux Grecs des premiers alphabets comportant, dans une unique série de signes, des consonnes et des voyelles. Les langues sémitiques elles-mêmes disposent depuis plus de deux mille ans d'une notation alphabétique complète du point de vue de la phonétique; mais avec, pour les voyelles brèves, des signes facultatifs qui trouvent place hors de la ligne des lettres proprement dites; et, pour les voyelles longues, des lettres dont le statut phonétique semi-consonantique est d'autant plus complexe que, dans la famille sémitique comme ailleurs, l'écriture assume outre sa fonction

phonétique, diverses fonctions grammaticales: notamment pour distinguer des formes homophones - i.e. coïncidant pour le son mais non pour le sens.

Par conséquent, et cela est essentiel dans l'analyse des textes arabes, une élaboration totalement automatique, ne requérant lors de la saisie aucune interprétation grammaticale, doit être fondée non sur la *scriptio plena*, mais, au contraire, sur une *scriptio minima*, d'où l'on a éliminé absolument tous les signes diacritiques (voyelles, redoublement de lettres); et même des marques de distinctions essentielles (telles que celles affectant les variantes de ÿ sans deux points), si celles-ci manquent de façon constante dans maintes éditions scientifiques de textes classiques; pour ne rien dire de la presse contemporaine.

Le passage même, à la *scriptio minima*, d'un fichier saisi avec des signes diacritiques offre encore quelques difficultés dont nous donnerons des exemples. Il ne suffit pas de supprimer les signes proprement dits; il faut encore prendre garde aux variantes de format requises pour une typographie élégante.

فَفَفَ فَ فَفَ فَ

GÿÇÇÇGÇÇÇµÇÇGÇÇÇí Gm GÿGµGí Gm

Partons d'une forme (imaginaire!) notée phonétiquement fafafa. Celle-ci pourrait être codée dans le fichier de texte par six octets correspondant à la séquence de caractères usuels : GÿGµGí ; où les minuscules désignent les diverses formes du f; et le G tient lieu du a bref: mais, pour éviter que le trait supérieur incliné, qui marque en arabe cette voyelle, n'adhère à la boucle du f, on doit allonger quelque peu des traits de liaison entre caractères; ce qui se commande par un octet qui, dans un alphabet latin, sert pour Ç.

Parfois, on doit, de même, introduire des espaces afin de placer une voyelle là où il n'y a pas de trait de liaison; mais ces espaces, commandés par un X, sont à distinguer du blanc usuel, en ce qu'ils n'interrompent pas le mot; et aussi du blanc insécable, dont la place dans le code ASCII est prise, en arabe sur Macintosh, par la forme finale de la lettre b; ce qui ne laisse pas d'engendrer des confusions - dont tous les traitements de texte ne s'accommodent pas, particulièrement en fin de ligne.

1.3 Du fichier de texte arabe à la liste des formes

Finalement, un programme 'triarab' dont nous ne croyons pas utile de publier le détail, traduit les formes successives d'un texte saisi selon les normes propres à l'arabe, dans un alphabet d'octets qui vont de 63 à 97 (i.e. de ? à a, selon le code ASCII); avec, pour afficher et imprimer cette

translittération, une police particulière, dessinée par nous en respectant, à des variantes mineures près (issues, pour la plupart, du dictionnaire de Hans WEHR), les normes de la transcription phonétique internationale; ainsi que l'ordre alphabétique propre à l'arabe.

{ᵓ ä a b p t t̲ ğ č ḥ k̲ d d̲ r z ž s š ṣ ḍ ṭ ẓ ᶜ ḡ f q̇ ġ k g l m h w y}

On voit que les fricatives {tha kha ghaïn} {t̲k̲ḡ}sont marquées d'un trait, supérieur ou inférieur; {shin jim} {šğ}, d'un petit v supérieur; les emphatiques {ṣḍṭẓq̇}, d'un point. On a réservé la place de signes ajoutés pour des transcriptions de mots non arabes - berbères, persans, urdus: G dur {ġg}, emphatique ou non; č = tch.

Notre alphabet commence par les trois signes {ᵓ ä a} qui rendent respectivement le hamza, écrit comme une lettre isolée; le ta marbouta - finale usuelle du féminin phonétiquement assez voisine du a, mais rattachée au t qui en prend la place quand une désinence possessive est adjointe au mot; et l'alif proprement dit, sans distinction de signes diacritiques, susceptibles d'en faire un i, ou un A long.

De même notre y peut correspondre au son a: en sorte que la préposition 'vers', prononcée ila, se trouve écrite: aly; en rangeant les lettres de gauche à droite, comme il est de règle dans les transcriptions latines.

Voici, par exemple, la transcription d'une phrase arabe en une suite de formes; comme dans nos publications antérieures citées plus haut relatives au grec, au latin etc, chaque forme est accompagnée de son adresse dans le texte, par n° de chapitre et de phrase (verset ou alinéa).

١هذه المقالة هي اول مقالة يفحص فيها عن انواع الموجود المقصود بالفحص عنها

اولا في هذا العلم وذلك ان هذا العلم

1 *Dans ce livre commence l'étude des types de l'être qui fait l'objet premier de cette science [la métaphysique]. Car celle-ci...*

Il s'agit du début du commentaire par Averroès (Ibn Roshd) du livre Z de la métaphysique d'Aristote.

Nous en donnons, d'une part, une traduction, rendant ce qui nous paraît être le sens; et, d'autre part, en face du listage des formes arabes, un essai de mot à mot français; où les formes du texte initial sont écrites en caractères gras; avec des adjonctions en maigre; et, entre parenthèses des parties à supprimer.

hḏh	cette
1 1	
almqalä	partie
1 1	
hy	(elle)
1 1	c'est la
awl	premier[e]
1 1	
mqalä	partie
1 1	où
yfḥṣ	(est étudié)
1 1	on s'enquiert
fyha	(dans elle)
1 1	
ᶜn	(sur)
1 1	des
anwaᶜ	sortes
1 1	
almwǧwd	de l'être
1 1	ce qui est l'objet
almqṣwd	(le) recherché
1 1	
balfḥṣ	dans l'étude
1 1	
ᶜnha	sur elles
1 1	
awla	premièrement
1 1	
fy	dans
1 1	
hḏa	cette
1 1	
alᶜlm	science
1 1	
wḏlk	et cela
1 1	
an	c'est que
1 1	
hḏa	cette
1 1	
alᶜlm	science
1 1	... ...

Il importe de noter que la transcription des formes arabes, fondée sur la *scriptio minima* n'est pas une romanisation fidèle, comme celle utilisée, dans sa thèse, par A. Abi FARAH: ici, le passif ne se distingue pas de l'actif, les désinences des cas manquent; et, répétons-le, les caractères a y ne correspondent pas à des timbres de voyelles longues.

Une dernière surprise est encore réservée au lecteur attentif: si, afin de ne déroger en rien à la transcription automatique, on respecte le découpage typographique de la chaîne écrite, on se trouve distinguer des segments qui ne sont pas strictement ce qu'un linguiste appellerait des formes de mots.

En effet, les conjonctions de coordination w et f (wa et fa) se lient au mot suivant à quelque catégorie que celui-ci appartienne; l'article défini se lie au nom; certaines prépositions se lient aux noms et pronoms, voire à des sortes de conjonctions de subordination.

Les expériences dont nous rendons compte montrent que l'analyse s'accommode du dénombrement de ces segments, au même titre que de formes de mots: *a posteriori*, on pourra assimiler les segments issus d'un mot à une sorte de paradigme, analogue d'une déclinaison; et on se souviendra qu' A. SALEM, a appelé l'attention des linguistes sur l'intérêt des statistiques de segments répétés.

2 Le corpus des textes arabes

Récemment, notre collègue Chr. RUTTEN a bien voulu nous inviter à participer à une recherche comparative portant sur les grands commentaires qu'Averroès, puis Saint Thomas d'Aquin, ont consacré au Livre Z de la métaphysique d'Aristote. Il nous a semblé que la statistique linguistique pourrait contribuer à cette recherche.

Saint Thomas disposait d'une traduction latine de l'original arabe d'Averroès: cette traduction mérite certes d'être comparée au latin de Saint Thomas; mais nous présumons, d'après ce que nous savons des traductions littérales utilisées au XIII-ème siècle, que se posera, de toute manière, le problème de la comparaison entre textes philosophiques écrits dans deux langues différentes.

L'occasion s'offrait donc d'entreprendre, préalablement, d'étendre à l'arabe la chaîne de traitement déjà mise en œuvre pour d'autres langues. La seule adaptation requise concerne la transcription d'un texte, saisi sur support informatique, en une suite normalisée de formes de mots. Le §1 du présent article présente le programme 'triarab' conçu à cet effet.

Avant d'exposer, au §3, les quelques résultats obtenus dans notre exploration, il reste à décrire le petit corpus des textes analysés. En bref, celui-ci comprend treize fragments ou chapitres d'auteurs divers; chaque fragment comptant environ 3000 caractères (3k).

2.1 Le commentaire de la Métaphysique par Averroès {rZt1... rZt4}

L'Andalou Ibn Roshd, dont le nom a été transcrit: Averroès, naît en 1126 dans une famille de Cordoue, déjà illustrée par son aïeul, jurisconsulte du rite Malékite. Il meurt à Marrakech en 1198.

Le grand commentaire du Livre Z par Averroès s'ouvre sur une brève introduction; puis offre une suite d'alinéas du texte traduit du grec, accompagnés chacun d'un commentaire long d'une page ou deux. Dans l'édition classique du Père Maurice Bouyges, les alinéas du texte sont désignés par les sigles T1, T2,...; et les commentaires afférents par C1, C2. Nous avons saisi quatre fragments, notés: {rZt1... rZt4}: le 1-er comprend l'introduction avec {T1, C1}; le 2-ème, {T2, C2};...; le 4-ème {T4, C4} .

2.2 Un texte apologétique chrétien isf†

Nous disposions d'autre part de quelques fragments déjà saisis afin de confronter à des parallèles philosophiques, un chapitre de théologie.

Il s'agit d'un texte apologétique chrétien de langue arabe intitulé:

"Epitre sur l'Unité et la Trinité"

et dû à Muhyî Al-Dîn Al-Içfahânî; auteur oriental qui, selon le P. Cheikho, a pu vivre vers l'an 1100 (de l'ère chrétienne). En 1962, est paru dans la collection des "Recherches publiées sous la direction de l'Institut de Lettres Orientales de Beyrouth", le texte arabe de l'Épitre, édité, traduit et annoté par M. Allard et G. Troupeau.

Le fragment isf† est le chapitre 1 de cette épitre. Les autres fragments, philosophiques, renferment diverses allusions à la doctrine Trinitaire qu'on peut trouver chez des auteurs arabo-islamiques, hors de tout contexte de controverse; voire, sans que le nom de chrétien ne soit prononcé. Il apparaît, en effet, que l'idée que Dieu est Un et Trine est connue de ces grands penseurs; et qu'il y a, attesté par l'usage des mots, un courant caché qui, aux XI-ème XII-ème siècles, va de l'Orient à l'Occident, de Ghazâlî à Ibn Roshd.

2.3 Le philosophe autodidacte, d' Ibn Tofayl {TfHy TfGh}

Ibn Tofayl (Muhammad Abu Bakr, cité par les scholastiques latins sous le nom d'Abubacer), naquit dans les environs de Grenade en 1100, et mourut à Marrakech en 1175. Ce savant aristotélicien, également versé dans la médecine, les mathématiques, la poésie..., a, devant la postérité, le mérite d'avoir présenté, au Calife almohade Abu Ia'kub Yusuf, le jeune Ibn Roshd (Averroès) qui fut chargé par le Calife de commenter l'œuvre d'Aristote.

D'Ibn Tofayl lui-même, l'ouvrage le plus connu aujourd'hui est une histoire philosophique dont le héros, Hay ben Yaqzân, d'abord allaité par une gazelle, grandit, loin des hommes, dans un île, où il découvre, seul, par étapes, l'ensemble des connaissances humaines telles qu'Ibn Tofayl en conçoit l'enchaînement: des sciences de la nature jusqu'à la théologie, jusqu'à la connaissance mystique.

Parvenu au terme de cinq semaines d'années, Hayy a découvert l'existence de Dieu comme cause première de l'univers. Il s'interroge, avec émerveillement, sur le moyen de sa propre découverte; sur ce qu'il est lui-même et que les animaux ne peuvent être... De son esprit, qui est, dit-il, à l'image de l'Esprit du Créateur, Hayy retient qu'il est à la fois: "la faculté de connaître, celui qui connaît et ce qui est connu". Le fragment TfHy suit quelques pas de Hayy, jusqu'à cette formule déjà considérée par l'Içfahânî.

Avant de mettre en scène Hayy, Ibn Tofayl propose une histoire animée des coryphées de la pensée. Deux pages, qui constituent notre fragment TfGh, sont consacrées aux livres de Ghazâlî (?-1111; l'Algazel des scholastiques). Y apparaît, en profondeur, une perspective de plans doctrinaux; et, finalement, entre des parenthèses et des exclamations, la croyance en la Trinité.

2.4 L'effondrement des philosophes, selon Ghazâlî {ghz1 ghz2 ghz3}

Ghazâlî mérite bien d'arrêter Ibn Toffayl. Né dans le Khorassan, brillant professeur à Bagdad, puis errant dix ans entre Damas, le Caire et la Mecque, à nouveau professeur..., Ghazâlî se présente vivant, brûlant plutôt, dans son autobiographie de "La délivrance de l'erreur"; itinéraire chevaleresque d'une âme dont la boulimie d'étude et de pensée alimente tantôt le doute, tantôt l'inspiration, tantôt l'anathème. Qu'a-t-il cru? de tant de formes qu'il nous montre de lui-même, en a-t-il choisi une pour y demeurer finalement?

À défaut de "livres cachés" qu'Ibn Toffayl dit avoir vainement cherchés en Andalousie, on a saisi du 'tahafot', où Ghazâlî fulmine contre les philosophes, une dissertation dans laquelle ceux-ci sont accusés de ne pouvoir, contrairement à ce qu'ils prétendent, démontrer l'unicité de Dieu; car il y a là, sinon une profession de foi, du moins, comme pour Hayy, des échos non équivoques de la doctrine trinitaire.

Cette dissertation nous a donné trois fragments {ghz1 ghz2 ghz3}.

2.5 Le 'Traité décisif', d'Averroès {Rmq1 Rmq2 Rmq3}

Avec Ibn Roshd la dévotion à Aristote fut à son zénith: jamais, ni avant ni après, elle ne s'éleva nulle part plus haut: et l'on trouve dans son œuvre l'archétype de la scholastique latine du XIII-ème siècle. Un tel rationalisme n'étant pas accepté de tous ses concitoyens, Averroès contre-attaque.

D'une part, il répond, point par point, au tahafot de Ghazâlî, par son تهافت التهافت, "l'effondrement de l'effondrement".

D'autre part, il s'exprime avec concision dans un: "Traité Décisif, où l'on établit en quoi s'accordent la loi {islamique} et la sagesse {des philosophes}". Au "Traité décisif" d'Averroès, copistes anciens et éditeurs modernes ont adjoint un "Exposé de méthodes de preuve des convictions de la foi". De cet "Exposé", nous reproduisons intégralement, en trois fragments {Rmq1 Rmq2 Rmq3}, le chapitre "sur les attributs divins", où la doctrine chrétienne de la Trinité est clairement, encore que prudemment, introduite.

2.6 Références bibliographiques des textes saisis

Muḥyī al-dīn al-īṣfahānī: "Épitre sur l'Unité et la Trinité"; texte arabe édité, traduit et annoté par M. Allard et G. Troupeau; in Recherches publiées sous la direction de l'Institut de Lettres Orientales de Beyrouth; 1962.

رسالة اشرف الحديث في شرفي التوحيد والتثليث

للشيخ الامام محيي الدين الاصفهاني رحمه الله تعالى

Ibn Tofayl: "Hayy bin Yaqzân": traduction de Léon Gauthier, S.P.A.G. (Papyrus), Paris.

ابن طفيل : حي بن يقظان ؛

١) خزانة الفكر العربي ، مؤسسة ناصر للثقافة ؛

٢) بتقديم وتحقيق فاروق سعد ، دار الافاق الجديدة ، بيروت ؛

Ghazâlî (Algazel)

تهافت الفلاسفة : L'effondrement des philosophes; texte arabe établi par Maurice Bouyges, S.J.; l'imprimerie catholique; Beyrouth, 1927. {À cette condamnation qui frappe, notamment, l'Aristotélisme arabe, Ibn Roshd a répondu par son تهافت التهافت, "l'effondrement de l'effondrement".}

ألمنقذ من الضلال : "La délivrance de l'erreur"; le titre arabe est diversement complété; édition bilingue, avec traduction française, introduction et notes, par Farid Jabre; Librairie Orientale, Beyrouth; 1969. {Ce livre évoque une multiplicité d'Écoles de pensée, dans un récit autobiographique de la recherche de la Vérité.}

Ibn Roshd (Averroès) : "Philosophie d'Ibn Roshd:

1) Traité Décisif, où l'on établit en quoi s'accordent la loi {islamique} et la sagesse {des philosophes} ;

2) Exposé de méthodes de preuve des convictions de la foi";

Müller, Munich, 1859; Le Caire, 1895; Beyrouth, دار الآفاق الجديدة, 1978.

إبن رشد : فلسفة إبن رشد :

١ .. فصل المقال وتقرير ما بين الشريعة والحكمة من الاتصال ..

٢ .. الكشف عن مناهج الأدلة في عقائد الملة ..

منشورات ... دار الآفاق الجديدة .. بيروت .. ١٩٧٨ ..

Grand commentaire de la métaphysique: texte arabe inédit établi par Maurice BOUYGES, S.J., Dar El-Machreq, Beyrouth; 2-ème éd., 1967

تفسير ما بعد الطبيعة ؛ دار المشرق ، بيروت .. ١٩٦٧ ..

N.B. L'ensemble des fragments saisis se trouve, accompagné de traductions françaises et de commentaires, dans une note, non publiée, intitulée: *La doctrine de la Trinité chez les philosophes du monde arabo-islamique aux XI-ème XII-ème siècles (de l'ère chrétienne).*

3 Elaboration et analyse du corpus

3.1 Les programmes utilisés

Le texte arabe a été saisi par 'Alkaatib'. Ce logiciel crée des fichiers de lignes successives qu'on peut afficher sur un traitement de texte latin usuel, qui range les caratères de gauche à droite, chaque ligne venant dans l'ordre de l'arabe; (et en caractères arabe si on a demandé la police appropriée). Mais pour le traitement statistique, on part du fichier même de texte arabe que crée Alkaatib: suite des octets correspondant aux caractères (cf. supra, §1.1), à deux particularités près: d'abord, une en-tête de 1154 octets pour le format du texte; d'autre part, à la fin de chaque séquence saisie de 126 caractères, une séparation marquée par deux occurrences de l'octet $7C (124); ce qui, dans le code ASCII, correspond à ||. Il va sans dire que ces repères ont été éliminés.

Ainsi qu'on l'a expliqué au §1, le programme 'triarab' transforme le fichier de texte saisi, en une suite de formes, chacune étiquetée par verset et chapitre, selon le numérotage introduit dans le texte. Le §1.3 offre un exemple de liste non triée, créée dans l'ordre du texte: on voit que chaque forme occupe une ligne impaire, l'étiquette étant sur la ligne suivante.

Aux particularités près qu'implique la structure de l'écriture de l'arabe, ce programme ne diffère pas essentiellement des programmes 'trigrec', 'trilat', 'trigalac', 'trihisp'... utilisés pour le grec, le latin, le français, l'espagnol..., dans de précédentes études. On signalera seulement que, dans leur version actuelle, les programmes 'trilng' remplissent, simultanément, pour une langue 'lng' donnée, les fonctions des programmes 'forlng' et 'trilng' utilisés antérieurement; i.e., créent, selon la demande, des listes dans l'ordre du texte et des listes triées alphabétiquement; i.e., pour un texte 'Disq:txt', un fichier non trié 'Disq:txt§', un fichier trié: 'Disq:txt§t'.

Sur les listes de formes, l'élaboration se poursuit sans distinction de langue: en effet, quels que soient l'alphabet propre à cette langue et la représentation qu'on en a choisie par des octets, une forme est désormais traitée comme une suite d'octets; et une liste de formes, comme une liste de telles suites; le tri alphabétique étant simplement un tri suivant l'ordre naturel des octets, sans référence à la valeur linguistique de ceux-ci.

Dans la présente étude, on a utilisé les quatre programmes 'qamus', 'trimu§', 'trimu' et 'tridic'; dont nous rappellerons les fonctions: 'qamus' crée, à partir d'une liste triée de formes, le dictionnaire du texte: i.e. une liste triée, 'Disq:txt$§', où chacune des formes distinctes attestées dans le texte se rencontre une fois, occupant une ligne de rang pair, avec, sur la ligne précédant chaque forme, le nombre des occurrences de celle-ci dans le texte.

Par 'trimu§' - programme de tri des lignes par paires, qui peut aussi servir à transformer, en liste triée, une liste créée dans l'ordre d'un texte - le dictionnaire se trouve trié dans l'ordre des fréquences croissantes: on obtient une liste Disq:txt$§t où, comme dans Disq:txt$§, chaque forme est sur une ligne de rang pair, avec sa fréquence écrite sur la ligne précédente. (On peut dire qu'ici, la forme est traitée comme une étiquette relativement à la fréquence sur laquelle porte le tri).

D'après ce dictionnaire complet trié, on choisit, en partant du bas de la liste, i.e. des formes les plus fréquentes, le sous-ensemble, ou lexique, des formes retenues afin de créer un tableau de correspondance entre formes et chapitres - ou fragments. Dans la présente étude (cf. infra §3.2), on a considéré principalement un lexique V de formes de mots outil; mais sous diverses variantes V, W, Y. Le choix fait, on dispose d'une liste 'Disq:V' dans laquelle les formes retenues ne sont pas dans l'ordre alphabétique, et se mêlent à des nombres de fréquence qu'on n'a pas supprimés.

Le programme 'trimu', qui trie une liste en considérant les lignes individuellement et non par paires, produit, à partir de 'Disq:V', une liste triée 'Disq:Vt' qui commence par les nombres qu'on avait laissés, rangés des plus petits au plus grands; et donne ensuite, dans l'ordre alphabétique, les formes retenues. (Plus précisément, si certaines formes commençaient par des octets précédant la suite {48...57} des codes des chiffres, ces formes sortiraient en tête). Désormais, le bloc des nombres étant éliminé, cette séquence, notée encore 'Disq:Vt', est prise pour lexique.

Le tableau de correspondance est créé en format de texte par le programme 'tridic'. De façon précise, on a un tableau dont les lignes renvoient aux formes du lexique et les colonnes à des fragments consécutifs du texte initial: alinéas (ou versets), chapitres, ou autres segments; ce dernier découpage étant défini par un fichier Disq:txt§Ax créé à cet effet; lequel délimite les segments et leur attribue des sigles.

Lors de la saisie, le corpus arabe s'est trouvé partagé en 10 chapitres; certains de ceux-ci avaient été convenablement délimités, compte tenu du sens, pour avoir une longueur de l'ordre de grandeur souhaité, soit quelque 3000 caractères: e.g. les deux fragments du 'Philosophe autodidacte' (cf. §2.3) TfGh et TfHy, relatifs, respectivement aux œuvres de Ghazâlî et à la méditation théologique de Hayy.

Mais le passage du tahafot de Ghâzalî, (cf. §2.4), et celui de l'appendice au Traité Décisif d'Averroès, (cf. §2.5), saisis sans subdivision en chapitres, ont été découpés ensuite, chacun en trois fragments.

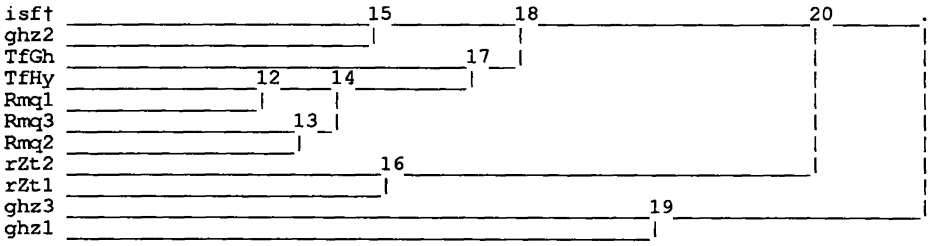
```
Subdivision du corpus arabe
isf†  isfahânî     1    9999
ghz1  tahafot      2       6
ghz2               2      15
ghz3               2    9999
TfGh  Tofayl:Ghz   3    9999
TfHy  Tofayl:Hay   4    9999
Rmq1  Rsh:maqal    5      10
Rmq2               5      15
Rmq3               5    9999
rZt1  Roshd:Zeta   7    9999
rZt2               8    9999
rZt3               9    9999
rZt4              10    9999
```

Sur le listage ci-joint, on voit une ligne par fragment: avec le sigle, un commentaire éventuel (dépourvu de chiffres) et l'indication du dernier verset pris (désigné par chapitre et numéro; 9999 étant mis pour 'dernier'): e.g. le 3-ème fragment du tahafot, ghz3, va de l'alinéa 16 (après la fin de ghz2) jusqu'à la fin du passage saisi.

On trouvera, au §3.2 un extrait du tableau de correspondance, 'Disq:txt§WcorA', ainsi créé.

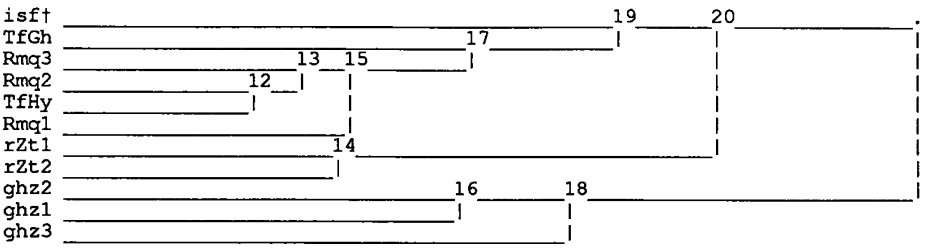
3.2 Choix du lexique et variantes des analyses

Une première analyse a été effectuée avec un lexique Δ de 89 formes, comprenant des mots pleins alğmhwr (le commun, le vulgaire)) alṣfat (les attributs; notamment de Dieu), aussi bien que des pronoms: hw(il) hy(elle).



La CAH des textes (qui porte sur 11 fragments seulement, rZt3 et rZt4 n'étant pas saisis lors de cette analyse) est assez satisfaisante: avec pour classes principales 19 ({ghz1 ghz3} du tahafot); 16 ({rZt1 rZt2} du commentaire de Zeta); et 17: qui associe, à {TfHy TfGh} d'Ibn Tufayl, les 3 fragments {Rmq1 Rmq2 Rmq3} de l'Exposé de méthodes, adjoint au traité décisif: association satisfaisante, car, d'une part, Ibn Roshd est disciple d'Ibn Tufayl; et, d'autre part, à la différence du commentaire de la Métaphysique, les deux œuvres ainsi réunies, relèvent de la vulgarisation philosophique. Mais la classe 15, {isf† ghz2}, est inattendue.

On reprend donc l'analyse, avec les mêmes 11 fragments et un lexique V de 62 formes de mots outil. Il faut s'interroger sur la valeur de cette locution: d'une part, on a dit, au §1.3, qu'il peut s'agir non de mots isolés mais de groupes de mots, liés dans l'écriture arabe: e.g. fyha (dans elle); ou même bḏath (en lui-même); d'autre part, dans les textes philosophiques, des mots outil du langage commun entrent dans des locutions techniques, et deviennent des marques du contenu plus encore que du style.



Cette réserve étant faite, on voit que le dénombrement des mots vides fournit une CAH satisfaisante, où est corrigée l'anomalie de celle obtenue avec Δ: car ghz2 rejoint {ghz1 ghz3}; et isf† est quasi isolé.

Quand les fragments {rZt3 rZt4} ont été saisis, on les a adjoints en supplément à l'analyse, et rattachés (par le programme 'discri'), chacun au fragment principal le plus proche: (rZt3->Rmq3)(rZt4->rZt2)
rZt4 va avec rZt2, autre fragment du commentaire de Zeta; et rZT3 va avec Rmq3; qui est d'une autre œuvre, mais du même auteur.

Dans les analyses ultérieures, les 13 fragments sont en principal; et l'on prend un lexique W de 71 formes sans mots pleins caractérisés. La classification obtenue pour les 13 fragments ne laisse rien à désirer; mais on doit supprimer k̲t̲rä et mn: k̲t̲rä, multiplicité, forme de faible fréquence, est associée à ghz3, du tahafot de Ghazâlî; mn est ambigu: pronom interrogatif, ou préposition (indiquant l'origine); et les emplois sont nombreux avec chacun des deux sens. On peut contester la présence du mot wạḥd, dont le sens usuel est 'un', mais qui est un maître mot de la théologie: on l'a conservé parce que sa distribution est très étendue; autant et plus que celle de kl, tout...

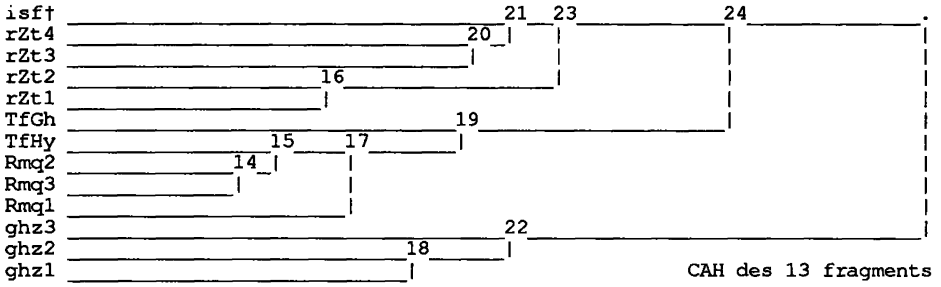
mots de W x parties de Un2:Darab:DZ:totd

13	isf†	ghz1	ghz2	ghz3	TfGh	TfHy	Rmq1	Rmq2	Rmq3	rZt1	rZt2	rZt3	rZt4
a̲k̲r	3	0	1	0	3	0	0	1	0	4	2	0	0
k̲t̲rä	0	0	1	7	1	0	0	0	3	0	0	0	0
kl	11	2	10	1	1	0	0	0	3	0	1	5	8
kma	2	0	1	1	0	2	3	1	1	0	0	0	0
la	6	24	14	7	4	9	6	1	7	2	3	6	6
lanh	1	0	1	0	0	1	0	0	2	0	0	2	3
mn	31	3	15	7	12	13	18	6	29	7	10	16	30
wạḥd	9	3	9	4	0	0	0	0	6	0	1	6	8

On s'arrête donc à un lexique Y de 69 formes, ou séquences de formes de mots que nous avons considérés comme des outils; et dont la liste apparaît dans un tableau de classification donné au §3.3.

Assurément, le choix de Y n'échappe pas à toute critique; mais la base étroite des textes saisis, n'encourage pas à poursuivre les essais.

3.3 Résultats obtenus en croisant 13 fragments de textes arabes avec un lexique de 69 mots

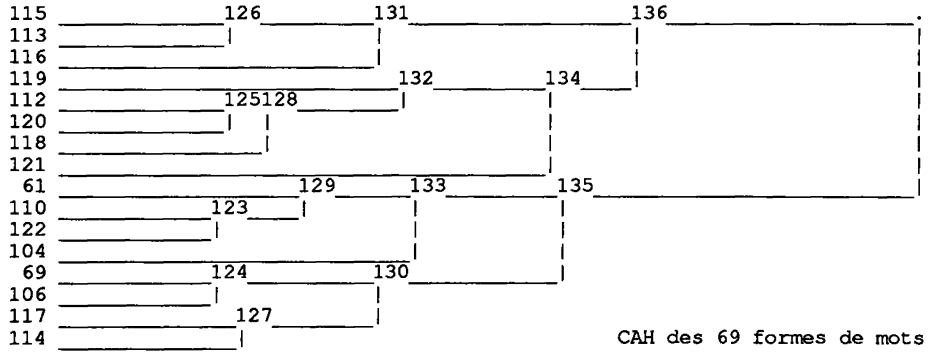


3.3.1 Classification de l'ensemble J des 13 fragments

Nous présentons d'abord cette classification dont l'interprétation est claire. Au sommet, la classe j22 des trois fragments du tahafot de Ghazâlî s'oppose à la classe j24, qui comprend les dix autres fragments; j24 se scinde en j23 et j19. Dans j23, le fragment isf†, de l'Épitre sur l'Unité et la Trinité, s'adjoint aux quatre fragments {rZt1...} saisis du grand commentaire de la Métaphysique d'Aristote par Averroès: isf† est assurément un texte à part, mais la précision technique de l'exposé fait accepter qu'il soit avec {rZt1...}. Enfin on trouve dans j19 des fragments d'Averroès et de son maître Ibn Tofayl, fragments qui, nous l'avons dit, relèvent de la vulgarisation philosophique.

```
Disq:txt$YcorAjVacoriq
j23: 119++          69=0        114-
j22: 119-   122+  104++  117+ 114++
j17: 112++  121-   69++
```

L'agrégation des fragments en classes, résulte de la disposition spatiale du nuage N(J); dont on donne, au §3.3.3, la projection sur les axes 1, 2 et 3. Mais le nuage N(J) lui-même, est un nuage de profils sur l'ensemble I des formes: les classes de fragments s'interprètent donc par leur association avec des formes, ou plus sommairement, des classes de formes (cf. listage Vacor).



c	Partition de Y en 16 classes : Sigles des formes de la classe c		
115	akr mnh fyha tm	TfGh+++++	rZt1+++
113	ma ᶜly aw		rZt1+++++
116	anha ydl wdlk bdath hdh hda		rZt3++++
119	wlma ḍd ᶜlyh dlk alyh alty wanma hy lanh		rZt4+++++
112	bh kma wla	TfHy+++++	
120	anh anma ada lys an	Rmq3+ Rmq1++	rZt1---
118	wama aldy bl whw whda fanh ay hw	isf†+++ Rmq2+	ghz1- ghz2-
121	lan ǧhà alawl waḥd kl	isf†+++++	
61	wḍt	Rmq1++++++	
110	fy šyᵓ		ghz2++++
122	aly ᶜn fan ḡyr kan		ghz3+++++
104	dath fhw wada ᶜnh		ghz3++++++
69	ykwn	Rmq3++ Rmq2+++	
106	bha wḍd lm ala	TfGh+++++	
117	fla ama ayda fyh ad		ghz2+++++
114	wlys lh wan la		ghz1++++++

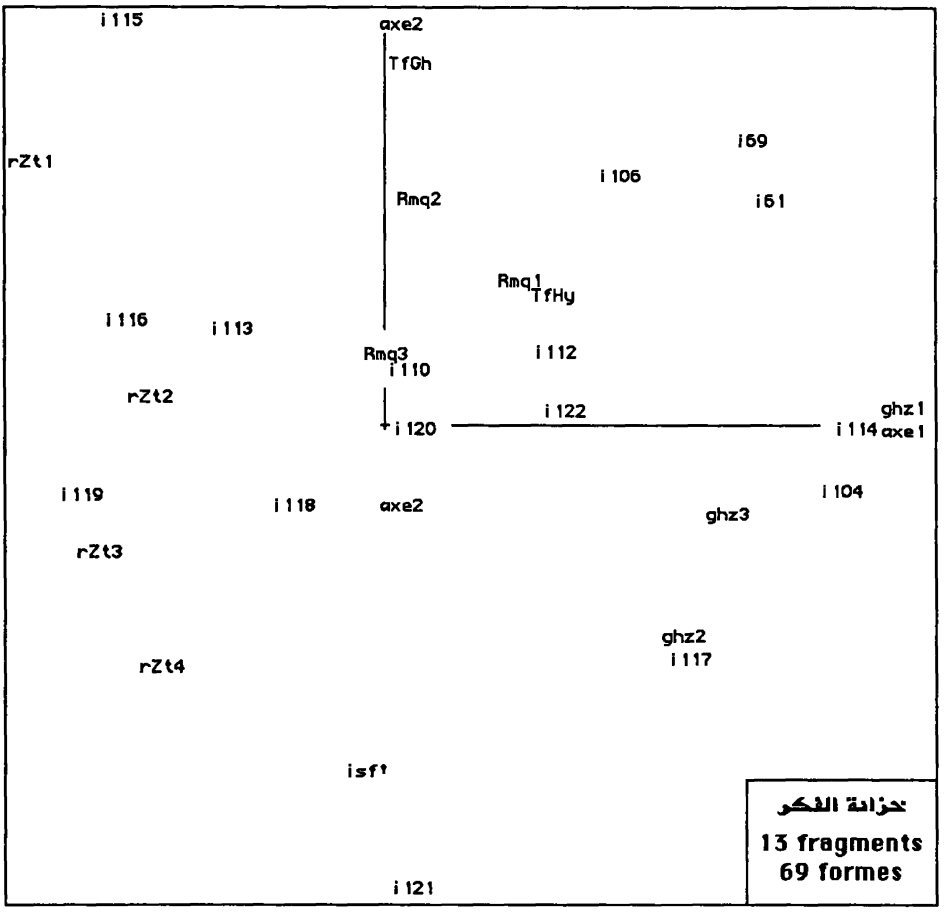
3.3.2 Classification de l'ensemble Y des 69 formes de mots outil

On a retenu une partition de Y en 16 classes.

Il est remarquable que la plupart de ces classes se caractérisent par une association très forte avec un fragment particulier. Pourtant, de liens moins forts avec d'autres fragments, résulte la disposition globale équilibrée qu'exprime, conjointement avec l'analyse factorielle, la classification des fragments.

Quant au contenu des classes, l'arabisant cherchera la place des négations - plus fréquentes dans i135 que dans i136; trouvera tel mot isolé, loin du mot lié à w=et; etc.

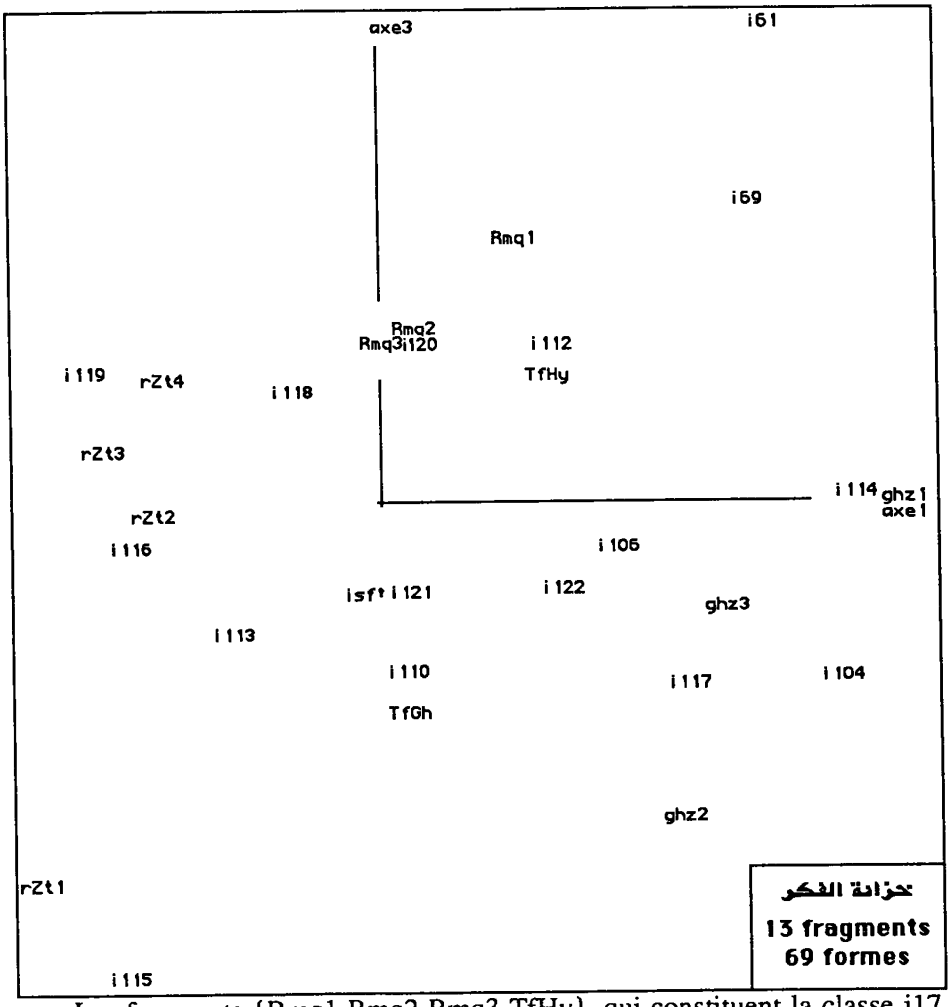
١١٥ اخر منه فيها ثم
١١٣ ما على او
١١٦ انها يدل وذلك بذاته هذه هذا
١١٩ ولما قد عليه ذلك اليه التى وانما هى لانه
١١٢ به كما ولا
١٢٠ انه انما اذا ليس ان
١١٨ واما الذى بل وهو وهذا فانه اى هو
١٢١ لان جهة الاول واحد كل
٦١ وقت
١١٠ فى شىء
١٢٢ الى عن فان غير كان
١٠٤ ذاته فهو واذا عنه
٦٩ يكون
١٠٦ بها وقد لم الا
١١٧ فلا اما ايضا فيه اذ
١١٤ وليس له وان لا



3.3.3 Analyse de la correspondance entre fragments et formes de mots

```
Disq:txt$YcorAcortx : mots de Y x fragments de textes
trace  :   8.021e-1
rang   :      1     2     3     4     5     6     7     8     9    10
lambda :   1650  1062   965   854   731   600   569   515   328   268 e-4
taux   :   2057  1323  1203  1065   911   748   710   643   409   334 e-4
cumul  :   2057  3381  4584  5649  6560  7308  8018  8661  9070  9404 e-4
```

L'axe 1 est créé par l'opposition entre les trois fragments {ghz1 ghz2 ghz3} du tahafot, et le commentaire {rZt1...rZt4} du livre Z par Averroès; isf† se place entre rZt4 et ghz2, mais plus écarté sur (F2<0): ce qui permet de critiquer l'agrégation de isf† aux {rZt...}. Les fragments de vulgarisation s'écartent peu de O sur l'axe1. Sur l'axe2, on note l'étalement de rZt1...rZt4; et l'opposition de isf† à TfGh (qui s'agrège à j17, à un haut niveau de la CAH).



Les fragments {Rmq1 Rmq2 Rmq3 TfHy}, qui constituent la classe j17, proches de l'origine sur l'axe1, sont dans le quadrant (F2>0; F3>0) du plan (2,3); et apparaissent, dans le plan (1,3), bien groupés et séparé du reste de J. On voit, sur le listage Disq:txt§YcorAjFacor, que TfGh s'écarte le plus de j17 dans le plan (3,8); c'est toutefois à j17 que s'agrège TfGh; passage de vulgarisation original, en ce qu'il décrit l'attitude du grand Ghazâlî, se dissimulant au vulgaire, aux disciples, proposant des énigmes, même aux vrais philosophes.

Quant aux classes de mots, on les trouvera proches des fragments, signalés au §3.3.2 pour en être les caractéristiques respectives.

4 Conclusions et perspectives

Le présent travail a montré que, nonobstant les particularités de l'arabe, le dénombrement automatique des formes - définies comme suites de caractères délimitées par des blancs ou des signes de ponctuation - offre, dans cette langue, matière à des études stylistiques cohérentes.

D'un corpus peu étendu, on ne peut extraire des structures globales universelles. Mais on retiendra que les œuvres d'Averrhoès ont été partagées entre deux classes, dont l'une contient des fragments dus à Ibn Tufayl. Ainsi, il apparaît qu'entre les ouvrages de vulgarisation de deux auteurs différents (mais, il est vrai, proches parents l'un de l'autre), la distance peut être moindre qu'entre la vulgarisation et l'exposé en forme, produits par un même auteur. L'hétérogénéité d'un recueil est donc compatible avec son attribution à un auteur unique.

Dans un domaine, comme la philosophie ou toute autre science, où le niveau de l'exposé peut varier par degrés, nous ne savons pas encore distinguer ces variations graduelles de celles qui dénotent un changement d'auteur. Nous savons encore moins ce que peut être la signature stylistique d'un auteur pratiquant plusieurs genres littéraires bien distincts: tels que prose historique, poésie lyrique, théâtre en vers ou en prose.

Mais nous espérons que l'observation patiente de corpus étendus et divers, comprenant des œuvres dont l'attribution à leur auteur est certaine, offrira une vue globale des manières d'écrire, dans toutes leurs nuances.