Deuxième Partie

NORME LEXICOLOGIQUE

« Le caractère le plus assuré d'une norme, c'est de ne satisfaire personne à commencer par son auteur »

(Charles Muller, 1968)

En traitement automatique des langues écrites et surtout dans les études quantitatives des textes, le point de départ est généralement les mots graphiques. Or le mot graphique ne correspond pas toujours, surtout dans les langues à systèmes d'écritures non segmentées, à l'unité lexicale de base, ce qui rend indispensable l'établissement d'une *norme lexicologique* qui a pour principale tâche de définir les unités de décompte, faciliter et harmoniser les opérations de dépouillement et préparer le corpus aux traitement et l'analyse statistiques. Cela se traduit par un certain nombre de décisions lexicographiques à prendre dont la principale caractéristique est qu'elles devraient être à la fois simples et durables.

Le prétraitement des données textuelles s'opère en deux moments pour chacun d'entre eux il faut établir une *norme* : le premier est le moment d'enregistrement du texte en machine, de son apurement et de son harmonisation, le second consiste 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.

Dans tout traitement automatique des textes de quelque nature que ce soit, pour obtenir des résultats sûrs et fiables, il faut veiller à ce que deux conditions soient parfaitement remplies : la première consiste à suivre rigoureusement les règles de saisie et d'harmonisation préétablies, la seconde soumet le corpus à une opération de dépouillement *stricto sensu* incluant quatre grandes étapes précédant tout traitement statistique ou autre, la segmentation, la lemmatisation, la désambiguïsation et la catégorisation.

Par conséquent, deux normes doivent nécessairement être établies : *une norme de saisie et d'harmonisation* et une *norme de dépouillement*.

CHAPITRE 4

Norme de saisie et d'harmonisation

Regroupant choix arrêtés dans le cadre de la pratique typographique de l'arabe et règles de saisie et d'harmonisation, cette *norme* doit répondre à une double exigence : d'une part, elle ne doit pas trop freiner l'opérateur dans sa tâche de saisie des textes et ne pas multiplier sinon minimiser les risques d'erreur et d'ambiguïté ; d'autre part, tout en restant aussi près que possible du texte saisi, elle doit garantir une meilleure uniformité graphique et une certaine constance qui permettront par la suite un meilleur passage des mots graphiques aux formes (items) et des formes aux listes, base de tout traitement lexicométrique. Elle devrait permettre, en outre, de minimiser au maximum les opérations de correction et de retour en arrière.

Nous allons donc présenter, dans les pages qui suivent, la *norme de saisie et d'harmonisation* que nous avons adoptée. Elle regroupe les différentes décisions arrêtées, leur fondement théorique et leur contexte d'application.

1. Norme de saisie des textes arabes

Le système d'écriture[110] de la langue arabe est bien complexe. Cette complexité a posé un certain nombre de problèmes depuis l'époque de l'imprimerie jusqu'aux nouvelles technologies actuelles en passant, bien sûr, par la machine à écrire. À la source de ces problèmes on peut trouver plusieurs facteurs :

- Les différentes formes que peuvent prendre les lettres de l'alphabet arabe selon leur position dans le mot (initiale, médiane, finale ou isolée).

- Le fait que les voyelles brèves n'ont pas le même statut que les consonnes puisqu'elles sont considérées, en quelque sorte, comme des « accessoires orthographiques » placés au-dessus et au-dessous des consonnes.

- Les ligatures[111] qu'elles soient lexicales (ممّيا , لا , الله, ...), contextuelles (la cursivité de l'écriture arabe عمع), ou esthétiques (« Fin, louange à Dieu », « Joyeuse fête », ...)

Il faut signaler qu'en français, par exemple, deux principales normes de saisie ont été adoptées : l'une au début des années soixante pour la confection, à Nancy, du

[110] Nous renvoyons le lecteur à la thèse de J. Dichy**,** *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*, 1990.

[111] Yannis Haralambous définit trois types de ligatures, linguistiques, esthétiques et contextuelles. Voir Y. Haralambous, *Tour du monde des ligatures*, 1995, pp. 85-97.

TLF[112], l'autre par l'équipe de Maurice Tournier au Laboratoire de lexicologie politique de Saint Cloud[113].

1.1. Le clavier arabe

L'ensemble des caractères composant le clavier arabe[114] sont :

1.1.1. Les lettres (40 caractères)

1.1.1.1. Les 29 consonnes de l'alphabet arabe

Il est vrai que les grammairiens arabes classiques s'étaient divisés sur le nombre exact des lettres de l'alphabet arabe ; certains disaient qu'elles sont au nombre de 28 considérant que le *alif* et le *hamza* ne font qu'une seule lettre (le *alif* n'étant qu'un simple support du *hamza*), d'autres, au contraire, considéraient que le *alif* et le *hamza* sont deux lettres différentes augmentant ainsi le nombre de l'alphabet à 29 lettres. A l'ère de l'informatique (et même du temps de la machine à écrire), le problème a été traité autrement, le point de vue pratique a pris le pas sur les considérations théoriques. Qu'il soit ou non une consonne, le *alif* a déjà un code machine propre à lui : celui de la voyelle longue. De ce fait il peut avoir le même statut que ces deux semblables le *wâw* et le *yâ'* c'est-à-dire celui de semi-voyelle : il peut être considéré tantôt comme voyelle, tantôt, le cas échéant, comme consonne. De plus, le *hamza* ayant l'*alif* comme support possède deux autres codes différents de celui qu'il a déjà sans support : ils

---

[112] Cette norme est décrite dans la préface du *Dictionnaire de la langue du XIXe et du XXe siecle*, Paris, 1971.

[113] Voir : P. Lafon, J. Levevre, A. Salem, M. Tournier, *Le Machinal. Principes d'enregistrement informatique des textes*, *op. cit*. Voir aussi D. Labbé, Normes de saisie et de dépouillement des textes politiques, Grenoble, 1990, 135 p.

[114] Clavier standard 101/102 touches.

correspondent à ses deux variantes, le « *alif hamza* en chef » et le « *alif hamza* souscrit ». L'*alif* , lui, a encore une autre forme variante dite « brève ou tordue » (*maqÒûra*).

'alif	*hamza*	*bâ'*	*tâ'*	*×â '*	*Jîm*	*Îâ'*	*Ìâ'*	*dâl*	*Æâ l*	*râ'*	*zây*	*sîn*	*šîn*	*Òâd*
ا	ء	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص

Ãâd	*Ôâ'*	*Ûâ'*	*Ýayn*	*Èayn*	*fâ'*	*qâf*	*kâf*	*lâm*	*mîm*	*nûn*	*hâ'*	*wâw*	*yâ'*
ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي

1.1.1.2. Variantes de consonnes (8 variantes)

Trois des 29 lettres de l'alphabet arabe présentent des variantes au niveau de l'écriture. Ces variantes sont au nombre de huit : une variante du *tâ'* (le *tâ' marbûÔa*) , une variante du *alif* (le *alif maqÒûra*), cinq variantes du *hamza* et le *alif mamdûda* qui est un caractère remplaçant soit la suite « *hamza* voyellé + *hamza* quiescent » soit la suite « *hamza* avec *alif* comme support + *fatÎa* + â »

Variantes de consonnes	
tâ' marbûÔa	ة
'alif maqÒûra	ى
'alif hamza **en chef**	أ
wâw hamza **en chef**	ؤ
yâ' hamza **en chef**	ئ
'alif hamza **souscrit**	إ
'alif waÒla **en chef**	ٱ
'alif madda **en chef**	آ

1.1.1.3. Voyelles brèves (3 voyelles)

Outre les voyelles brèves, il existe bien, en arabe, des voyelles longues correspondant à l'allongement phonétique de celles-ci mais il est à noter ici que ces voyelles longues correspondent, en terme de caractères machine, aux mêmes caractères que les consonnes *alif* « ا », *wâw* « و » et *yâ'* « ي ».

Voyelles brèves	
FatÎa	◌َ
Âamma	◌ُ
Kasra	◌ِ

1.1.2. Les signes diacritiques

- La *šadda*

Marque de la gémination consonantique, la *šadda* est ce signe diacritique qui ressemble à un petit 3 couché au-dessus d'une consonne pour remplacer une suite d'une consonne quiescente (sans voyelle) et de la même consonne mûe (vocalisée).

šadda	◌ّ

- Le *sukûn*

Appelé aussi, en morphologie, voyelle zéro, le *sukûn* qui veut dire quiescence est, par définition, l'absence de voyelle. Ce signe diacritique en forme de zéro est toujours placé au-dessus de la consonne qui n'est pas vocalisée.

sukûn	◌ْ

- Le *tanwîn*

Bien que ce signe diacritique soit considéré, scripturalement, comme le redoublement de la graphie d'une voyelle brève, il a cependant un code informatique différent qui n'est absolument pas le redoublement du code de la voyelle brève redoublée.

Tanwîn	
Tanwîn al-FatÎa	◌ً
Tanwîn al-Âamma	◌ٌ
Tanwîn al-Kasra	◌ٍ

- La *madda* et la *waÒla*

Bien que chacune d'entre elles soit considérée, du point de vue scriptural, comme un signe diacritique autonome, la *madda*, le petit *alif* couché, et la *waÒla*, le petit signe ressemblant à la *Ãamma*[115], sont codées, en caractères machine, soudées en un seul "bloc" à leur support, le *alif*. Ces "blocs" sont donc considérés, à juste titre, comme des consonnes à part entière et sont codés ainsi. Ce dont il est question, dans la nomenclature de la norme ISO-8859-6, ce n'est pas la *madda* ou la *waÒla* mais le « *' alif madda* en chef » et le « *Þalif waÒla* en chef ». Ces deux caractères ne sont en fait que des variantes de consonnes comme indiqué *supra*.

Þalif madda en chef	آ
Þalif waÒla en chef	أ

1.1.3. Les signes de ponctuation

1.1.3.1. Les signes logiques

Il est à noter que la langue arabe utilise deux types de virgules codées différemment : une virgule numérique identique à la virgule en caractères latins et une autre textuelle inversée et orientée à droite.

Point	Point virgule	Virgule textuelle	Virgule numérique
.	؛	،	,

Point d'interrogation	Point d'exclamation	Points de suspension	Deux points
؟	!	…	:

1.1.3.2. Les signes séquentiels

Il est à noter que l'apostrophe n'existe pas en arabe.

---

[115] A l'origine, ce signe diacritique est la lettre ص « *Ò* » qui est l'initiale de l'impératif صِلْ « *Òil* » ([ici] tu dois faire une liaison). Cette annotation est empruntée au système d'écriture du Coran et qui s'y trouve pour en faciliter la lecture et la psalmodie التجويد.

Tout comme en français, le caractère qui représente le tiret est le même correspondant au signe mathématique de soustraction « - ».

Parenthèse ouvrante	Parenthèse fermante	Crochet ouvrant	Crochet fermant	Accolade ouvrante	Accolade fermante
(	)	[	]	{	}

Guillemet simple ouvrant	Guillemet simple fermant	Guillemet double	Chevron ouvrant	Chevron fermant
’	‘	"	<	>

1.1.4. Les caractères spéciaux

1.1.4.1. La *kašida*[116]

La *kašida* est un caractère spécial qui est utilisé pour étendre le joint entre deux caractères arabes. De ce fait, ce caractère n'a aucune valeur, ni phonétique, ni même morphologique.

kašida	**ـ**

1.1.4.1. Les opérateurs mathématiques

Il est à noter que, dans les langues à caractères latins, du fait de l'existence de la consonne « x », l'on a l'habitude de noter, seulement en machine, la multiplication par le caractère « * ». Cette homographie n'existant pas en arabe, la multiplication est dons notée en utilisant la caractère normalement créé à cet effet, le signe « × ».

Addition	Soustraction	Multiplication	Division	Egalité
+	-	×	÷	=

1.1.4.2. Les autres caractères spéciaux

Barre verticale	|
Barre oblique (slash)	/
Barre oblique inversée (antislash)	\
Espace souligné	_
Esperluette ("et" commercial)	&
Astérisque	*
Pour cent	%
Dollar	$
Dièse	#
Arrobas	@

[116] Voir *infra* p. 173

Tilde	~

1.1.5. Les chiffres

Contrairement à ce qui est répandu, même dans certains pays arabes, les chiffres arabes sont ceux utilisés en occident et dans tout le Maghreb « 1, 2, 3, ... ». Les chiffres qu'on utilise dans certains pays arabes du Moyen-Orient « ١, ٢, ٣, ... », sont en fait des chiffres indiens.

	arabe / indien
Un	**1 / ١**
Deux	**2 / ٢**
Trois	**3 / ٣**
Quatre	**4 / ٤**
Cinq	**5 / ٥**
Six	**6 / ٦**
Sept	**7 / ٧**
Huit	**8 / ٨**
Neuf	**9 / ٩**
Zéro	**0 / ٠**

1.2. Les irrégularités orthographiques de l'arabe

La langue arabe est une langue quasiment phonogrammique dans la mesure où chaque phonème est pratiquement toujours transcrit par un seul graphème : « L'orthographe arabe étant (...) relativement saine, le passage du plan de la phonie à celui de la graphie ne présente pas de grosses difficultés »[117]

Il y a, cependant, certains mots qui présentent une inadéquation entre les graphèmes et les phonèmes. Ce phénomène se traduit par deux cas de figure :

- Soit par élision d'un graphème toujours prononcé.
- Soit par ajout d'un graphème orthographique non prononcé.

1.2.1. Ce qui est élidé mais toujours prononcé

- Les pronoms relatifs de la troisième personne du singulier masculin **الذي** *'allaÆî* et féminin **الّتِي** *'allatî*, et de la troisième personne du pluriel masculin **الـذين** *'allaÆîna*, s'écrivent avec une seule lettre *lâm* (celle de l'article) mais se prononcent avec deux *lâm*. Il est à noter que le même relatif au duel et au pluriel féminin, n'est pas dans ce cas, l'adéquation entre graphèmes et phonèmes y est respectée : **اللّذان – اللّذيْن – اللّتَيْنِ – اللّتَيَانِ** *'allaÆâni - 'allaÆayni -'allatâni - 'allatayni*, ainsi que **اللاّتِي – اللّوَاتِي** *'allawâtî - 'allâtî*.

- Les mots qui commencent par un *lâm* auquel s'ajoute le *lâm* de l'article et auxquels est préfixé un troisième *lâm*. Dans ce cas, deux *lâm* seulement sont écrits et non trois. En fait le troisième n'est pas totalement absent puisqu'il est remplacé par la *šadda* : **لبس** *labs* (ambiguïté) → **اللّبس** *'allabs* (l'ambiguïté) → **للّبس**

[117] D. Cohen, *Etudes de linguistique sémitique et arabe*, 1970, pp. 55-56.

li-llabs (à, pour l'ambiguïté) ; il est à noter que cette graphie est différente de **للبس** *li-labs* (à, pour une ambiguïté).

- Certains mots sont toujours prononcés avec un *alif* (voyelle longue « â ») qui n'est pas écrit, à savoir :
  - الله « *allâh* » (Allah)
  - إله « *Þilâh* » (un dieu)
  - لكنْ « *lâkin* » (mais, devant une phrase)
  - لكنّ « *lâkinna* » (mais, devant un nom ou un pronom)

- L'*alif* de la particule d'attention « **ها** » est élidé quand celle-ci est préfixée aux démonstratifs :
  - هذا « *hâÆâ* » (ce, ceci, celui-ci)
  - هذه « *hâÆihi* » (cette, celle-ci)
  - هؤلاء « *hâÞulâÞi* » (ceux-ci, celles-ci)

- L'*alif* du démonstratif « **ذا** » est élidé quand est suffixé à ce dernier un *lâm* comme dans :
  - ذلك « *Æâlika* » (cela, celui-là), ذلكما « *Æâlikumâ* », ذلكم « *Æâlikum* », ذلكنّ « *Æâlikunna* » (même sens : cela, celui-là ; mais en s'adressant à deux ou plusieurs personnes).

1.2.2. Ce qui est écrit mais non prononcé

- Le *wâw* orthographique de **عَمْرو** *Ýamr*. En effet, pour qu'il soit distingué du prénom **عُمَر** *Ýumar* (Omar), le prénom **عَمْرو** *Ýamr* (Amr) se voit adjoindre, au cas sujet (**عَمْرٌو**) et au cas indirect (**عَمْرٍو**), un *wâw* (la lettre « **و** ») orthographique qui n'est pas prononcé. Au cas direct le problème ne se pose pas puisque **عُمَر** *Ýumar* est un diptote qui ne peut prendre de *tanwîn* et donc pas de *alif* final. De ce fait la distinction est facilement établie entre les deux prénoms au cas direct puisqu'on aura : **عَمْرًا** (sans le *wâw* orthographique) et **عُمَرَ** (sans *tanwîn* ni *alif* final).

- Le numéral *cent* مائة *miÞa* peut s'écrire avec un *alif* orthographique au singulier, au duel مائتـان *miÞatâni* et agglutiné à d'autres numéraux, les unités en l'occurrence, comme dans ثلاثمـائة ×*alâ*×*umiÞat-in* (trois cents). Il y a d'autres graphies possibles de ce mot comme : منة (cette graphie est en fait sa graphie originelle) ou, très rarement, مأة .

- L'*alif* orthographique dans certaines formes verbales[118] **:**
  Certaines formes verbales conjuguées à l'accompli ou à l'impératif se voient adjoindre à la fin un *alif* orthographique non prononcé : كَتَبُوا « *katabû* » ou أُكْتُبُوا « *Þuktubû* ».

- Le démonstratif أُولائك « *Þulâ*Þ*ika* » est écrit avec une voyelle longue « *û* », mais prononcé avec une voyelle brève « *u* ».

1.3. Saisie des signes de ponctuation

Tout comme les symboles, les chiffres ou tout autre système de renfort de l'écriture, la ponctuation qui fait partie aujourd'hui de la grammaire de toute langue, présente une importance capitale dans l'acte d'écriture de tel ou tel auteur et un intérêt certain pour tous ceux désirant étudier le style d'un auteur, les caractéristiques, quantitatives ou autres, d'un genre ou d'une époque. Ceci est envisageable car les signes de ponctuation, au même titre que les mots ou les constructions qui agrémentent une œuvre, ne sont pas choisis arbitrairement : « Le choix des ponctuations dépend, dans le passé comme aujourd'hui, des situations, des genres (…), de l'auteur et des styles. »[119]

---

[118] Voir *infra*, p. 193.

[119] Nina Catach, *La ponctuation. Histoire et système*, 1994, p. 113.

Comme la ponctuation marque « le passage de la pensée non langagière à la pensée langagière (du non-linguistique au linguistique) »[120], toute étude linguistique quelle soit quantitative, stylistique ou d'autre nature, doit obligatoirement réserver aux signes de ponctuation tout l'intérêt qui leur revient.

Il est donc **impératif de saisir tous les signes de ponctuation** présents dans le corpus à traiter.

1.3.1. Le statut particulier de la ponctuation dans notre corpus

Compte tenu du fait que, dans notre édition de référence, les signes de ponctuation ont été ajoutés par les éditeurs critiques du texte, ʿAḥmad az-Zayn et ʿAḥmad ʿAmîn, nous avons pris le parti de les considérer comme de simples signes délimiteurs et non comme faisant partie du vocabulaire de Tawḥîdî. Ce parti pris s'explique par deux raisons, une raison de principe : l'authenticité non établie de la ponctuation présente, et une raison de fait : son irrégularité et son hétérogénéité, comme nous l'expliquons ci-après :

1.3.1.1. Le système de ponctuation entre énonciation et interprétation

Il est vrai que les éditeurs scientifiques ont eu le mérite d'introduire, selon leur interprétation du discours et du contexte discursif, les signes de ponctuation pour, entre autres, faciliter la lecture et la compréhension du texte.

---

[120] Michel Gautier, Etude de l'acquisition et théorie linguistique : actions en retour, in *Mélanges linguistiques offerts à E. Benveniste*, SLP, 1975, cité par N. Catach, *idem*, p. 118

Cependant l'utilité de ces signes, surtout les signes de base, se limite à cette fonction-là. Ils ne peuvent en aucun cas nous renseigner sur le style de l'auteur, énonciateur du discours.

Définie comme étant « l'ensemble des signes visuels d'organisation et de présentation accompagnant le texte écrit, *intérieurs* au texte et *communs* au manuscrit et à l'imprimé »[121], la ponctuation, pour être considérée comme faisant partie du message linguistique de l'auteur, doit impérativement répondre à une condition *sine qua non* : être commune au manuscrit et à l'imprimé. Ce qui est loin d'être le cas pour ce qui concerne notre corpus. C'est, en effet, ce genre de situation qui a poussé Nina Catach à affirmer que « dire que la ponctuation appartient en principe à l'auteur ne s'applique évidemment pas aux textes du passé. »[122]

De ce fait, les signes de ponctuation ne peuvent pas faire partie des éléments constituants du discours de l'auteur. Malgré leur importance dans les textes arabes classiques édités tardivement (à une époque où la ponctuation est devenue partie intégrante de l'acte d'écriture), les signes de ponctuation font seulement partie de l'acte d'interprétation. Mais restent tout de même étrangers à l'acte d'énonciation. Ils ne reflètent, en effet, qu'une simple lecture, aussi fidèle soit-elle, d'un discours produit quelques siècles auparavant.

1.3.1.2. L'irrégularité de la ponctuation présente dans le corpus

Si les signes de base tels que le point « . », les deux points « : » et le point d'interrogation « ؟ » ont été introduits plus ou moins avec succès, certains signes de ponctuation, en revanche, ont été placés d'une façon un peu capricieuse et fantasque. En

[121] Nina Catach, *idem*, p. 9.

[122] *Ibididem*, p. 9.

effet, pour les citations, par exemple, elles sont encadrées, quand elles le sont, tantôt par des guillemets tantôt par des parenthèses. En ce qui concerne les digressions, parfois elles sont encadrées par des tirets cadratins et parfois par des tirets semi-cadratins.

Aussi, une partie infime des noms propres de lieu et un certain nombre de noms d'œuvres sont-ils placés entre parenthèses. Mais comme ce traitement appliqué à certains noms propres n'est pas généralisé, il est alors difficile d'admettre cette hétérogénéité et cette irrégularité dans la ponctuation.

À cela s'ajoute les signes propres au domaine de l'édition scientifique et introduits par les deux éditeurs comme, par exemple, les crochets qui marquent un mot ou un groupe de mots initialement absents du manuscrit principal et ajoutés par les éditeurs à partir d'autres manuscrits secondaires incomplets. Il est indiscutable que ces signes ne sont pas et ne peuvent en aucun cas faire partie du vocabulaire de l'auteur.

1.4. Saisie des voyelles

1.4.1. Voyelles brèves

Inventées, postérieurement aux consonnes, vers le VIII^e^ siècle, les voyelles brèves, la *fatÎa* « a », la *kasra* « i » et la *Ãamma* « u », ne sont régulièrement écrites que dans le texte coranique, dans un cadre scolaire ou dans la littérature pour enfants.

Une voyellation totale reste toujours, aussi bien pour le récepteur humain que pour le récepteur machine, le meilleur moyen garantissant une interprétation et donc une analyse des textes directe, rapide et fiable avec un minimum d'ambiguïtés morpho-lexicales et syntaxiques. Malheureusement la réalité est toute autre, et les textes arabes courants ne sont jamais totalement voyellés. Ils ne sont voyellés, quand ils le sont, et c'est très rare, que partiellement et dans la plupart du temps d'une façon très arbitraire.

Ce choix nuisible à l'alphabétisation et désastreux pour le traitement automatique de l'arabe peut être expliqué, aussi bien du côté du scripteur que de l'opérateur de saisie, par ce que l'on a l'habitude d'appeler « la loi du moindre effort » énoncée par Georges Kingsley Zipf dans son livre *Human Behavior and the Principle of the Least Effort*. Il peut aussi s'expliquer, du côté de certains émetteurs, par une sorte de « point d'honneur aristocratique »[123] ou par une espèce de codage stylistique : « Une réflexion (en 1958) du poète libanais *Yûsuf Ta'âlb* (« il faut savoir intriguer le lecteur... ») montre bien le côté *énigmatique de l'arabe* »[124].

En effet, vu l'organisation du système d'écriture arabe, la saisie directe sur clavier (ou sur machine à écrire) devient, si l'on décide de saisir toutes les voyelles brèves, une opération pénible, fastidieuse et coûteuse en terme de temps et d'investissement. La question de la perte de rapidité due à la vocalisation a déjà été soulevée en 1971 par Roland Meynet dans son livre *L'écriture arabe en question* : « l'usage de la vocalisation réduit considérablement la rapidité de la composition typographique qui ne dépasse pas 60 mots à la minute, alors qu'avec les machines à caractères latins on dépasse les 100 mots à la minute »[125]

Pour résoudre le problème de la voyellation, deux solutions sont à envisager. La première consiste à utiliser un outil de voyellation automatique des textes arabes qui soit à la fois fiable et rapide. Malheureusement, à notre connaissance, cet outil, remplissant ces conditions, n'existe pas encore. Un prototype, cependant, a été réalisé il y a quelques années par Malek Ghénima[126], mais nous n'avons pas eu la possibilité de l'essayer sur notre corpus puisqu'il n'y a pas eu de suites données à ce prototype qui présente, nous semble-t-il, un inconvénient : il fonctionne sous MS-DOS® , ce qui n'est

---

[123] G. Lecomte, *Grammaire de l'arabe*, 1968, p.16.

[124] V. Monteil, *L'arabe moderne*, 1960, p. 42.

[125] Cité par R. Zghibi, Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, (pp. 155-182), in : *Document numérique*, vol. 6 n° 3-4/2002, *Unicode, écriture du monde ?*, Hermes, 2002, p. 164.

[126] Voir : Ghénima M., *Un système de voyellation de textes arabes*, Lyon, 1998.

plus du tout pratique de nos jours, de plus, il a, apparemment, une faible cadence. Mais cette expérience mérite d'être poursuivie et améliorée.

La deuxième solution consiste à procéder à une voyellation partielle en amont c'est-à-dire au moment-même de la saisie du texte et/ou, le cas échéant, en aval, en procédant à une harmonisation vocalique du texte déjà saisi sans voyelles (ou voyellé arbitrairement). Par la force des choses c'est la deuxième solution que nous avons adoptée et que nous présentons ici dans le cadre de la *norme lexicologique*.

Pour cela, des règles de saisie des voyelles doivent être établies définissant quelles voyelles exactement devront être saisies, à quels endroits et pourquoi.

Selon leur nature, morphologique, syntaxique ou de conditionnement phonologique, les voyelles brèves sont de trois types : **voyelles morphologiques**, **voyelles casuelles** et **voyelles d'appui** (dites aussi voyelles de liaison).

1.4.1.1. Voyelles morphologiques

Les voyelles morphologiques sont celles qui appartiennent au schéma vocalique du mot, c'est-à-dire celles qui font partie de sa structure morphologique.

Ces voyelles ne sont pas toutes nécessaires à la lecture ou à l'interprétation et à l'analyse des mots. Cependant, pour réduire les cas d'ambiguïté virtuelle[127], certaines voyelles doivent impérativement être saisies, comme, par exemple, la *Ãamma* sur le *yâÞ al-muÃâraŕa* du verbe de IV^e^ forme conjugué à la troisième personne du singulier masculin au présent يُكتب « *yuktibu* » pour le distinguer du verbe de première forme conjugué à la troisième personne du singulier masculin au présent يَكتب « *yaktubu* ». D'un autre côté,

[127] Voir la définition du terme plus loin.

la forme يُكتب peut être analysée de deux manière : le même verbe conjugué à la même personne soit à la voix active soit à la voix passive ; c'est pourquoi on ajoutera, pour distinguer ces deux formes, sur la lettre médiane du verbe (ici le *tâÞ*) une *fatÎa* pour la voix passive يُكتَب « *yuktabu* » ou une *kasra* pour la voix active يُكتِب « *yuktibu* ».

Nous appellerons ce type de voyelles **les voyelles distinctives**. Dans le cas d'une homographie virtuelle, nous appelons donc **voyelle distinctive** la voyelle qui, à une position donnée du schéma vocalique, marque la différence entre un vocalisme et un autre. C'est la (les) voyelle(s) qui nous permet(tent) de distinguer deux (ou plusieurs) homographes consonantiques[128].

Vu leur importance et leur rôle considérable dans le filtrage des ambiguïtés virtuelles, nous préconisons de saisir **les voyelles morphologiques distinctives**.

Toutes les voyelles morphologiques ne sont donc pas nécessaires à saisir et encore moins certaines voyelles "inutiles". En effet, il n'y a aucun intérêt à saisir, par exemple, la *fatÎa* devant le *alif* « ...ـَا », voyelle longue qui a le même timbre, ou devant le *tâ' marbûÔa* « ...ـَة » puisque dans les deux cas la voyelle précédant ces deux lettres ne peut être que la *fatÎa*[129]. De même, il est inutile de saisir la *kasra* au-dessous du *hamza*, lui-même écrit au-dessous du *alif* « إ », puisque cette graphie n'est possible que si la voyelle du *hamza* est la *kasra* « ـِ » sinon il serait écrit au-dessus du *alif* « أ ».

---

[128] En ce qui concerne l'homographie consonantique voir Chapitre 5

[129] C'est, d'ailleurs, pour cette raison que, dans pratiquement tous les systèmes de translittération issus de la tradition orientaliste, le *tâ' marbûta* « ة ou ـة » est transcrit par / *at* / et non pas seulement par la consonne / *t* /.

Comme les règles régissant la saisie des voyelles morphologiques, étape qui se fait en amont, sont quasiment les mêmes règles gérant l'harmonisation vocalique qui, elle, se fait le cas échéant en aval, toutes ces règles seront présentées plus loin quand nous traiterons de l'harmonisation[130].

1.4.1.2. Voyelles casuelles

Comme l'arabe est une langue flexionnelle, non seulement les verbes sont conjugués, mais les noms et les adjectifs sont aussi déclinés. La déclinaison des noms et des adjectifs est une déclinaison à trois cas[131] : le cas sujet (nominatif) dont la marque est la *Ãamma* « ُ », le cas direct (accusatif) dont la marque est la *fatÎa* « َ » et le cas indirect (génitif) dont la marque est la *kasra* « ِ ». Ces marques sont placées au-dessus ou au-dessous de la dernière consonne du nom ou de l'adjectif[132].

Nous proposons de **ne pas saisir les voyelles casuelles** car elles ne sont nécessaires, ni à l'opération de segmentation ni à la catégorisation lexicale. En revanche, pour un traitement automatique à visée d'analyse syntaxique, il peut y avoir des cas d'ambiguïté pour lesquels la saisie de certaines voyelles casuelles s'avère nécessaire.

---

[130] Voir chapitre 5.

[131] Il existe certains noms et adjectifs qui, s'ils ne sont ni définis par l'article ni premiers termes d'une annexion, n'ont que deux voyelles de déclinaison, la *fatÎa* et la *Ãamma* ; ils ne peuvent non plus avoir de *tanwîn*. Ce sont ce que l'on appelle les diptotes.

[132] Il faut remarquer que les voyelles finales dans les noms et adjectfs au pluriel externe masculin (ex. معلّمونَ – معلّمينَ) et les noms et adjectifs au duel, masculin et féminin, (ex. معلّمانِ – معلّمَيْنِ ؛ معلّمتانِ – معلّمتَيْنِ) ne sont pas des voyelles casuelles.

1.4.1.3. Voyelles d'appui

Appelées aussi voyelles de liaison, les voyelles d'appui sont les trois voyelles brèves (la *Ãamma*, la *fatÎa* et la *kasra*) utilisées, dans un cadre de conditionnement phonologique (syllabique), pour remplacer un *sukûn* « ْ » à la fin d'un mot-outil ou d'un verbe suivis d'un mot commençant par un *hamza* instable, et ce dans le but de résoudre un problème phonologique.

En effet, il est impossible, en arabe, d'avoir deux *sukûn* qui se suivent immédiatement. Alors, quand on se trouve devant le cas où un mot se terminant par un *sukûn* (مِنْ *min* « de », مَنْ *man* « qui ?, celui qui », هُمْ *hum* « eux », قَالَتْ *qâlat* « elle a dit ») et suivi par un mot commençant par un *hamza* instable, on remplace le *sukûn* du premier mot par une voyelle d'appui ; cette voyelle est en général une *kasra* sauf dans les cas suivants :

- une *fatÎa* quand il s'agit de مِنْ suivie de l'article ال.
- une *Ãamma* quand il s'agit des pronoms pluriels masculins كُمْ , هُمْ ou du suffixe verbal du pluriel masculin ...تُمْ.

Les exemples qui suivent enregistrent bien le remplacement du *sukûn* dans chacun des cas énumérés (ceci représente le cas général, mais il y a quelques exceptions) :

مِنْ + الْكتاب	→	مِنَ ٱلْكتاب	« du livre »
مَنْ + الْولد ؟	→	مَنِ + ٱلْولد ؟	« Qui est l'enfant ? »
قَالَتْ + الْبنت	→	قَالَتِ + ٱلْبنت	« La fille a dit »
هُمْ + الناس	→	هُمُ + ٱلناس	« Ce sont les gens »

Étant donné que le seul rôle de ces voyelles est phonologique (faciliter la prononciation), nous suggérons de **ne pas saisir les voyelles d'appui.** De plus, leur saisie va à l'encontre du principe-même d'harmonisation.

1.4.2. Les deux principes de la voyellation

En plus de ce qui a été présenté jusque là concernant la saisie des voyelles, nous avons pu poser deux principes qui nous ont guidé dans l'établissement de cette *norme de saisie et d'harmonisation* que nous présentons ici : le principe de l'économie de voyelles et celui des degrés de voyellation .

- Principe de l'économie de voyelles

Partant du constat présenté plus haut selon lequel les textes arabes ne sont jamais entièrement voyellés et considération faite de la complexité et du coût élevé en termes de temps et d'investissement qu'engendre une opération de voyellation totale, la meilleur solution pour le traitement automatique de l'arabe reste une opération de voyellation partielle normalisée, bien menée et combinée, à terme, à une opération d'harmonisation corrigeant ce qui pourrait échapper à l'opérateur de saisie comme erreurs ou maladresses. Cette opération doit viser un double objectif : une harmonisation totale de toutes les occurrences d'un même vocable, et un degré zéro d'ambiguïté virtuelle[133].

Pour ce faire, un principe à la fois rigoureux et simple peut faciliter la tâche et garantir de bons résultats : il s'agit du principe de l'économie des voyelles. En effet, pour voyeller, par exemple, une unité lexicale simple de six consonnes on commence d'abord par voir si une non-voyellation totale (c'est-à-dire aucune consonne n'est voyellée) est susceptible d'engendrer des ambiguïtés virtuelles. Si tel est le cas, on

[133] Pour la notion d'ambiguïté virtuelle, voir chapitre 5.

essaye d'abord de voyeller une seule consonne (la consonne distinctive), puis s'il y a toujours une quelconque ambiguïté, deux consonnes distinctives, puis trois et ainsi de suite jusqu'à éventuellement voyellation totale de l'unité lexicale si nécessaire.

Il est bien évident que cette règle ne va pas être testée aveuglément sur toutes les unités lexicales une par une. Elle sera appliquée à des catégories, des sous-catégories, des groupes, des schèmes... C'est ce que nous avons tenté de faire pour la saisie et pour l'harmonisation primaire.

- Principe des degrés de voyellation

Nous avons adopté, pour les mots lexicaux, une règle qui gère le degré de voyellation. Celle-ci doit, bien évidemment, être combinée avec la règle d'économie de voyelles évoquée plus haut.

En effet, cette règle attribue à des catégories de mots lexicaux, à des sous-catégories ou à des schèmes un degré de vocalisation allant, dans un ordre décroissant, de la vocalisation totale, par exemple, pour les verbes à la non vocalisation, par exemple, pour les noms, passant par les adjectifs, les participes et les *maÒdar*.

Ces deux règles de voyellation sont incorporées dans les règles d'harmonisation présentées *infra*.

1.4.3. Voyelles longues

1.4.3.1. Le *Þalif maqÒûra*

Certains mots se terminant par une voyelle longue « â » ne sont pas écrits, en finale, avec un *alif* "normal" mais avec un *alif* "tordu" appelé *ÞalifmaqÒûra* qui a la forme du *yâÞ* final ou isolé, selon la lettre qui le précède, mais sans les points du *yâÞ*.

Ce qui est à noter, ici, c'est que quand ces mots sont agglutinés à des pronoms suffixes, le *ÞalifmaqÒûra* est transformé soit en *yâÞ*, uniquement dans certains cas comme les prépositions إلى ou هـ + إلى (على إليه « à lui »), soit en *Þalif* normal (أبى + ها أباها « il l'a refusée »).

Ce phénomène doit attirer l'attention de l'opérateur humain (ou de l'informaticien au moment de l'écriture de son programme automatique) pour qu'il prévoie de rendre, après l'opération de segmentation, au mot sa graphie d'origine (avec le *ÞalifmaqÒûra*) dans le cadre de ce que nous avons appelé l'**harmonisation régulatrice**[134].

1.4.3.2. Le alif suscrit

La présentation de l'*alif* suscrit a été faite ci-dessus au paragraphe « Les irrégularités de l'arabe ».

En dehors du texte coranique, **l'*alif* suscrit n'est pratiquement jamais écrit**. Cependant sous MS Word®, il est automatiquement ajouté au-dessus du mot الله , il suffit, en effet, de taper هـ + ل + ل + ا pour que le logiciel de traitement de texte transforme le tout en cette graphie : الله en ajoutant automatiquement non seulement le *Þalif* suscrit mais aussi la *šadda*.

---

[134] Voir p. 147.

1.5. Saisie de quelques consonnes

1.5.1.Le *hamza*

L'écriture du *hamza* constitue le seul problème majeur que l'orthographe arabe connaisse.

Le *hamza* s'écrit soit seul, directement sur la ligne (ء), soit au-dessus ou au-dessous d'un support qui peut être le *Þalif* (أ), le *wâw* (ؤ) ou le *yâÞ* sans les points qu'on appelle la *nibra* (ئ ئـ - ـئـ) et ce en fonction de la voyelle du hamza et de celle le précédant.

- Au début du mot :

Au début du mot, le *hamza* a toujours l'*alif* comme support. Si la voyelle du *hamza* est une *fatÎa* ou une *Ãamma*, le *hamza* est écrit au-dessus de l'*alif* (أ). Si sa voyelle est une *kasra*, le *hamza* est écrit au-dessous de l'*alif* (إ).

- Au milieu du mot :

Les règles d'écriture du *hamza* sont basées sur un certain « rapport de force » des voyelles qui sont classées, de la plus forte à la plus faible, comme suit : la *kasra* « ِ », la *Ãamma* « ُ », la *fatÎa* « َ » et en fin le *sukûn* « ْ » (i > u > a > ∅).

Le *hamza* est écrit sur le support qui correspond à la voyelle la plus forte entre celle du *hamza* et celle qui le précède. Toutefois, une seule exception est à noter ici : le *hamza* n'a aucun support s'il est précédé d'un « â » et porte une *fatÎa*.

Le tableau suivant résume ces règles :

Au milieu du mot				
Voyelle avant le *hamza* ➔ / ➘ Voyelle du *hamza*	*kasra* « ِ »	*Ãamma* « ُ »	*la fatÎa* « َ »	*sukûn* « ْ »
kasra « ِ »	ئ	ئ	ئ	ئ
Ãamma « ُ »	ئ	ؤ	ؤ	ؤ
la fatÎa « َ »	ئ	ؤ	أ	أ
sukûn « ْ »	ئ	ؤ	أ	ء

Tableau 3 :
Les différents supports du *hamza* au milieu du mot

➘ A la fin du mot :

A la fin du mot, le support du *hamza* dépend uniquement de la voyelle le précédant :

À la fin du mot				
Voyelle avant le hamza ➔	*kasra* « ِ »	*Ãamma* « ُ »	*la fatÎa* « َ »	*sukûn* « ْ »
	ئ	ؤ	أ	ء

Tableau 4
Les différents supports du *hamza* à la fin du mot

En fait, nous avons tenu à rappeler ici les règles d'écriture du *hamza* dans le but principal d'insister sur la nécessité de régulariser, après la segmentation, la graphie des mots comportant un *hamza*.

1.6. Saisie des signes diacritiques :

1.6.1. La *šadda* : (marque de gémination consonantique)

En traitement automatique de l'arabe, du fait de son statut particulier différent des autres signes diacritiques, la *šadda* doit être considérée, informatiquement parlant, comme une consonne à part entière. En effet, elle n'est, théoriquement mais aussi dans la pratique, que le remplacement d'un redoublement de consonnes dont la première est quiescente, c'est-à-dire sans voyelle. C'est pour cette raison que dans les entrées des dictionnaires, lexiques, manuels de conjugaison, etc., toujours basées sur la racine, la *šadda* est remplacée par la lettre à laquelle elle est censée se substituer. Dans la plupart des dictionnaires arabes, pour trouver par exemple, le verbe شَدَّ *šadda* « attacher, serrer », il faut aller chercher la racine شدد *šdd*.

Ne pas saisir la *šadda*, c'est priver le récepteur humain, mais surtout le récepteur machine[135], d'un moyen très important d'interprétation et donc d'analyse lui permettant de filtrer et de réduire les ambiguïtés. C'est ce qui met en évidence, outre sa valeur morphologique, la valeur **distinctive** de la *šadda*. En effet, en écrivant la *šadda* sur la

[135] Selon le mode d'existence des ambiguïtés, C. Fuchs énumère trois types de récepteurs : le récepteur humain, le récepteur linguiste et le récepteur machine. Catherine Fuchs, *Les ambiguïtés du français*, 1996, pp. 49-53.

deuxième consonne, par exemple, de la graphie non voyellée ambiguë كتب *ktb* cela nous permet d'éliminer d'emblée treize interprétations parmi les seize envisageables[136] de ce mot en ne gardant, à partir de la graphie non voyellée ainsi obtenue (كتّب *kttb*), que trois analyses possibles كَتَّبَ *kattaba* « faire écrire, *à la voix active* », كُتِّبَ *kuttiba* « faire écrire, *à la voix passive* » et كَتِّبْ *kattib* « fais écrire, *(impératif)* ». Cette valeur distinctive est explicite en transcription phonologique où le verbe *kattaba* est nettement distinguable du verbe *kataba*. Il n'est pas inutile de rappeler que la valeur distinctive de la consonne géminée est aussi observable dans d'autres langues : ainsi en italien, par exemple, la gémination de la lettre « n » permet de distinguer le verbe *dona* « il donne, il fait cadeau » du nom *donna* « dame ».

En plus de ce statut particulier qu'a la *šadda*, son traitement informatique en tant que consonne à part entière devrait nous conduire d'une façon très efficace à avoir, au moins, deux atouts majeurs :

- faciliter énormément la désambiguïsation, en synthèse[137] ;
- diminuer considérablement l'explosion combinatoire, en analyse.

Il est donc **impératif que la *šadda* soit toujours saisie**. Une exception doit, cependant, être signalée : c'est le cas où la *šadda* est placée après l'article suivi d'une

---

[136] Sur les 16 analyses possibles du mot كتب voir : R. Ouersighni, *La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : Utilisation pour la détection et le diagnostic des fautes d'accord*, Thèse de doctorat, 2002, p.19.

[137] Dans sa thèse, Riadh Ouersighni a calculé le taux d'ambiguïté lexicale liée à la *šadda* : « Nous pouvons dire que la « *šadda* », à elle seule, est responsable de 39% de l'ambiguïté lexicale totale ». ibid, p. 64.

lettre dite solaire[138]. Dans ce cas, la *šadda* est due à un conditionnement phonologique causé par l'assimilation du *lâm* de l'article.

1.6.2. Le *tanwîn*[139]

Le *tanwîn* est la marque de l'indétermination des noms et des adjectifs. Par définition, c'est l'adjonction, à la voyelle casuelle, de la valeur phonétique d'un *nûn* (la lettre "n"). Il est effectué en redoublant la graphie de la voyelle brève marquant le cas du nom ou de l'adjectif. La marque du nom ou de l'adjectif indéterminé nominatif est « ٌ », celui du nom ou de l'adjectif indéterminé accusatif est « ً » et celui du nom ou de l'adjectif indéterminé génitif est « ٍ ». À l'accusatif, un *Þalif* orthographique est ajouté à la fin du mot après le *tanwîn* sauf quand ce mot se termine par un *tâÞ marbûÔa* « ةً », un *Þalif maqÒûra* « ىً », un *hamza* sur un *Þalif* « أً » ou un *hamza* précédé d'un *Þalif* « ءً ».

Le *tanwîn* ne sera pas saisi sauf celui de l'accusatif suivi du *Þalif* orthographique et ce pour distinguer, éventuellement, le nom ou le *maÒdar* indéfini accusatif du verbe conjugué à la 3e personne du duel masculin :

كَتَبَا ≠ كُتُبًا

---

[138] Les lettres solaires sont : ("ت" ، "ث" ، "د" ، "ذ" ، "ر" ، "ز" ، "س" ، "ش" ، "ص" ، "ض" ، "ط" ، "ظ" ، "ل" ، "ن").

[139] Nous ne rentrons pas ici dans les détails de la classification que fait la grammaire classique du *tanwîn* en *tanwîn ÞaÒîl* « authentique » et *tanwîn Èayr ÞaÒîl* « non- authentique » et de celle de la première classe en *tanwîn at-tamakkun* « d'affermissement », *tanwîn at-tankîr* « d'indéfinition », *tanwîn al-ÝiwaÃ* « de compensation » et *tanwîn al-muqâbala* « de correspondance ». Pour plus de détail voir : G. Ayoub, La nominalité du nom ou la question du *tanwîn*, in *Arabica*, tome 38, Brill, Leiden, 1991, pp. 151-213 et D. E. Kouloughli, Sur le statut linguistique du *tanwîn*. Contribution à l'étude du système déterminatif de l'arabe, in *Arabica*, tome 48, Brill, Leiden, 2001, pp. 21-50.

Des livres *(accusatif)*		Ils ont écrit *(duel)*
هَرَبًا	≠	هَرَبَا
Le fait de fuir *(accusatif)*		Ils ont fui *(duel)*

Ceci a pour but principal de réduire les ambiguïtés virtuelles.

Il est important de noter ici qu'il ne faut pas saisir le *tanwîn* au-dessus du *þalif* orthographique comme le font, à tort, certains imprimeurs dans certains pays arabes. Il faut, au contraire, suivre la tradition typographique arabe en saisissant le *tanwîn* avant le *þalif* orthographique, c'est-à-dire au-dessus de la consonne qui précède ce dernier.

1.6.3. Le *sukûn* (ou « voyelle zéro »)

L'absence de voyelles brèves est marquée, en arabe, par un signe diacritique placé toujours au-dessus de la consonne quiescente.

« Le *sukûn*, réalisé « ◌ْ » - le terme signifie littéralement « quiescence » -, note l'absence de / *a* /, de / *i* /, de / *u* /. De ce fait, il peut être dit « voyelle zéro » ; au demeurant c'est ce chiffre que son dessin reproduit. Et il peut être dit aussi « marque d'implosion » puisqu'il marque toujours une consonne post-vocalique et donc implosive »[140].

Sur le plan morphologique, le *sukûn* fait partie du schéma vocalique du mot. Cependant, l'absence de voyelles au niveau de la prononciation n'a pas toujours ce statut morphologique permanent. Elle peut, en effet, parfois traduire un état phonologique transitoire où, à la fin du mot, la voyelle zéro a une valeur pausale. Partant de la règle phonologique de l'arabe littéral qui stipule que l'arabe ne commence

[140] A. Roman, *Grammaire de l'arabe*, 1980, p. 13.

jamais par un *sukûn* et ne s'arrête jamais (au niveau de la prononciation) sur une voyelle, la voyelle finale du dernier mot de la chaîne parlée (c'est-à-dire à la pause), même si elle est écrite, n'est jamais prononcée[141]. Mais cette valeur pausale ne doit évidemment pas être traduite, graphiquement, par le *sukûn*. On peut, à la limite, ne pas écrire la voyelle finale ; mais la remplacer par un *sukûn* serait, en prose, une faute.

En dehors de certains mots-outils dont la forme d'harmonisation choisie laisse apparaître le *sukûn* final ou, à la limite, certaines formes verbales au jussif, on ne doit donc **pas écrire, à la fin des mots, un *sukûn*** à valeur pausale.

Une autre mauvaise habitude typographique chez certains imprimeurs est aussi à proscrire, il s'agit du *sukûn* placé sur le *wâw* ou le *yâÞ* quand ces deux derniers sont des voyelles longues.

En effet, partant du fait que « le "و" et le "ي" sont des voyelles longues : û […] et î, quand elles ne sont pas articulées avec une autre voyelle qui suit »[142], certains imprimeurs du moyen orient (c'est malheureusement aussi une habitude au niveau de l'écriture manuscrite), pour montrer que ces deux semi-voyelles prennent la valeur de voyelles longues, placent dessus un *sukûn*. Outre le fait que ce choix n'a aucun fondement théorique, il peut être à l'origine, pour le récepteur machine, de certaines ambiguïtés virtuelles dont on peut se passer volontiers. De ce fait, **placer un *sukûn* sur une voyelle longue est une chose à bannir.**

---

[141] Exception, bien sûr, faite des quelques mots-outils qui se terminent par un *sukûn*. Mais dans ce cas, le *sukûn* fait partie de la structure morphologique de ces mots-outils.

[142] R. Blachère, Eléments de l'arabe classique, p. 11.

1.6.4. La *waÒla*

Par opposition au *hamza* de coupure (ou *hamza stable*) همزة القطع *hamzat al-qaÔ'* qui fait partie intégrante du mot et est toujours prononcé, le *hamza* de liaison (ou *hamza instable*) همزة الوصل *hamzat al-waÒl* est un *hamza* initial qui est écrit (il ne l'est pas systématiquement) avec son support (le *alif* de l'article) et prononcé quand il commence une phrase (الوصل **...** *al- waÒlu...* , ... أُكْتُبْ *Þuktub...*), alors qu'il n'est pas prononcé, puisqu'assimilé à la lettre qui le suit, et seulement son support (le *alif*) est écrit quand il est au milieu d'une phrase, c'est-à-dire quand il est précédé d'un mot ou d'une particule (همزة ٱلوصل *hamzatu-l-waÒli*, إقرأ درسك وٱكْتُبْ فرضك *ÞiqraÞ darsa-ka wa-ktub farÃa-ka*). Pour marquer graphiquement l'assimilation du hamza on utilise, parfois, un signe diacritique spécial écrit au-dessus du *Þalif* (ٱ) et c'est, en fait, ce signe-là que l'on appelle la *waÒla*.

Comme l'utilisation de ce signe diacritique n'est pas systématique surtout dans les tapuscrits, nous préconisons, dans un souci d'harmonisation, de ne pas saisir la *waÒla*. L'*alif* , seul, est donc saisi à la place du « *Þalif waÒla* en chef »

1.6.5. La *madda*

Remplaçant un *hamza* écrit sur un *alif* suivie de la voyelle longue « â » (أا) ou un *hamza* écrit sur un *alif* avec, comme voyelle, la *fatÎa*, suivie d'un autre *hamza* quiescent également écrit sur un *alif* (أْأ) ; le *alif* allongé ألف ممدودة *Þalif mamdûda* est marqué graphiquement par un *alif* surmonté par un signe diacritique bien particulier ressemblant au tilde (آ), ce qui donne l'image en quelque sorte d'un *alif* couché. Compte tenu de

son caractère morphologique**, nous préconisons que la *madda* ou plus précisément le « *alif madda en chef* » soit systématique saisi.**

1.7. Saisie des caractères spéciaux :

Les caractères spéciaux de mise en forme ainsi que les ligatures esthétiques sont à proscrire totalement de tout enregistrement informatique des textes à visée de traitement automatique et surtout de lexicométrie.

1.7.1. La *kašida*

Les *kašîdas* sont des caractères spéciaux utilisés d'abord dans un but esthétique dans la calligraphie arabe, puis dans un but pratique pour étendre le joint entre deux caractères arabes. Elles sont donc utilisées pour améliorer l'aspect d'un texte justifié en allongeant les mots plutôt qu'en augmentant l'espacement entre les mots.

Il ne faut donc pas utiliser les signes kachidés parce que la *kašida* est conservée lors de la segmentation et des étapes suivantes. Ceci va engendrer l'éparpillement des formes sur plusieurs vocables alors qu'elles ne sont, en réalité, que les occurrences d'un même et unique vocable. On peut, en effet, avoir la même forme écrite tantôt sans aucune *kašida*, tantôt avec une seule, tantôt avec deux *kašîdas* ou plus.

Exemple :

كتاب peut contenir une ou plusieurs *kašîdas* :

كتاب =

كتـاب =

كتـــاب =

كـتـــاب =

1.7.2. Les ligatures esthétiques

Utilisées dans un but exclusivement esthétique, ces ligatures sont des formes calligraphiques représentant des phrases entières dont les éléments sont entrelacés d'une façon très harmonieuse. Elles ne sont donc pas indispensables sous cette forme et doivent, de ce fait, être remplacées par leurs composantes, non-ligaturées sans changer la validité grammaticale ou le sens du texte à traiter automatiquement.

Exemple :

Caractère	Doit être remplacé par :	Traduction
ﷺ	عيد سعيد	« Joyeuse fête »
ﷺ	كلّ عام وأنتم بخير	« Bonne année »
ﷺ	بسم الله الرحمان الرحيم	« Au nom de Dieu, clément et miséricordieux »
ﷺ	تمّ بحمد الله	« Fin, louange à Dieu »,
ﷺ	الله أكبر	« Dieu est le plus grand »

1.7.3. La ponctuation esthétique

Dans certains textes arabes, comme le texte coranique ou autres, on peut trouver des signes de ponctuation sous des formes esthétiques différentes de la forme conventionnelle et ont donc des codes machine différents du code usuel.

Il est donc nécessaire de remplacer ces formes esthétiques par les formes usuelles. Les parenthèses, par exemple, du type « ﴿ ﴾ », « ﴿ ﴾ », ou « ﴿﴾ » doivent être remplacées par « ( ) » ou par des guillemets « " » quand ils renferment une citation coranique.

1.7.4. Les chiffres esthétiques

Comme les signes de ponctuation, on peut trouver dans certains textes arabes, comme le texte coranique, des chiffres sous des formes esthétiques non-conventionnelles et ayant des codes machine non-usuels.

Il est donc nécessaire de remplacer ces chiffres esthétiques par les formes usuelles. Les chiffres, par exemples, du types « ❶, ❷, ❸... » ou « ①, ②, ③... » doivent être remplacés par « 1, 2, 3... ». Il est nécessaire de noter ici que la saisie directe des chiffres ou le remplacement des chiffres esthétiques doit se faire d'une façon homogène en utilisant soit les chiffres arabes (ce que nous suggérons fortement) soit les chiffres indiens et non pas les deux. Ceci, bien entendu, dans le cas où l'on décide de saisir les nombres en chiffres et non en toutes lettres. On pourrait, le cas échéant, faire le choix de convertir, au moment de l'harmonisation, ces chiffres en lettres comme nous le préconisons plus loin dans la *norme d'harmonisation*.[143]

[143] Voir chapitre 5.

2. Norme d'harmonisation des textes arabes

L'harmonisation graphique est une étape très importante dans toute opération d'enregistrement de corpus sur support informatique. Accompagnant ou se succédant à la saisie, cette opération doit garantir une meilleure uniformité graphique et une certaine constance qui permettront par la suite un meilleur passage des mots graphiques aux formes (items) et des formes aux listes, base de tout traitement lexicométrique. Elle devrait permettre, en outre, de minimiser au maximum les opérations de correction et de retour en arrière.

L'harmonisation graphique doit avoir comme objectifs de réduire sinon d'annuler les ambiguïtés virtuelles, de permettre l'homogénéité des listes de fréquences c'est-à-dire d'éviter la dispersion d'un même lexème dans des listes des fréquences sous différentes formes et de garantir par conséquent des calculs fiables.

En outre, une bonne harmonisation devrait faciliter les étapes suivantes, à savoir la segmentation (harmonisation primaire), la lemmatisation et la catégorisation (harmonisation régulatrice). Pour ce faire, la *norme d'harmonisation*, dans le cas de choix multiples, doit opter pour des choix, bien évidemment fondés théoriquement, mais surtout facilitant les étapes suivantes du dépouillement du corpus.

2.1. Les règles d'harmonisation (et de saisie)

2.1.1. Harmonisation primaire (avant segmentation)

Nous tenons d'abord à signaler ici que bon nombre de ces règles d'harmonisation sont aussi des règles de saisie. De ce fait, et dans le cas de la constitution de corpus par saisie directe sur clavier et non pas par scannérisation ou autre moyen, une bonne saisie, c'est-à-dire une saisie respectant toute les règles énumérées, réduirait d'une façon considérable les opérations d'harmonisation et permettrait un gain de temps inestimable. Dans le cas contraire, une application rigoureuse des règles d'harmonisation serait donc nécessaire.

Ce type d'harmonisation que nous appelons harmonisation primaire est opéré en amont, avant toute opération de segmentation ou de lemmatisation. L'autre type d'harmonisation que nous appelons harmonisation régulatrice est éventuellement nécessaire en aval, après la segmentation pour réparer d'éventuels "dégâts" orthographiques causés par la rupture de la cursivité de l'écriture des mots graphiques et donc des changements de règles d'orthographe.

2.1.1.1. Harmonisation des Verbes

Les verbes doivent être entièrement vocalisés, surtout le lemme, c'est-à-dire le verbe correspondant à la troisième personne du singulier masculin de l'accompli actif. En effet, non vocalisée, cette forme produit une grande partie des homographes consonantiques et représente ainsi une source importante d'ambiguïté polycatégorielle (Verbe/Participe actif, Verbe/Élatif, Verbe/*MaÒdar*,...) et même monocatégorielle (Verbe/Verbe).

2.1.1.2. Harmonisation des Nombres

Les nombres sont, eux aussi, à vocaliser entièrement qu'ils soient cardinaux ou ordinaux. Cependant, un cas spécifique est à noter en ce qui concerne l'ordinal "Deuxième" « ثَانٍ ». Cet ordinal, défini, s'écrit toujours avec un *yâÞ* final « الثانِي », et indéfini s'écrit sans ce *yâÞ* aux cas sujet et indirect « ثَانٍ » mais retrouvant son *yâÞ* final au cas direct « ثَانِيًا ». Ce comportement n'est pas propre à ce nombre, il caractérise tout un groupe de noms réunissant ce que l'on appelle les noms "incomplets" « الأسماء المنقوصة » et une partie des pluriels des noms se terminant par un *Þalif maqÒûra* « الأسماء المقصورة ».

Outre l'harmonisation vocalique des nombres saisis en toutes lettres et l'harmonisation lexicale de l'ordinal « ثَانٍ », un autre choix doit cependant être arrêté concernant les nombres en chiffres. Une conversion de ces chiffres en lettres doit, en effet, être effectuée pour garantir une harmonisation et une homogénéité totale de tous les nombres du corpus. Ce choix est justifié au niveau théorique puisqu'aucun dictionnaire de langue ne comporte d'entrée à "73", par exemple, mais à "trois" et à "sept".

2.1.1.3. Harmonisation des Noms et des *MaÒdars* primitifs

Pour tous les noms et *maÒdars* qui ont pour schème *fiÝl*, *fuÝl*, ou *faÝl* فَعِل – فُعل – فَعل, le *sukûn* écrit sur la lettre médiane permet de les distinguer, d'une part des verbes

trilitères simples *faʕala- faʕula- faʕila* فَعَلَ- فَعُلَ- فَعِلَ , et d'autre part, d'autres noms et *maÒdars* de schème *fiʕil*, *fiʕal*, *fuʕal*, *faʕul*, *faʕil*, ou *faʕal* فَعَل فَعِل فَعُل فُعَل فِعَل فِعِل:

قلْب	**Nom** *(cœur)*	≠	قَلَبَ	**Verbe** *(inverser)*
كبْر	***MaÒdar*** *(arrogance)*	≠	كَبُرَ	**Verbe** *(grandir)*
علْم	**Nom** *(science)*	≠	عَلِمَ	**Verbe** *(savoir)*
نفْس	**Nom** *(âme)*	≠	نَفَس	**Nom** *(souffle)*
مثْل	**Nom** et **Adj**. *(analogue)*	≠	مَثَل	**Nom** *(exemple, parabole)*
رِجْل	**Nom** *(pied)*	≠	رَجُل	**Nom** *(homme)*

Dans le cas d'ambiguïté lexicale entre deux noms ou *maÒdars* ayant un *sukûn* sur la médiane, ce qui est très rare pour ces schèmes, la voyelle de la première consonne est alors ajoutée.

2.1.1.4. Harmonisation de certains Noms et *MaÒdars* dérivés

Quatre *maÒdars* dérivés (trois trilitères dérivés et un quadrilitère dérivé) sont homographes consonantiques soit du participe actif féminin soit des verbes dérivés correspondants. En effet, le *maÒdar* trilitère dérivé de troisième forme مُفَاعَلَـةٌ est homographe consonantique du participe actif féminin de la même forme مُفَاعِلَةٌ . Il est donc nécessaire, pour pouvoir les distinguer de n'écrire que la « *fatÎa* » pour le *maÒdar* et la « *kasra* » pour le participe actif féminin. Quant aux trois autres *maÒdars*, il s'agit des deux *maÒdars* trilitères dérivés de cinquième et de sixième formes تَفَعُّـل et تَفَاعُـل homographes consonantiques, respectivement, des verbes de cinquième et de sixième

formes تَفَعَّلَ et تَفَاعَلَ . Les verbes étant, selon notre *norme*, entièrement vocalisés, il suffit donc de saisir, pour les *maÒdars*, seulement la « *Ãamma* » au-dessus de la troisième lettre de la racine pour la sixième forme : تفاعُل , et la « *Ãamma* » au-dessus de la troisième lettre de la racine et de la *šadda* pour la cinquième forme : تفعُّل . Le dernier *maÒdar*, quadrilitère, est تَفَعْلُل *tafaÝlul* qui est homographe consonantique du verbe quadrilitère dont il dérive تَفَعْلَلَ *tafaÝlala* ; une *Ãamma* sur le premier *lâm* du *maÒdar* permet de lever l'ambiguïté.

En ce qui concerne les noms dérivés, une forte homographie est observable pour deux formes graphiques مفعل et مفعلة réparties entre noms de lieu et de temps, nom d'instrument et *maÒdar mîmî* et qui, combinées à des vocalismes différents (jouant notamment sur la première lettre م et la troisième lettre ع), engendrent douze homographes consonantiques dont huit appartenant à ces trois catégories de noms dérivés et quatre à des participes (actif et passif, masculin et féminin). Le tableau suivant regroupe tous les homographes consonantiques engendrés par ces deux seules formes :

	مفعل *mf Ýl*	مفعلة *mf Ýla*
Noms de temps et de lieu **اسما الزمان والمكان**	**مَفعَل** (مَكتَب *maf Ýal*) **مَفعِل** (مَجلِس *maf Ýil*)	**مَفعَلة** (مَكتَبة *maf Ýala*)
Noms d'instrument **اسم الآلة**	**مِفعَل** (مِصعَد *mif Ýal*)	**مِفعَلة** (مِلعَقة *mif Ýala*)
Masdar mîmî **المصدر الميمي**	**مَفعَل** (مَدخَل *maf Ýal*) **مَفعِل** (مَوعِد *maf Ýil*)	**مَفعِلة** (مَعرِفة *maf Ýila*)
Participe actif **اسم الفاعِل**	**مُفعِل** (مُكتِب *muf Ýil*)	**مُفعِلة** (مُكتِبة *muf Ýila*)
Participe passif **اسم المفعول**	**مُفعَل** (مُدخَل *muf Ýal*)	**مُفعَلة** (مُدخَلة *muf Ýala*)

Tableau 5 :
L'homographie consonantique des deux schèmes *mafÝal* et *mafÝala*

Il est donc impératif, pour pouvoir distinguer ces différents homographes consonantiques, de saisir les deux voyelles distinctives : celle placée sur la première lettre « م » et celle placée sur la troisième lettre « ع ».

Cependant, il est à noter que, parmi ces homographes, deux sont des homographes globaux présentant ainsi une ambiguïté effective polycatégorielle : مَفعَل (entre "Noms de temps et de lieu" et "*MaÒdar mîmî*") et مَفعِل (entre "Noms de temps et de lieu" et "*MaÒdar mîmî*"). Cette ambiguïté ne peut, en effet, être levée que par une opération de catégorisation affectant à chacun de ces homographes soit la catégorie des noms de temps et de lieu soit celle des *MaÒdar mîmî*.

2.1.1.5. Harmonisation des Participes

Le participe actif des verbes simples qui est construit sur le schème فاعِل *fâŕil* est un homographe consonantique du verbe dérivé de troisième forme فَاعَلَ *fâŕala*. Pour distinguer ces deux homographes consonantiques nous préconisons de n'écrire pour le participe actif que la « *kasra* » (la voyelle « i ») au-dessous de la consonne médiane de la racine. Le verbe, lui, est entièrement vocalisé.

قاتِل **Participe actif** *(tuant, assassin)* ≠ قَاتَلَ **Verbe** *(combattre)*

En ce qui concerne les participes des formes dérivées, pour distinguer le participe actif du participe passif, la voyelle de la consonne médiane doit être écrite. Il s'agit de la « *kasra* » pour le participe actif, et de la « *fatÎa* » pour le participe passif.

De plus, pour distinguer ces mêmes participes de certains mots lexicaux (nom de lieu, ...), la « *Ãamma*» (la voyelle « u ») doit être écrite sur la première lettre « م ». On pourra ainsi facilement distinguer par exemple مُكسِب « rentable, profitable » de مَكسب « gain, avantage », ou مُفاخِر « vantard » de مَفاخِر « exploits ».

	Participe actif	**Participe passif**
Deuxième forme	مُفعِّل *mufaŕŕil*	مُفعَّل *mufaŕŕal*
Troisième forme	مُفاعِل *mufâŕil*	مُفاعَل *mufâŕal*
Quatrième forme	مُفعِل *mufŕil*	مُفعَل *mufŕal*
Cinquième forme	مُتفعِّل *mutafaŕŕil*	مُتفعَّل *mutafaŕŕal*
Sixième forme	مُتفاعِل *mutfâŕil*	مُتفاعَل *mutfâŕal*
Septième forme	مُنفعِل *munfaŕil*	مُنفعَل *munfaŕal*
Huitième forme	مُفتعِل *mufŕaŕil*	مُفتعَل *mufŕaŕal*

Neuvième forme	مُفعَلّ *mufʕall*	
Dixième forme	مُستفعِل *mustafʕil*	مُستفعَل *mustafʕal*

Tableau 6 :
Homographie des participes

2.1.1.6. Harmonisation des élatifs

Pour distinguer les élatifs أفعل التفضيل *'afÝal at-tafÃîl* qui ont pour schème أَفْعَـل *'afÝal* des verbes de quatrième forme construits sur le schème أَفْعَـلَ *ÞafÝala* (qui, normalement, sont entièrement vocalisés), nous avons convenu de ne marquer des voyelles de ces élatifs que le *sukûn* sur la deuxième lettre et la *fatÎa* sur la troisième lettre.

Cas particulier : Certains élatifs ont un *hamza* comme deuxième lettre, ce *hamza* quiescent précédé d'un autre *hamza* ayant la *fatÎa* comme voyelle se transforme donc en *madda*[144]. Ne restant plus, de ce fait, que deux consonnes pouvant porter des voyelles, nous avons convenu de ne mettre pour ce type d'élatif aucune voyelle. Le verbe correspondant, quant à lui, prendra les deux voyelles :

آثر **Élatif** *(plus attirant)* ≠ آثَرَ **Verbe** *(préférer, favoriser)*

2.1.1.7. Harmonisation des « أسماء منقوصة »

Les noms "incomplets" **الأسـماء المنقوصـة** sont des noms déclinables dont la dernière consonne, quand ils sont définis ou premiers termes d'une annexion, est un *yâÞ* tels :

المحامي	« *al-muÎâmî* »	(L'avocat)
القاضي	« *al-qâÃî* »	(Le juge, le cadi)
الوادي	« *al-wâdî* »	(Le fleuve, l'oued)

[144] Voir § 1.6.5. la *madda*, p. 172.

La spécificité de ces noms est qu'ils perdent leur *yâÞ* final quand ils sont indéfinis et ne sont pas premiers termes d'une annexion et ce seulement aux cas sujet et indirect. De ce fait l'avant dernière consonne, désormais finale, se voit adjoindre une double *kasra*, تنوين الكسرة *tanwîn al-kasra*. Au cas direct, le *yâÞ* final se maintient. Il est à noter qu'une partie des pluriels des noms se terminant par un *Þalif maqÒûra*, « **الأسماء المقصورة** » sont aussi dans ce cas, comme, par exemple, المقهى « *al-maqhâ* » (Le café) qui forme son pluriel en المقاهي « *al-maqâhî* ».

Nom défini	Nom indéfini cas sujet	Nom indéfini cas indirect	Nom indéfini cas direct
المحامي	محامٍ	محامٍ	محاميًا
القاضي	قاضٍ	قاضٍ	قاضيًا
الوادي	وادٍ	وادٍ	واديًا
المقاهي	مقاهٍ	مقاهٍ	مقاهيَ

Dans un souci d'harmonisation d'abord, et pour faciliter la lemmatisation ensuite, nous suggérons donc de saisir le *tanwîn al-kasra* pour ce type de noms quand ils sont indéfinis. C'est d'ailleurs cette forme qui sera le lemme des **أسماء منقوصة** comme nous allons le voir plus loin dans la *norme de lemmatisation*. Quoi qu'il en soit, c'est au niveau de l'harmonisation régulatrice (voir *infra*) qu'il faudra faire très attention à ce type de noms quant à l'adjonction ou la perte du *yâÞ*.

2.1.1.8. Harmonisation des noms propres

En arabe classique, et encore de nos jours dans certains pays arabes, les noms propres de personne sont composés d'un grand nombre d'éléments appellatifs et identificatoires : la *kunya* (père de …), l'*ism* (prénom), le *nasab* (composé du nom du père et d'un grand nombre d'aïeux), la *nisba* (lieux de naissance et de résidence), le *laqab* (surnom tiré d'un trait physique ou d'un titre honorifique)…

Du fait donc de la longueur hors-pair de ces noms propres, il est très fréquent de trouver dans des textes arabes, surtout classiques, une même personne désignée à des endroits différents d'un texte par une multitude d'appellations toutes tirées de la longue chaîne d'éléments composant son nom. En effet, l'on peut désigner une personne par sa *nisba* à un endroit du texte, par sa *kunya* à un autre endroit, par son *laqab* à un troisième, par deux ou trois éléments à un quatrième endroit… et ainsi de suite.

Le résultat en est que lors du passage des formes aux listes, les occurrences du nom propre de telle ou telle personne se trouvent éparpillées sur plusieurs lemmes. Il est également possible de trouver des occurrences de deux ou trois formes nominales désignant des personnes différentes ayant la même *nisba* ou la même *kunya* groupées sous le lemme nominal d'une seule personne. Outre des ambiguïtés inopportunes, ceci présente l'inconvénient majeur de corrompre les données et de fausser les calculs.

Exemple de notre corpus :

Il s'agit de :	**ذو الكفايتيْن أبو الفتح علي بن أبي الفضل محمّد بن العميد**		
Cité en tant que :	ذي الكفايتين	ابن العميد	ابن العميد أبا الفضل
Référence :	p. 3 et p. 66	p. 16, p. 54, p. 61 et p.66	p. 132

Il s'agit de :	**أبو عبد الله العارض الحسين بن أحمد بن سعدان الوزير**	
Cité en tant que :	الحسين	أبي عبد الله العارض
Référence :	p. 139	p. 4

Pour éviter ces dégâts regrettables et contribuer à lever les ambiguïtés, nous préconisons d'harmoniser les parties d'un même nom propre à tous les endroits du corpus en ajoutant, là où il en manque, un, deux ou trois éléments identificatoires de façon à permettre facilement la reconnaissance d'une personne à partir d'un nombre minimal mais distinctif des éléments de son nom complet.

2.1.1.9. Harmonisation des dates

Comme dans la langue parlée, il arrive à un locuteur d'évoquer une date du vingtième siècle, par exemple, en ne citant que les dizaines et les unités de cette date composée normalement de quatre chiffres. On peut donc facilement dire « J'ai visité tel pays en 98 », pour parler bien évidement de l'année 1998.

Exemple de notre corpus :

Ce qui est écrit :	سنة سبعين l'année soixante-dix	سنة أربعين l'année quarante	سنة أربع وستين l'année soixante-quatre
Ce qu'il faut lire :	سنة سبعين وثلاثمئة l'année trois cent soixante-dix	سنة أربعين وثلاثمئة l'année trois cent quarante	سنة أربع وستين وثلاثمئة l'année trois cent soixante-quatre
Référence :	p. 105	p. 129	p. 137

Une harmonisation des dates dans ce cas est fortement conseillée, pour ne pas dire obligatoire, pour éviter d'avoir des dates erronées. Les circonstances historiques du discours devront théoriquement permettre de lever ce genre d'ambiguïté et d'harmoniser toutes les dates d'un corpus en ajoutant les milliers et les centaines pour une date millénaire et les centaines pour une date centenaire. Cette harmonisation est d'autant plus nécessaire que certains éditeurs critiques, en l'occurrence les deux éditeurs de notre

corpus, font parfois des notes de bas de page pour dire qu'il faut lire la date de telle ou telle façon.

Mais l'harmonisation des dates en arabe ne concerne pas seulement ce type de problème lié à la troncation des dates. Il y a également une autre constatation liée aux différentes tournures en arabe classique pour dire ou écrire une date concernant un jour du mois. En effet, en plus de la méthode habituelle qui consiste à écrire, par exemple, ليلة العاشر من شوّال (*la nuit du 10 du mois de Šawwâl*), il est fréquent de trouver dans des textes classiques la formulation suivante, لِخمسٍ خَلَتْ من شهرِ رجب (*à cinq nuits passées du mois de Rajab*), ou encore la tournure لِثلاثٍ بَقِينَ من شهرِ محرّم (*à trois nuits restant du mois de MuÎarram*)

Quelque soit donc la tournure utilisée ou le style préféré à tel ou tel auteur, il va falloir harmoniser toutes ces dates pour pouvoir les traiter d'une façon homogène et faciliter par là même leur traitement automatique.

En plus de ces deux problèmes qui nécessitent une harmonisation des dates, un troisième cas sollicite, lui aussi, une opération d'harmonisation : il s'agit de la façon de lire et surtout d'écrire, en arabe classique, les nombres au-delà des centaines. En effet, pour écrire en toute lettres, par exemple, 1998, il y a deux façon de le faire : soit on commence, de droite à gauche, par les unités puis les dizaines, les centaines et en fin les milliers, ce qui est plus logique et plus conforme au sens de l'écriture arabe : ثَمَانِيَةٌ وتِسْعُونَ وتِسْعُمِئَةٍ وأَلْفٌ , soit la façon dont on a l'habitude en arabe moderne et dont l'ordre de prononciation des chiffres est, comme en français, les milliers, les centaines, les unités et en fin les dizaines : أَلْفٌ وتِسْعُمِئَةٍ وثَمَانِيَةٌ وتِسْعُونَ . Si l'on choisit de convertir tous les nombres écrits en chiffres, en toutes lettres, une harmonisation est donc nécessaire pour ne choisir qu'une seule façon d'écrire les nombres en toutes lettres.

2.1.1.10. Harmonisation des signes de ponctuation

Dans le but de faciliter l'opération de segmentation, tous les signes de ponctuation doivent être précédés et suivis d'un espace. Ceci évitera d'avoir des cas de figures où l'on a des formes graphiques représentant des lexèmes liés à des virgules, points ou autres. Il est donc **préférable d'insérer avant chaque signe de ponctuation un espace**. Chaque signe de ponctuation sera ainsi encadré par deux espaces.

2.1.1.11. Harmonisation des mots-outils

Comme les mots-outils forment des listes fermées et qu'ils représentent tout de même 60 %[145] de l'ensemble des occurrences des textes arabes, il est important, et en même temps aisé, de dresser le tableau de toutes les formes pour chaque mot-outil et de choisir, parmi elles, en toute connaissance de cause la graphie d'harmonisation la plus adéquate respectant les règles de saisie et d'harmonisation énoncées plus haut.

[145] Voir notre chapitre "*Les catégories lexicales*" pp. 562-627

Différentes formes	Graphie d'harmonisation	Remarques
أ / أَ	**أ**	
إذا / إذَا / إذا / إذَا	**إذا**	
إذا / إذًا / إذا / إذًا	**إذًا**	
إذن / إذَنْ / إذن / إذَن/ إذنْ / إذَن / إذنْ/ إذَنْ	**إذَنْ**	
إلا / إلا / إلاّ / إلاَّ/ إلاّ / إلاَّ	**إلاّ**	
الآنَ / الآن	**الآنَ**	
أما / أَمَا / أَما / أمَا	**أما**	
أمَا / أَمَّا / أَمّا / أمَّا / أما ؟ / أَما ؟	**أمّا**	
أن / أَنْ / أَن / أنْ /	**أنْ**	
أنّ / أَنَّ / أَنّ / أنَّ / أن ؟ / أَن ؟	**أنّ**	
إن / إِنْ / إِن / إنْ /	**إنْ**	
إنّ / إِنَّ / إِنّ / إنَّ / إن ؟ / إِن ؟	**إنّ**	
أنا / أَنَا / أَنا / أنَا	**أنا**	
أنت / أَنْتَ / أَنْت / أَنتَ / أنْتَ / أَنت / أنْت / أنتَ	**أنتَ**	
أنت / أَنْتِ / أَنْت / أَنتِ / أنْتِ / أَنت / أنْت / أنتِ	**أنتِ**	
أي / أَيْ / أَي / أيْ	**أيْ**	
أي / أَي / أيّ / أَيّ / أَيُّ / أَيَّ / أَيِّ / أيُّ / أيَّ / أيِّ / أيُ / أيَ	**أيّ**	
أيضا / أَيْضًا / أَيْضا / أَيضًا / أيْضًا / أَيضا / أيْضا / أيضًا	**أيضًا**	
بـ... / بِـ...	**بـ...**	
بئس / بِئْسَ / بِئس / بئْس / بئسَ / بِئْس / بِئسَ / بئْسَ	**بِئْسَ**	
بعد / بَعْدَ / بَعد / بعْد / بعدَ / بَعْد / بَعدَ / بعْدَ / (مِن) بَعْدِ / (مِن) بَعدِ / (مِن) بعْدِ / (مِن) بعدِ	**بعْد**	
بعد / بَعْدُ / بَعْد / بَعدُ / بعْدُ / بَعد / بعْد / بعدُ / من بَعْدُ / من بعد / من بَعْد / من بَعدُ / من بعْدُ / من بَعد / من بعْد / من بعدُ / مِنْ بَعْدُ / مِنْ بعد / مِنْ بَعْد / مِنْ بَعْدُ / مِنْ بعْدُ / مِنْ بعْد / مِنْ بعدُ / مِن بَعْدُ / مِن بعد / مِن بَعْد / مِن بعْدُ / مِن بعْد / مِن بَعدُ / مِن بعدُ / مِن بَعد / مِن بعْد / مِن بعدُ / مِنْ بَعْدُ / مِنْ بعد / مِنْ بَعْد / مِنْ بَعدُ / مِنْ بعْدُ / مِنْ بعْد / مِنْ بعد	**بَعْدُ / مِن بَعْدُ**	Cet exemple de « مِن بَعْدُ » montre bien l'explosion combinatoire des graphies différentes d'un même mot vocalisé partiellement d'une façon arbitraire. « بَعْدُ » comptabilise, seul, huit graphies différentes, « مِن » quatre, quant à leur combinaison, elle passe à 32 graphies.
بل / بَلْ / بَل / بلْ / بَلِ الـ... / بلِ الـ...	**بل**	
بين / بَيْنَ / بَين / بيْن / بينَ / بَيْن / بَينَ / بيْنَ/ (مِن) بَيْنِ / (مِن) بَينِ / (مِن) بيْنِ / (مِن) بينِ	**بيْن**	
ثَمّ / ثَمَّ / ثُمّ / ثُمَّ	**ثَمّ**	
ثُمّ / ثُمَّ / ثُمّ / ثُمَّ	**ثُمّ**	
حتى / حتّى / حَتَّى / حَتّ / حتَّى / حَتى / حتَى	**حتّى**	

حذو / حِذْوَ / حِذو / حذْو / حذوَ / حِذْو / حِذوَ / حذْوَ	**حِذْوَ**	
حيث / حَيْثُ / حَيث / حيْث / حيثُ / حَيْث / حَيثُ / حيْثُ	**حَيْثُ**	
حين / حِينَ / حِين / حينَ	**حِينَ**	حِينَ (Adv.) , à ne pas confondre avec حِين (Nom)
عن / عَنْ / عَن / عنْ / عَنِ الـ.../ عنِ الـ...	**عن**	
عند / عِنْدَ / عِنْد / عِندَ / عنْدَ / عِند / عنْد / عندَ	**عِنْدَ**	
غير / غَيْرَ / غَيْرُ / غَيْرٍ / غَيْرِ / غيْرَ / غيْرُ / غيْرِ / غَيرَ / غَيرُ / غَيرِ / غَير / غيْر / غيرَ / غيرُ / غيرِ	**غيْر**	
فـ... / فَـ...	**فـ...**	
فوق / فَوْقَ / فَوق / فوْق / فوقَ / فَوْق / فوْقَ / فَوقَ	**فوْق**	Nom — كما أنّه ليس فَوْقَ الجنْس فوْق (*'Al-'Imtâ'Ý* p.213)
فوق / فَوْقَ / فَوق / فوْق / فوقَ / فَوْق / فوْقَ / فَوقَ	**فَوْقَ**	Adverbe
فيم / فِيمَ / فِيم / فيمَ	**فِيمَ**	
قد / قَدْ / قَد / قدْ	**قَدْ**	
قطّ / قَطُّ / قطُّ / قَطّ	**قَطُّ**	
ك / كَ	**كـ**	Particule de comparaison
ك / كَ	**كَ**	Pronom suffixe (2e p. sing. masc.)
ك / كِ	**كِ**	Pronom suffixe (2e p. sing. fem.)
كلّ / كُلّ / كُلُّ / كُلَّ / كُلِّ / كلُّ / كلَّ / كلِّ	**كُلّ**	
كم / كَمْ / كَم / كمْ	**كَم**	(combien ?)
كم / كُمْ / كُم / كمْ	**كُم**	Pronom suffixe (2e p. pl. masc.)
كما / كَمَا / كَما / كمَا	**كَما**	(comme)
كما / كُمَا / كُما / كمَا	**كُما**	Pronom suffixe (2e p. duel)
كيف / كَيْفَ / كَيْف / كَيفَ / كيْفَ / كَيف / كيْف / كيفَ	**كيْفَ**	
لـ... / لِـ... / لَـ... (لَه – لَك – لَنا – ...)	**لـ...**	Préposition
لـ... / لَـ...	**لَـ...**	Particule de corroboration
لقد / لَقَدْ / لَقد / لقَد / لقدْ / لَقَد / لَقدْ / لقَدْ	**لَقَدْ**	
لكن / لَكِنْ / لَكن / لكِن / لكنْ / لَكِن / لَكنْ / لكِنْ	**لكنْ**	
لكنّ / لَكِنَّ / لَكنّ / لكِنّ / لكنَّ / لَكِنّ / لَكنَّ / لكِنَّ	**لكنّ**	
لم / لَمْ / لَم / لمْ	**لم**	
لم / لِمْ / لِم / لمْ	**لِمَ**	
لما / لَمَّا / لَمّا / لمّا	**لَمّا**	
لو / لَوْ / لَو / لوْ / لَوِ الـ... / لوِ الـ...	**لو**	
مع / مَعَ / مَع / معَ	**مع**	
مما / ممّا / مِمَّا / مِمّا / ممَّا / مِمَا / مِما / ممَا	**مِمّا**	
ممّن / مِمَّنْ / مِمَّن / مِمّنْ / ممَّنْ / مِمّن / ممَّن / ممّنْ	**مِمَّن**	
من / مَنْ / مَن / منْ / مَنِ الـ...	**مَن**	

مِن	من / مِنْ / مِن / منْ / مِنَ الـ...	
منذ	منذ / مُنْذُ / مُنْذ / مُنذُ / منْذُ / مُنذ / منْذ / منذُ	
نَحْوَ	نحو / نَحْوَ / نَحْو / نَحوَ / نَحْوَ / نَحو / نَحْو / نحوَ	
نَعَمْ	نعم / نَعَمْ / نَعم / نعَم / نعمْ / نَعَم / نَعمْ / نعَمْ	
نِعْمَ	نعم / نِعْمَ / نِعم / نعْم / نعمَ / نِعْم / نِعمَ / نعْمَ	
هـ	هـ / هُ / هِ	(فيهِ / بِهِ)
هم	هم / هُمْ / هُم / همْ / هِمْ / هِم / هِمْ / همْ	(فيهِم / بِهِم)
هما	هما / هُما / هُما / همَا / هِمَا / هِما / همَا	(فيهِما / بِهِما)
هنّ	هنّ / هُنَّ / هُنّ / هنَّ / هِنَّ / هِنّ / هِنَّ / هنَّ	(فيهِنّ / بِهِنّ)
هَهُنا	ههنا / هَهُنَا / هَهُنا / هَهنَا / هَهنَا / هَهُنَا / هَهنا / هههُنا / ههنَا / هاهنا / هَاهُنَا / هَاهُنا / هَاهنَا / هَاهنَا / هاهُنَا / هَاهنا / هاهُنا / هاهنَا / ها هنا / هَا هُنَا / هَا هُنا / هَا هنَا / ها هُنَا / هَا هنا / ها هُنا / ها هنَا	Outre ce problème d'harmonisation graphique, ce mot-outil a une spécificité qui nécessite un autre choix à faire, il appartient, en effet, à la liste des mots à graphies multiples présentée dans le tableau p. 223
هيهاتَ	هيهات / هَيْهَاتَ / هَيْهَات / هَيْهاتَ / هَيهَاتَ / هيْهَاتَ / هَيْهات / هَيهَات / هَيهاتَ / هيْهَات / هيْهاتَ / هيهَاتَ / هَيهات / هيْهات / هيهَات / هيهاتَ	
و	و / وَ	Coordonnant. Ne s'applique pas au « و » commençant un verbe (وَقَفَ – وَلَعَ), un adjectif (وَجِيه – وَضِيع) ou un nom (وَجْه – وَهْم).

2.1.1.12. Harmonisation des mots à graphies multiples

En arabe, comme dans pratiquement toutes les langues, il existe un certain nombre de mots qui sont à même de s'écrire de différentes façons. Ce phénomène puise, en arabe, une spécificité additionnelle dans la disjonction totale entre l'ensemble des consonnes et celui des voyelles. De ce fait, une partie de ces mots à graphies multiples ont une même structure consonantique mais leur variation graphique résulte des vocalismes différents que peut prendre ce support consonantique.

Ces variations graphiques se manifestent à deux niveaux : un niveau purement orthographique ou scriptural, c'est ce que nous avons appelé les variantes graphiques,

et un niveau lexical, c'est ce que nous avons appelé les variantes lexicales . Ce dernier type de variantes peut résulter soit d'alternances vocaliques soit d'alternances consonantiques.

Nous sommes donc en présence d'un phénomène combinatoire de quatre variables : graphie unique, graphies multiples, prononciation unique, prononciations multiples qui doivent former des couplets représentant l'aspect graphonologique des mots. Ces couplets sont : (graphie unique, prononciation unique), (graphie unique, prononciations multiples), (graphies multiples, prononciation unique), (graphies multiples, prononciations multiples). Ces quatre couplets représentent respectivement : un mot unique – et c'est le cas général des mots -, des variantes lexicales (à alternances vocaliques) d'un même mot, des variantes graphiques d'un même mot et des variantes lexicales (à alternances consonantiques et, par conséquent, vocaliques) d'un même mot. Le tableau qui suit résume ces différents couplets :

	Graphie unique	**Graphies multiples**
Prononciation unique	***Mot unique*** *Exemple :* كتاب	***Variantes graphiques*** *Exemple :* يقرؤون \ يقرأون
Prononciations multiples	***Variantes lexicales*** **à** alternances vocaliques *Exemple :* كِخْ \ كَخْ \ كُخْ	***Variantes lexicales*** **à** alternances consonantiques *Exemple :* اللاّتِي \ اللاّتِ

Nous allons par la suite définir et détailler ces cas de figures de la variation graphique, en donner des exemples, soit tirés directement de notre corpus, soit résultant de nos recherches antérieures, et arrêter, pour chaque mot à graphies multiples, un choix conscient et réfléchi d'une forme graphique servant de graphie d'harmonisation.

2.1.1.12.1. Les variantes graphiques

Les variantes graphiques sont les graphies multiples d'un même mot qui se prononcent de la même manière. Autrement dit, ce sont les mots qui présentent une variété graphique mais une unicité phonétique comme, par exemple, : **ها هنا / هاهنا / ههنا**.

Mots à graphies multiples	Graphie d'harmonisation	Remarques
ههنا \ هاهنا \ ها هنا	**هَهُنا**	Démonstratif (Ici)
أُولاَلِكَ \ أُولَى لِكَ	**أُولاَلِكَ**	Démonstratif
طاووس \ طاوس	**طاووس**	Paon
داوود \ داود	**داوود**	Nom propre (David)
ناووس \ ناوس	**ناووس**	Sarcophage
شاوول \ شاول	**شاوول**	Nom propre (Shaoul)
يقرؤون \ يقرأون	**يقرؤون**	Ils lisent
مئة \ مائة \ مأة	**مئة**	Cent
ثلاثمئة \ ثلاثمائة \ ثلثمئة \ ثلثمائة \ ثلاث مئة \ ثلاث مائة	**ثلاثمئة**	Numéral composé
ياسين \ يسين \ يس	**ياسين**	Nom propre (sauf dans le Coran : يس)
هُوَيْنَى \ هُوَيْنَا	**هُوَيْنَا**	Douceur, lenteur
ياأَيُّ \ يَأَيُّ	**ياأَيُّ**	

2.1.1.12.2. Les variantes lexicales

Les variantes lexicales sont les graphies multiples d'un même mot qui se ne prononcent pas de la même manière. Autrement dit, ce sont les mots qui présentent à la fois une variété graphique et une variété phonétique.

Les variantes lexicales peuvent être dues à des changements vocaliques ou à des changements consonantiques.

2.1.1.12.2.1. Les alternances vocaliques

Les variantes lexicales à variation vocalique sont les mots ayant le même agencement consonantique mais dont le schéma vocalique varie d'une forme à l'autre. Exemple : **هَيْتَ** / **هِيتَ** .

Mots à graphies multiples	Graphie d'harmonisation	Remarques
كِخْ \ كَخْ \ كُخْ	**كِخْ**	Onomatopée
أُفٍّ \ أُفًّا \ أُفَّ \ أُفٍّ \ أُفُّ \ أُفْ	**أُفٍّ**	Ouf !, Zut !, Maudit soit...
وِشْكانَ \ وُشْكانَ \ وَشْكانَ	**وِشْكانَ**	Nom de verbe
حَيْصَ بَيْصَ \ حِيصَ بِيصَ	**حَيْصَ بَيْصَ**	Locution grammaticale
تِلْكِ \ تَلْكِ	**تِلْكِ**	Démonstratif
ذِهِ \ ذِهْ	**ذِهِ**	Démonstratif (Ceci, celui-ci)
هاءَ \ هاءِ	**هاءَ**	Nom de verbe
هِيتَ \ هَيْتَ	**هِيتَ**	

2.1.1.12.2.2. Les alternances consonantiques :

Les variantes lexicales à variation consonantique sont les mots dont la structure consonantique varie d'une forme à l'autre entrainant, ce faisant, une variation phonétique. Exemple : **نِعِمَّا \ نِعْمَ**.

Mots à graphies multiples	Graphie d'harmonisation	Remarques
رِضا \ رِضى \ رُضا \ رُضى	**رِضا**	(Satisfaction)

تِهْ \ تِهِ \ تِهِي	**تِهِ**	Démonstratif (celle-ci)
حَيَّهَلَ \ حَيَّ هَلاَ	**حَيَّهَلَ**	Nom de verbe
هَأْ هَأَ \ هِئْ هِئْ	**هَأْ هَأْ**	Onomatopée
نِعْمَ \ نِعِمَّا	**نِعْمَ**	
اللاّتِي \ اللاّتِ	**اللاّتِي**	Relatif
اللاّئِي \ اللاّءِ	**اللاّئِي**	Relatif
بَيْنَما \ بَيْنَا	**بَيْنَما**	Adverbe
ذَيْتَ \ ذَيَّةَ	**ذَيْتَ**	Circonlocutif
كَيْتَ \ كَيَّةَ	**كَيْتَ**	Circonlocutif
كَأَيٍّ \ كَأَيَّنْ \ كَأَيْنْ \ كَأْيِنْ \ كَائِنْ \ كَيِنٍ	**كَأَيٍّ**	Circonlocutif
لَعَلَّ \ عَلَّ	**لَعَلَّ**	Peut-être
أَنَّنَا \ أَنَّا	**أَنَّنَا**	
إِنَّنَا \ إِنَّا	**إِنَّنَا**	
لَكِنَّنَا \ لَكِنَّا	**لَكِنَّنَا**	
تَتَلَذَّذُ \ تَلَذَّذُ	**تَتَلَذَّذُ**	Contexte poétique ou coranique
فم \ فو (\ فا \ في)	**فم**	

2.1.2. Harmonisation régulatrice (après segmentation)

Intervenant en aval, c'est-à-dire après la segmentation, cette opération a pour but la régularisation de la graphie de certains mots ayant subi des transformations graphiques dues à des considérations diverses : scripturales (écriture cursive :ة ت), orthographiques (أ ؤ), morphologiques (ي ِني), syntaxiques (ون و) ou phonétiques (عِنْدَ عِنْدِ).

2.1.2.1. Le *ÞalifÞ* orthographique[146]

Les verbes conjugués à la deuxième personne du pluriel masculin à l'impératif se terminant toujours par « *û* » اُكْتُبُوا *Þuktubû* « écrivez », et à la troisième personne du pluriel masculin à l'accompli se terminant toujours par « *û* » ou « *aw* » كَتَبُوا *katabû* « ils ont écrit », رَمَوْا *ramaw* « ils ont lancé », se voient systématiquement adjoindre un *Þalif* qui ne se prononce pas. Ce *Þalif* disparaît lorsque la forme verbale est liée à un pronom personnel suffixe اُكْتُبُوها *Þuktubû-hâ* « écrivez-la », كَتَبُوه *katabû-hu* « ils l'ont écrit », رَمَوْهُمْ *ramaw-hum* « ils leur ont lancé ».

Ceci ne pose finalement aucun problème au niveau de la saisie, mais c'est à la suite de la segmentation que nous obtiendrons deux unités lexicales dont l'une est une forme verbale incorrectement orthographiée (كَتَبُو + ه). Il va donc falloir procéder à une opération d'harmonisation régulatrice pour rétablir la forme graphique initiale des mots qui pourraient subir ce genre de détriment suite à la segmentation.

---

[146] Pour plus de détail sur le *Þalif* orthographique, voir l'article de H. Hamzé, *'alif al-faòl*, dans « *Îawliyyât 'al-JâmiŶa t-tûnisiyya* », N° 32, Tunis, 1991, pp. 23-52.

2.1.2.2. Le *tâÞ marbûÔa*

Le *tâÞ marbûÔa* (la lettre « *t* » liée) qui se trouve à la fin de certain mots se transforme en *tâÞ maftûÎa* (la lettre « *t* » ouverte) quand ces derniers sont agglutinés à des pronoms personnels suffixes مدرسة مدرستكم .

Dans ce cas aussi et après l'opération de segmentation, on obtient (مدرست + كم) où le mot مدرسة « école » est mal orthographié مدرست*. Là également, l'harmonisation régulatrice est une étape qui s'avère nécessaire pour pouvoir redonner sa graphie d'origine à ce type de mots.

2.1.2.3. L'élision du *Þalif* de l'article

En présence d'une agglutination de la préposition لِ *li-* « à, pour » ou de la particule corroborative لَ *la-* avec un mot commençant par l'article défini الـ *Þal-* « le, la, les » le *alif* de l'article est élidé (voir *supra* : Les irrégularités orthographiques de l'arabe). Ainsi, la séquence لـ + الكتاب *li + Þal + kitâb* « au/pour le livre » s'écrit للكتاب *li-l-kitâb*. Après segmentation on pourrait avoir la séquence erronée suivante : ل \ ل \ كتاب . Ce qui nécessite le recours à une harmonisation régulatrice pour rendre à l'article défini son *alif* perdu en cours de route.

2.1.2.4. Le *yâÞ* des « **أسماء منقوصة** »

Quand ils sont définis, les noms "incomplets" se voient adjoindre un *yâÞ* final. Après segmentation, dépourvus de l'article défini, ces noms doivent normalement retrouver leur graphie d'origine et donc se débarrasser de cet augment. C'est dans le

cadre donc d'une harmonisation régulatrice que l'opérateur humain ou le programme informatique doive faire le nécessaire pour ôter à ces noms le *yâÞ* et le remplacer par le *tanwîn* de la *kasra*.

Le tableau qui suit résume, si ce n'est pas tous, du moins la majorité des cas de figures où une harmonisation régulatrice est indispensable, soit par l'opérateur humain soit par le programme informatique, pour redonner leurs graphies d'origine aux unités lexicales ayant subi des modifications ou des anomalies suite au processus de segmentation :

Type	Avant segmentation	Après segmentation	Harmonisation régulatrice
أ □ إ	وعلى هذا كُلّ أمّة في **مبدإ** سعادتها	و\على\هذا\كُلّ\أمّة\في**مبدإ**\سعادة\ها	و\على\هذا\كُلّ\أمّة\في**مبدأ**\سعادة\ها
ء □ ئ	وفقره **وغنائه** ، وشِدّته **ورخائه** ، **وسرّائه وضرّائه** ، وخِيفَتِه **ورجائه**	و\فقر\ه\و**غنائ**\ه\و\شِدّة\ه\و**رخائ**\ه\و**سرّائ**\ه\و**ضرّائ**\ه\و\خِيفَة\ه\و**رجائ**\ه	و\فقْر\ه\و**غناء**\ه\و\شِدّة\ه\و**رخاء**\ه\و**سرّاء**\ه\و**ضرّاء**\ه\و\خِيفَة\ه\و**رجاء**\ه
أ □ ئ	ليس في فطرته ولا عادته ولا **مَنشئه**	ليس\في\فطرة\ه\و\لا\عادة\ه\و\لا**مَنشئ**\ه	ليس\في\فطرة\ه\و\لا\عادة\ه\و\لا**مَنشأ**\ه
أ □ ؤ	ما يُعتَقَد صوابه **وخطؤه**	ما\يُعتَقَد\صواب\ه\و**خطؤ**\ه	ما\يُعتَقَد\صواب\ه\و**خطأ**\ه
ء □ ؤ	ومنْ أَسَرَه **رجاؤه** ، طال **عناؤه** ، وعظُم **بلاؤه**	و\مَن\أَسَرَ رَ\ه**رجاؤ**\ه\طال**عناؤ**\ه\و\عَظُمَ**بلاؤ**\ه	و\مَن\أَسَرَ رَ\ه**رجاء**\ه\طال**عناء**\ه\و\عَظُمَ**بلاء**\ه
ي □ ني	طَالَبْتَنِي	طَالَبْتَ**ني**	طَالَبْتَ**ي**
ى □ا	**وكفاه**	و**كفا**\ه	و**كفى**\ه
ى □ ي	لدينا	**لدي**\نا	**لدى**\نا
□ ا	بقرة واحدة **وعدموها** طلبوا سائر البقر **وفقدوها**	بقرة\واحدة\و**عدمو**\ها\طلبوا\سائر\ال\بقر\و**فقدو**\ها	بقرة\واحدة\و**عدموا**\ها\طلبوا\سائر\ال\بقر\و**فقدوا**\ها
→ ُ و	أَنُلْزِمُكُمُوهَا	أَ\نُلْزِمُ\كُمُو\هَا	أَ\نُلْزِمُ\كُم\هَا
ة □ ت	حياتي	حيا**ت**\ي	حيا**ة**\ي
يْن □ يْ	ومُثولي بيْن **يديكَ**	و\مُثول\ي\بيْن**يديْ**\كَ	و\مُثول\ي\بيْن**يديْن**\كَ
ان □ ا	**كتابا** التلميذ	**كتابا**\ال\تلميذ	**كتابان**\ال\تلميذ
ون □ و	**محامو** المتّهم	**محامو**\ال\متّهم	**محامون** \ال\متّهم

على\قدْر\ما\يكون\عدد**سنين**\ها	على\قدْر\ما\يكون\عدد**سني**\ها	على قدر ما يكون عدد **سنيها**	**ين □ ي**
قال\ال**محامٍ**	قال\ال**محامي**	قال **المحامي**	**ٍ □ ي**
ذلك\تنبيه\ل**ال**\نائم\و\إيقاظ\ل**ال**\ساهي	ذلك\تنبيه\ل**ل**\نائم\و\إيقاظ\ل**ل**\ساهي	ذلك تنبيه **للنّائم** وإيقاظ **للسّاهي**	**ال □ ل**
عِنْدَ\ي	**عِنْدِ**\ي	عِنْدِي	**عِنْدَ □ عِنْدِ**

Tableau 7 :
Récapitulatif des différents cas d'harmonisation régulatrice
que nous avons relevés, avec exemples en contexte

CHAPITRE 5

Norme de dépouillement

Après le premier moment qui est l'enregistrement du corpus en machine, son apurement et son harmonisation, le dépouillement du corpus constitue le deuxième moment important du prétraitement des données textuelles avant de les livrer au traitement et à l'analyse quantitatifs.

Étroitement liée aux autres étapes de dépouillement des textes, la segmentation prépare le terrain aux autres tâches principales qui sont la lemmatisation et la catégorisation passant par la désambiguïsation. La lemmatisation consiste à reconnaître pour chaque mot sa forme de base, que l'on appelle le lemme. La catégorisation consiste à choisir parmi les différentes catégories possibles du lemme, la bonne catégorie en fonction du contexte du mot ; cette dernière étape n'est possible que suite à une opération de désambiguïsation.

Ces étapes composent ce que l'on peut appeler l'analyse lexicale : c'est le même processus qui, analysant le texte, va définir les limites des unités de décompte (segmentation), assigner le lemme (lemmatisation) et la catégorie grammaticale (catégorisation) à chaque unité en fonction de ses voisins (désambiguïsation).

Nous allons donc définir une *norme de dépouillement* regroupant les différents choix adoptés et les différentes décisions arrêtées, leurs fondements théoriques et méthodologiques et leur contexte d'application.

1. Segmentation

La segmentation est une étape fondamentale dans le traitement automatique d'un texte, son rôle est de découper un texte en unités d'un certain type qu'on aura définies et repérées préalablement.

La segmentation d'un texte informatisé est l'opération de délimitation de segments de ses éléments de base qui sont les caractères, en éléments constituants de différents niveaux structurels : paragraphe, phrase, syntagme, mot graphique, forme, morphème, ...

C'est une opération consistant à structurer le texte en passant d'un ensemble continu de caractères à une suite discrète de mots communément appelés, en Traitement Automatique des Langues et en particulier en Lexicométrie, segments.

1.1. Ecritures segmentées *VS* Ecritures non segmentées

En Traitement Automatique des Langues, on présente généralement les langues, quant à leur système d'écriture, comme appartenant à deux familles différentes : les langues « avec séparateurs » et les langues « sans séparateurs ».

Les langues dites « avec séparateurs » sont celles qui ont des systèmes d'écritures segmentées c'est-à-dire des écritures délimitées par des espaces (*spacedelimited writings*) et où les mots sont nettement séparés par des délimiteurs (espace, signes de ponctuation, caractères spéciaux, ...). Le français et l'anglais sont des langues typiquement représentatives de cette famille.

A ce type de langues on oppose les langues dites « sans séparateurs ». Ce sont celles qui présentent des systèmes d'écritures non segmentées (*unsegmented writings*) où les mots ne sont pas séparés par des espaces et où les frontières des mots ne sont pas nettes. Le japonais, le chinois et surtout le thaï sont les représentants parfaits de cette famille de langues.

Mais qu'en est-il de la langue arabe ?

La langue arabe présente un système d'écriture à l'intersection des deux familles. C'est un système d'écriture qui combine une écriture segmentée et une écriture non segmentée. En effet, une partie des mots graphiques arabes correspondent à des mots minimaux séparés par des délimiteurs. En revanche, une bonne partie des mots graphiques arabes sont composés d'une suite d'unités lexicales agglutinées analysable en termes de mots minimaux et de clitiques et qu'il faut donc segmenter si l'on veut arriver aux unités de base les composant et qui constitueront les unités de décompte.

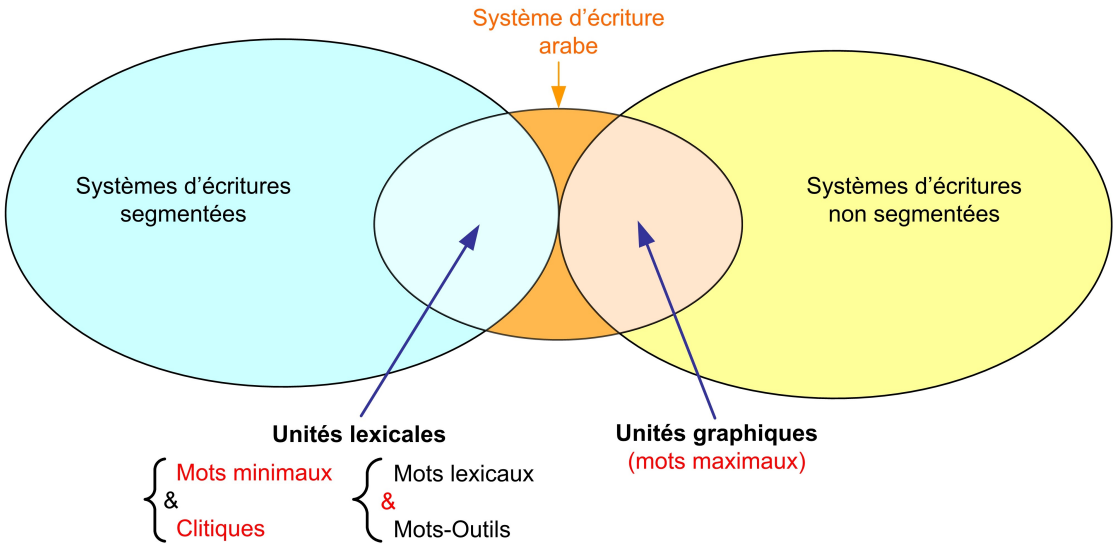

Figure 39
Systèmes segmentés et systèmes non segmentés

1.2. Les types de segmentation

S'agissant de la segmentation d'un texte, il existe plusieurs niveaux d'analyse auxquels on peut s'arrêter pour repérer les différents éléments constituant le texte et en définir les frontières. On peut s'arrêter au niveau de la phrase, au niveau de la proposition ou à celui du syntagme. Mais on peut arriver aussi au niveau du mot graphique, au niveau des unités lexicales ou aller au delà de celles-ci pour arriver aux unités de base les composant : les morphèmes.

Selon la visée de l'analyse à entreprendre : lexicale, morphologique ou syntaxique, on peut généralement parler de trois grands types d'application de la segmentation :

- ***L'itémisation*** (en anglais *tokenization* ou *word segmentation*) qui est la segmentation d'un texte en mots ou items lexicaux (*tokens*). Ce type de segmentation est aussi appelé ***segmentation lexicale***.
- ***La segmentation morphologique*** qui va plus loin que la segmentation lexicale en cherchant à isoler les différents constituants des items lexicaux en unités distinctes, plus petites, qui sont les morphèmes.
- ***Le chunking*** qui consiste à isoler les différents constituants du texte en unités indépendantes, supérieures aux mots, comme les propositions, les syntagmes etc. Ce type de segmentation est aussi appelé ***segmentation syntaxique***.

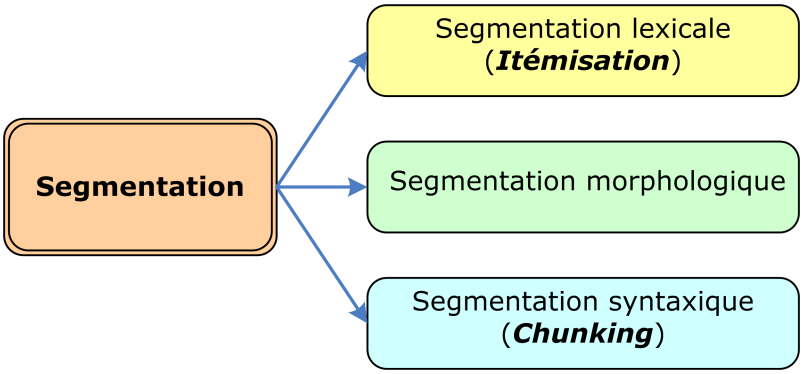

Figure 40
Les types de segmentation

Alors que l'itémisation s'attaque directement aux unités graphiques d'un texte pour en faire une suite d'unités lexicales, la segmentation morphologique se charge, elle, de ces unités lexicales pour arriver à une suite de constituants de niveau inférieur, quant au chunking s'intéresse à des groupes de mots en les préparant à une analyse des relations qui pourraient exister entre eux, donc à une analyse syntaxique.

Par segmentation de texte nous entendons, dans ce travail, l'analyse des mots graphiques en mots minimaux et clitiques[147]. Autrement dit, l'analyse des unités graphiques en unités lexicales. Ces unités lexicales sont des lexèmes et des clitiques (proclitiques et enclitiques), en d'autres termes , des mots lexicaux et des mots-outils[148]. Il s'agit donc de segmentation lexicale ou itémisation. C'est pourquoi dorénavant quand nous utilisons le terme de segmentation, nous entendons bien évidemment segmentation au sens d'itémisation.

En lexicométrie en général et en lexicométrie arabe en particulier, le point de départ pour le découpage de textes et pour toutes les autres phases de dépouillement

[147] Voir § concernant l'analyse du mot graphique en arabe pp. 207-211

[148] Tous les mot-outils ne sont pas des clitiques. Les mots-outils qui peuvent être soit des proclitiques soit des enclitiques, simples ou composés, sont présentés dans les listes fermées établies à l'annexe A

c'est toujours le mot graphique en tant qu'unité primaire et ce en dépit de l'inadéquation qui puisse exister entre l'unité graphique et l'unité lexicale.

1.3. Le mot entre réalité linguistique et manifestation graphique

Le *mot* est défini, en linguistique traditionnelle, comme étant « un élément linguistique significatif composé d'un ou de plusieurs phonèmes ; cette séquence est susceptible d'une transcription écrite (idéogrammatique, syllabaire ou alphabétique) comprise entre deux blancs ; dans ses divers emplois syntaxiques, elle garde sa forme, soit totalement, soit partiellement (dans le cas de la flexion). Sur le plan sémantique, le mot dénote un objet (substantif), une action ou un état (verbe), une qualité (adjectif), une relation (préposition), etc. C'est cette définition qui est retenue en lexicographie »[149].

Mais cette définition du « *mot* » pose problème. En effet, pour la linguistique contemporaine, le terme de « *mot* », terme « décrié et irremplaçable dans les niveaux de l'analyse linguistique »[150] est loin d'avoir une définition satisfaisante. Cette entité qui semble être « assez claire pour le sens commun, est scientifiquement suspecte »[151] et même si elle paraît « familière et évidente pour le grand public, constitue pour le linguiste une source de difficultés théoriques considérable »[152]. Mais malgré cette difficulté à le définir, écrivit le père fondateur de la linguistique contemporaine, le *mot* « est une unité qui s'impose à l'esprit, quelque chose de central dans le mécanisme de la

---

[149] Jean Dubois, *Dictionnaire de linguistique et des sciences du langage*, Paris 1994, p. 312

[150] Emile Nenveniste, *Problèmes de linguistique générale*, p. 123

[151] Charles Muller, *Langue française. Débats et bilans. Recueil d'articles*, 1993, p. 3.

[152] Niklas-Salminen Aïno, *La lexicologie*, 1997, p.13.

langue »[153]. Il serait oiseux de se livrer ici à de longues discussions sur le *mot* et d'exposer toutes les théories et les tentatives de le définir. Nous nous contentons des quelques définitions suivantes pour se convaincre du flou linguistique qui accompagne cette notion :

- Pour André Martinet par exemple, qui définit le monème comme étant l'unité significative minimale que l'on peut dégager dans la chaîne parlée, le *mot* est « un syntagme autonome formé de monèmes non séparables »[154]. L'unité minimale pourvue d'un sens est donc, pour Martinet, le *monème* (l'unité minimale de première articulation). Les *monèmes* comprennent les *lexèmes*, unités du lexique appartenant à une liste ouverte, et les *morphèmes*, unités de la grammaire appartenant à un inventaire fermé. Mais pour délimiter ces unités, A. Martinet n'offre que le critère de partition en classes ouvertes et classes fermées, critère basé sur l'opposition entre lexique et grammaire. Mais pour certain nombre de linguistes ce critère n'est pas tout à fait satisfaisant : « Le critère de classe ouverte ou fermée conduit donc à des regroupements quelque peu artificiels, car on rassemble ainsi des unités dont les statuts sont fort différents. En effets, les possibilités de créativité lexicale en matière d'affixes sont beaucoup plus réduites qu'ailleurs ; ils tendent à fonctionner comme des classes fermées »[155].

Aux *syntagmes* (mots graphiques "simples"), A. Martinet oppose les *synthèmes* qu'il définit comme « unités linguistiques dont le comportement syntaxique est strictement identique à celui des monèmes avec lesquels ils commutent, mais qui peuvent être conçus comme formés d'éléments sémantiquement identifiables »[156]. Quoique cohérent, le modèle de Martinet reste quelque peu flou quand il s'agit de trancher « à partir de quand un synthème "fonctionne comme" un monème (…) Par

---

[153] Ferdinand de Saussure, *Cours de linguistique générale*, p. 154

[154] André Martinet, *Eléments de linguistique générale*, Armand Colin, 1970, p. 114

[155] François Gaudin et Louis Guespin, *Initiation à la lexicologie française. De la néologie aux dictionnaires*, 2000, p. 212

[156] André Martinet, 1967, p.12

ailleurs la définition du *synthème* suppose résolue la question des "éléments sémantiquement identifiables". »[157]

- Quant à Bernard Pottier, il considère que les unités inférieures au mot graphique sont des morphèmes, au sens traditionnel. Mais il apporte des précisions importantes au moins sur deux plans :

  - Il oppose, au sein des morphèmes, les *lexèmes*, ou morphèmes lexicaux, et les *grammèmes*, ou morphèmes grammaticaux. Les *lexèmes* constituent les éléments lexicaux les plus simples, racines ou mots simples dépourvus de leurs flexions. Les *grammèmes* sont ces formants qui contiennent les affixes, les marquent d'accords, ...
  - Il fait aussi la distinction entre les unités de langue et les unités de discours. Les unités de langue sont les *lexèmes* et les *grammèmes* ; alors que les unités de discours sont les *lexies*. La lexie étant définie comme « *l'unité minimale significative de discours* ». De plus, B. Pottier distingue trois types de lexies : les lexies simples (*fleur*), les lexies composées (*chou-fleur*) et les lexies complexes (*pomme de terre*).

- Après avoir banni dans un premier temps, le terme *mot* (tout court), Igor Mel'čuk, distingue deux acceptions différentes du *mot* : mot_1 = "mot-forme" qui traite le mot comme une entité « CONCRÈTE », et mot_2 = "lexème" qui traite le mot comme une entité « ABSTRAITE ». Dans la première acception, « le mot est un ÉLÉMENT SPÉCIFIQUE, une unité textuelle, que nous appellerons *mot-forme* »[158] alors que dans la deuxième acception, « le mot est un ENSEMBLE d'éléments spécifiques ayant un "noyau" commun sur le plan sémantique. C'est une unité lexicographique, qu'on appellera *lexème* »[159]. Puis après avoir remarqué que « notre caractérisation du mot-

---

[157] François Gaudin et Louis Guespin, *op. cit*, p.213

[158] Igor Mel'čuk, Cours de Morphologie Générale (Théorique et descriptive), 1993, p. 99

[159] Igor Mel'čuk, *idem*, p. 99

forme coïncide avec la formulation succincte et claire d'Antoine Meillet : "Un mot [=mot-forme — I. M.] résulte de l'association d'un sens donnée à un ensemble de sons donnés susceptible d'un emploi grammatical donné" (Meillet 1921 : 30) »[160], Mel'čuk donne la définition du *mot* comme suit : « Un "mot" est soit un *mot-forme*, soit un *lexème*. Un *mot-forme* est un cas particulier du signe linguistique$_1$ (l'ensemble d'un signifié, d'un signifiant et d'un syntactique). Un *lexème* est un ensemble de mots-formes et de syntagmes$_1$ ne différant que par leurs significations flexionnelles. »[161]. Il faut savoir que Mel'čuk entend par syntactique l'ensemble des informations d'un côté, spécifiant la paire (signifié, signifiant) et de l'autre côté, spécifiant le comportement du signifiant donné dans de telles combinaisons. Mais notons que bien avant Mel'čuk, Charles Muller, grand spécialiste de la statistique lexicale, avait donné une définition et une analyse du mot comme unité du texte et unité du lexique dans un article devenu célèbre et, depuis, une référence dans les études lexicométrique[162].

Ce flou linguistique et cette difficulté théorique à définir le « *mot* » pour ne pas parler, comme Charles Muller, de « suspicion scientifique » dans le domaine de la linguistique ont été, en partie, hérités par le TALN. Domaine pour lequel le *mot*, parce que quasiment indéfinissable, est assez souvent difficilement identifiable à l'intérieur d'une phrase, donc d'un texte, en particulier pour les langues à systèmes d'écritures non segmentées.

En TALN, on a le plus souvent tendance, à donner au *mot* une définition qui se situe au niveau de la manifestation graphique et qui est plutôt dictée par des considérations pratiques et pragmatiques. Il est vrai qu'au stade de l'enregistrement du texte en machine, sur support informatique, le *mot* « ne se distingue pas de l'unité graphique définie comme une suite de caractères alphabétiques et diacritiques isolée de

[160] Igor Mel'čuk, *idem*, p. 102

[161] *ibidem*, p. 103

[162] Charles Muller, *Le mot: unité du texte et unité du lexique*, dans (Muller 1979)

ses co-occurrents par des espaces blancs, par un ou plusieurs signes de ponctuation ou par une apostrophe ou encore par un trait d'union »[163]. Cependant, nous pensons que cette adéquation entre le *mot* comme « élément linguistique » et l'unité graphique s'arrête précisément à cette phase d'enregistrement et de conditionnement informatique perçue comme point de départ et comme étape préparatoire à la phase de dépouillement du corpus qui nécessite une prise de position linguistique et l'adoption d'une certaine définition du lexique et de ses éléments. Cette phase de dépouillement aboutit à la constitution de listes de formes et de lemmes qui seront la base de l'analyse quantitative.

Aussi, la nécessité de découper le texte, ou plus précisément les unités graphiques le composant, en unités plus petites repose-t-elle sur le fait que dans un processus de collecte des unités de décompte, éléments de base de toute étude statistique, les unités graphiques ne sont pas des unités « atomiques » insegmentables. Ces unités « atomiques » recherchées sont en deçà des unités graphiques : se sont, en lexicométrie, les unités lexicales. Elles sont incluses dans les unités graphiques, elles en sont en quelque sorte « prisonnières ». Et le processus de segmentation n'est, en fait, qu'une opération de « libération » de ces unités atomiques et une opération de définition de leurs frontières.

Cependant, cette adéquation entre le mot en général et l'unité graphique trouve justement sa pleine justification dans les études sur l'analyse du mot graphique en arabe. Cette analyse initiée au début des années soixante par David Cohen a été poursuivie et améliorée par d'autres travaux à savoir le *rapport Desclés*[164], le programme de recherche *SAMIA*[165] ou l'équipe constituée autour de *DIINAR*[166].

---

[163] Cossette André, *La richesse lexicale et sa mesure*, 1994, p. 49

[164] Desclés J-P et al., *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement Assisté par Ordinateur*, 1983.

[165] Dichy J., Hassoun M., (éd.), *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe, travaux SAMIA.1*, 1989.

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

1.4. L'analyse du mot graphique en arabe

1.4.1. Le mot graphique arabe : représentation en constituants immédiats

Axés sur la notion du mot graphique, les travaux sur l'analyse automatique de l'arabe ont commencé vers le début des années soixante. Le texte qui a donné le « coup d'envoi » de ces travaux fut l'article de David Cohen paru en 1961 dans la revue de l'Association pour le Traitement Automatique des LAngues naturelles (ATALA) et intitulé "*Essai d'une analyse automatique de l'arabe*".

Dans ce travail, repris et révisé près de dix ans plus tard[167], D. Cohen propose un schéma général des mots graphiques maximaux entièrement vocalisés analysables en constituants morphologiques ultimes. Cette analyse consiste à décomposer le mot graphique en racine, schème, base, préfixes, suffixes, antéfixes, postfixes[168]. Toute base est analysable, selon D. Cohen, en une racine et un schème qu'il définit par ailleurs comme « deux entités formelles discontinues, la première constituée par une succession d'éléments phoniques dont la nature, le nombre et l'ordre sont constants, la seconde par une sorte de "moule" de forme également constante, mais admettant comme ses éléments constitutifs n'importe quelle suite ordonnée de phonèmes qui définit une racine »[169]

Le *mot maximal* est donc cette unité décomposable en proclitiques, préfixes, base, suffixes et enclitiques. La concaténation des préfixes, de la base et des suffixes

---

[167] D. Cohen, *Etudes de linguistique sémitique et arabe*, 1970.

[168] Les termes antéfixes et postfixes ont été remplacés par la suite par, respectivement, proclitiques et enclitiques. Voir à ce sujet : J. Dichy et M. O. Hassoun (éd.), *op. cit.*

[169] (Cohen 1970)

formant ce que D. Cohen appelle *mot minimal*, le *mot maximal* peut par conséquent être analysé en proclitiques, mot minimal et enclitiques.

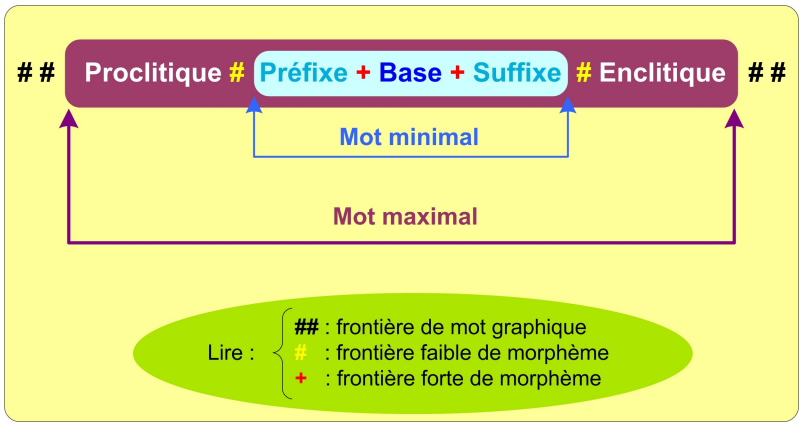

Figure 41
Schéma général du mot graphique en arabe

Le mot minimal est une unité libre minimale au sens de Bloomfield, c'est-à-dire qu'elle peut exister d'une façon autonome sans avoir besoin des clitiques. En revanche, la base ne peut pas se passer des affixes et n'est, de ce fait, pas considérée comme une forme libre minimale.

Cependant, bien que précurseurs et productifs, les travaux de D. Cohen restent marqués par deux limites frappantes : la première est que ces travaux ont traité seulement de la langue arabe moderne écrite entièrement vocalisée, et la deuxième est qu'ils se situaient exclusivement au niveau de l'analyse et ne s'intéressaient pas à la synthèse dans le sens de la génération de mots minimaux ou maximaux à partir des leurs constituants. Ces deux limites ont d'ailleurs été évitées dans les travaux qui ont suivi et qui ont été, en quelque sorte, le prolongement des travaux de D. Cohen, c'est-à-dire le rapport Desclés, les travaux de SAMIA et les travaux de DIINAR.

C'est dans la perspective de répondre aux exigences de la conception d'un dictionnaire informatisé que la description du modèle linguistique a été élaborée depuis les travaux du programme SAMIA et jusqu'à la confection de DIINAR.1. Dans ce

modèle linguistique de l'analyse, le schéma général du mot graphique ne contient pas uniquement les formants constitutifs du mot mais aussi des informations d'ordre syntaxique et sémantique associées à chacun des éléments qui le composent. Ces informations sont couplées avec des règles qui assurent la bonne formation structurelle du mot. Les règles de bonne formation structurelle du mot sont basées sur les propriétés distributionnelles de ses différents formants. Elles assurent la compatibilité entre ces formants.

Nous présentons dans l'annexe A des tableaux contenant les listes des proclitiques et enclitiques, simples et composés qui peuvent être agglutinés à des mots minimaux. Mais nous donnons ici leur nombre et la règle de combinaison des enclitiques :

- Les proclitiques

  Les proclitiques simples sont au nombre de 19, alors que les proclitiques composés (ou agglutinés) sont au nombre de 64. Dans les 64 proclitiques composés, il y a 34 composés de deux proclitiques simples, 24 sont composés de trois proclitiques simples et seulement 6 sont composés de quatre proclitiques simples.

- Les enclitiques

  Selon que l'on choisisse ou non de faire abstraction des modalités du genre et du nombre pour les enclitiques homographes (comme le ي *y*, 1e P. M/F ou نا *nâ*, 1e P. Duel/Pl - M/F), leur nombre varie entre 12 et 18 pour les enclitiques simples, et entre 40 et 102 pour les enclitiques composés.

- La combinatoire des enclitiques

  On ne peut pas avoir plus de deux enclitiques simples entrant en combinaison. De plus cette combinaison est régie par une règle sémantico-syntaxique concernant les personnes représentées par ces pronoms clitiques, dans un ordre

croissant de "spécification" (*taÌÒîÒ*) allant de la 1ère personne à la 3ème personne. La 1ère personne est donc plus spécifiée que la 2ème personne qui, elle, est plus spécifiée que la 3ème personne. Selon cette combinatoire, l'on ne peut donc avoir que les combinaisons suivantes : 1ère P-2ème P, 1ère P-3ème P ou 2ème P-3ème P (il est donc interdit d'avoir 2ème P-1ère P ni 3ème P-2ème P ni 3ème P-1ère P).

1.4.2. Le mot graphique arabe : formants-noyau et formants-extensions

L'intérêt de cette approche de l'analyse du mot graphique arabe, présentée dans (Dichy 1997-b), est qu'elle mette en évidence l'unité lexicale latente en faisant apparaître notamment la saillance du noyau lexical. Le souci principal de cette analyse est de donner « un fondement formel à la grammaire qui régit les relations entre les morphèmes constitutifs du mot graphique »[170]. Cette "formalisation" de la grammaire est basée sur une convention de définition des morphèmes constitutifs de l'unité-mot, appelés *formants de mot*. Ces *formants de mot*, qui peuvent être répartis en *formants-noyau (Fn)* et *formants-extensions (Fe)*, sont « des signes linguistiques minimaux dont les relations de contextualisation sont limitées aux autres morphèmes inclus dans l'unité composée que constitue le mot dans sa manifestation graphique »[171].

Les relations de contextualisation qui caractérisent les *formants de mot* sont de deux types : une relation d'ordre et un ensemble de relations de collocation. Les positions occupées par les formants sur le vecteur *ordonné* de représentation du mot (Dichy 1990), illustre bien cette relation d'ordre. Les relations de collocation quant à elles, ont pour objet :

---

[170] Dichy Joseph, Pour une lexicomatique de l'arabe, 1997, p. 297

[171] *ibidem*, p. 295

« - de valider ou d'invalider la compatibilité des formants entre eux (par exemple, le proclitique (سَ , /sa-/), qui est la marque du futur n'est compatible qu'avec des préfixes, bases et suffixes du paradigme verbal de l'inachevé.

- de permettre les modifications affectant un nombre important de formants, en fonction de leur contexte (par exemple : modifications de bases verbales relevant de racines anomales selon la personne, le genre et le nombre). »[172]

Une compatibilité doit être respectée, entre la grammaire régissant ces relations d'un côté, et les deux fonctionnements dissymétriques de la *synthèse* et de l'*analyse*, de l'autre côté.

Cette approche aboutit à considérer la présence dans l'unité-mot, d'un formant-noyau (Fn) correspondant à une catégorie lexicale (la base) appartenant à une liste ouverte, et celle d'un ensemble de formants-extensions (Fe) appartenant à une liste fermée. Cela a pour résultat de distinguer à l'intérieur des mots graphiques : les mots-lemmes, ayant un noyau lexical, et les mots qui en sont dépourvus.

Cependant, le formant-noyau (Fn) « ne coïncide pas nécessairement avec une unité lexicale (UL). Certains formants-extensions (Fe) sont en effet susceptibles, lorsqu'ils sont associés à une base nominale, de se trouver, pour ainsi dire, pris avec elle dans un processus de lexicalisation. Un formant-extension sera dit lexicalisé (appelé formant-extension lexicalisé - Fel) lorsque l'unité <Fn, Fel> résultant de son association à un formant-noyau donné constitue une unité du lexique (UL) indépendante. »[173]

Ces considérations faites, cette approche permet d'analyser une unité lexicale soit comme un formant-noyau (UL = <Fn>) à l'instar des bases verbales ou nominales dépourvues de Fel (نادر *nâdir* [rare]), soit comme un ensemble UL = <Fn, Fel> où Fel

---

[172] *ibidem*, p. 304

[173] Dichy Joseph, *idem*, p. 299

peut inclure plus d'un formant, exemple نَـادِرَة [Plur. نَوَادِر, *nawâdir*] analysable en UL = <Fn = (نادر *nâdir*), Fel = (ـة *at*)>.

Joseph Dichy conclue cette présentation de l'analyse du mot graphique en *formants-noyau et formants-extensions* en formulant l'hypothèse selon laquelle « la relation d'ordre au sens strict et la présence d'un noyau lexical constituent deux traits universaux de la définition du mot, dont la manifestation la plus directement observable est le *mot graphique* »[174]

1.5. Nos unités de décompte

Comme nous venons de le voir, cette analyse du mot graphique arabe est certainement très féconde au niveau morphologique ou syntaxique. Cependant, cette décomposition poussée de l'unité graphique jusqu'à arriver aux morphèmes (base, affixes) n'est pas nécessaire pour la phase de dépouillement lexicométrique qui n'exige qu'une segmentation lexicale, à moins que l'objectif de l'étude quantitative ne soit le décompte des morphèmes et non celui des unités lexicales. Si l'itémisation du mot maximal en mot minimal et clitiques (proclitiques et enclitiques) est pertinente en lexicométrie, celle du mot minimal en Base et affixes (préfixes et suffixes) ne l'est pas pour autant, du moins dans cette perspective qui est la nôtre. Décomposer, par exemple, يَكْتُبُونَ *yaktubûna* [Ils écrivent] en يَ + كْتُبْ + ُونَ ne serait d'aucune utilité pour une analyse lexicométrique où les unités de décompte sont les unités lexicales et non les morphèmes. D'autant plus que ni la Base, c'est à dire le mot minimal dépourvu de ses affixes, ni ces affixes eux-mêmes ne constituent des unités lexicales autonomes (ou formes libres minimales, selon la terminologie de Bloomfield et Lyons).

---

[174] *ibidem*, p. 304-305

En outre, cette inadéquation entre l'unité graphique et de l'unité lexicale ne se manifeste pas qu'au niveau de l'agglutination, elle peut aussi se traduire au niveau compositionnel où plusieurs unités graphiques doivent être regroupées pour former un seul élément du lexique.

L'unité graphique ne correspond donc pas toujours à l'unité lexicale. L'unité graphique peut équivaloir à une unité lexicale comme elle peut équivaloir à plusieurs unités lexicales ; et inversement, plusieurs unités graphiques peuvent équivaloir à une seule unité lexicale. Voici, ci-après, quelques exemples de ces cas de figures :

1.5.1. Une unité graphique = une unité lexicale

Ce cas de figure représente plus de 36 % des formes graphiques différentes et près de 47 % des occurrences des unités graphiques de notre corpus[175].

C'est le cas des noms indéfinis simples sans clitiques :

Exemples : شيء *šayÞ* [une chose], qui a une fréquence de 73 occurrences sous cette forme-là, c'est-à-dire dénudée de tous clitiques y compris l'article défini.

C'est aussi le cas des verbes sans clitiques :

Exemples : قَالَ *qâla* [dire], qui a la fréquence 259 ou كان *kâna* [être au passé], avec une fréquence de 220 sous la forme non agglutinée.

Tous les mots-outils simples non agglutinés rentrent aussi dans ce cadre :

Exemples : في *fî* [dans], مِن *min* [de], على *Ýalâ* [sur], qui ont, dans notre corpus, respectivement 1050, 915 et 518 occurrences sous ces formes non agglutinées.

---

[175] Voir nos données statistiques concernant la segmentation, chapitre 2

1.5.2. Une unité graphique = plusieurs unités lexicales

C'est le cas le plus rencontré en arabe, il représente plus de 53 % des mots graphiques en arabe. C'est le cas du mot maximal évoqué plus haut et où l'on trouve parfois des phrases-mêmes contenues dans une seule unité graphique comme dans le verset XI-28 du Coran :

أَنُلْزِمُكُمُوهَا *'anulzimukumûhâ* [irions-nous vous les imposer ?][176].

C'est le cas aussi d'un certain nombre de mots-outils qui "se soudent" entre eux en provoquant une assimilation de certaines consonnes en contact :

أَلاَّ	=	أَنْ + لاَ	*'allâ = 'an + lâ*	[que…ne…pas]
مِمَّا	=	مِنْ + مَا	*mimmâ = min + mâ*	[de ce que]
هَهُنَا	=	هُنَا هَا +	*hahunâ = hâ + hunâ*	[ici, il y a]
لِئَلاَّ	=	لِ + أَنْ + لاَ	*li'allâ = li + 'an + lâ*	[pour que…ne…pas]

Le même phénomène se produit aussi avec certains interrogatifs. Dans ce cas, outre la mutation et l'assimilation des consonnes (pour certains), l'*alif* final disparaît :

عَمَّ ؟	=	عَنْ + مَا ؟	*Ýamma = Ýan + mâ*	[à quel sujet ?]
مِمَّ ؟	=	مِنْ + مَا ؟	*mimma = min + mâ*	[de quoi… ?]
بِمَ ؟	=	بِ + مَا ؟	*bima = bi + mâ*	[avec quoi ?]
فِيمَ ؟	=	فِي + مَا ؟	*fîma = fî + mâ*	[dans quoi ?, concernant quoi ?]

[176] La traduction est de J. Berque, voir : *Le Coran*, Essai de traduction de l'arabe par Jacques Berque, 1990, p. 233.

1.5.3. Plusieurs unités graphiques = une unité lexicale

On peut effectivement se trouver, devant des cas où plusieurs unités graphiques ne forment en réalité qu'une seule unité lexicale ou plus précisément, une unité polylexicale :

1.5.3.1. Les locutions grammaticales

Composées de mots-outils simples, ces unités polylexicales ont un sens différent de celui de chaque mot-outil ayant contribué à sa composition :

إِلاَّ أَنَّ	*'illâ 'anna*	[cependant]
عَلَى أَنَّ	*Ýalâ 'anna*	[bien que, néanmoins]
عَلَى أَنْ	*Ýalâ 'an*	[à condition que, sous réserve que]

1.5.3.2. Les noms communs composés

Comme en français ou dans d'autres langues, il y a des noms communs qui sont composés de plusieurs unités graphiques :

مالك الحزين	*mâliku l-Îazîn*	[héron]
شقائق النعمان	*šaqâ'iqu 'al-nuÝmân*	[coquelicots]

1.5.3.3. Les noms propres composés

Certains noms de personnes ou même de lieux sont formés de deux ou plusieurs unités graphiques :

عَبْدُ اللهِ	*Ýabdu l-llahi*	[Abdallah (esclave de Dieu)]
دَار السَّلاَمِ	*dâru s-salâmi*	[Capitale de Tanzanie]

Il y a même des noms de personnes qui sont composés d'une phrase complète :

تَأَبَّطَ شَرًّا	*taÞabbaÔa šarr-an*	[littéralement : Il a pris le mal sous le bras]. Nom d'un célèbre poète arabe de l'époque antéislamique.
شَابَ قَرْنَاهَا	*šâba qarnâhâ*	[littéralement : Elle s'est faite des tresses (cheveux) blanches]. Nom d'une femme de l'époque antéislamique.

1.5.3.4. Les nombres composés :

ثَلاَثَةَ عَشَر	*×alâ×ata Ýašara*	[treize]
ثَمَانِيَةٌ وتِسْعُونَ وتِسْعُمِئَةٍ وأَلْفٌ	*×amâniyatun wa tisÝûna wa tisÝumi'atin wa 'alfun*	[1998]

Nous allons donc voir plus en détail ces cas de figures, et d'autres, pour savoir exactement où commence et où se termine telle ou telle unité lexicale c'est-à-dire en définir les frontières.

1.6. Les frontières entre les mots

1.6.1. Les unités segmentables

1.6.1.1. Le mot graphique en général

C'est le cas le plus général des mots graphiques en arabe dont l'analyse a été présentée plus haut. Le mot graphique correspond soit à un mot maximal segmentable en mot minimal et clitiques, soit à un mot minimal et dans ce cas, les clitiques correspondent à l'ensemble vide : on pourrait parler de *clitiques zéro*.

1.6.1.2. Les formes verbales « composées » !

Outre l'accompli "absolu" الماضي المطلق *'al-mâÃî l-muÔlaq* appelé aussi par certains grammairiens modernes الماضي الكامل *'al-mâÃî l-kâmil*[177] qui exprime le procès achevé dans le passé, l'on peut trouver, en arabe, l'accompli duratif ou d'habitude الماضي الاستمراريّ أو التعوّديّ *'al-mâÃî l-istimrâriyy* ou *'at-taÝawwudiyy* appelé aussi الماضي الناقص *'al-mâÃî n-nâqiÒ*[178] qui exprime l'inaccompli localisé dans le passé et est rendu par *kâna* + un verbe à l'inaccompli, l'accompli lointain الماضي البعيد *'al-mâÃî l-baÝîd* appelé aussi الماضي الأكمل *'al-mâÃî l-'akmal*[179] qui exprime le passé dans le passé et est rendu par *kâna* + (*qad*) + un verbe à l'accompli, et l'accompli futuratif[180] الماضي الاستقباليّ *'al-mâÃî l-istiqbâliyy* appelé

---

[177] Rachîd aš-Šartûnî, *MabâdiÞ al-Ýarabiyya*, tome 4, 1951, p.10

[178] *ibid*, *idem*, p.10

[179] *ibid*, *idem*, p.10-11

[180] Voir : Rached Hamzaoui, *L'académie de langue arabe du Caire. Histoire et œuvre*, 1975, p. 381

aussi المستقبل السابق *al-mustaqbalu s-sâbiq*[181] qui indique que deux procès auront lieu successivement dans le futur et est rendu par *kâna* (à l'inaccompli) + *qad* + un verbe à l'accompli.

Peut-on réellement parler, avec ces différents accomplis, de verbes composés ou de formes verbales composées en arabe ?

Il n'existe pas, en arabe, à proprement parler de verbes composés tels qu'on les connaît, par exemple, en français. L'arabe ne connaît pas les formes composées pour la simple raison qu'il n'a pas d'auxiliaire de conjugaison. Le verbe كَانَ\يَكُونُ *kâna/yakûnu* n'est absolument pas un auxiliaire de conjugaison, il est considéré par la Tradition grammaticale arabe comme un verbe opérateur appliqué à toute une phrase. Il a, selon l'expression de Blachère et Gaudefroy-Demombynes[182], une valeur d'*exposant temporel* devant un verbe à l'indicatif. Les rapprochements que font parfois certains arabisants, dans un souci purement didactique, entre les temps composés en français comme le plus-que-parfait ou le futur antérieur, et la structure arabe utilisant *kâna/yakûnu* + un autre verbe placé dans son champ n'ont, nous semble-t-il, aucun bien-fondé théorique. C'est ce qui a poussé R. Blachère et M. Gaudefroy-Demombynes à objecter qu'« il serait faux de voir dans les combinaisons *kâna* + *accompli* ou *kâna* + *inacc. Indic.*, des complexes analogues aux temps composés du français. En réalité chaque élément de la combinaison, en arabe, garde son sens et sa valeur propre »[183]. De même, bien avant Blachère et Gaudefroy-Demombynes, Antoine-Isaac Silvestre De Sacy était fort catégorique quant à la considération du verbe *kâna/yakûnu* comme un verbe auxiliaire : « L'usage que font les Arabes du verbe كَانَ pour modifier la valeur du prétérit et des aoristes, pourrait faire envisager ce verbe comme un verbe auxiliaire, mais ce serait une

[181] *Rachîd aš-šartûnî*, op. cit., p.11

[182] Régis Blachère et Maurice Gaudefroy-Demombynes, *Grammaire de l'arabe classique (Morphologie et syntaxe)*, 1952, p.253

[183] R. Blachère et M. Gaudefroy-Demombynes, *op. cit.*, p.271

erreur. L'emploi du prétérit ou de l'aoriste du verbe كَانَ n'a réellement pour objet que d'exprimer, par la réunion de deux verbes (...) cette double relation de temps »[184]. D'autant plus que le verbe *kâna/yakûnu* et le verbe placé dans son champ peuvent avoir des sujets différents. En outre, un seul et même *exposant temporel*, au commencement du récit, peut régir plusieurs autres verbes placés dans son champ pour décrire un(des) événement(s) passé(s), par exemple. Cet *exposant temporel* à l'accompli « donne la sphère du passé qui vaut pour toute la suite ; les inaccomplis se succèdent en coordination, tant que l'on a besoin d'exprimer un inaccompli dans le passé »[185] comme dans l'exemple suivant que donne Silvestre De Sacy :

«**وكان يُحبّ** الشعر والشعراء **ويَميل** إلى أهل الأَدب والفقه **ويَكرَه** المِراء في الدين »[186]

(C'est nous qui soulignons)

« *Wa* ***kâna yuÎibbu*** *š-šiŕra wa š-šuŕarâ'a wa* ***yamîlu*** *'ilâ 'ahli l-'adabi wa l-fiqhi wa* ***yakrahu*** *l-mirâ'a fi d-dîni* »

« **Il *aimait*** la poésie et les poètes, ***avait de l'inclinaison*** pour les hommes de lettres et les jurisconsultes, et ***détestait*** les contestations en matière de religion ».

En effet, après cette assise théorique, nous considérons qu'il est tout à fait évident voire même nécessaire de segmenter la combinaison de ces verbes, c'est-à-dire, séparer l'exposant temporel *kâna* et le(s) verbe(s) qui est (sont) dans son champ et de compter chaque verbe à part puisqu'ils sont, comme nous venons de le voir, des verbes "autonomes" dont chacun garde son sens et sa valeur propre.

Cette segmentation est très importante en elle-même certes, mais elle a aussi une répercussion non moins inestimable lors de la lemmatisation comme nous le verrons plus loin.

---

[184] Antoine-Isaac Silvestre De Sacy, *Grammaire arabe à l'usage des élèves de l'École spéciale des langues orientales vivantes*, 1831, p. 213.

[185] Henri Fleisch, "Études sur le verbe arabe", dans *Mélanges Louis Massignon*, 1957, p. 174.

[186] Antoine-Isaac Silvestre De Sacy, *idem*, p. 208.

1.6.1.3. Les mots-outils agglutinés

Les mots-outils suivants ont été segmentés :

Mot-Outil graphique		Mots-Outils simples
مِمَّا	➔	مِن\ما
مِمَّن	➔	مِن\مَن
عَمَّا	➔	عن\ما
عَمَّن	➔	عن\مَن
فيما	➔	في\ما
كيْلا	➔	كيْ\لا
ألاّ	➔	أنْ\لا
لئلاّ	➔	ل\ أنْ\لا

Il serait légitime de se poser la question de savoir pourquoi avons-nous décidé de segmenter des mots-outils graphiques tels que : عَمَّن , مِمَّا ou لئلاّ bien que les mots-outils simples entrant dans leur composition (exception faite de كيْلا) ne soient pas graphiquement directement isolables en raison des phénomènes d'assimilation إدغـام *ÞidÈâm* et de mutation قلـب *qalb*, alors que, d'un autre côté, nous n'avons pas vu nécessaire de segmenter, comme nous le montrons plus loin, des mots-outils graphiques tels que : بمَ ؟ (avec quoi ?), فِيمَ ؟ (en quoi ?) ou عَلاَمَ ؟ (sur quoi ?), bien que les mots-outils simples qui les composent soient graphiquement directement isolables. Nous présentons les justifications de notre choix plus loin dans la section consacrée aux interrogatifs agglutinés.

1.6.1.4. Le cas de إلاّ

Le mot-outil إلاّ *Þillâ* a quatre acceptions différentes en arabe :

- particule d'exception استثنائيّة *Þisti×nâÞiyya*,
- particule de restriction حصريّة *ÎaÒriyya*,
- particule "nominale" اسميّة *Þismiyya* (ayant le sens de غَيْر *Èayr* [autre]),
- particule "amalgamée" وكّبة *murakkaba* (amalgame de إنْ et de لا).

Les trois premières formes sont fortement lexicalisées et ne doivent donc pas être segmentées. La quatrième elle, n'est que ponctuellement amalgamée (ou agglutinée) de إنْ (الشـرطيّة الجازمـة [particule de condition régissant l'apocopé]) et de لا (particule de négation), et ne jouit pas d'un degré de lexicalisation lui permettant de former une nouvelle unité indépendante des deux particules la composant ; elle doit de ce fait être segmentée. Notons par ailleurs que dans ce cas de figure, les opérations de segmentation et de désambiguïsation s'entremêlent : on peut commencer par segmenter pour filtrer les ambiguïtés, ou désambiguïser d'abord, pour décider de segmenter dans un deuxième temps.

1.6.2. Les unités non segmentables

1.6.2.1. Les unités polylexicales

1.6.2.1.1. Les mots composés

- **Les noms communs composés**

La composition est utilisée ici dans le sens de « la formation d'une unité sémantique à partir d'éléments lexicaux susceptibles d'avoir par eux-mêmes une autonomie dans la langue »[187]. Du fait que cette composition soit durable et non occasionnelle, les unités résultantes de cette association ne doivent pas être segmentées.

Nom composé	Traduction
ابن أخ	Neveu
بنت عمّ	Cousine
ابن عِرس	Belette
حمار الوحْش	Zèbre
داء الثعلب	Alopécie
دودة القزّ	Ver à soie
ربيع الأوّل	Mois du calendrier de l'Hégire
عِنَب الحيّة (الحنظل)	coloquinte
كلْب الماء	Phoque
ابن آوى	Chacal
بنت وردان	Blatte, Cafard
أمّ أربع وأربعين (حَرِيش)	Mille-pattes
مالِك الحزين	Héron
شقائق النعمان	Coquelicots
أمير المؤمنين	Commandeur des croyants

- **Les noms propres composés**

De la même manière que les noms communs composés, les noms propres composés ne doivent pas non plus être segmentés.

NP de personnes	NP de lieux	Noms de tribus, groupes et nations	Noms des œuvres

[187] Dubois J. et al., *Dictionnaire de linguistique et des sciences du langage*, *op. cit.*, p. 106.

إصلاح المنطق	بنو عبْد المطّلب	باب الجسر	ابن عبّاد
إنقاذ البَشَر مِن الجبْر والقدَر	آل سامان	البحْر الميّت	الحجّاج بن يوسف أبو محمّد
صفْو الشرْح	أهل الكتاب	البيْت العتيق	ذو الكفايتيْن
كتاب إقليدس	بنو أسد	دُومة الجنْدل	رسول الله
كتاب الحيوان	أصحاب رسول الله	سُرّ مَن رَأَى	سيْف الدولة

- **Les nombres composés**

Qu'ils soient de composition directe comme اثنا عشر *Þi×nâ Ýašar* [douze] ou خمسَ عشـرة *Ìamsa Ýašrata* [quinze], ou de composition avec coordination comme خمسـة وأربعـون *Ìamsatun wa ÞarbaÝûna* [quarante-cinq] ou ثلاثة وثلاثون *×alâ×atun wa ×alâ×ûna* [trente-trois], ces nombres composés ne sont pas à segmenter.

- **Les deux formes exclamatives**

Étant donné que ces deux formes d'exclamation أَفْعِلْ بِـ… – مَا أَفْعَلَ sont des mots-outils composés, il ne faut pas qu'ils soient segmentés pour ne pas éparpiller les éléments entrant en leur composition.

Solution de commodité :

Quand la segmentation se fait à la main ou quand le segmenteur automatique ou semi-automatique utilisé n'est pas doté, pour une raison ou une autre, d'un inventaire d'unités polylexicales, il faut détecter manuellement ces unités complexes et leur adjoindre, le cas échéant, un caractère spécial déterminé pour garantir leur unité et éviter ainsi l'émiettement des unités les composant.

C’est d’ailleurs le cas auquel nous étions confronté dans ce travail. En effet, notre segmenteur n’étant pas encore doté d’une liste d’unités polylexicales, nous avons fait le choix, avant de passer le texte au segmenteur, de procéder à une opération de repérage manuel de ces unités composées puis de procéder à une ligature, entre elles, des unités lexicales composant des unités polylexicales. Cette opération de ligature ou d’agglutination forcée a été possible par l’adjonction entre les unités lexicales simples du caractère spécial d’addition « + » remplaçant ainsi l’espace qui y existait. Une règle a été ajoutée par la suite au programme du segmenteur[188] pour qu’il reconnaisse comme une seule unité les suites de caractères alphabétiques séparées par « + ».

1.6.2.2. Les unités agglutinées

1.6.2.2.1. Les interrogatifs agglutinés

N’ont pas été segmentés les interrogatifs suivants :

Interrogatif agglutiné	À ne pas segmenter
بِمَ ؟	*ب\مَا
لِمَ ؟	*لِ\مَا
فِيمَ ؟	*فِي\مَا
مِمَّ ؟	*مِن\ما
عَمَّ ؟	*عَن\ما
عَلاَمَ ؟	*عَلى\ما
حَتَّامَ ؟	*حَتَّى\ما

Contrairement à la liste des mots-outils agglutinés vus précédemment (§ 1.6.1.3), ces mots-outils sont fortement lexicalisés comme interrogatifs agglutinés et ce notamment pour marquer une distinction entre eux et des mots-outils agglutinés d’autre

[188] Voir le chapitre 1 : *Constitution et dépouillement du corpus*.

nature tels que : بِمَا *bimâ* [en ce que / avec ce que], مِمَّا *mimmâ* [de ce que], لِمَا *limâ* [à ce que / pour ce que / pour ce qui est], عَمَّا *Ýammâ* [sur ce que / de ce que], etc. qui eux, sont segmentable respectivement en ب\ما , مِن\ما , لِ\ما , عَن\ما , alors que les interrogatifs بِمَ *bima* [avec quoi ? / à quoi ?], مِمَّ *mimma* [de quoi ?], لِمَ *lima* [pourquoi ?], عَمَّ *Ýamma* [sur quoi ? / de quoi ? / qu'est-ce que ?], etc. ne le sont pas, du fait de leur utilisation exclusivement interrogative. Aussi, le choix de ne pas segmenter va-t-il contribuer à la désambiguïsation et faciliter de ce fait, l'opération de catégorisation. En effet, au cas où l'on décide de segmenter ces interrogatifs agglutinés, on va se demander à chaque fois si ما , issu de la segmentation doit être catégorisé comme interrogatif, relatif, particule de *maÒdarité*, nom de condition ou particule de négation.

1.6.2.2.1. Les mots-outils lexicalisés

Il est à noter que la question de segmenter ou non certains démonstratifs comme, par exemple, ذاك *Æâka* [celui-ci] (démonstratif de mitoyenneté) composé, à l'origine, du démonstratif ذا *Æâ* et de كاف المخاطبة *kâf al-muÌâÔaba* [pronom personnel affixe de 2e personne] ou ذلــك *Æâlika* [celui-là] (démonstratif d'éloignement) composé du démonstratif ذا *Æâ*, de لام البُعد *lâm al-buÝd* [le "l" d'éloignement] et de كاف المخاطبة *kâf al-muÌâÔaba* [pronom personnel affixe de 2e personne] , ne se pose même plus tellement ces démonstratifs agglutinés sont fortement lexicalisés et appartenant, de ce fait, à la même liste que les démonstratifs simples les composant.

Du fait de leur forte lexicalisation, les mots-outils présentés dans la liste suivante n'ont pas été segmentés :

Mot-outil agglutiné	À ne pas segmenter
كَيْفَما	*كَيْفَ\ما
قَلَّما	*قَلَّ\ما
كَثُرَما	*كَثُرَ\ما
عِنْدَما	*عِنْدَ\ما
طَالَما	*طَالَ\ما
إنَّما	*إنَّ\ما
كُلَّما	*كُلَّ\ما
كأنَّما	*ك\أنَّ\ما* ni كأنَّ\ما
أيُّما	*أيُّ\ما
كَما	*كَ\ما
كَيْما	*كَيْ\ما
لَعَلَّما	*لَعَلَّ\ما
حَيْثُما	*حَيْثُ\ما
رَيْثَما	*رَيْثَ\ما
مِثْلَما	*مِثْلَ\ما
حِينَما	*حِينَ\ما
بَيْنَما	*بَيْنَ\ما
اللّهمّ	*اللّه\مّ ni *يا\اللّه
وَقْتَئِذٍ	*وَقْتَ\إذٍ
حِينَئِذٍ	*حِينَ\إذٍ
ساعَتَئِذٍ	*ساعَةَ\إذٍ
لَقَدْ	*لَ\قَدْ
هكذا	*ه\ك\ذا
كأنّ	*ك\أنّ
سِيَّما	*سِيَّ\ما
نِعِمَّا	*نِعِمْ\مَا[189]

C’est le cas aussi du mot-outil agglutiné كأنَّ *kaÞanna* [comme si] qui, composé au départ, de la préposition كَ *ka* [comme] et du mot-outil أنَّ *’anna* [que, parce que], est considéré par la suite, en raison de sa forte lexicalisation, en tant que mot-outil simple. Il a rejoint, de ce fait, la liste fermée des mots-outils appelés إنَّ وأخواتها *’inna wa-’aÌawâtuhâ* [*’inna* et ses sœurs].

1.6.2.3. Les mots commençant par "al"

1.6.2.3.1. Les noms propres commençant par Þal

Un grand nombre de noms propres arabes sont définis par l’article. Ce dernier fait partie intégrante de l’identité de ces mots et leur donne même, dans certains cas, la qualité de nom propre.

Nom propre commençant par "al"	Traduction
الله	*Allah*
البندقيّة	*Venise*
القرآن	*Le coran*
الريّ	*Rayy* (ville en Iran)
الجاحظ	*JâhiÛ*
المحمّدون	*Les Mohamed*

Il est à noter, ici, que le pluriel des noms propres n’est pas défini par nature et n’est plus spécifié sémantiquement ; et c’est pour cette raison que l’on doit le définir par

[189] Ces deux mots-outils s’écrivent agglutinés *niÝimmâ* du fait que le *Ýayn* de *niÝma* ait une *kasra* à la place du *sukûn*. Ils s’écrivent, en revanche, en deux mots distincts dans le cas contraire : نِعْمَ مَا *niÝma mâ*.

l'article pour qu'il ne perde pas son statut de nom propre. Le pluriel d'un nom propre défini par nature est donc un nom propre défini par l'article. Le pluriel, par exemple, de محمّد *MuÎammad* [Mohamed] qui est défini par nature est المحمّدون *al-MuÎammadûna* [Les Mohamed], défini par l'article.

1.6.2.3.2. Les relatifs

La grande majorité des relatifs comme الذِي *Þallaæî* - التِي *Þallatî* - الذِينَ *Þallaæîna* - اللاَّتِي *Þallâtî* - اللَّذَانِ *Þallaæâni* - اللَّتَانِ *Þallatâni*, commencent par l'article défini *Þal*, devenu partie intégrante de ces unités. Segmenter ces unités lexicales, c'est toucher à leur intégrité et leur ôter le statut de relatifs qu'ils ont acquis en partie grâce à l'article parce que c'est celui-ci qui les spécifie sémantiquement.

1.6.2.3.3. Les jours de la semaine

À part quelques rares contextes où les jours de la semaine peuvent être utilisés indéfinis, ces unités lexicales sont, la plupart du temps, utilisées définies par l'article *Þal*. Ainsi, lundi الاثنين *Þal-Þi×nayni*, mardi الثلاثاء *Þa×-×ulâ×âÞa*, mercredi (الإربعاء الأربعاء) *Þal-ÞirbiʕâÞa (Þal-ÞarbiʕâÞa ou Þal-ÞarbuʕâÞa)*, jeudi الخميس *Þal-Ìamîsa*, vendredi الجمعة *Þal-jumuʕata*, samedi السبت *Þas-sabta*, et dimanche الأحد *Þal-ÞaÎada*, sont presque toujours écrits avec l'article[190] ; il ne faut donc pas segmenter ces unités lexicales.

[190] Sauf dans des tournures du type : اليَوْمَ يَوْمُ جُمُعَةٍ *Þal-yawma yawmu jumuʕat-in*.

1.6.2.4. Les unités discontinues : la négation bi-segmentale

Des tournures telles que ليس…بعدُ *laysa…baʕdu* [ne...pas encore (ph. nominale)], لم…بعدُ *lam…baʕdu* [ne...pas encore (ph. verbale)], لا…أبدًا *lâ…Þabadan* [ne...jamais (présent)], لم…أبدًا *lam…Þabadan* [ne...jamais (passé)], لم…قطُّ *lam…qaÔÔu* [ne...jamais (passé)], لن…أبدًا *lan…Þabadan* [ne...jamais (futur)], sont parfois bien utiles pour étudier le style d'au auteur. Pour les étudier, il est inévitable de les repérer dans le discours. Pour ce faire, faut-il les regrouper ou les segmenter ?

Nous évoquons, en fait, ces unités discontinues parce que notre propos ici est bien l'établissement d'une *norme de segmentation* ; et comme toute norme, elle doit aspirer à l'exhaustivité et prévoir toutes les possibilités envisageables. Ceci étant, nous avouons que ce cas de la négation bi-segmentale est un cas très délicat quant au choix à faire et aux décisions à arrêter pour regrouper ou segmenter.

On peut en effet, choisir de ne pas segmenter, donc de regrouper ensemble ces unités même si elles sont séparées par une suite de mots plus ou moins longue. Comme on peut décider de segmenter dans un premier temps et puis de faire des regroupements après coup, pour pouvoir éventuellement étudier ces unités discontinues plus en détail à part. C'est cette deuxième solution que nous suggérons dans ce cadre de la *norme de segmentation*.

2. Désambiguïsation

2.1. L'homographie en arabe

L'homographie est la conformité, au niveau de la manifestation écrite, de deux ou plusieurs unités linguistiques distinctes.

En général, dans les langues isolantes, toute forme graphique appartenant à deux ou plusieurs lemmes est une forme homographe. Mais dans les langues agglutinantes ou flexionnelles et surtout en arabe, la question est beaucoup plus complexe. Cette assertion nécessite donc davantage de précisions quand il s'agit du mot graphique en général et du mot graphique arabe[191] en particulier.

En effet, compte tenu de la distinction fondamentale qui existe, en arabe, entre mot minimal et mot maximal, et sachant que l'identité graphique peut avoir lieu entre deux mots minimaux, deux mots maximaux ou un mot minimal et un mot maximal, nous devons apporter des précisions en définissant l'homographie de la façon suivante :

Toute forme graphique appartenant à deux ou plusieurs lemmes ou pouvant être segmentée de deux ou plusieurs façons différentes donnant, chacune d'entre elles, des vocables se rattachant à un ou plusieurs lemmes différents est une forme homographe.

En d'autres termes, est considéré comme homographe :

- Tout mot minimal se rattachant à deux ou plusieurs lemmes,

  Exemples **:**

[191] Voir la notion du mot graphique en arabe (p. 207-211), et pour plus de détail voir (Dichy 1990) et (Hassoun 1987).

دار → دَارٌ [une maison]
→ دَارَ [tourner]

حديث → حَدِيثٌ [conversation]
→ حَدِيثٌ [récent]

- Tout mot maximal ayant deux ou plusieurs segmentations possibles donnant chacune des vocables se rattachant à un ou plusieurs lemmes différents,

Exemples **:**

فتحته → فَتَحْتُ\هُ – فَتَحْتَ\هُ – فَتَحَتْ\هُ [je l'ai - tu l'as - elle l'a ouvert]
→ فَ\تَحْتَ\هُ [et au-dessous de lui]
→ فَتْحَةُ\هُ [sa fente]

أباه → أَبا\هُ [son père, *à l'accusatif*]
→ أَبَى\هُ [il l'a refusé]

- Tout mot maximal, d'une part se rattachant, tel quel, à un ou deux lemmes et, d'autre part pouvant être segmenté, de deux ou plusieurs façons, en mot minimal et clitiques se rattachant, chacun d'entre eux, à un, deux ou plusieurs lemmes.

Exemples **:**

	→	أَفْعَالٌ	[verbes, actions]
أفعال	→	أ\فِعَالٌ ؟	[est-ce un accomplissement ?]
	→	أ\ف\عَالٍ ؟	[alors, est-il haut ?]

	→	وَلَهَ	[perdre la tête, se passionner]
وله	→	وَلَهٌ	[agitation, stupeur, passion]
	→	و\ل\ هـ	[et / à - pour / lui]

2.1.1. Les types d'homographie

2.1.1.1. Homographie consonantique

Nous appelons homographie consonantique toute conformité, au niveau de la manifestation écrite, entre deux ou plusieurs structures consonantiques de mots entièrement non vocalisés. Cette homographie peut disparaître dès lors qu'on vocalise partiellement (en plaçant les voyelles distinctives) ou totalement le mot homographe.

Exemples : كتب

2.1.1.1. Homographie globale

Nous appelons homographie globale toute conformité, au niveau de la manifestation écrite, de deux ou plusieurs mots entièrement vocalisés. Contrairement à l'homographie consonantique, la conformité n'est pas seulement entre les structures consonantiques, mais aussi entre les structures vocaliques des mots homographes. C'est-

à-dire que les mots homographes ont le même agencement de consonnes et le même vocalisme.

Exemples :

قَالَ → قَالَ \ يَقُولُ [Dire]
→ قَالَ \ يَقِيلُ [Faire la sieste]

حديث → حَدِيثٌ Nom
→ حَدِيثٌ Adjectif

2.1.1.2.1. Un premier cas d'homographie globale : les verbes concaves.

Il existe en arabe un certains nombre de verbes concaves qui présentent la particularité d'être homographes à l'accompli (la forme équivalente à l'infinitif français) et qui ne peuvent être différenciés qu'à l'inaccompli.

Les verbes concaves sont de deux catégories : des verbes concaves en « y » (*'ajwaf yâ'iyy*) c'est-à-dire des verbes dont la troisième consonne de la racine est un *yâ'* « y » et des verbes concaves en « w » (*'ajwaf wâwiyy*) c'est-à-dire des verbes dont la troisième consonne de la racine est un *wâw* « w ». De ce fait, nous avons trois sortes d'homographies : une homographie interne à la catégorie des verbes concaves en « y », une homographie interne à la catégorie des verbes concaves en « w » et une homographie mixte entre verbes concaves en « y » et verbes concaves en « w ». Dans les deux premiers types d'homographies on est en présence d'une conformité de deux verbes ayant la même racine, alors que dans le troisième type c'est une homographie entre deux verbes ayant deux racines différentes.

À partir de la base de connaissances DIINAR, nous avons répertorié tous les verbes concaves qui sont homographes, ils sont au nombre de 347 verbes concaves[192] répartis comme suit :

- 6 verbes homographes concaves en « y » (*'ajwaf yâ'iyy*)
- 12 verbes homographes concaves en « w » (*'ajwaf wâwiyy*)
- 329 verbes homographes entre verbes concaves en « y » et verbes concaves en « w ». Ce dernier type d'homographes se répartit ainsi :
  - 269 verbes de la forme I (construite sur le schème ***fâla*** فَالَ)

[192] Voir la liste complète de ces verbes à l'Annexe B

- ❖ 36 verbes de la forme IV (construite sur le schème ***'afâla*** أَفَالَ)
- ❖ 12 verbes de la forme VII (construite sur le schème ***'infâla*** انْفَالَ)
- ❖ 10 verbes de la forme VIII (construite sur le schème ***'iftâla*** افْتَالَ)
- ❖ 2 verbes de la forme X (construite sur le schème ***'istafâla*** اسْتَفَالَ)

Nous présentons ici quelques exemples de chacune des trois catégories des verbes concaves homographes :

↳ Homographie entre verbes concaves en « w » et verbes concaves en « y » :

Homographe	Concave en "w" / Racine	Concave en "y" / Racine
حَالَ	يَحُولُ) حَالَ / حول)	يَحِيلُ) حَالَ / حيل)
غَارَ	يَغُورُ) غَارَ / غور)	يَغِيرُ) غَارَ / غير)
قَالَ	يَقُولُ) قَالَ / قول)	يَقِيلُ) قَالَ / قيل)

↳ Homographie interne des verbes concaves en « w » :

Racine	**Homographes**
دود	يَدُودُ) دَادَ)
	يَدَادُ) دَادَ)
روح	يَرُوحُ) رَاحَ)
	يَرَاحُ) رَاحَ)

↳ Homographie interne des verbes concaves en « y » :

Racine	**Homographes**

بيت	(بَاتَ يَبِيتُ)
	(بَاتَ يَبَاتُ)
هيب	(هَابَ يَهِيبُ)
	(هَابَ يَهَابُ)

Il faut remarquer que les verbes homographes de formes IV, VII, VIII et X, même issus de deux racines différentes, présentent une homographie aussi bien à l'accompli qu'à l'inaccompli. Le seul moyen dans ce cas de les différencier c'est de revenir à la racine. Exemple :

أَطَافَ / يُطِيفُ de racine طوف ≠ أَطَافَ / يُطِيفُ de racine طيف

2.1.1.2.2. Un deuxième cas d'homographie globale :

Pluriel / MaÒdar de schème فُعُول fuÝûl.

Parmi les schèmes qui ont la particularité d'être homographes en arabe, nous citons un schème particulier qui est à la fois un schème de *maÒdar* de première forme et un schème de pluriel d'un certain nombre de noms primitifs : il s'agit du schème فُعُول *fuÝûl*. L'homographie relative à ce schème entre *maÒdar* et pluriel est une homographie globale engendrant une ambiguïté effective (voir *infra*) dans ce sens où, les deux formes présentent une conformité non seulement entre leurs structures consonantiques, mais aussi entre leurs structures vocaliques. Nous présentons, dans le tableau suivant, quelques exemples de ces homographes (Voir Annexe C, pour la liste complète).

Exemples :

Verbe	Homographes *MaÒdar* / Pluriel	Nom au singulier
[s'effaroucher] أَسِدَ / يَأْسَدُ	**أُسُود**	أَسَدٌ *[Lion]*

[piquer une colère] حَرِبَ / يَحْرَبُ	**حُرُوب**	حَرْبٌ *[Guerre]*
[être authentique] حَقَّ / يَحُقّ	**حُقُوق**	حَقٌّ *[Droit]*
[être facile] سَهُلَ / يَسْهُلُ	**سُهُول**	سَهْلٌ *[Plaine]*
[jouer – se distraire] شَمَعَ / يَشْمَعُ	**شُمُوع**	شَمْعٌ *[Cire]*
[être généreux] كَرُمَ / يَكْرُمُ	**كُرُوم**	كَرْمٌ *[Vigne]*

Nous avons compté, à partir de la base de donnée DIINAR, 4348 *maÒdar* et 530 pluriels ayant pour schème فُعُول parmi eux il y a 430 homographes.

La particularité de ce schème est qu'il est fortement homographe parmi les pluriels de noms puisque les homographes représentent tout de même 81,13 % des 530 pluriels ayant pour schème فُعُول . Les *maÒdar* homographes ne représentent que 9,89 % des 4348 *maÒdar* ayant pour schème فُعُول .

2.2. Les ambiguïtés de l'arabe écrit

L'ambiguïté peut être détectée à plusieurs niveaux d'analyse différents du texte ou du discours. Le niveau d'analyse lexicale, le niveau d'analyse morphologique, le niveau d'analyse syntaxique, le niveau d'analyse sémantique ou celui d'analyse pragmatique.

2.2.1. Les niveaux d'ambiguïtés

Partant de la définition déjà présentée plus haut stipulant que l'ambiguïté se définit par le fait que « une même forme se voit associer plusieurs significations

disjointes et mutuellement exclusives »[193], il est clair que le côté sémantique est présent à tous les niveaux d'analyse, du niveau lexical au niveau pragmatique passant par le niveau syntaxique.

De ce fait, il serait oiseux de parler d'ambiguïté sémantique proprement dite, puisque « toute ambiguïté est effectivement un phénomène sémantique »[194].

Nous pouvons, par conséquent, classer les ambiguïtés en trois grands types selon que l'on s'intéresse à l'identification des unités de base, à la constitution des structures et relations syntaxiques ou au repérage des valeurs énonciatives et des situations discursives :

- **Les ambiguïtés morpho-lexicales :** Se situent, d'une part au niveau des mots et se rapportent à l'identification des mots graphiques et à leur analyse en mots minimaux et clitiques (ambiguïtés morphologiques) et, d'autre part au niveau des unités lexicales et se rapportent à la catégorisation des unités lexicales en général (simples et complexes), à l'identification des unités lexicales complexes et à la distinction entre mots lexicaux et mots-outils (ambiguïtés lexicales).

  Exemples :

  حديث → حَدِيثٌ [conversation]
  حديث → حَدِيثٌ [récent]

- **Les ambiguïtés syntaxiques :** Se situent au niveau des structures syntaxiques. Elles concernent l'identification des structures propositionnelles et syntagmatiques et des relations prédicatives.

  Exemples :

---

[193] Catherine Fuchs, *Les ambiguïtés du français*, 1996, p. 139.

[194] *Ibidem*, p.139.

رأيت أستاذة الحضارات القديمة

→ V + (*ʻIÃâfa*) + Adj

→ V + (*MuÃâf* + (*MuÃâf ʻIlayh* + Adj))

- **Les ambiguïtés pragmatiques :** Se situant au niveau du discours, ces ambiguïtés qui ne peuvent être levées que par des connaissances extra-linguistiques concernent l'identification des actants et des valeurs énonciatives (valeurs référentielles et valeurs interlocutives).

Exemples :

إنّما يخشى الله من عباده العلماء[195]
ÞInnamâ yaÌšâ l-llâh min Ýibâdihi l-Ýulamâþ

→ إنّما يخشى اللّهَ من عباده العلماءُ
ÞInnamâ yaÌšâ l-llâha min Ýibâdihi l-Ýulamâþu
[Craignent Dieu, parmi ses esclaves, ceux qui savent.]

→ إنّما يخشى اللّهُ من عباده العلماءَ
ÞInnamâ yaÌšâ l-llâhu min Ýibâdihi l-Ýulamâþa
[Dieu craint, parmi ses esclaves, ceux qui savent.]

وبشريّ هؤلاء قد شابه إلهيّ هؤلاء[196]
wa bašariyy hâÞulâÞi qad šâbah Þilâhiyy hâÞulâÞi

→ وبشريّ هؤلاء قد شَابَهَ إلهيَّ هؤلاء
wa bašariyyu hâÞulâiÞ qad šâbaha Þilâhiyya hâÞulâÞi
[Et l'humain de ceux-ci ressembla au divin de ceux-là.]

→ وبشريّ هؤلاء قد شَابَهُ إلهيُّ هؤلاء
wa bašariyyu hâÞulâiÞ qad šâbahu Þilâhiyyu hâÞulâÞi
[Et l'humain de ceux-ci fut altéré par le divin de ceux-là.]

2.2.2. Les types d'ambiguïté

Nous avons vu plus haut les deux types d'homographie, l'homographie consonantique et l'homographie globale. À l'homographie consonantique correspond l'ambiguïté virtuelle et à l'homographie globale correspond l'ambiguïté effective[197]

[195] Coran : XXXV-28.

[196] *Al-ÞImtâÝ wa-l-MuÞânasa*.

2.2.2.1. Ambiguïté virtuelle

Nous appelons ambiguïté virtuelle toute ambiguïté engendrée par l'absence, partielle ou totale, de voyellation dans l'écriture des mots arabes. Autrement dit, toute ambiguïté pouvant être levée par simple voyellation, totale ou partielle, est une ambiguïté virtuelle.

Exemple : كتب qui peut être analysé d'au moins quatre manières[198] :

كتب
- → كَتَبَ [écrire]
- → كُتِبَ [il a été écrit]
- → كُتُب [des livres]
- → كَتْب [le fait d'écrire]

دين
- → دِين [religion]
- → دَيْن [dette]
- → دِينَ [a été condamné]

---

[197] Nous empruntons à Catherine Fuchs les termes d'« ambiguïté virtuelle » et « ambiguïté effective », mais en leur donnant des définitions différentes émanant des spécificités de la langue arabe. Voir : Catherine Fuchs, *op. cit.*, 1996.

[198] Nous ne considérons ici ni les désinences casuelles ni la *šadda*. Si cela était pris en compte, les analyses possibles du mot كتب se verraient multipliées par quatre. Sur les 16 analyses possibles de ce mot voir : Ouersighni, *op. cit.*, p. 19.

2.2.2.2. Ambiguïté effective

Nous appelons ambiguïté effective toute ambiguïté qui ne peut être levée, même après voyellation, que dans un contexte phrastique ou discursif faisant appel soit aux cooccurrences, immédiates ou médiates, du mot ambigu soit aux connaissances extra-linguistiques. Autrement dit, il existe, entre les mots homographes, une adéquation totale non seulement au niveau des consonnes composant chacun des mots, mais aussi au niveau de leurs schémas vocaliques.

Exemple :

- قال qui peut être analysé en (قَالَ يقول) [dire] ou (قَالَ يقيل) [faire la sieste]
- ظهور qui peut être analysé en ظُهُورٌ [des dos] ou ظُهُورٌ [le fait d'apparaître]

2.2.3. Ambiguïté et agglutination

2.2.3.1. Ambiguïté agglutinante

Nous appelons ambiguïté agglutinante toute ambiguïté se rapportant à des mots maximaux, c'est-à-dire à des mots pouvant être segmentés en mots minimaux et clitiques. En d'autres termes, l'ambiguïté agglutinante est celle résultant de l'agglutination d'un mot minimal et de clitiques. De cette agglutination naît un mot maximal homographe soit à un mot minimal soit à un autre mot maximal de composition différente.

Exemple :

أفعال → أ\فِعَالٌ
أفعال → أ\ف\عَالٍ

2.2.3.2. Ambiguïté segmentale

Nous appelons ambiguïté segmentale toute ambiguïté se rapportant à des mots minimaux qu'ils soient le résultat d'une segmentation ou qu'ils soient utilisés dépourvus de tous clitiques.

2.3. La désambiguïsation

La désambiguïsation peut se faire en deux étapes. Une première étape se déroulant pendant le prétraitement et consistant à filtrer et à réduire au maximum les **ambiguïtés virtuelles** par deux procédés complémentaires, mais disjoints, qui sont la voyellation et l'harmonisation vocalique et/ou consonantique, c'est ce que nous appelons **les filtres réducteurs d'ambiguïté**. Et une deuxième étape intervenant au moment du dépouillement lexicologique et servant à lever **les ambiguïtés effectives** par le biais de la segmentation et de la catégorisation, c'est ce que nous appelons **la levée d'ambiguïtés effective**.

2.3.1. Les filtres réducteurs d'ambiguïté

Il y a des filtres qui peuvent contribuer à réduire sinon à lever les ambiguïtés virtuelles. Deux de ces filtres sont d'ordre scriptural et obéissent à un certain nombre de règles que nous avons établies dans le cadre de la *norme lexicologique* proposée. Il est évident que c'est le contexte phrastique voire même extra-linguistique qui permet à ces filtres de jouer, en matière d'interprétation, un rôle réducteur d'ambiguïté.

2.3.1.1. Le filtrage des ambiguïtés par la voyellation

Dans le cas d'une voyellation partielle, bien que les voyelles, si elles sont introduites arbitrairement, puissent être, elles-mêmes, un facteur révélateur d'ambiguïtés, elles peuvent, en revanche, contribuer très efficacement à réduire les ambiguïtés si, toutefois, elles sont introduites d'une manière cohérente et codifiée.

Ainsi un *sukûn* « ْ » bien placé sur la lettre médiane des noms et *maÒdarS* de schèmes « *faЎl* », « *fiЎl* » ou « *fuЎl* » permet-il de dissiper toute ambiguïté virtuelle éventuelle avec des noms et *maÒdars* de schèmes « *faЎal* », « *faЎil* » ou « *faЎul* » (نَفْس ≠ نَفَس) ou avec des verbes de schèmes « *faЎila* », « *faЎula* » ou « *faЎala* » (عَلِمَ ≠ عَلَّمَ ؛ جَسُمَ ≠ جَسْم).

- Le cas de la *šadda* : Comme nous l'avons signalé plus haut[199], veiller à ce que la *šadda* soit systématiquement saisie (à l'exception de celle suivant l'article défini) c'est réduire de près de 40% le nombre des ambiguïtés virtuelles. De plus, sa saisie quasi-systématique peut même permettre, dans certains cas, de filtrer des ambiguïtés agglutinantes ; notamment dans le cas de la combinaison (*li* + *'al* + mot commençant par la lettre « *l* ») qui aurait la même graphie que la combinaison (*li* + mot commençant par la lettre « *l* ») si la *šadda* n'était pas saisie.

للبس	→	ل \ لبس	[à - pour une ambiguïté]
	→	ل \ ال \ لبس	[à - pour l'ambiguïté]

[199] Voir le chapitre 4 : *Norme de saisie et d'harmonisation*

En saisissant la *šadda*, le mot n'est plus ambigü puisqu'il n'accepte désormais qu'une seule analyse :

للّبس → ل \ ال \ لبس [à - pour l'ambiguïté]

2.3.1.2. Le filtrage des ambiguïtés par l'harmonisation

Si elle est bien appliquée selon une opération stable et codifiée, l'harmonisation peut contribuer sûrement et efficacement au filtrage des ambiguïtés.

Pour s'en convaincre, deux exemples tirés de la *norme d'harmonisation* suffisent, il s'agit de :

- L'harmonisation des noms propres (voir "*La norme de saisie et d'harmonisation*").
- L'harmonisation des dates (*idem*).

2.3.2. La levée d'ambiguïtés effective

2.3.2.1. La désambiguïsation par segmentation

Il est à noter que cette opération s'attaque exclusivement aux ambiguïtés agglutinantes.

فلك → فَلَكٌ [orbite]
فلك → ف \ ل \ كَ [alors, tu as]

مكروه → مَكْرُوهٌ [déplaisant]

→ مَكَّرُوا \ هُ [ils ont joué un mauvais tour]

→ شَابَهَ [ressembler à]

شَابَه → شَابَ \ هُ [alors, tu as]

Ces exemples illustrent bien comment la segmentation contribue à lever l'ambiguïté et, ce faisant, facilite la tâche à l'étape suivante qui est la lemmatisation.

2.3.2.2. La désambiguïsation par catégorisation

Elle intervient après la lemmatisation proprement dite. Coïncident avec la deuxième étape de la lemmatisation au sens large du terme dont la première étape est l'identification, cette opération de désambiguïsation s'attaque principalement aux **ambiguïtés polycatégorielles**.

Étant donné que la lemmatisation est une étape se déroulant après la segmentation, et que l'on a, par conséquent, affaire à des unités lexicales de base, à des mots minimaux, désambiguïser revient à résoudre le problème de l'homographie des vocables et de leurs lemmes de rattachement.

En effet, à ce stade, un vocable est reconnu homographe :

- lorsqu'il est rattaché à deux ou plusieurs lemmes graphiquement différents et de catégorie grammaticale différente ou identique :

  - Catégorie grammaticale différente :

تقوى → تقوى *Nom*

تقوى → قَوِيَ *Verbe*

❖ Catégorie grammaticale identique :

أقبل → أَقْبَلَ *Verbe*
أقبل → قَبِلَ *Verbe*

↳ lorsqu'il est rattaché à deux ou plusieurs lemmes graphiquement identiques mais de catégorie grammaticale différente :

خير → خَيْرٌ *Nom*
خير → خَيِّرٌ *Adjectif (élatif)*

أفعى → أفعى *Nom*
أفعى → أفعى *Verbe* (devenir méchant)

3. Lemmatisation

3.1. *Lemme*, *lemmatiser*, *lemmatisation*… : proposition de traduction

Jusqu'au jour d'aujourd'hui, il n'existe aucun équivalent en arabe des termes *lemme(s)*, *lemmatiser*, *lemmatisé*, *lemmatisation*, *lemmatiseur* ; ce qui ajoute un frein supplémentaire au développement et à la diffusion des études lexicométriques arabes.

On pourrait penser à traduire tout simplement *lemme* par le mot لام *lâm*, qui se trouve être d'un côté le nom de la lettre ل *l*, et de l'autre, un homographe du verbe لاَمَ يَلُـومُ \ مَ *lâma / yalûmu* [reprocher]. Cette proposition a le seul avantage d'établir une équivalence phonologique presque parfaite entre le mot-source et le mot-cible. Une autre proposition serait de traduire *lemme* par le mot ليـم *lîm* qui, lui aussi, est déjà utilisé en arabe et réservé à la fois aux mots *sosie* et *conciliation*. Ces deux propositions ne sont pas valables non seulement parce que les mots proposés sont déjà réservés, mais aussi et surtout parce qu'ils ne sont pas productifs car ils ne permettent pas la dérivation. Les termes dérivés tels que le pluriel *lemmes*, le verbe *lemmatiser*, l'adjectif *lemmatisé*, le substantif *lemmatisation* ou l'outil *lemmatiseur*, pourraient rester sans équivalents.

Nous nous sommes donc orienté vers un néologisme qui exprimerait, en arabe, la définition du terme *lemme* et qui serait capable de produire les dérivés souhaités. Étant donné que la définition de *lemme* tourne autour de la notion de forme canonique, de forme générique, d'adresse lexicale, donc d'une certaine *origine* ou *source* morpho-lexicale commune à toutes les formes fléchies qui en sont les actualisations dans le discours, nous avons pensé rendre cette idée centrale par la notion de أصل *ÞaÒl* [origine, source, racine]. Ainsi, trouve-t-on dans *Al-Ëalîl : MuÝjam muÒÔalaÎât an-naÎw al-Ýarabî* une définition du terme أصل qui s'apparente avec la signification du terme *lemme* :

« تسمية تعني الغالب أو ما ينبغي أن يكون الشيء عليه، أو الأسبقيّة في المرتبة (يقابله الفرع) »[200]

« *tasmiyat-un taÝnî l-Èâliba Þaw mâ yanbaÈî Þan yakûna š-šayÞu Ýalayhi, Þawi l-Þasbaqiyyata fi l-martabati (yuqâbiluhu l-farÝu)* »

[200] *Georges Mitrî ʿAbd al-Masîḥ et Hânî Georges Ṭâbrî, Al-Ëalîl : MuÝjam muÒÔalaÎât an-naÎw al-Ýarabî*, 1990, p. 80

« une dénomination qui désigne [l'état] le plus fréquent, [la forme] que devrait avoir une chose, ou la primauté (opposée à branche, rameau) »
(la traduction est de nous).

Suite à cette définition du terme *ÞaÒl*, les deux auteurs donnent la liste de tout ce qui a été considéré, dans les sciences du langage arabes, comme *ÞaÒl* (origine) et de tout ce qui a été considéré comme *farÝ* (branche) ; on y trouve par exemple : le singulier est le *ÞaÒl* du pluriel et du duel, le masculin est le *ÞaÒl* du féminin, l'indéfini est le *ÞaÒl* du défini, le verbe à l'accompli est le *ÞaÒl* de celui à l'inaccompli ou à l'impératif, etc. Dans cet esprit, revenir du *farÝ* au *ÞaÒl*, est passer de la forme fléchie au "lemme", c'est le principe-même de la lemmatisation.

Cependant, même si l'acception du mot *lemme* gravite autour de cette notion d'*origine* morpho-lexicale, les deux termes ne sont tout de même pas synonymes en français. D'autant plus que le mot أصل est polysémique en arabe, renfermant plusieurs significations dans la langue générale et ayant de nombreuses applications dans de multiples domaines de spécialité : linguistique, philosophie, jurisprudence, anthropologie, etc. Nous étions donc conduit à faire appel à un procédé néologique élaboré pour la première fois en 1966 par Salah Garmadi dans sa traduction du livre de Jean Cantineau, *Cours de phonétique arabe*, pour traduire le terme *phonème*. Ce procédé, repris tout de suite après par Abdessalem Mseddi (Mseddi, 1984 : 76) pour traduire les termes *syntagme*, *glossème*, *sémème*, *morphème*, *monème*, *morphonème* et *tonème*, est un procédé faisant intervenir à la fois la néologie de forme, en l'occurrence la dérivation et l'emprunt, et la néologie de sens. Il consiste à suffixer au mot (trilitère) représentant le noyau de la définition, un ميم *mîm* qui serait l'équivalent du suffixe "*me*" (syntag**me**) ou "*ème*" (morph**ème**) ; le terme ainsi obtenu aura le schème فَعْلَيم *faÝlam*. Ainsi, après la traduction de Garmadi de *phonème* par صَوْتَم *Òawtam*, Mseddi a-t-il pu traduire *syntagme* par مَنْظَـم *manÛam*, *glossème* par مَعْلَـم *maÝlam*, *sémème* par مَفْهَـم

mafham, ***morphème*** par صَيْغَم *ÒayÈam*, ***monème*** par لَفْظَم *lafÛam*, ***morphonème*** par صَرْفَم *Òarfam* et ***tonème*** par مَنْغَم *manÈam*. Nous nous sommes donc basé sur ce procédé néologique pour traduire *lemme* en formant, à partir du noyau أصل *ÞaÒl,* le terme **أَصْلَم** ***ÞaÒlam***, construit donc sur le schème فَعْلَم *faÝlam*.

Il n'est pas certain, en fait, qu'il y ait dans le terme-source *lemme*, le même suffixe que celui dans *syntagme* ou dans *morphème*, mais quoi qu'il en soit, la similitude, au niveau morpho-lexical de la langue de départ, entre tous ces termes appartenant au domaine linguistique, est telle qu'il serait regrettable de ne pas en tirer avantage pour proposer un terme-cible qui reproduirait cette similitude dans la langue d'arrivée et qui, au demeurant, traduit parfaitement la signification du terme-source. Par ailleurs, il existe en arabe classique, un homographe du mot أَصْيَلَم ; c'est un adjectif dérivé du verbe صَلَمَ qui signifie *couper à la racine*. L'adjectif أَصْلَم est utilisé dans le sens d'un « homme dont les oreilles ont été coupées ». Le fait que ce mot existe déjà ne pose absolument aucun problème, et ce au moins pour deux raisons : la première est que l'adjectif est un vieux mot très rarement utilisé dans la langue générale ou, comme terme, en métrique classique ; la seconde et la plus importante est que les deux mots, même s'ils sont homographes, n'ont de commun ni la racine (صلم pour l'adjectif et أصل pour le nom que nous proposons) ni le schème (أَفْعَل pour l'adjectif et فَعْلَم pour le nom proposé) ni, par conséquent, l'augment (le préfixe أ *Þa* pour l'adjectif et le suffixe م *m* pour le terme *lemme*).

La proposition que nous faisons de traduire *lemme* par أَصْيَلَم est également justifiée par le fait que ce terme permette aisément la dérivation. En effet, ce terme

forme naturellement son pluriel en أَصَالِم *ÞaÒâlim*, le verbe *lemmatiser* peut facilement être rendu par le verbe quadrilitère simple (de schème يُفَعْلِـلُ / فَعْلَـلَ) يُؤَصْـلِمُ / أَصْـلَمَ *ÞaÒlama / yuÞaÒlimu*, l'adjectif *lemmatisé* par le participe passif مُؤَصْـلَم *muÞaÒlam*, le substantif *lemmatisation* aura comme équivalent en arabe, le nom d'action (*maÒdar*) أَصْلَمَة *ÞaÒlama* ; et nous proposons en fin, de rendre les termes *lemmatiseur* (humain ou machine) par مُؤَصْلِم *muÞaÒlim* et *lemmatiseur automatique* par مُؤَصْلِم آلِي *muÞaÒlim Þâlî*.

Terme	Traduction proposée
lemme	أَصْلَم
lemmes	أَصَالِم
lemmatiser	يُؤَصْلِمُ / أَصْلَمَ
lemmatisé	مُؤَصْلَم
lemmatisation	أَصْلَمَة
lemmatiseur	مُؤَصْلِم
lemmatiseur automatique	مُؤَصْلِم آلِي

Tableau 8
Récapitulatif des traductions proposées

3.2. Critères de lemmatisation

À la différence du français où la lemmatisation des noms et des adjectifs se fait selon le nombre et/ou le genre, la lemmatisation en arabe est réalisée selon quatre critères : le genre, le nombre, le cas et/ou l'annexion :

- ***Selon le genre :*** en général, les adjectifs au féminin sont lemmatisés au masculin.

  Exemple : جميلة *jamîla* [belle] → جميل *jamîl* [beau].

↳ ***Selon le nombre :*** en général, les noms et les adjectifs au pluriel ou au duel sont lemmatisés sous leurs formes au singulier.

Exemple 1 : كِتابانِ *kitâbâni* [deux livres] → كِتاب *kitâb* [livre].

Exemple 2 : مَدارِس *madâris* [écoles] → مَدْرَسة *madrasa* [école]

↳ ***Selon le cas :*** à part quelques rares exceptions de mots figés au cas direct (accusatif) ou indirect (génitif), le cas par défaut, hors contexte phrastique, pour les noms et les adjectifs est le cas sujet (nominatif). Le lemme des noms et des adjectifs doit donc être lemmatisé au cas sujet, sans marquer les voyelles casuelles.

Exemple 1 : رجلاً *rajul-an* [homme (à l'accusatif)] → رجل *rajul* [homme (au nominatif)]

Exemple 2 : كِلَيْ *kilay* [tous les deux (à l'accusatif ou au génitif)] → كِلا *kilâ* [tous les deux (au nominatif)]

Exemple 3 : بَنِي أَسَد *Banî ÞAsad* [nom d'une tribu (à l'accusatif ou au génitif)] → بَنُو أَسَد *Banû ÞAsad* [nom d'une tribu (au nominatif)]

↳ ***Selon l'annexion :*** quand ils sont premiers termes d'une annexion, certains mots subissent des changements. Les cinq noms par exemple, se voient adjoindre un *wâw* final و, un *Þalif* final ا ou un *yâÞ* final ي, selon qu'ils soient au cas sujet, direct ou indirect. Les noms et les adjectifs au pluriel externe masculin (se terminant par ...ـونَ *ûna* ou par ...ـينَ *îna*) se voient inévitablement amputer le نَ *na* final quand ils sont premiers termes d'une annexion. Sont également dans ce cas de figure, les noms et les adjectifs au duel (se terminant par ...ـانِ *âni* ou par ...ـيْنِ *ayni*) avec la disparition du نِ *ni* final. Comme, en dehors de toute lexicalisation, l'état d'annexion constitue une composition occasionnelle, les termes de cette composition sont dissociés lors de la

segmentation et sont lemmatisés chacun de son côté. Le lemme sera donc basé sur la forme isolée et non sur la forme annexée.

Exemple 1 :

أبو *Þabû* [père de *(nominatif)*]
أبا *Þabâ* [père de *(accusatif)*] → أب *Þab* [père]
أبي *Þabî* [père de *(génitif)*]

Exemple 2 :

مُعَلِّمُو *muʕallimû* [les instituteurs de *(nominatif)*] → (مُعَلِّمُونَ *muʕallimûna*)
مُعَلِّمِي *muʕallimî* [les instituteurs de *(accus./gén.)*] → (مُعَلِّمِينَ *muʕallimîna*)
→ مُعَلِّم *muʕallim* [instituteur]

L'opération de lemmatisation est parfois réalisée en combinant plusieurs étapes faisant intervenir des critères différents (par exemple le nombre, le cas et l'annexion dans n'importe quel ordre), confondues dans la pratique, mais qui constituent, méthodologiquement, des étapes distinctes.

3.3. Considérations linguistiques pour une lemmatisation réussie

3.3.1. Les types de pluriel en arabe

Il ne s'agit pas ici de la distinction à faire entre pluriel externe et pluriel brisé ; distinction, du reste, importante aussi bien au niveau du programme informatique écrit dans un objectif de lemmatisation automatique, qu'au niveau de la structure de la base de données lexicale sous-jacente à ce programme. En effet, pour lemmatiser un mot reconnu (par le programme) comme pluriel externe, il suffit d'écrire une règle lui ôtant

le suffixe ـونَ ou ـينَ, pour le masculin, et le suffixe ـات (en le remplaçant dans la majorité des cas par un *tâÞ marbûÔa* ـة), pour le féminin. Une opération similaire n'est pas possible pour le pluriel brisé (ou interne) qui est formé au moyen d'une modification de la morphologie interne du singulier par ajout d'affixes, suppression de lettres et/ou modification de son vocalisme. Dans le cas du pluriel brisé, une opération de fléchage est nécessaire entre chaque mot singulier et son (ses) pluriel(s) brisé(s) pour pouvoir lemmatiser d'une manière automatique ou semi-automatique. Ce fléchage traduit, en fait, une structure relationnelle de la base de données nominale selon le critère de nombre. Par ailleurs, il est à noter que certains noms peuvent avoir à la fois un pluriel externe et un (des) pluriel(s) brisé(s). Notons également que les adjectifs ont, dans la grande majorité des cas, un pluriel externe.

Il est question, en revanche, de types au sein du pluriel brisé lui-même où l'on trouve par exemple, le pluriel de petit nombre جمع القلّة *jamÝ al-qilla*, le pluriel de grand nombre جمع الكثرة *jamÝ al-ka×ra*, etc.

3.3.1.1. Pluriel de petit nombre et pluriel de grand nombre

Le pluriel de petit nombre جمع القلّة *jamÝ al-qilla* est un pluriel désignant un nombre compris entre trois et dix ; il comporte quatre schèmes (أَفْعُل – أَفْعَال – أَفْعِلَةٌ – فِعْلَة). Le pluriel de grand nombre جمع الكثرة *jamÝ al-ka×ra* quant à lui, désigne un nombre compris entre trois et l'infini[201] ; on en compte plus d'une trentaine de schèmes dont au

[201] Certains grammairiens considèrent que le pluriel de grand nombre désigne un nombre compris entre onze et l'infini, et que c'est le pluriel extrême مُنتهى الجموع *muntahâ l-jumûÝ* qui désigne un nombre entre trois et l'infini.

moins une quinzaine sont des pluriels du type مُنتهى الجموع *muntahâ l-jumûY* [pluriel extrême]. Le pluriel extrême est un pluriel de grand nombre comportant un *Þalif* (appelé ألف التكسير *Þalif at-taksîr* [*Þalif* de brisure]) suivi de deux ou de trois lettres à condition, dans ce dernier cas, que celle de milieu soit un ي *î* [i long]. Nous présentons dans le tableau de la page suivante, les schèmes les plus attestés du pluriel brisé : ceux du pluriel de petit nombre et ceux du pluriel de grand nombre y compris ceux du pluriel extrême.

Un certain nombre de mots arabes ont à la fois un pluriel de petit nombre et un pluriel de grand nombre, auquel cas, lemmatiser ces mots c'est ramener à la fois les deux pluriels, s'ils sont rencontrés dans un corpus, au lemme commun qui est le singulier. Les deux pluriels par exemple, أَنْفُس *Þanfus* (pluriel de petit nombre) et نُفُوس *nufûs* (pluriel de grand nombre) seront ramenés à leur lemme commun : le singulier نَفْس *nafs* [âme].

Schème du Pluriel	Exemple	Singulier	Traduction
Pluriel de petit nombre جمع القلّة			
أَفْعُل	أَنْفُس	نَفْس	âme
أَفْعَال	أَجْنَاس	جِنْس	genre
أَفْعِلَة	أَطْعِمَة	طَعَام	nourriture
فِعْلَة	صِبْيَة	صَبِيّ	enfant
Pluriel de grand nombre جمع الكثرة			
فُعْل	حُمْر	أَحْمَر	rouge
فُعُل	كُتُب	كِتَاب	livre
فُعَل	غُرَف	غُرْفَة	chambre
فِعَل	قِطَع	قِطْعَة	pièce
فُعَلَة	قُضَاة (à l'origine قُضَيَة)	قَاضٍ	juge
فَعَلَة	كَتَبَة	كَاتِب	écrivain
فِعَلَة	قِرَدَة	قِرْد	singe
فَعْلَى	مَرْضَى	مَرِيض	malade
فُعَّل	رُكَّع	رَاكِع	prosterné
فُعَّال	كُتَّاب	كَاتِب	écrivain
فِعَال	جِبَال	جَبَل	montagne
فُعُول	قُلُوب	قَلْب	cœur
فِعْلَان	غِرْبَان	غُرَاب	corbeau
فُعْلَان	قُضْبَان	قَضِيب	bâton
فُعَلَاء	عُقَلَاء	عَاقِل	sensé
أَفْعِلَاء	أَقْرِبَاء	قَرِيب	proche
Pluriel extrême مُنتهى الجموع			
فَوَاعِل	قَوَاعِد	قَاعِدَة	base
فَوَاعِيل	قَوَارِير	قَارُورَة	bouteille
مَفَاعِل	مَسَاجِد	مَسْجِد	mosquée
مَفَاعِيل	مَكَاتِيب	مَكْتُوب	correspondance
فَعَالِل	دَرَاهِم	دِرْهَم	dirham
فَعَالِيل	دَنَانِير	دِينَار	dinar
أَفَاعِل	أَصَابِع	إِصْبَع	doigt
أَفَاعِيل	أَسَالِيب	أُسْلُوب	style
تَفَاعِل	تَجَارِب	تَجْرِبَة	expérience
تَفَاعِيل	تَقَاسِيم	تَقْسِيم	découpage
فَيَاعِل	صَيَارِف	صَيْرَف	changeur
فَيَاعِيل	دَيَاجِير	دَيْجُور	ténèbres
فَعَائِل	قَبَائِل	قَبِيلَة	tribu

فَعَالَى\فَعَالٍ	فَتَاوَى\فَتَاوٍ	فَتْوَى	*fatwa*
فَعَالِيُّ	كَرَاسِيُّ	كُرْسِيّ	chaise

Tableau 9
Les différents schèmes du pluriel brisé

3.3.1.2. Le "pluriel de pluriel"

Le pluriel peut lui-même avoir un pluriel en arabe. Ainsi, بَيْت *bayt* [maison] par exemple, peut-il avoir, en plus de son pluriel بُيُوت *buyût* [maisons], un pluriel de pluriel qui est بُيُوتَات *buyûtât* [maisons] ; ou كَلْب *kalb* [chien] → pluriel كِلاَب *kilâb* [chiens] → pluriel de pluriel أَكَالِب *Þakâlib* [chiens] ; ou encore ظُفْر *Ûufr* [ongle] → pluriel أَظَافِر *ÞaÛâfir* [ongles] → pluriel de pluriel أَظَافِير *ÞaÛâfîr* [ongles].

Le lemme du pluriel de pluriel n'est pas le pluriel mais le singulier. Ainsi, quand ils figurent dans un même corpus, le pluriel et son éventuel pluriel seront-ils ramenés, l'un et l'autre, au même lemme : le singulier.

3.3.1.3. Le "nom de pluriel" اسم الجمع

Le "nom de pluriel" est un pluriel dont le singulier n'a pas la même structure morpho-lexicale, autrement dit, le pluriel et le singulier n'ont en commun ni le radical, ni-même la racine. Seule une relation sémantique lie ce type de pluriel à son singulier. Le singulier par exemple, de جَيْش *jayš* [armée] est جُنْدِي *jundî* [soldat], celui de نِسَاء *nisâÞ* [femmes] est اِمْرَأَة *imraÞa* [femme], celui de خَيْل *Ìayl* [chevaux] est فَرَس *faras* [cheval ou jument], celui de شَـعْب *šaÝb* [peuple] est رَجُـل\اِمْـرَأَة *rajul/imraÞa*

[homme/femme], celui de إِبِـــل *Þibil* [camélidés] est جَمَـــل\نَاقَـــة *jamal/nâqa* [chameau/chamelle] ou celui de غَنَـم *Èanam* [espèce ovine] est شَـاة *šât* [mouton ou brebis]. Il est à noter que le "nom de pluriel" est un mot singulier lexicalement, mais pluriel sémantiquement. Étant donné qu'il est lexicalement singulier, il peut avoir un pluriel (شُعُوب *šuÝûb* [peuples]) et un duel (شَعْبَان *šaÝbâni* [deux peuples]).

Considération faite que ce pluriel et son singulier sont intégralement différents sur le plan morpho-lexical, le "nom de pluriel" ne doit pas être lemmatisé sous sa forme du singulier. Le lemme du singulier sera la forme elle-même ; et le lemme du "nom de pluriel" (et éventuellement de son pluriel) sera le "nom de pluriel" lui-même.

3.3.1.4. Le collectif **اسم الجنس الجمعي**

Le nom collectif est un nom qui désigne soit l'espèce d'un ensemble d'individus distinct, dont on peut dériver le nom d'unité (اسم الوحدة), il s'agit de ce que l'on appelle en arabe : اسـم الجنس الجمعي *Þism al-jins al-jamÝî*, comme le mot سَمَك *samak* [(du) poisson] dont le nom d'unité est سَمَـكَة *samaka* [un poisson] ; soit la masse individuée ne permettant pas d'obtenir le nom d'unité, il s'agit de ce que l'on appelle اسـم الجنـس الإفـرادي *Þism al-jins al-Þifrâdî*, comme le mot زَيْـت *zayt* [huile] qui n'a pas de singulatif. Ce dernier cas de collectif ne pose aucun problème au niveau de la lemmatisation. En revanche, *Þism al-jins al-jamÝî* peut exister, dans un corpus, en même temps que son éventuel singulatif, voire même du pluriel du singulatif (qui est différent du collectif). Il n'est pas commode d'avoir comme entrées, dans un lexique-index (lemmatisé), à la fois le collectif (سَمَك), le singulatif (سَمَكَة) et le pluriel du singulatif

(سَمَكَات). Étant donné qu'il est primaire par rapport au singulatif, c'est bien le collectif que nous avons choisi comme lemme, et du singulatif et de son pluriel. De ce fait, سَمَكَة et سَمَكَات par exemple, seront lemmatisés sous le lemme سَمَك .

3.3.1.5. Le "singulier-pluriel"

Il existe en arabe, un certain nombre de mots qui sont à la fois singulier et pluriel. Dans cette catégorie, on trouve par exemple, فُلْك *fulk* [navire(s), vaisseau(x) (c'est ce mot qui a donné, en français, la *felouque*)], عَدُوّ *Ýaduww* [ennemi(s)], ضَيْف *Ãayf* [hôte(s)], وَلَـد *walad* [enfant(s)], دِلاَص *dilâÒ* [poli(s) et luisant(s)], جُنُـب *junub* [personne(s) ayant fait un rêve mouillé ou ayant eu un rapport sexuel], etc. Certains de ces mots peuvent évidemment avoir d'autres formes de pluriel, comme ضَيْف qui peut avoir ضُيُوف comme pluriel, ou أَعْداء pluriel de عَدُوّ ou encore وَلَد ayant le pluriel أَوْلاَد.

Qu'ils aient ou non, dans un corpus, d'autres formes de pluriel (auquel cas, celles-ci seront ramenées au singulier), ces singuliers-pluriels seront lemmatisés sous cette-même forme qui, même si elle exprime un pluriel, est également la forme du singulier.

Cas particulier : le cas de تَبَع *tabaÝ*

Le cas de تَبَع *tabaÝ* est assez particulier puisque ce "singulier-pluriel" a un pluriel أَتْبَاع *Þatbâ Ý* qui, en plus de son singulier تَبَع *tabaÝ* , peut également avoir un autre singulier : تَابِع *tâbiÝ* . Nous avons décidé, pour ce cas, de lemmatiser le pluriel أَتْبَاع

sous la forme du singulier "pur" (تَابِع). Et pour harmoniser le lexique-index, nous avons considéré le "côté pluriel" du "singulier-pluriel" تَبَع et avons décidé de lemmatiser aussi تَبَع sous la forme du singulier "pur" (تَابِع).

3.3.1.6. Le pluriel sans singulier

Il existe en arabe, un petit nombre de noms au pluriel qui n'ont pas (ou plus) de singulier et qui ne sont attestés dans la langue que sous cette forme au pluriel. Parmi ces noms, l'on trouve تَعَاشِيب *taʕâšîb* [morceaux de pâturages herbeux clairsemés], تَعَاجِيب *taʕâjîb* [choses étonnantes], تَجَاوِيد *tajâwîd* [pluies abondantes], (تَبَاشِير (الفجر *tabâšîr (al-fajr)* [premières lueurs (du matin)], (شَمَـاطِيط (ثـوب *(×awb) šamâÔîÔ* [(vêtement) déchiqueté], (عَبَابِيـد\عَبَادِيـد (خيـل *(Ìayl) ʕabâbîd / ʕabâdîd* [chevauchée répartie sur les quatre coins du monde] et أَبَابِيل *Þabâbîl* [volées/troupes d'oiseaux].

Étant donné que le singulier n'existe pas, il est évident que le lemme de chacun de ces noms soit la forme elle-même, c'est-à-dire le pluriel.

3.3.2. Les verbes exclusivement à la voix passive

Une dizaine de verbe en arabe, ne sont attestés (dans le sens indiqué dans le tableau suivant) qu'au passif, à l'image du verbe يُحْتَضَرُ *yuÎtaÃaru* [agoniser], de أُغْمِيَ

عليه)) *ÞuÈmiya (Ýalayhi)* [s'évanouir] ou de شُـدِهَ *šudiha* [être stupéfait]. Il est donc inévitable que le lemme soit celui correspondant à la voix passive.

Verbe au passif	Signification
يُحْتَضَرُ	agoniser
شُدِهَ	être stupéfait
عُنِيَ (بالشيء)	être préoccupé de quelque chose
زُهِيَ (على)	regarder avec dédain
فُلِجَ	être frappé de paralysie
حُمَّ	avoir la fièvre
سُلَّ	être atteint de tuberculose
جُنَّ	être/devenir fou
غُمَّ (الهلال)	être masqué par un nuage (lune)
أُغْمِيَ (عليه)	s'évanouir
انْتُقِعَ (لونه)	changer de couleur

Tableau 10
Liste des verbes qui sont toujours au passif

3.3.3. Les adjectifs exclusivement au féminin

Il existe un certain nombre d'adjectifs qui sont exclusivement réservés au sexe féminin. Une partie de ces adjectifs sont construits sur des schèmes du féminin comme عَـذْرَاء *ÝaÆrâÞ* [vierge] et نُفَسَـاء *nufasâÞ* [femme qui vient d'accoucher], d'autres, sémantiquement féminins, sont construits sur des schèmes du masculin comme حَامِـل *Îâmil* [(femme) enceinte] et حَـائِض *ÎâÞiÃ* [femme ayant ses menstrues]. Le lemme sera donc la forme au singulier (l'adjectif féminin singulier).

adjectif féminin	Signification
Adjectifs féminins morphologiquement et sémantiquement	

عَذْرَاء	femme vierge
نُفَسَاء	femme qui vient d'accoucher
Adjectifs sémantiquement féminins mais morphologiquement masculins	
حَائِض	femme ayant ses menstrues
حَامِل	femme enceinte
مُرْضِع	femme qui donne à téter
بَغِيّ	prostituée

Tableau 11
Liste des adjectifs exclusivement au féminin

3.3.4. Les adjectifs mixtes

Il y a très peu d'adjectifs qui sont à la fois masculins et féminins, à savoir les deux adjectifs : جُنُب *junub* [personne(s) ayant fait un rêve mouillé ou ayant eu un rapport sexuel] et عَقِيم *Ýaqîm* [stérile]. Il est à noter que ces deux adjectifs qui sont mixtes en genre, sont également invariables en nombre, c'est-à-dire que la même forme est commune au singulier, au duel et au pluriel. Le lemme est naturellement la forme elle-même.

3.4. Les lemmes adoptés

En nous basant sur les considérations linguistiques citées ci-avant et sur d'autres considérations d'ordre théorique, méthodologique ou pratique, nous présentons dans ce qui suit les différents lemmes que nous avons adoptés et qui ont servi d'adresses lexicales pour les différentes formes fléchies de notre corpus.

- Pour les **verbes** : le lemme est le verbe conjugué à la 3ème personne du singulier masculin de l'accompli actif, exception faite d'une dizaine de verbes

qui ne peuvent être qu'à la voix passive comme par exemple, يُحْتَضَرُ *yuÎtaÃaru* [agoniser].

↳ Pour les **noms** : le lemme est le singulier, exception faite des pluriels qui n'ont pas de singulier, comme أَبَابِيل *Þabâbîl* [volées/troupes d'oiseaux], ou de ceux dont le singulier n'a pas la même racine comme نِسَاء *nisâÞ* [femmes] qui est le pluriel de اِمْرَأَة *imraÞat*. Il est à noter que même le pluriel de pluriel est ramené au singulier ; sous le lemme par exemple, بَيْت *bayt* [maison] seront rassemblés le pluriel بُيُوت *buyût* et le pluriel de pluriel بُيُوتَات *buyûtât*.

↳ En ce qui concerne les noms singuliers qui ont deux pluriels, un pluriel de petit nombre (exemple : أَنْفُس *Þanfus*) et un pluriel de grand nombre (exemple : نُفُوس *nufûs*), les deux sont ramenés au seul lemme : le singulier (exemple : نَفْس *nafs* [âme]).

↳ Le lemme des **noms "incomplets"** **الأسماء المنقوصة** sera la forme indéfinie de ces noms, saisie avec *tanwîn al-kasra*. Le lemme, par exemple du nom "incomplet" محامي (ainsi que de son pluriel محامون – محامين , et de son duel محاميْن – محاميان) est la forme محامٍ .

↳ Pour les **adjectifs** : le lemme est le singulier masculin. Il faudra se garder d'un petit nombre d'adjectifs masculins se terminant par un تاء مربوطة *tâÞ marbûÔa* [t fermé final] (qui est le plus souvent la marque du féminin) comme فَرُوقَة *farûqa* [peureux], et de certains intensifs, beaucoup plus nombreux, qui sont construits sur le schème فَعَّالَة *faÝÝâla* comme par exemple, رَحَّالَة *raÎÎâla* [très grand voyageur], ou عَلاَّمَة *Ýallâma* [très grand savant] qui sont, en fait, des adjectifs masculins.

↳ En ce qui concerne les adjectifs qui n'ont pas de masculin, sémantiquement, comme حَامِل *Îâmil* [(femme) enceinte] et حَائِض *ÎâÞiÃ* [femme ayant ses menstrues] ou à la fois sémantiquement et morphologiquement comme عَذْرَاء *ÝaÆrâÞ* [vierge] et نُفَسَاء *nufasâÞ* [femme qui vient d'accoucher], le lemme est, bien entendu, la forme elle-même au singulier.

↳ Pour les **noms d'unité** : le lemme est le collectif اسم الجنس الجمعي *ism al-jins al-jamŶiyy* à partir duquel a été formé le nom d'unité (singulatif) اسم الوحدة *ism al-waÎdat* , comme par exemple, pour سَمَكَة *samaka* [un poisson] le lemme est سَمَك *samak* [(du) poisson]. Le lemme du pluriel (du singulatif) سَمَكَات *samakât* [des poissons] est également le collectif.

↳ Pour les **cinq noms** : le lemme est la forme tronquée (أَب et non أَبَا , أَبُو ni أَبِي), au cas sujet (ذُو et non ذَا ni ذِي)

↳ Pour les **noms propres** : le lemme est le nom propre au cas sujet (أبو الوفاء et non أبا الوفاء ni أبي الوفاء), (اِمْرُؤ القَيْس et non اِمْرِئ القَيْس ...)

↳ Pour certains **duelatifs** : le lemme est la même forme qui est au duel et qui ne doit surtout pas être ramenée au singulier. Le duelatif par exemple أَبَوَانِ *Þabawâni* [les (deux) parents], littéralement en arabe « *les deux pères* » n'est pas l'addition d'un père et d'un autre père, mais bien d'un père et d'une mère.

↳ Pour les **Mots-outils** : généralement le lemme est la forme elle-même sauf pour quelques mots-outils à savoir :

- ❖ Les **cardinaux** : le lemme est le nombre cardinal qui correspond à un nombré masculin, au cas sujet et isolé (c'est-à-dire non annexé, exemple أَلْفَانِ et non أَلْفَا).

- ❖ Les **ordinaux** : le lemme est le nombre ordinal masculin (étant donné qu'ils se comportent comme des adjectifs). Exception faite des dizaines, des centaines et des milliers où c'est un cardinal qui est utilisé pour exprimer un adjectif ordinal, exemple الليلة الأربعون–المسألة المئة – الرجل الألف.

- ❖ Les **verbes figés** : comme ils sont figés soit à l'accompli, soit à l'impératif, soit à l'inaccompli (un seul verbe), le lemme sera le verbe lui-même pour les premiers, le verbe à l'impératif pour les deuxièmes et, enfin, la même forme pour le dernier.

- ❖ Les **annectifs** كِلاَ et كِلْتَا : ces deux annectifs ont la particularité d'être toujours au cas sujet quand ils sont premier terme d'une annexion, mais deviennent déclinables quand des pronoms

personnels leur sont suffixés. Auquel cas, c'est toujours la forme au cas sujet qui est le lemme (كِلاَ et كِلْتَا et non كِلَيْ et كِلْتَيْ)

↳ Nous avons choisi de lemmatiser des lettres de l'alphabet qui figurent isolées dans le corpus (dans un cadre de démonstration mathématique), sous leurs noms respectifs. Les lettres par exemple, ج et ب ont été lemmatisées sous جِيم et بَاء. Ce choix a été fait pour éviter d'éventuelles ambiguïtés entre, par exemple, la lettre ب et la préposition ب, ou entre la lettre أ et l'interrogatif أ.

↳ Le pluriel des noms d'animaux ou d'inanimés qui sont formés annectivement avec, comme premier terme, ابـن *Þibn* [fils de] diffère d'un singulier à un autre. En effet, ces noms peuvent avoir, comme premier terme, au pluriel, أبناء *ÞabnâÞ* [fils de (pl.)] ou بنات *banât* [filles de] ; ainsi, le pluriel de ابن آوى *Þibn Þâwâ* [chacal] est-il أبناء آوى *ÞabnâÞ Þâwâ* [chacals] alors que le pluriel de ا بن عِرس *Þibn Ýirs* [belette] est بنات عِرس *banât Ýirs* [belettes]. En outre, le pluriel de بنت وَرْدان *bint wardân* [cafard] est بنات وَرْدان *banât wardân* ; et celui de ذو القِعْدة *Æû l-qiÝda* [mois de l'année lunaire] est ذَوات القِعْدة *Æawât l-qiÝda*. Ce qui signifie qu'il n'est pas systématique, lors de la lemmatisation, que l'opérateur (humain ou machine) lemmatise un pluriel dont le premier terme est أبناء, sous un lemme dont le premier terme est ابن, ni un pluriel commençant par بنات, sous un lemme commençant par بنت, ni un pluriel dont le premier terme est ذَوات, sous un lemme dont le premier terme est ذَات *Æât*. Chaque pluriel doit donc être lemmatisé sous le singulier dont il dépend effectivement selon les relations singulier-pluriel établies clairement. Nous donnons les différents cas de figures de ces irrégularités dans le tableau suivant :

Pluriel	Lemme	Signification
أبناء آوى	ابن آوى	chacal
بنات عِرس	ابن عِرس	belette
بنات وَرْدان	بنت وَرْدان	cafard, blatte
ذَوات القِعْدة	ذو القِعْدة	11ème mois de l'année lunaire
أمهاتّ أربع وأربَعين	أمّ أربع وأربَعين	mille-pattes

Tableau 12
Cas d'irrégularités des pluriels au niveau du premier terme de l'annexion

↳ Quand un nom au pluriel a deux singuliers différents, l'un masculin, l'autre féminin, à l'instar de بَواعِث *bawâʕi×* [causes, attraits] qui peut avoir, comme singulier, باعِث *bâʕi×* et باعِثة *bâʕi×a* ; ou مَعايِب *maʕâyib* [défauts] dont le singulier est, soit مَعاب *maʕâb*, soit مَعابة *maʕâba* ; nous avons décidé de retenir le singulier masculin comme lemme du pluriel en question.

↳ Le cas de عُلَماء : le singulier de عُلَماء *ʕulamâʔ* [savants] est, normalement, عَليم *ʕalîm*, mais ce pluriel est fortement usité, dans notre corpus mais également en arabe moderne, comme étant le pluriel de عالِم *ʕâlim*. Compte tenu de cette quasi "lexicalisation" et pour cadrer avec le contexte discursif, nous avons pris la décision de lemmatiser عُلَماء en عالِم .

4. catégorisation

« *Procéder à l'indexation d'un texte*, nous dit Charles Bernet, *c'est opter pour une certaine définition du lexique et de ses éléments* »[202], mais cette relation entre les éléments du texte, et ceux du lexique n'est pas une relation directe dans ce sens où l'objet de la lexicométrie n'est pas le lexique mais le(s) vocabulaire(s). Néanmoins, à une certaine étape du cheminement de l'analyse lexicométrique, l'étape de catégorisation, l'on est amené à faire correspondre aux unités du vocabulaire, dotées d'une dimension réelle, des unités virtuelles du lexique. Ce lexique qui est considéré, selon l'approche adoptée, comme :

1. la somme, théorique, de tous les vocabulaires,
2. l'ensemble des unités lexicales, c'est-à-dire des mots lexicaux et des mots-outils,
3. l'ensemble des morphèmes d'une langue,
4. l'ensemble des mots lexicaux et locutions.

Seulement, la statistique lexicale depuis Pierre Guiraud et Charles Muller, a toujours favorisé la deuxième définition du lexique.

4.1. Principes de catégorisation

Deux démarches différentes parfois même contradictoires sur certains points, s'opposent quant à la classification des unités lexicales de la langue arabe : la Tradition grammaticale arabe aspirant à l'exhaustivité et donnant la prééminence du sens sur la forme, et la linguistique moderne soucieuse de systématisation et insistant sur l'analyse

[202] Charles Bernet, *Le vocabulaire des tragédies de Jean Racine. Analyse statistique*, 1983, p.17

de la forme qui est « une démarche contemporaine, issue du positivisme logique et de la pensée structurale »[203].

Mis devant les problèmes d'ordre méthodologique qu'il rencontre très souvent dans son domaine et qui sont, entre autres, la définition des unités, le repérage des occurrences relevant d'une même unité, la levée d'ambiguïtés polycatégorielles, etc., le chercheur en lexicométrie se trouve confronté à une double obsession : d'un côté, faire des inventaires les plus exhaustifs possible des unités du lexique pour qu'elles puissent correspondre aux unités du (des) vocabulaire(s) étudié(s) et de l'autre côté, tenir compte des exigences du TAL et de ses formalismes avec toute la rigueur que cela implique notamment au niveau des critères de catégorisation pour obtenir en fin de compte que chaque unité lexicale puisse être affectée à une seule et unique catégorie lexicale.

4.1.1. Mots lexicaux vs mots-outils

Notons d'abord que c'est le poète Zhang Yan, au XIII^{ème} siècle, qui « avait inventé les notions de *mot plein* (*Shi Zi*) et de *mot vide* (*Xu Zi*) – pour opérateur grammatical – qui ont fait dans nos grammaires contemporaines une si belle carrière »[204]. Même si, à la suite de B. Pottier, « nous nous refusons à croire que la langue puisse posséder des mots vides »[205], qu'on les appelle *mots vides*, *mots de relation*, *mots grammaticaux* ou *mots-outils*, ces unités sont en quelque sorte comparables à du ciment qui lie entre eux les *mots pleins*, *mots sémantiques*, *mots lexicaux*. Ce sont des mots « dont le rôle est plus syntaxique que sémantique et qui servent davantage à établir une relation qu'à définir une substance »[206].

---

[203] Joseph Dichy, *L'écriture...*, *op. cit.*, 1990, p. 494

[204] Henri Meschonnic, *Des mots et des mondes. Dictionnaire, encyclopédies, grammaires, nomenclatures*, 1991, p. 20

[205] Bernard Pottier, *Systématique des éléments de relation*, p. 95

[206] Etienne Brunet, *Le vocabulaire français de 1789 à nos jours*, p. 361

Les mots-outils servent donc à faire fonctionner le discours, à construire des phrases et à les organiser en un texte cohérent. Même s'ils contribuent certainement à la signification générale du contexte discursif dont ils font partie, les mots-outils n'ont cependant pas de signification en dehors de ce contexte. Contrairement aux mots lexicaux, ils ne font pas penser à une image mentale correspondant à une part du réel. Le contenu de ces unités à fonction syntaxique, ne peut être appréhendé conceptuellement hors contexte. Ils sont « sémantiquement déterminés par leur fonction, qui reflète la structure immanente de la langue. Leur sens est entièrement investi par le code syntactique ou par la situation de communication, comme on le voit pour les pronoms personnels, pour les adverbes de lieu et de temps »[207].

Cette dépendance des mots-outils vis-à-vis de la structure syntaxique et du contexte discursif, les mots lexicaux eux, ne la subissent pas. Au contraire, « les signifiés des mots "lexicaux" sont relativement indépendants du système abstrait de la langue et sont sémantiquement analysables hors contexte, d'où la possibilité de périphrase synonymique : la définition du dictionnaire »[208]. Ces mots lexicaux sont doués d'une valeur dénominative leur permettant de référer aux choses du monde.

Au niveau de la répartition des unités lexicales entre ces deux ensembles de classes, la langue arabe ne fait pas exception et ses unités lexicales se prêtent également à cette opposition entre *mots lexicaux* ayant une signification en soi et *mots-outils* servant de liant au discours et recevant leur signification de cette fonction syntaxique qu'ils sont censés remplir.

---

[207] Alain Rey, *Le lexique : Images et modèles. Du dictionnaire à la lexicologie*, 1977, p. 165

[208] A. Rey, *idem*, p. 165

4.1.2. Qu'appelle-t-on mot-outil en arabe ?

Souvent, dans un contexte linguistique arabe, le terme أداة *Þadât* (pl. أدوات *Þadawât*) [outil] est utilisé comme simplification de أداة نحويّـة *Þadât naÎwiyya* [outil grammatical]. Contrairement à ce qu'on trouve dans certains écrits linguistiques, commentaires de textes de grammaire ou traductions de citations de livres de grammaire arabe, le terme d'*outil grammatical*, ou de *mot-outil*, n'est en aucun cas synonyme de *particule*, chère à la Tradition grammaticale arabe. Le terme *particule* est proposé par la tradition orientaliste pour traduire le terme حرف *Îarf* [lettre, mot[209], litt. : extrémité]. Dans la division tripartite opérée par la Tradition grammaticale, la *particule* est définie par opposition au nom et au verbe ; tout ce qui n'est ni nom ni verbe est *particule*. Le *mot-outil*, lui, est défini par rapport à la relation syntaxique dont il acquiert sa signification ; il est dépendant du contexte. Le critère qui permet de définir le *mot-outil*, et donc de le distinguer du *mot lexical*, est cette opposition : fonction syntaxique *vs* valeur dénominative. Le *mot-outil* n'est donc pas la *particule*. Le premier englobe la deuxième. Toute *particule* est *mot-outil* mais tout *mot-outil* n'est pas *particule*. Cette vision du *mot-outil*, souvent implicitement adoptée ou appliquée, n'a été que rarement précisée ou énoncée d'une manière explicite. Il est frappant de voir que, sur le plan de la catégorisation classique, les unités qui sont dépourvues de toute valeur dénominative et vouées à des relations syntaxiques, et qui auraient dû être regroupées dans un seul et même ensemble, celui des mots-outils, ont été éparpillées entre les trois classes, la classe des verbes (pour les verbes figés), celle des noms (pour les noms-outils) et celle des particules. Les deux figures suivantes montrent la classe des verbes dans laquelle on trouve les verbes figés (ou fonctionnalisés), et celle des noms où l'on trouve un ensemble de mots-outils (relatifs, démonstratifs, interrogatifs, ...) formant ce que l'on appelle les noms-outils.

---

[209] Voir ces notions dans (Dichy, 1990).

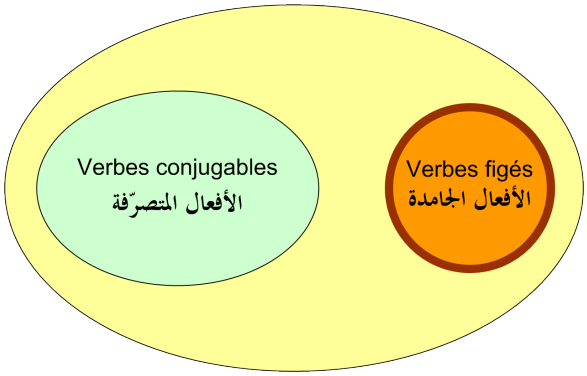

Figure 42
Catégorisation classique des Verbes contenant les Verbes figés

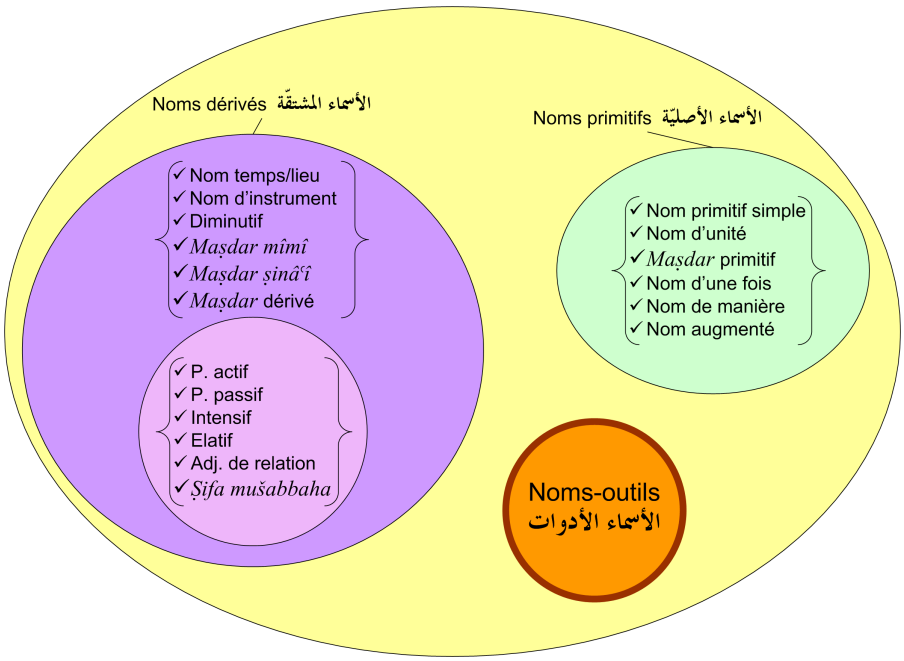

Figure 43
Catégorisation classique des Noms contenant les Noms-outils

Nous énumérons dans ce qui suit, quelques définitions, usages ou indications qui corroborent l'idée que, malgré la confusion qui a accompagné les notions de *mot-outil* et de *particule*, les deux termes ne sont aucunement synonymes et que la *particule* n'est qu'un élément parmi d'autres du *mot-outil* :

↳ Dans *MiftâÎ as-saÝâda wa-miÒbâÎ as-siyâda* مفتاح السعادة ومصباح السيادة , *Óâš Kubrâ Zâda* (1495-1561) définit les mots-outils comme suit :

» والمراد بالأدوات : الحروف ، وما شابهها من الأسماء والأفعال والظروف «[210]

« *Wa-l-murâdu bi-l-Þadawâti : al-Îurûfu, wa-mâ šâbahahâ mina l-ÞasmâÞi wa-l-ÞafÝâli wa-Û-Ûurûfi* »

« Et par mots-*outils*, il faut entendre : les particules et leurs semblables parmi les noms, les verbes et les adverbes » (la traduction est de nous).

↳ De la même manière, *Georges Mitrî ÝAbd al-Masîÿ* et *Hânî Georges Tâbrî*, dans leur *Lexique des termes de la grammaire arabe*, *Al-Ëalîl : MuÝjam muÒÔalaÎât an-naÎw al-Ýarabî* الخليل : معجم مصطلحات النحو العربي , définissent le mot-outil comme suit :

» الأداة : الحرف ، وما تضمّن معناها من الأسماء والأفعال والظروف ، نحو : مِنْ – سِوَى – حاشا – أمسِ «[211]

« Al-Þadâtu : al-Îarfu, wa-mâ taÃammana maÝnâhâ mina l-ÞasmâÞi wa-l-ÞafÝâli wa-Û-Ûurûfi, naÎwa : min - siwâ - Îâšâ - Þamsi »

« Mot-outil : c'est la particule et tout ce qui en a la même signification [syntaxique] parmi les noms, les verbes et les adverbes, comme *min* [de, à partir de] - *siwâ* [autre que] - *Îâšâ* [sauf, en dehors de] - *Þamsi* [hier] » (la traduction est de nous).

Par la suite, les auteurs de ce lexique présentent plusieurs sous-groupes de mots-outils en précisant à chaque fois, la nature des éléments qui les composent. Ils présentent, par exemple, les mots-outils d'exception أدوات الاستثناء *Þadawât al-Þisti×nâÞ* comme étant composés de *particules* حروف *Îurûf* comme إلاّ *Þillâ* (...), de *noms* أسماء *ÞasmâÞ* comme غيْر – سِوَى *Èayr - siwâ* (...), de *verbes* أفعال *ÞafÝâl*

[210] *Óâš Kubrâ Zâda, MiftâÎ as-saÝâda wa-miÒbâÎ as-siyâda fi mawÃûÝât al-Ýulûm*, p. 379

[211] Georges Mitrî *ÝAbd al-Masîÿ* et Hânî Georges Tâbrî, *op. cit.*, p. 38

comme ليس – لا يكون *laysa - lâ yakûnu* (...), de *verbes*/*particules* comme عدا – خلا – حاشا *Ýadâ - Ìalâ - Îâšâ* (...) et en fin, de لا سيّما *lâ siyyamâ* et ses sœurs لا مثل ما *lâ mi×la mâ* (...).

- Dans le livre d'*Az-Zajjâjî*[212] (m. 340/951), كتاب حروف المعاني *Kitâb Îurûf al-maÝânî*, sur les 137 mots-outils recensés (ou 146 si l'on sépare les quelques entrées groupées), près de la moitié, 62 (ou 68) mots-outils, ne sont point des *particules* au sens de la Tradition grammaticale arabe. Nous y repérons des verbes (ليس , أصبح , أضحى , أمسى , بات , ظلّ , تعال , هات ...), des adverbes (الآن , بعد , بين , ثَمَّ , تحت , حَنَانَيْك , حيّهل , رويداً , صددك , لعمرك , ويْكأنّ , صه , ...), des interjections (أمام , حيث ...), des démonstratifs (أولى), des circonlocutions (كَأَيّن), des noms (غَيْرٍ , كُلّ , بعض.), voire même des expressions figées (سبحان الله , غفرانك لا كفرانك , لبيك وسعديك , معاذ الله , ...).
- Parmi les 105 mots-outils qu'*Ibn Hišâm al-PAnÒârî* (1309-1360) recense et analyse dans son livre, مغني اللبيب عن كتب الأعاريب *MuÈnî al-labîb Ýan kutub al-ÞaÝârîb*[213], consacré en grande partie au sujet des mots-outils, une trentaine ne sont pas des particules au sens de la Tradition grammaticale arabe. Parmi eux, on trouve des adverbes (عَلُ , عَوْضُ ...), des pronoms personnels (هُوَ ...), des verbes figés (لَيْسَ , عَسَـــى ...), des circonlocutions (كَـــذَا , كَـــأَيٍّ ...), des interjections (بَلْهَ , بَجَلْ ...) ou d'autres noms comme كِلاَ , كُلّ , غَيْر ...

---

[212] Az-Zajjâjî, *Kitâb Îurûf al-maÝânî*, 1986

[213] *PIbn Hišâm Al-PAnÒârî, MuÈnî al-labîb Ýan kutub al-ÞaÝârîb*, 1997

- *Muhammad At-Tûnjî* inclut dans les 156 mots-outils que comporte son dictionnaire des mots-outils[214] معجـم الأدوات النحويّـة *MuÝjam al-Þadawât an-naÎwiyya* [litt. : Dictionnaire des outils grammaticaux], une soixantaine de mots-outils autres que particules. Ce sont des relatifs (الأُلَى), des adverbes (عَلٍ , عَوْضُ , الآنَ ...), des pronoms personnels (إيّـاك , هُـوَ...), des verbes figés (لَيْـسَ ...), des circonlocutions (كَـذَا , كَـأَيِّ ...), des interjections (بَلْـهَ , هَيَّـا , هَيْـهَـاتَ , مَـهْ ...) ou d'autres noms comme كِلاَ , كُلّ , غَيْر...

- *Dans son livre* الجنى الداني في حروف المعاني *Al-Janâ ad-dânî fî Îurûf al-maÝânî*[215], *Al-Ýasan Þbn Al-Qâsim Al-Murâdî* (m. 749/1348) recense quant à lui, 105 mots-outils ; parmi eux nous repérons une trentaine qui ne sont pas des particules au sens de la Tradition grammaticale arabe. Nous y trouvons tous les pronoms personnels isolés (أَنَا ... هُنَّ), des interrogatifs (مَنْ , كَمْ...), des adverbes (مُذْ , إذ , إذَا ...), des démonstratifs (ذا), des verbes figés (لَيْسَ , عَسَى ...), des interjections (بَلْهَ , بَجَلْ ...), etc.

- Comme beaucoup d'autres, Donnat Vernier tombe lui aussi, dans la confusion entre *particules* et *mots-outils*. Il déclare que « les particules en arabe se divisent, comme dans les autres langues, en adverbes, prépositions, conjonctions et interjections »[216].

- À plusieurs endroits de son livre *JâmiÝ ad-durûs al-Ýarabiyya* où il traite des mots-outils, *MuÒÔafâ Al-Ðalâyînî* fait systématiquement la distinction entre ceux qui

---

[214] *Muhammad at-Tûnjî, MuÝjam al-Þadawât an-naÎwiyya*, 1995

[215] *Al-Ýasan Þbn al-Qâsim al-Murâdî, Al-Janâ ad-dânî fî Îurûf al-maÝânî*, 1992

[216] Donat Vernier S. J., *Grammaire arabe, composée d'après les sources primitives*, 1892, t.1, p. 393

sont des particules et ceux qui sont des noms, des adverbes ou des verbes, comme en témoignent les citations suivantes :

« حرفا الاستفهام هما : (هل والهمزة) وبقيّة أدوات الاستفهام أسماء »[217]

« Îarfâ l-istifhâm humâ : (hal wa-l-hamza) wa-baqiyyatu Þadawâti l-istifhâmi asmûÞ-un »

« Les deux particules interrogatives sont *hal* et la *hamza*. Le reste des mots-outils interrogatifs sont des noms » (la traduction est de nous) ;

Ou :

« أدوات الشرط : منها ما هو حرف، وهما: "إنْ وإذْ ما" [...]. ومنها ما هو اسم [...] وهي "مَنْ وما ومهما وأيٌّ وكيفما" ومنها ما هو ظرفُ زمانٍ [...] وهي "أَينَ وأَنَّى وأَيَّانَ ومتى وإذ". ومنها ما هو ظرفُ مكانٍ [...] وهي "حيثما" »[218]

« ÞAdawâtu š-šarÔi : minhâ mâ huwa Îarf-un, wa-humâ "Þin wa-ÞiÆ mâ" […]. Wa minhâ mâ huwa Þism-un […], wa-hiya "man wa-mâ wa-mahmâ wa-Þayyu wa-kayfamâ" wa-minhâ mâ huwa Ûarfu zamân-in […] wa-hiya "Þayna wa-Þannâ wa-Þayyâna wa-matâ wa-ÞiÆ". wa-minhâ mâ huwa Ûarfu makân-in[…] wa-hiya "Îay×umâ" »

« Les mots-outils de condition sont : soit une particule, "*Þin* et *ÞiÆ mâ*" [...] ; soit un nom [...], "*man*, *mâ*, *mahmâ*, *Þayyu* et *kayfamâ*" ; soit un adverbe de temps [...], "*Þayna*, *Þannâ*, *Þayyâna*, *matâ* et *ÞiÆ*" ; soit un adverbe de lieu [...], "*Îay×umâ*" » (la traduction est de nous) ;

Ou encore :

« وبقيّة الأدوات التي تجزم فعلَيْن أسماءٌ لا حروف، كمَن وما ومهما ومتى وأخواتها »[219]

« Wa-baqiyyatu l-Þadawâti l-latî tajzimu fiÝlayni ÞasmâÞ-un lâ Îurûf-un, ka-man wa-mâ wa-mahmâ wa-matâ wa-ÞaÌawâtihâ »

« Et le reste des mots-outils qui régissent l'apocopé pour deux verbes successifs, sont des noms[-outils] et non des particules comme *man*, *mâ*, *mahmâ*, *matâ* et leurs sœurs » (la traduction est de nous).

---

[217] *MuÒÔafà al-Ṣalâyînî, Jâmiʿ ad-durûs al-ʿarabiyya*, t. 1, p.12, note 4.

[218] *Ibid*, *idem*, t. 2, p.203.

[219] *Ibid*, *idem*, t. 3, p.253, note 1.

- Même s'il n'échappe pas, lui non plus, à cet amalgame entre *particule* et *mot-outil*, Henri Fleish se pose tout de même la question des frontières de la notion de "particule" *Îarf* et ce du fait de la définition fourre-tout qu'on a voulu donner à ce terme : « On traduit habituellement par "particule" *Îarf* ainsi défini. Mais jusqu'où fallait-il étendre l'extension de ce *Îarf* ? Beaucoup d'instruments grammaticaux (ces *adawât* [sing. *adât*], terme d'*al-Farrâþ*, [...]) se présentaient d'eux-mêmes. Des cas étaient moins clairs [...] »[220]
- En complément à la célèbre concordance du Coran *Al-MuÝjam al-mufahras li-þalfâÛ al-QurÞân al-karîm* المعجم المفهرس لألفاظ القرآن الكريم faite par *Mohammad FuPâd PAbd Al-Bâqi*[221] et dans laquelle, à part quelques rares mots-outils, il n'est principalement question que des noms et des verbes contenus dans le Coran, *PIsmâÝîl PAÎmad ÝAmâyra* et *ÝAbd al-Ĵamîd* MuÒtafâ as-Sayyid ont réalisé la concordance des mots-outils et des pronoms personnels utilisés dans le Coran : معجم الأدوات والضمائر في القرآن الكريم *MuÝjam al-Þadawât wa-Ã-Ãamâþir fi-l-QurÞân al-karîm*. Même si les auteurs de cette concordance placent les pronoms personnels en dehors des mots-outils (l'on remarque ici qu'en dépit de la distinction faite entre les deux groupes, le fait de les traiter ensemble traduit bien leur grande similitude fonctionnelle), ils insistent bien dans l'introduction, sur la définition et les raisons du choix du terme *mots-outils* أدوات et non de celui de *particules* :

> « ونستميح القارئ عذرًا عن استخدام كلمة الأدوات بدلاً من المصطلح الشائع "حروف المعاني"، فهذه الكلمة أوفى بالحاجة من المصطلح المركّب من كلمتين "حروف المعاني"؛ فإنّ من الحروف ما هو خالص في

---

[220] Fleisch Henri, *Encyclopédie de l'Islam*, article ĴARF, p. 210

[221] *Mohammad FuPâd Pabd al-Bâqî, Al-MuÝjam al-mufahras li-þalfâÛ al-QurÞân al-kaîm*, 1987

الحرفيّة كالباء والفاء وبل... ومنها ما يجمع بين الاسميّة والحرفيّة والفعليّة كـ"ما" و"حاشا" و"عدا". وهو على أيّ حال مصطلح كوفيّ قديم فضلاً عن تجدّد استعماله لدى المحدثين «[222]

« Wa-nastmîÎu l-qâriÞa ÝuÆr-an Ýani stiÌdâmi kalimati l-Þadawâti badal-an mina l-muÒÔalaÎi š-šâÞiÝi "Îurûf al-maÝânî", fa-hâÆihi l-kalimatu Þawfâ bi-l-Îâjati mina l-muÒÔalaÎi l-murakkabi min kalimatayni "Îurûf al-maÝânî" ; fa-Þinna mina l-Îurûfi mâ huwa ÌâliÒ-un fî l-Îarfiyyati ka-l-bâÞi wa-l-fâÞi wa-bal… Wa-minhâ mâ yajmaÝu bayna l-ismiyyati wa-l-Îarfiyyati wa-l-fiÝliyyati ka-"mâ" wa-"Îâšâ" wa-"Ýadâ". Wa-huwa Ýalâ Þayyi Îâlin muÒÔalaÎ-un kûfiyy-un qadîm-un faÃl-an Ýan tajaddudi stiÝmâlihi ladâ l-muÎdaÆîna. »

« Le lecteur est prié d'excuser l'utilisation du mot [mots-]outils au lieu du terme courant "particules" *Îurûf al-maÝânî*. Ce mot est plus adéquat que le terme composé "*Îurûf al-maÝânî*" ; aussi certains de ces *Îurûf* sont-ils exclusivement de véritables particules comme le "*bâÞ*", le "*fâÞ*", "*bal"*... D'autres en revanche, peuvent être des noms, des particules ou des verbes comme "*mâ*", "*Îâšâ*", ou "*Ýadâ*". Quoi qu'il en soit, c'est un ancien terme coufique qui de plus, est réactualisé par les [linguistes] contemporains ». (La traduction est de nous)

Par ailleurs, parmi les mots-outils pour lesquels les auteurs établissent la concordance, l'on trouve des adverbes (إِذَا , إِذْ , ثَمَّ , حَيْثُ ...), des relatifs (, التِي الذِي , الذِينَ ...), des démonstratifs (أُولاَءِ , ذَا , ذَانِكَ ...), des circonlocutions (كَأَيِّنْ), des interrogatifs (مَتَى , مَاذَا , كَيْفَ , أَنَّى ...), etc.

- Fidèle à son engagement de réorganiser la classification des unités lexicales de l'arabe et de les définir sur la base de la complémentarité de la forme et du sens (voir *infra*), *Tammâm Ĥassân* définit clairement le *mot-outil* sous cet angle de vue :

» الأداة مبنى تقسيمي يؤدّي معنى التعليق والعلاقة التي تعبّر عنها الأداة إنّما تكون بالضرورة بين الأجزاء المختلفة من الجملة «[223]

« ÞAl-Þadâtu mabn-an taqsîmiyy-un yuÞaddî maÝnâ at-taÝlîqi, wa-l-Ýalâqatu l-latî tuÝabbiru Ýanhâ l-Þadâtu Þinnamâ takûnu bi-Ã-Ãarûrati bayna l-ÞajzâÞi l-muÌtalifati mina l-jumlati »

---

[222] *ÞIsmâÝîl ÞAÎmad ÝAmâyra et ÝAbd al-Ĥamîd MuÒtafâ as-Sayyid, MuÝjam al-Þadawât wa-Ã-ÃamâÞir fi-l-QurÞân al-kaîm*, 1988, p. 10 de l'introduction.

[223] *Tammâm Ĥassân, Al-luÈa l-Ýarabiyya. MaÝnâhâ wa mabnâhâ*, sans date, p.123.

« Le mot-outil est une catégorie [grammaticale] dont le rôle est d'établir une relation. La relation [syntaxique] exprimée par le mot-outil est forcément [établie] entre les différents constituants de la phrase » (la traduction est de nous).

À la suite de cette définition claire et en parfait accord avec la vision, quasi-universelle, qu'on a aujourd'hui du *mot-outil*, *Tammâm Ḥassân* présente le *mot-outil* comme étant composé de deux types : 1) le **mot-outil primitif** *al-ÞadâtalÒliyya* الأداة الأصليّة qui correspond aux *particules* ; et 2) le **mot-outil converti** *al-Þadât al-muÎawwala* الأداة المحوّلة correspondant aux adverbes, aux noms-outils (interrogatifs, démonstratifs, relatifs, noms de condition...), aux verbes (figés ou incomplets) et aux pronoms personnels.

4.1.3. Faut-il réorganiser le lexique arabe ?

Même si, depuis *Sîbawayhi* et jusqu'au début du VIII/XIV[e] siècle, la tendance générale était de considérer que le discours tout entier est Verbe, Nom et Particule, cette division tripartite du discours n'a pas toujours fait l'unanimité entre les grammairiens arabes classiques ou plus tardifs, et encore moins chez les linguistes contemporains. Il y a toujours eu, tout au long de l'histoire de la grammaire arabe, des remises en question de la rigidité de cette classification, et une insistante dénonciation de son hétérogénéité. Plusieurs études critiques autour des sciences du langage arabes le signalent[224]. Il n'est pas dans notre propos ici d'exposer toutes les tentatives qui ont été faites pour sortir de cette rigidité qui caractérise la division tripartite des classes de mots de la langue arabe. Mais, en plus de ce qui a été cité plus haut concernant les mots-outils, nous évoquons ici à titre d'exemple quelques propositions allant dans le sens de la dénonciation de la rigidité et de l'hétérogénéité de la division tripartite : celle par exemple, de *ÞAÎmad Ibn*

[224] Voir entre autres, les travaux de *Tammâm Ḥassân*, d'André Roman et (Gabr 1980)

Nâßir, grammairien du VIII/XIV[e] siècle, d'ajouter une quatrième classe, celle du *nom de verbe* اسم الفعل *Þism al-fiÝl [interjection]*. *ÝAÌud ad-Dîn al-PÎjî* (m. 756/1355) a proposé, quant à lui, une typologie en neuf classes. *PÎbn as-Sarrâj* (m. 316/928-9) donne, dans son livre *Kitâb al-ÞuÒûl fi n-naÎw*, une classification qui trouble la typologie tripartite. Parmi les linguistes contemporains, nous citons *Tammâm Ísassân* qui déclare, suite à une étude critique des fondements théoriques et méthodologiques de la classification des grammairiens arabes classiques, que :

« التقسيم الذي جاء به النحاة بحاجة إلى إعادة النظر ومحاولة التعديل بإنشاء تقسيم آخر جديد مبنيّ على استخدام أكثر دقّة لاعتباري المبنى والمعنى »[225]

« At-taqsîmu l-laÆî jâÞa bi-hi n-nuÎâtu bi-Îâjat-in Þilâ ÞiÝâdati n-naÛari wa-muÎâwalati t-taÝdîli bi-Þinšâßi taqsîm-in ÞâÌara jadîd-in mabniyy-in Ýalâ stiÌdâm-in Þak×ara diqqat-an li-Ýtibâray al-mabnâ wa-l-maÝnâ »

« La classification opérée par les grammairiens [arabes] a besoin d'être revue et remaniée de façon à aboutir à une nouvelle classification basée sur un emploi plus rigoureux des notions de forme et de sens » (la traduction est de nous)

C'est dans la continuité de son analyse des unités lexicales de l'arabe et de sa ligne de pensée insistant sur la complémentarité de la forme et du sens (complétées d'ailleurs et approfondies dans son livre *Manâhij al-baÎ× fî al-luÈa*), que *Tammâm Ísassân* propose sept catégories lexicales pour classer les éléments du lexique arabe : Nom, Adjectif, Verbe, Pronom, *Ëâlifa*[226], Adverbe et Mot-outil.

André Roman, qui se place, quant à lui, dans une perspective de rupture totale avec la Tradition grammaticale arabe, met l'accent sur « l'échec » des grammairiens arabes classiques à définir les unités lexicales sans pour autant que cet échec ne les incite à repenser la classification tripartite : « Les grammairiens, inlassablement,

---

[225] *Tammâm Ísassân, Al-luÈa l-Ýarabiyya*, op. cit., p.88

[226] Dans cette catégorie des *Ìawâlif* (pl. de *Ìâlifat*), *Tammâm Ísassân* classe les interjections (أسماء الأفعال), les onomatopées (أسماء الأصوات) et les deux formes d'exclamation (مَا أَفْعَلَ – أَفْعِلْ بِـ…).

s'attacheront, sans les reclasser, à définir ces données traditionnelles. Jamais ils n'aboutiront. Jamais les quelques définitions proposées par eux ne seront opératoires, régulièrement. Jamais, cependant, ils ne concluront de leur échec millénaire à la nécessité de rompre avec la Tradition »[227].

Que l'on adopte la position appelant à la rupture avec la division tripartite du discours considérée trop rigide, peu cohérente ou hétéroclite, ou que l'on considère que cette division tripartite a fait tout au long de son histoire l'objet d'un consensus quasi-total et que les différentes propositions, dans le passé, voulant la réformer « restent en deçà d'une rupture théorique avec Sibawayhi. Ils [les changements] présentent, si l'on peut dire, une diversité dans l'unité et un changement dans la continuité »[228], il est évident, du point de vue du TAL arabe, que la classification de la Tradition grammaticale fait fonctionner un système de catégories, par endroits, lexicalement hétérogènes ne permettant aucun gain de clarté pour la répartition catégorielle de certaines unités lexicales. Certains "réaménagements" s'imposent donc pour avoir une organisation des éléments du lexique s'adaptant aux exigences de l'analyse lexicométrique du discours et du TAL arabe en général. L'opposition, par exemple, entre lexicalité et fonctionnalité des unités du lexique est devenue une pratique inéluctable, pour toutes les langues, dans les domaines de l'ingénierie linguistique et du traitement de l'information textuelle.

Le but visé dans la proposition d'une telle organisation des unités lexicales n'est évidemment pas l'originalité. Nous ne prétendons pas proposer une typologie nouvelle et définitive du Lexique arabe ; tel n'est aucunement notre propos. Nous avons simplement proposé une organisation ponctuelle des unités lexicales, dictée par les exigences de l'approche adoptée, et appliquant, loin d'une démarche purement

---

[227] André Roman, *Genèse et typologie des unités de la langue arabe*, 1994, p. 118

[228] Hassan Hamzé, *Les parties du discours*, *op. cit.*, 1994, p. 97

éclectique, certains changements et aménagements proposés par des linguistes arabes et arabisants (certaines de ces propositions, comme nous l'avons vu plus haut, ont même été faites par des grammairiens classiques).

Nous présentons dans la page suivante le schéma général de l'organisation des unités lexicales de l'arabe avec les ramifications sur trois niveaux (quatre si l'on considère l'opposition mots lexicaux/mots-outils) ; suite à quoi, nous présenterons en détail, toutes les catégories lexicales de base, les catégories lexicales, les sous-catégories et les sous sous-catégories que nous avons retenues pour notre travail et pour la *norme lexicologique* que nous proposons dans une perspective de lexicométrie arabe.

Organisation du **lexique arabe** dans une perspective de lexicométrie

- Mots lexicaux
  - Verbes
    - Verbes simples
      - Trilitères simples
      - Quadrilitères simples
    - Verbes augmantés
      - Trilitères augmentés
      - Quadrilitères augmantés
  - Noms primitifs
    - Primitifs simples
      - Trilitères simples
      - Quadrilitères simples
      - Pentalitères simples
    - Noms d'unité
    - *MaÒdar* primitif
    - Noms d'une fois
    - Noms de manière
    - Noms primitifs augmentés
  - Noms dérivés
    - Noms de temps et de lieu
    - Noms d'instrument
    - Diminutifs
    - *MaÒdar mîmî*
    - *MaÒdar Ôinâ Yî*
    - *MaÒdar* dérivé
      - *MaÒdar* trilitère augmenté
      - *MaÒdar* quadrilitère simple
      - *MaÒdar* quadrilitère augmenté
  - Adjectifs
    - Participes actifs
    - Participes passifs
    - Intensifs
    - Elatifs
    - *Ñifa mušabbaha*
    - Adjectifs de relation
- Mots-outils
  - Particules
    - Article défini
    - Prépositions
    - Coordonnants
    - Particules de négation
    - Particules de comparaison
    - Particules d'attente ou de certitude
    - Particules du futur
    - Particules de condition
    - Particules de corroboration
    - Particule d'espérance et d'apitoiement
    - Particules de maÒdarité
    - Particules d'appel
    - Particules de réponse
    - Particules interrogatives
    - Particules d'exception
    - Particules d'ouverture
    - Particules d'incitation et de remords
    - Particules assimilées au verbe
    - Particules d'alternatives
    - Particules d'explication
  - Noms-Outils
    - Démonstratifs
    - Interrogatifs
    - Noms de condition
    - Relatifs
    - Noms de verbes
    - Onomatopées
    - Noms de nombres
      - Nombres cardinaux
      - Nombres ordinaux
      - Nombres fractionnaires
    - Circonlocutions
    - Adverbes
    - Pronoms personnels
    - Annectifs
  - Verbes fonctionnalisés
    - Verbes figés à l'accompli
    - Verbes figés à l'impératif
    - Verbes figés à l'inaccompli
  - Mots-outils composés

Figure 44

4.2. Les catégories lexicales retenues et leur code

Nous présentons ci-après en détail, toutes les catégories lexicales retenues. Elles sont classées hiérarchiquement : les catégories de base (verbes, noms primitifs, noms dérivés, adjectifs et noms composés pour les mots lexicaux, et mots-outils ; sans oublier les noms propres qui sont eux, à la périphérie du lexique), les catégories (verbes simples, noms primitifs simples, diminutif, participe actif, etc. pour les mots lexicaux, et particules, noms-outils, verbes fonctionnalisés, etc. pour les mots-outils ; noms propres de personnes, etc.), les sous-catégories (verbe trilitère simple, nom trilitère simple etc. pour les mots lexicaux et prépositions, démonstratifs, verbes figés à l'accompli, etc. pour les mots-outils) , les sous sous-catégories (verbe augmenté de forme II, etc. pour les mots lexicaux et nombres cardinaux, etc. pour les mots-outils). Devant chaque catégorie lexicale nous donnons le code que nous avons choisi pour elle et sous lequel elle a été enregistrée.

- **Verbes الأفعال Code : 1**

  Cette catégorie de base regroupe tous les éléments du lexique en relation au temps et dont les formes fléchies sont conjuguées, autrement dit, elles varient en personne, en "temps" (aspect), en mode et en voix[229]. De cette définition, il est clair que les "verbes" qui ne remplissent pas cette condition ne sont pas admis dans cette classe ; c'est pour cette raison que les verbes fonctionnalisés (verbes figés الأفعال الجامدة dans le Tradition grammaticale arabe) ne sont pas classés parmi

[229] Une présentation claire et bien structurée des paradigmes de conjugaison des verbes arabes est donnée dans l'introduction de : Joseph Dichy et Sam Ammar, *Les verbes arabes*, 1999, p. 11-20

les verbes de cette catégorie de base, mais plutôt dans les mots fonctionnels (mots-outils).

- **Verbe simple الفعل المجرّد Code : 11**

➢ **Verbe trilitère simple** (الفعل الثلاثي المجرّد فَعَلَ\فَعُلَ\فَعِلَ) **→ Code : 111**

➢ **Verbe quadrilitère simple** (الفعل الرباعي المجرّد فَعْلَلَ) **→ Code : 112**

- **Verbe augmenté الفعل المزيد Code : 12**

➢ **Verbe trilitère augmenté** الفعل الثلاثي المزيد **Code : 121**

- ❖ Forme **II** (*faŶŶala* - فَعَّلَ) → **Code : 1210**
- ❖ Forme **III** (*fâŶala* - فَاعَلَ) → **Code : 1211**
- ❖ Forme **IV** (*'afŶala* - أَفْعَلَ) → **Code : 1212**
- ❖ Forme **V** (*tafaŶŶala* - تَفَعَّلَ) → **Code : 1213**
- ❖ Forme **VI** (*tafâŶala* - تَفَاعَلَ) → **Code : 1214**
- ❖ Forme **VII** (*'infaŶala* - اِنْفَعَلَ) → **Code : 1215**
- ❖ Forme **VIII** (*'iftaŶala* - اِفْتَعَلَ) → **Code : 1216**
- ❖ Forme **IX** (*'ifŶalla* - اِفْعَلَّ) → **Code : 1217**
- ❖ Forme **X** (*'istafŶala* - اِسْتَفْعَلَ) → **Code : 1218**
- ❖ Autres → **Code : 1219**

اِفْعَالَّ – اِفعَوْعَلَ – اِفْعَوَّلَ – اِفْعَنْلَى – اِفْتَعْلَى

➢ **Verbe quadrilitère augmenté** الفعل الرباعي المزيد **Code : 122**

- ❖ (*tafaŶlala* - تَفَعْلَلَ) → **Code : 1220**
- ❖ (*'ifŶanlala* - اِفْعَنْلَلَ) → **Code : 1221**
- ❖ (*'ifŶalalla* - اِفْعَلَلَّ) → **Code : 1222**

- **Noms primitifs الأسماء الأصليّة Code : 2**

Les noms primitifs sont des noms qui ne sont pas dérivés d'un verbe. Ils désignent des êtres animés ou des choses, entités concrètes ou abstraites. En font partie les collectifs et les singulatifs (individués).

↳ Nom primitif simple الاسم الأصليّ المجرّد Code : 21

Les noms primitifs simples peuvent être composés de trois consonnes initiales (trilitères) de quatre (quadrilitères) ou de cinq consonnes initiales (pentalitères).

- **Nom trilitère simple** الاسم الثلاثي المجرّد **Code : 211**

فَعْل (شَمْس) \ **فَعَل** (بَصَل) \ **فَعِل** (كَبِد) \ **فَعُل** (رَجُل) \ **فُعَل** (صُرَد) \ **فِعْل** (رِجْل) \ **فِعَل** (عِنَب) \ **فِعِل** (إِبِل) \ **فُعْل** (قُفْل) \ **فُعُل** (عُنُق)

- **Nom quadrilitère simple** الاسم الرباعي المجرّد **Code : 212**

فَعْلَل (جَعْفَر) \ **فِعْلِل** (زِبْرِج) \ **فِعْلَل** (دِرْهَم) \ **فَعْلُل** (بَرْشُن) \ **فِعَلّ** (سِبَطْر) \ **فُعْلُل** (جُحْدُب)

- **Nom pentalitère simple** الاسم الخماسي المجرّد **Code : 213**

فَعَلَّل (سَفَرْجَل) \ **فَعْلَلِل** (جَحْمَرِش) \ **فُعَلِّل** (خُزَعْبِل) \ **فِعْلَلّ** (زِبْخَفْر)

↳ Nom d'unité اسم الوَحْدة Code : 22

C'est un singulatif (individué) formé par l'ajout d'un ة at à un collectif pour former un individu de l'espèce ou de la masse désignée par le collectif. Exemple, du collectif سمك *samak* [du poisson], nous obtenons le nom d'unité سمكة *samaka* [un poisson].

...ة (سمك : سمكة)

↳ *MaÒdar* primitif المصدر الأصلي Code : 23

En plus du *maÒdar* [nom d'action] primitif lui-même, nous classons dans cette catégorie le nom de *maÒdar* اسم المصدر comme وضوء *wuÃûÞ* [ablutions] ainsi que le

maÒdar intensif مَصْدَر المُبَالَغَة construit sur le schème تَفْعَال*tafŶâl* (تَضْرَاب \ تَعْدَاد \ تَطْوَاف \ تَسْآل).

فَعْل (أَكْل) \ **فَعَل** (فَرَح) \ **فُعُول** (جُلُوس) \ **فِعَال** (جِماح) \ **فَعَلان** (غَلَيان) \ **فُعَال** (سُعال) \ **فَعِيل** (رَحِيل) \ **فِعَالة** (إمارة) \ **فُعْلة** (سُمْرة) \ **فُعُولة** (سُهُولة) \ **فَعَالة** (فَصَاحة) \ **تَفْعَال** (تَعْدَاد)

↳ **Nom d'une fois اسم المرّة Code : 24**

C'est un nom formé par l'ajout d'un ة *at* au nom d'action des verbes simples (il aura, dans ce cas, le schème فَعْلة) et dérivés (sauf, pour ces derniers ceux qui se terminent déjà par ة).

فَعْلة (وَقْفة) \ **انْفِعالة** (انطلاقة) \ **افْتِعالة** (ابتسامة)

↳ **Nom de manière اسم الهيئة Code : 25**

Appelé aussi en arabe, اسم النوع *ism an-nawŶ*, ce nom est construit exclusivement, à partir du verbe trilitère simple, sur le schème فِعْلة et indique la manière d'accomplir l'action.

فِعْلة (جِلْسة)

↳ **Nom primitif augmenté الاسم الأصلي المزيد Code : 26**

(حِصَان – قِنْدِيل – إنسان – عصفور –...)

- **Noms dérivés الأسماء المشتقّة Code : 3**

Contrairement aux noms primitifs, les noms dérivés eux, sont soit des noms ou *maÒdars* dérivés d'un verbe (déverbaux), soit des noms dérivés du nom primitif (comme le diminutif). Les formes déverbales sont « celles des formes non

conjuguées qui dénotent un procès, c'est-à-dire une action ou une actualisation, exprimé dans son déroulement »[230].

↳ Noms de temps et de lieu اسما الزمان والمكان Code : 31

مَفْعِل (مَجْلِس – مَوْقِع – مَبِيت) \ **مَفْعَل** (مَكْتَب – مَرْمَى – مَقام) \ **مَفْعَلة** (مَزرعة – مَنامة – مَزَلّة) \ **مُفْعَل** (مُقام) \ **مُنْفَعَل** (مُنْصَرَف) \ **مُسْتَفْعَل** (مُسْتَقْبَل)

↳ Nom d'instrument اسم الآلة Code : 32

C'est un nom généralement dérivé du verbe trilitère simple transitif pour désigner un instrument utilisé pour accomplir l'action du verbe. Il y a quelques cas où le nom d'instrument est dérivé d'un verbe augmenté (comme مِزْمار *mizmâr* [trompette], du verbe de la forme II), d'un verbe intransitif (comme مِدْخَنَة *midÌana* [cheminée], du verbe intransitif دَخِنَ) voir même, et c'est rare, d'un nom primitif (comme مِحْبَرَة *miÎbara* [encrier] du nom primitif حِبْر *Îibr* [encre]).

مِفْعال (مِزْمار) \ **مِفْعَل** (مِصْعَد) \ **مِفْعَلة** (مِلْعَقة) \ **فاعِلة** (ساقِية) \ **فاعُول** (ساطُور) \ **فَعَّالة** (ثلاّجة) \ **فِعَّال** (إِزَّاث)

↳ Diminutif اسم التصغير Code : 33

فُعَيْل (قُلَيْم) \ **فُعَيِّل** (كُتَيِّب) \ **فُعَيْعِل** (جُعَيْفِر) \ **فُعَيْعِيل** (عُصَيْفِير)

↳ *MaÒdar mîmî* المصدر الميمي Code : 34

Le *maÒdar mîmî* est un *maÒdar* d'un type particulier. Sa particularité est qu'il commence par la lettre م *m*. Outre les schèmes que nous présentons dans l'encadré suivant correspondant à des *maÒâdir* de verbes trilitères simples, le *maÒdar mîmî*, quand il correspond à des verbes trilitères augmentés, a le même schème que le participe passif. Exemple : le *maÒdar* مُعْتَقَد *muÝtaqad* [croyance, conviction] de schème مُفْتَعَل *muftaÝal* et qui correspond au verbe de Forme VIII إِعْتَقَدَ *ÞiÝtaqada* [croire] de schème إِفْتَعَلَ *'iftaÝala*.

مَفْعَل (مَدْخَل) \ **مَفْعِل** (مَوْعِد) \ **مَفْعِلة** (مَعْرِفة) \ **مَفْعَلة** (مَفْسَدة)

[230] André Roman, *Etude de la phonologie et la morphologie de la Koinè arabe*, 1983, p. 1002

↳ *MaÒdar ÒinâÝî* المصدر الصناعي **Code : 35**

Le *maÒdar ÒinâÝî* est obtenu par l'ajout d'un ة *at* final à un adjectif de relation, lui-même obtenu par l'adjonction à un nom, d'un يّ *iyy* final. Exemple : إنسان *Þinsân* [Homme] → إنسانيّ *Þinsâniyy* [humain] → إنسانيّة *Þinsâniyya* [humanité].

...يّة (إنسانيّة)

↳ *MaÒdar dérivé* المصدر المشتقّ **Code : 36**

- ***MaÒdar* trilitère augmenté** المصدر الثلاثي المزيد **Code : 361**

تَفْعِيل – تَفْعِلة (تقسيم – تسمية) \ **فِعَال – مُفَاعَلة** (خِصام – مُخاصَمة) \ **إفْعال – إفَالة** (إحسان – إعانة) \ **تَفَعُّل** (تحسُّن) \ **تَفَاعُل** (تَخاصُم) \ **اِنْفِعَال** (انطلاق) \ **اِفْتِعَال** (اقتراب) \ **اِفْعِلال** (احمرار) \ **اِسْتِفْعَال – اِسْتِفَالة** (استقبال – استقامة)

- ***MaÒdar* quadrilitère simple** المصدر الرباعي المجرّد **Code : 362**

فَعْلَلة (دحرجة) \ **فِعْلال** (زلزال)

➢ ***MaÒdar*** **quadrilitère augmenté** المصدر الرباعي المزيد **Code : 363**

تَفَعْلُل (تَزَلْزُل) \ **اِفْعِنْلال** (احرنجام) \ **اِفْعِلّال** (اقشعرار)

- **Adjectifs الصفات Code : 4**

Il est à noter que les formes déverbales telles que les adjectifs, sont sujettes à une éventuelle nominalisation et constituent de ce fait un réservoir de formes nominales.

↳ **Participe actif اسم الفاعل Code : 41**

Le participe actif est un adjectif qui exprime celui qui fait l'action du verbe.

فاعِل \ مُفَعِّل \ مُفاعِل \ مُفْعِل \ مُتَفَعِّل \ مُتَفاعِل \ مُنْفَعِل \ مُفْتَعِل \ مُفْعَلّ \ مُسْتَفْعِل \ مُفَعْلِل \ مُتَفَعْلِل \ مُفْعَنْلِل \ مُفْعَلِلّ

↳ **Participe passif اسم المفعول Code : 42**

Le participe passif est un adjectif exprimant celui qui subit l'action.

مَفْعُول \ مُفَعَّل \ مُفاعَل \ مُفْعَل \ مُتَفَعَّل \ مُتَفاعَل \ مُنْفَعَل \ مُفْتَعَل \ مُفْعَلّ \ مُسْتَفْعَل \ مُفَعْلَل \ مُتَفَعْلَل \ مُفْعَنْلَل \ مُفْعَلَلّ

↳ **Intensif اسم المبالغة Code : 43**

Il sert à exprimer l'intensité ou la répétition. L'un des schèmes de l'intensif (فَعَّال) est notamment utilisé pour la formation d'un grand nombre de noms de métier.

فَعَّال (كذّاب) \ **مِفْعَال** (مِفْضال) \ **فَعُول** (ضَرُوب) \ **فَعِيل** (عَلِيم) \ **فِعِّيل** (سِكِّير) \ **فُعُّول** (قُدُّوس) \ **فَعَّالة** (علاّمة) \ **مِفْعِيل** (مِعْطِير) \ **فَيْعُول** (قَيُّوم) \ **فُعَّال** (كُبّار) \ **فَاعُول** (فاروق)

↳ **Elatif اسم التفضيل Code : 44**

Servant à exprimer la comparaison, l'élatif est un adjectif construit sur le schème أفْعَل *ÞafŶal* (فُعْلَى *fuŶlâ* pour le féminin), dérivé du verbe trilitère conjugable actif. Il a la valeur d'un comparatif quand il est indéfini suivi de la préposition مِن *min* [de], et d'un superlatif quand il est défini par l'article ou premier terme d'une annexion (que le deuxième terme de l'annexion soit défini ou non).

أفْعَل (أحْسَن)

↳ *Ñifa mušabbaha* الصفة المشبّهة **Code : 45**

La *Ñifa mušabbaha* (*bi-sm al-fâŶil*) باسم الفاعل (الصفة المشبّهة [litt. l'adjectif ressemblant ou assimilé (au participe actif)] est un adjectif dérivé du verbe intransitif pour exprimer une qualité durable, comme أعمى *ÞaŶmâ* [aveugle].

أفعل – فَعْلاء (أعمى – عمياء) \ **فَعْلان – فَعْلى** (شَبعان – شَبْعى) \ **فَعِل** (فَرِح) \ **فَعِيل** (كَرِيم – عَلِيّ) \ **فَعْل** (شَهْم) \ **فُعْل** (صُلْب) \ **فَعَل** (حَسَن) \ **فُعَال** (شُجاع) \ **فَعَال** (جَبان) \ **فَاعِل** (طاهر) \ **فَيْعِل** (طَيِّب) \ **فَيْعَل** (فَيْصَل)

↳ **Adjectif de relation** الاسم المنسوب **Code : 46**

L'adjectif de relation est obtenu par l'ajout d'un يّ *iyy* final à un nom. Il sert à exprimer la relation, la matière ou l'origine.

يّ... (عربيّ – فرنسيّ)

• **Mots-Outils** الأدوات **Code : 5**

Les mots-outils regroupent les particules, les noms-outils et les verbes fonctionnalisés. Vu leur spécificité et pour permettre, le cas échéant, de les étudier isolément, nous avons décidé de classer dans une catégorie à part, les ***mots-outils composés***.

↳ **Particules** الحروف **Code : 51**

- **Article défini** حرف التعريف **Code : 510**

اَلْ

- **Prépositions** حروف الجرّ **Code : 511**

إِلَى – بِ – تَ – حَتَّى – رُبَّ – عَلَى – عَنْ – فَ – فِي – كَ – لِ – مُذْ – مِنْ – مُنْذُ – وَ (واو رُبَّ)

- **Coordonnants** حروف العطف **Code : 512**

أَمْ – أَوْ – بَلْ – ثُمَّ – حَتَّى – فَ – وَ – لَكِنْ

- **Particules de négation** حروف النفي **Code : 513**

إِنْ – لاَ – لاَتَ – لَمْ – لَمَّا – لَنْ – مَا

- **Particules de comparaison** حرفا التشبيه **Code : 514**

ك – كَأَنَّ

- **Particule d'attente ou de certitude** حرف التوقّع أو التحقيق **Code : 515**

قَدْ – لَقَدْ

- **Particules du futur** حرفا الاستقبال **Code : 516**

سَ – سَوْفَ

- **Particules de condition** حروف الشرط **Code : 517**

إِذْمَا – أَمَّا – إِنْ – لَمَّا – لَوْ – لَوْلاَ – لَوْمَا – كَيْفَمَا

- **Particules de corroboration** حروف التوكيد **Code : 518**

أَنَّ – إِنَّ – قَدْ – لَ – نَّ – نْ

- **Particule d'espérance et d'apitoiement** حرف الترجّي والإشفاق **Code : 519**

لَعَلَّ

- **Particules de "*maÒdarité*"** حروف المصدريّة **Code : 5191**

أَ – أَنَّ – أَنْ – كَيْ – لَوْ – مَا – كَمَا

- **Particules d'appel** حروف النداء **Code : 5192**

آ – أَ – أَيْ – أَيَا – أَيَتُهَا – أَيُّهَا – هَيَا – وَا – يَا – اللَّهُمَّ

- **Particules de réponse** حروف الجواب **Code : 5193**

أَجَلْ – إِنَّ – إِيْ – بَجَلْ – بَلَى – جَلَلْ – جَيْرِ – كَلاَّ – لاَ – نَعَمْ – إِذًا

- **Particules interrogatives** حرفا الاستفهام **Code : 5194**

أَ – هَلْ

- **Particules d'exception** حروف الاستثناء **Code : 5195**

إِلاَّ – حَاشَا – خَلاَ – عَدَا – سِيَّمَا – لاَسِيَّمَا

- **Particules d'ouverture** حرفا الاستفتاح **Code : 5196**

أَلاَ – أَمَا

- **Particules d'incitation et de remords** حروف التحضيض والتنديم **Code : 5197**

أَلاَ – أَلاَّ – لَوْلاَ – لَوْمَا – هَلاَّ

- **Particules assimilées au verbe** الحروف المشبّهة بالفعل **Code : 5198**

أَنَّ – إِنَّ – كَأَنَّ – لَعَلَّ – لَكِنَّ – لَيْتَ

- **Particules d'alternative** حرفا التفصيل **Code : 5199**

أَمَّا – إِمَّا

- **Particules d'explication** حرفا التفسير **Code : 51991**

أَنْ – أَيْ

↳ **Noms-Outils الأسماء الأدوات** **Code : 52**

Nous avons proposé, en premier temps, cette appellation sans savoir qu'elle avait déjà été utilisée auparavant ; nous avons découvert par la suite que Régis Blachère l'avait déjà proposée dans sa *Grammaire de l'arabe classique*. Il définit les noms-outils comme suit : « Sous cette appellation, seront désignés des thèmes nominaux ou autres qui ont perdu leur valeur primitive ou qui, s'ils l'ont conservée, en ont pris parallèlement une autre qui permet de les utiliser comme de véritables outils grammaticaux »[231].

[231] R. Blachère, *Grammaire de l'arabe classique*, 1952, p.277.

- **Démonstratifs** أسماء الإشارة **Code : 520**

أُولاَءِ – أُولاَئِكَ – أُولاَكَ – أُولاَلِكَ – أُولَى – أُولَى لِكَ – تَا – تَاكِ – تَانِ – تَانِكَ – تَانِّكَ – تَلْكِ – تِلْكَ – تِلْكِ – تِهِ – تِهْ – تِي – تِيكَ – تِيكِ – تَيْنِ – تَيْنِكَ – تَيْنِّكَ – ذَا – ذَاكَ – ذَانِ – ذَانِكَ – ذَانِّكَ – ذَلِكَ – ذِلْكِ – ذِهِ – ذِهْ – ذِي – ذِيكِ – ذَيْنِ – ذَيْنِكَ – ذَيْنِّكَ – هَؤُلاَءِ – هَاتَانِ – هَاتِهِ – هَاتِيكَ – هَاتَيْنِ – هَذَا – هَذَاكَ – هَذَانِ – هَذِهِ – هَذَيْنِ – هُنَا – هُنَاكَ – هُنَالِكَ – هَهُنَا

- **Interrogatifs** أسماء الاستفهام **Code : 521**

أَنَّى – أَيُّ – أَيَّانَ – أَيْنَ – كَمْ – كَيْفَ – مَا – مَاذَا – مَتَى – مَنْ – مَنْ ذَا

- **Noms de condition** أسماء الشرط **Code : 522**

أَنَّى – أَيُّ – أَيَّانَ – أَيْنَ – أَيْنَمَا – حَيْثُمَا – كَيْفَمَا – مَا – مَتَى – مَنْ – مَهْمَا

- **Relatifs** الأسماء الموصولة **Code : 523**

أَيُّ – اَلْ – الأُلاَءِ – الأُلَى – التِي – الذِي – الذِينَ – اللاَّءِ – اللاَّئِي – اللاَّتِ – اللاَّتِي – اللَّتَانِ – اللَّتَيْنِ – اللَّذَانِ – اللَّذَيْنِ – اللَّوَاتِي – ذَا – مَا – مَنْ

- **Interjections (Noms de verbes)** أسماء الأفعال **Code : 524**

أُفٍّ – إِلَيْكَ – أَمَامَكَ – آمِينَ – آهِ – أَوَّهْ – إِيهِ – بَجَلْ – بَخٍ – بُطْآنَ – بَلْهَ – حَذَارِ – حَيَّ – حَيَّهَلَ – دُونَكَ – رُوَيْدَ – زِهْ – سُرْعَانَ – شَتَّانَ – شُكَانَ – صَهْ – عَلَيْكَ – مَكَانَكَ – مَهْ – نَزَالِ – هَا – هَاءَ – هَاءِ – هَاؤُمُ – هَاؤُمَا – هَاؤُنَّ – هَيَّا – هَيْهَاتَ – وَا – وَاهَا – وَرَاءَكَ – وَيْ

- **Onomatopées** أسماء الأصوات **Code : 525**

سَأْ – طَقْ – عَدَسْ – غَاقِ – قَبْ – كَخْ – كِخْ – نِخْ – هَلاَ – وَيْهِ

- **Noms de nombre** أسماء العدد **Code : 526**

  - Nombres cardinaux الأعداد الأصليّة Code : 5261

عدد مفرد : أحد – واحد – واحدة – اثنان – اثنين – اثنتان – اثنتين – ثلاثة – ثلاث – أربعة – أربع – خمسة – خمس – ستّة – ستّ – سبعة – سبع – ثمانية – ثمانٍ – ثمانيَ – تسعة – تسع – عشرة – عشر – مئة – مائة – ألف – مليون – مليار – بليون – بليار – ترليون

عدد مركّب : أحدَ عشر – إحدى عشرة – اثنا عشر – اثنيْ عشر – اثنتا عشرة – اثنتيْ عشرة – ثلاثَ عشرة – ثلاثةَ عشر – أربعَ عشرة – أربعةَ عشر – خمسَ عشرة – خمسةَ عشر – ستَّ عشرة – ستّةَ عشر – سبعَ عشرة – سبعةَ عشر – ثمانيَ عشرة – ثمانيةَ عشر – تسعَ عشرة – تسعةَ عشر

عقود : عشرون – عشرين – ثلاثون – ثلاثين – أربعون – أربعين – خمسون – خمسين – ستّون – ستّين – سبعون – سبعين – ثمانون – ثمانين – تسعون – تسعين

عدد معطوف : واحد وعشرون – … – اثنيْن وعشرين – … – ثلاثة وثلاثون – … – أربعة وثلاثين – … – خمسة وأربعون – … – ستّة وخمسين – … – سبعة وثمانون – … – ثمانية وتسعين – …

❖ Nombres ordinaux الأعداد الترتيبيّة Code : 5262

أَوَّلٌ – أُولَى – ثَانٍ – ثَانِيَةٌ – ثَالِثٌ – … – ثَامِنٌ – … – عَاشِرٌ – … – حَادِي عَشَر – حَادِيَة عَشْرَة – … – تَاسِعَ عَشَر – تَاسِعَة عَشْرَة – عِشْرُونَ – … – مِئَة – … – أَلْف – …

❖ Nombres fractionnaires الأعداد الجزئيّة Code : 5263

ثُلْثٌ – ثُلُثٌ – رُبْعٌ – رُبُعٌ – خُمْسٌ – خُمُسٌ – … – عُشْرٌ – عُشُرٌ

➢ **Circonlocutifs** أسماء الكناية **Code : 527**

ذَيْتَ – ذَيَّةَ – كَيْتَ – كَيَّةَ – كَذَا – هَكَذَا – كَمْ – فُلاَن – بِضْع – كَئِنٍ – كَأَيٍّ – كَأَيِّنْ – كَأَيْنْ – كَأْيِنْ – كَائِنْ

➢ **Adverbes** الظروف **Code : 528**

Sans entrer dans les détails des divergences terminologiques de la Tradition grammaticale arabe concernant les adverbes[232] (*mafÝûl fîh*, *Ûarf* pl. *Ûurûf*, *maÎall* pl. *maÎâll*, ou *Òifa* pl. *Òifât*), nous avons adopté une position d'ordre morpho-lexical. En effet, par adverbe, *Ûarf*, nous entendons tout

[232] Voir à cet effet, Moussaoui M., *Le circonstant de temps et de lieu dans le Coran*, Lyon, 2005, p. 25-42.

nom-outil non déclinable (figé) *Ûarf mabniyy* et ayant une utilisation exclusivement adverbiale *lâ yufâriqu Û-Ûarfiyya* [ne pouvant être utilisé qu'adverbialement] et ce qu'il soit strictement figé au cas direct *mabniyy Ýalâ n-naÒb* (قَطُّ – عَوْضُ – أَنَّ -...) ou pouvant être précédé des prépositions مِن ou إِلَى comme (قَبْلَ – بَعْدَ – فَوْقَ – مَتَى -...).

: Adverbes de temps

إِذْ – إِذَا – أَيَّانَ – مَتَى – أَمْسِ – الآنَ – مُذْ – قَطُّ – عَوْضُ – بَيْنَا – بَيْنَمَا – رَيْثَ – رَيْثَمَا – لَمَّا – خِلاَلَ – كُلَّمَا

: Adverbes de lieu

ثَمَّ – ثَمَّةَ – حَيْثُ – أَمَامَ – قُدَّامَ – بُحَاهَ – قُبَالَةَ – إِزَاءَ – وَرَاءَ – خَلْفَ – تَحْتَ – دُونَ – فَوْقَ – حِينَ – عَلُ – عِنْدَ – نَحْوَ

: Adverbes communs

بَيْنَ –لَدَى – لَدُنْ – أَنَّى – قَبْلَ – قَبْلُ – بَعْدَ – بَعْدُ

: Adverbes composés

بَيْنَ بَيْنَ – صَبَاحَ مَسَاءَ – لَيْلَ نَهَارَ

- **Pronoms personnels** الضمائر **Code : 529**

« La classe des pronoms, *nous dit Henri Fleisch*, est un domaine particulier, en dehors du comportement normal de la langue et dans la morphologie nominale et dans la morphologie verbale »[233]

أَنَا – ي – إِيَّايَ – أَنْتَ – كَ – إِيَّاكَ – أَنْتِ – كِ – إِيَّاكِ – هُوَ – هُ – إِيَّاهُ – هِيَ – هَا – إِيَّاهَا

نَحْنُ – نَا – إِيَّانَا – أَنْتُمْ – كُمْ – إِيَّاكُمْ – أَنْتُنَّ – كُنَّ – إِيَّاكُنَّ – هُمْ – هُمْ – إِيَّاهُمْ – هُنَّ – هُنَّ – إِيَّاهُنَّ

أَنْتُمَا – كُمَا – إِيَّاكُمَا – هُمَا – هُمَا – إِيَّاهُمَا

- **Annectifs** الأسماء الملازمة للإضافة **Code : 530**

En arabe, la relation de subordination entre deux noms (un nom et son complément déterminatif) s'exprime par l'adjonction de l'élément subordonné (le complément de nom) à l'élément qu'il complète ; on dit alors que les deux termes sont en état d'annexion. Il existe en arabe un

[233] Henri Fleisch, *Sur les pronoms personnels en arabe classique*, 1968, pp. 65-73, p. 70.

petit nombre de noms qui ne peuvent être utilisés que premier terme d'une annexion (il faut rappeler ici que la suffixation, pour les noms, est considérée comme un cas particulier de l'annexion). Les grammairiens arabes appellent ce type de noms الأسماء الملازمة للإضافة *al-ÞasmâÞ al-mulâzima li-l-ÞiÃâfa* [les noms faisant toujours partie d'une annexion]. Nous proposons d'appeler ***annectifs***, les noms ayant cette particularité.

كِلاَ – كِلْتَا – سِوَى – ذُو – أُولُو – وَحْدَ – كُلّ – بَعْض – غَيْر

↳ Verbes fonctionnalisés الأفعال الجامدة Code : 54

- **Verbes figés à l'accompli** أفعال ملازمة لصيغة الماضي **Code : 541**

Ce sont des verbes qui sont figés à l'accompli et ne peuvent être conjugués ni à l'inaccompli ni à l'impératif.

بِئْسَ – تَبَارَكَ – حَبَّ – حَبَّذَا – سَاءَ – سُقِطَ – شَدَّمَا – طَالَمَا – عَسَى – قَصُرَمَا – قَلَّمَا – كَثُرَمَا – لَيْسَ – نِعْمَ – نِعِمَّا – هَدَّ

- **Verbes figés à l'impératif** أفعال ملازمة لصيغة الأمر **Code : 542**

Ces verbes sont figés à l'impératif.

تَعَالَ – تَعَالَيْ – تَعَالَيَا – تَعَالَوْا – تَعَالَيْنَ – هَاتِ – هَاتِي – هَاتِيَا – هَاتُوا – هَاتِينَ – هَبْ – هَلُمَّ – هَلُمَّا – هَلُمُّوا – هَلُمِّي – هَلُمِّينَ – هَيْتَ/هِيتَ

- **Verbes figés à l'inaccompli** أفعال ملازمة لصيغة المضارع **Code : 543**

Il n'y a qu'un seul verbe dans cette catégorie et qui est figé à inaccompli.

يَهِيطُ

↳ Mots-Outils composés الأدوات المركّبة Code : 55

La composition peut être avec ou sans agglutination (ou amalgame). Dans cette catégorie, nous avons donc regroupé trois ensembles de mots-outils composés ; un groupe sans agglutination réunissant les deux formes

d'exclamation أَفْعِلْ بِـ... – مَا أَفْعَلَ , et deux groupes de mots-outils composés avec agglutination c'est-à-dire des composés dont les composants sont graphiquement soudés :

- le premier groupe comporte les mots-outils résultant de l'agglutination de certains mots-outils simples avec la particule مَا empêchant la rection مَا الكَافَّة عَنِ العَمَل *mâ al-kâffa Ýan al Ýamal* (– أَيُّمَا – رُبَّمَا – كَأَنَّمَا – أَنَّمَا – إِنَّمَا)

- le deuxième groupe comporte les interrogatifs composés résultant de l'agglutination de certaines prépositions avec l'interrogatif simple مَا qui, de ce fait, subit une transformation en perdant son *Þalif* final (لِمَ – بِمَ – عَلاَمَ – فِيمَ – عَمَّ – مِمَّ – ...)

أَفْعِلْ بِـ... – مَا أَفْعَلَ
أَيُّمَا – رُبَّمَا – كَأَنَّمَا – أَنَّمَا – إِنَّمَا
لِمَ – لِمَهْ – بِمَ – بِمَهْ – مِمَّ – مِمَّهْ – عَمَّ – عَمَّهْ – فِيمَ – فِيمَهْ – عَلاَمَ – عَلاَمَهْ – حَتَّامَ – حَتَّامَهْ – إِلاَمَ – إِلاَمَهْ – كَيْمَ – كَيْمَهْ

- **Noms propres أسماء الأعلام Code : 6**

Le nom propre est un mot qui désigne autoréférentiellement un individu donné sous une appellation spécifique qui n'est pas commune aux individus de la même espèce.

↳ **Noms Propres de personnes أسماء الأشخاص Code : 61**

Le nom propre de personne est traditionnellement composé, en arabe, de cinq éléments : كُنْيَة *kunya* [nom patronymique], اسْم *Þism* [prénom], نَسَب *nasab* [filiation], نِسْبَة *nisba* [adj. de relation] et لَقَب *laqab* [surnom]. La *kunya* est composée de *Þabû* أبو [père de] ou de *Þumm* أمّ [mère de] et d'un nom propre, généralement celui de la fille ou du fils aîné. L'*Þism* est le prénom de la personne proprement dite, il peut être composé de *Ýabd* عبد [serviteur, esclave] suivi de l'un des 99 qualificatifs de Dieu. Le *nasab* est la partie dans laquelle

sont mentionnés le père et les aïeux qui se suivent liés par le mot ابن *Þibn* [fils de] (quand il y en a plusieurs, à partir du deuxième, ابن *Þibn* est écrit بن *bn*). La *nisba* est un adjectif de relation se terminant par ي *î* et qui indique le pays, la ville ou le village, origine de la personne ; il peut aussi indiquer le métier. Enfin, le *laqab* qui se place généralement après la *nisba* est un surnom tiré d'un trait physique ou d'un titre honorifique.

أبو جعفر محمّد بن أحمد بن محمّد الصَيْمَرِيّ *Exemple* :

↳ **Noms Propres de lieux أسماء الأماكن Code : 62**

سُرّ مَن رَأَى (سَامَرَّاء) *Exemple* :

↳ **Noms de tribus, groupes et nations أسماء القبائل والأمم والفرق Code : 63**

بنو تميم *Exemple* :

↳ **Théonymes الأسماء الدينيّة Code : 64**

الله – الكريم *Exemple* :

↳ **Noms des œuvres أسماء الكتب Code : 65**

Ce sont des noms simples ou composés de plusieurs unités qui correspondent à des entités culturelles ou à des productions littéraires, théologiques, scientifiques, philosophiques ou autres.

إصلاح المنطق *Exemple* :

- **Noms composés الأسماء المركّبة Code : 7**

شقائق النعمان *Exemple* :

4.2.1. La base de données des mots-outils dans DIINAR[234]

La base de connaissances DIINAR était composée seulement de la base de données verbale et de la base de données nominale. Les mots-outils étaient quasiment absents de cette base de connaissances et les différentes versions d'analyseurs morphologiques ou morpho-syntaxiques basés sur DIINAR.0 utilisaient une liste ne regroupant qu'une infime partie des mots-outils. Ce qui avait pour conséquence de freiner l'analyse et de donner des résultats peu satisfaisants. En effet, on ne peut prétendre procéder à une analyse syntaxique des textes si l'on ignore la hiérarchisation et l'enchaînement syntagmatique qui rend compte des relations qui peuvent exister, dans un texte, entre ses mots lexicaux. Et l'un des repères indiscutables de cette structure syntaxique dans les textes, ce sont particulièrement les mots-outils, véritable ciment qui lie, entre eux, les éléments du discours.

C'est pour répondre à cette nécessité, et pour couronner la base de données DIINAR.1 en complétant le troisième volet de ses entrées (après les entrées nominales et verbales), qu'en octobre 1998, la tâche de constituer une base de données des mots-outils nous a été confiée, en collaboration avec Joseph Dichy. Nous avons donc entrepris ce travail linguistique qui avait pour objectifs de dresser et de saisir la liste complète des mots-outils, de définir une catégorisation de ces derniers et d'établir une liste des spécificateurs du niveau du mot associés aux mots-outils. Quelque temps plus tard, DIINAR s'est vu doter de sa troisième base de données, celle des mots-outils, regroupant quelques 452 mots-outils (après élimination d'un peu plus d'une dizaine pas très attestés dans la langue arabe). En même temps que la saisie directe de ces 452 mots-outils simples, l'interface de saisie nous a permis de décider des possibilités d'agglutination de ces mots-outils simples, entre eux ou à des clitiques et affixes. Quelques 11731 mots-outils agglutinés à des proclitiques et/ou des suffixes ont de ce fait été générés. La réalisation, informatique, de l'interface graphique de saisie et de

[234] **DI**ctionnaire **IN**formatisé de l'**AR**abe

mise à jour des mots-outils à été faite par R. Zaafrani[235]. Notre propre tâche consistait à fournir le contenu de la base de données des mots-outils, à saisir toutes les informations morpho-lexicales associées à ceux-ci, à déterminer les contextes gauche et droit de chacun d'entre eux et à envisager toutes les agglutinations possibles entre mots-outils ou entre un mot-outil et d'éventuels préfixes et/ou suffixes. Cette expertise linguistique qui était la nôtre nous a permis d'apporter plusieurs améliorations, en interaction avec R. Zaafrani, à l'interface graphique et donc à la modélisation et à la structure de la base de données qui est conçue sous MS ACCESS®. Nous présentons ci-dessous quelques captures d'écran de l'interface de saisie et de mise à jour des mots-outils de DIINAR.

Figure 45
Capture d'écran de l'interface de saisie et de mise à jour des mots-outils de DIINAR : Choix de la catégorie

[235] Voir (Zaafrani 2002)

Figure 46
Choix de la sous-catégorie

Figure 47
Choix de la sous-catégorie de niveau 2

Figure 48
Choix du mot-outil

Figure 49
Le mot-outil avec les spécificateurs du niveau
du mot qui lui sont associés

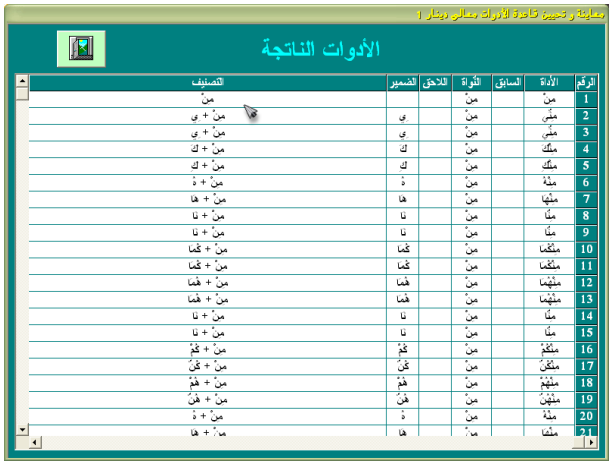

Figure 50
Le mot-outil simple avec ses différentes agglutinations possibles

CATEGORISATION DES MOTS-OUTILS

Ayant servi à la constitution de la base de données des mots-outils de DIINAR

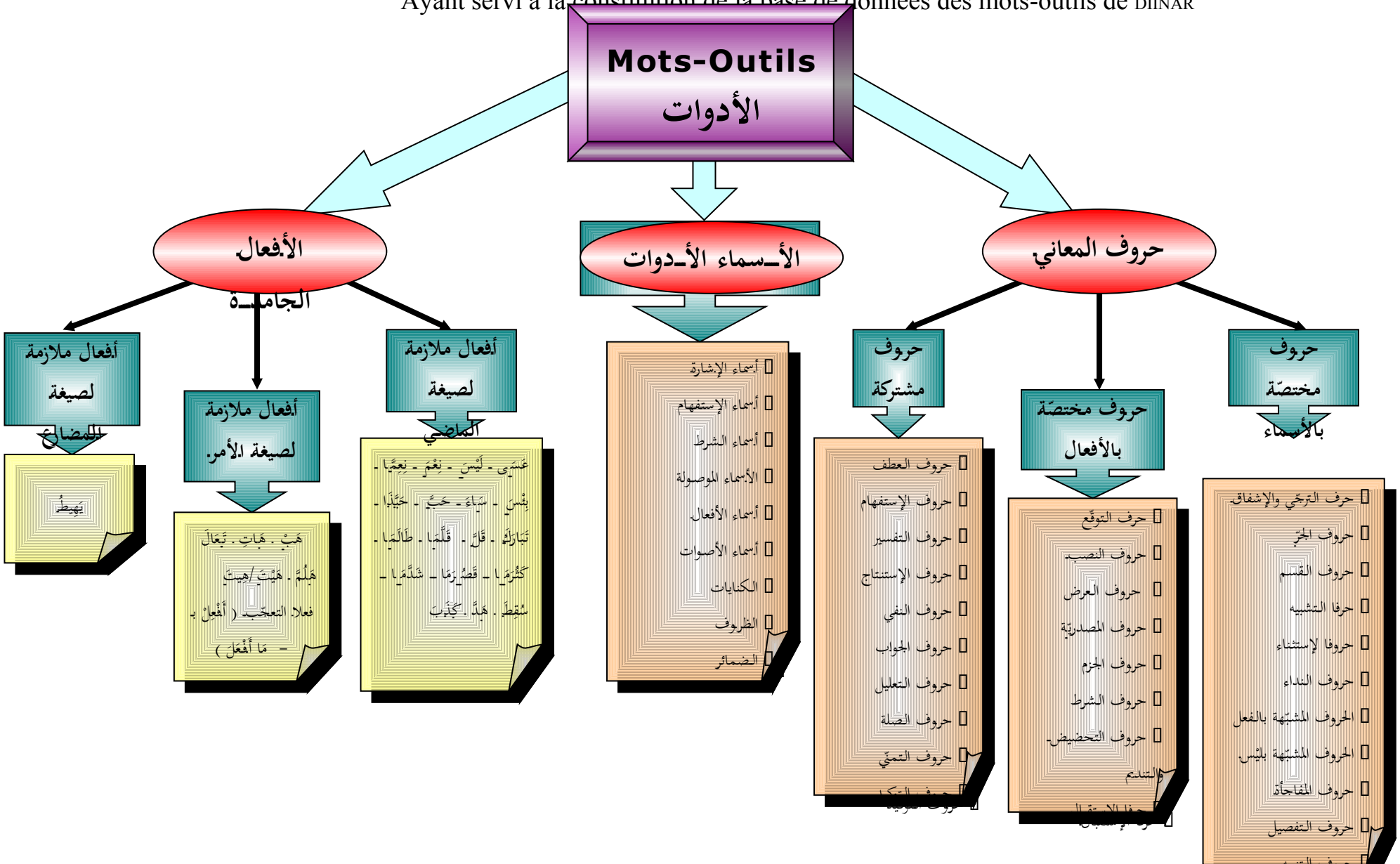

Figure 51

4.2.2. Les mots-outils entre DIINAR et la catégorisation retenue

Comme la base de données des mots-outils de DIINAR aspirait à l'exhaustivité, nous avons fait en sorte que toutes les sous-catégories et toutes les subdivisions, reflétant parfois des divergences entre grammairiens arabes classiques, soient présentes afin que tous les cas de figures soient prévus dans l'analyse des textes. Ce qui avait pour conséquence certaines redondances où des particules sont enregistrées dans plus d'une sous-catégorie. Cette situation, nous l'avons délibérément évitée dans la catégorisation de notre corpus afin que chaque mot-outil de celui-ci soit affecté à une seule et unique catégorie lexicale. Des simplifications de nature à alléger et faciliter l'opération de catégorisation ont également été faites.

En effet, au niveau des particules par exemple, nous n'avons pas retenu la distinction faite dans DIINAR entre particules affectées aux verbes, celles affectées aux noms et celles communes aux deux. Nous n'avons pas retenu non plus le critère de rection conduisant à des regroupements de particules déjà regroupées selon d'autres critères, ce qui a pour but d'avoir parfois les mêmes mots-outils appartenant à plus d'une sous-catégorie selon tel critère ou tel autre. Certaines sous-catégories sont donc en moins par rapport à DIINAR, mais d'autres sous-catégories ou mots-outils ont été ajoutés alors qu'ils n'existent pas dans DIINAR.

4.2.2.1. Ce qui est en moins par rapport à DIINAR

Nous n'avons pas retenu la sous-catégorie des particules d'alerte حروف التنبيه parce que les particules les composant sont déjà enregistrées dans d'autres sous-

catégories : أَلاَ et أَمَا dans la sous-catégorie des particules d'ouverture حرفا الاستفتاح ; et يَا dans la sous-catégorie des particules d'appel حروف النداء.

Également, la sous-catégorie les particules assimilées à *laysa* الحروف المشبّهة بليس n'a pas été retenue car toutes ses particules sont aussi incluses dans la sous-catégorie des particules de négation حروف النفي et qui sont مَا – لاَتَ – لاَ – إِنْ.

Les particules de surprise حرفا المفاجأة en l'occurrence, إِذْ et إِذَا sont aussi des adverbes et figurent donc en tant que tels dans la catégorie des noms-outils, sous-catégorie des adverbes الظروف.

Deux des particules du subjonctif حروف النصب c'est-à-dire أَنْ et كَيْ appartiennent déjà à la sous-catégorie des particules de *maÒdarité* حروف المصدريّة ; la troisième لَنْ à celle des particules de négation. Nous n'avons donc pas retenu cette sous-catégorie.

Les particules de l'apocopé حروف الجزم sont réparties entre la sous-catégorie des particules de négation (إِنْ – لاَ – لَمْ – لَمَّا – مَا), celle des adverbes (إِذْمَا – إِنْ – أَنَّى – أَيَّانَ – حَيْثُمَا) et celle des particules de condition (كَيْفَمَا). Il est donc inutile de garder cette sous-catégorie.

La sous-catégorie des particules de proposition حروف العرض est composée des trois particules أَلاَ – أَمَا – لَوْ qui sont déjà réparties entre la sous-catégorie des

particules d'ouverture (أَلاَ – أَمَا) et celle des particules de *maÒdarité* (لَوْ). Pour éviter les redondances inutiles, cette sous-catégorie n'a pas été retenue.

Les particules de motivation حروف التعليل sont au nombre de quatre ; كَيْ est déjà classée comme particule de *maÒdarité* (devant un verbe). Quant aux trois autres, فِي – لِ – مِنْ , nous les avons considérées toutes comme des prépositions bien que parmi elles, *lâm at-taÝlîl* (لِ لام التعليل) soit utilisée exclusivement devant un verbe. La sous-catégorie des particules de motivation n'a donc pas été retenue.

La sous-catégorie des particules de liaison حروف الصلة regroupe des particules appartenant déjà à celle des prépositions (بِ مِنْ -), à celle des particules de *maÒdarité* (مَا أَنْ -) et à celle des particules de négation (إِنْ).

Nous n'avons pas retenu non plus la sous-catégorie des particules du souhait حروف التمنّي car la particule هَلْ est déjà classée parmi les particules interrogatives et لَوْ parmi les particules de *maÒdarité*.

Outre cet allègement au niveau des sous-catégories qui a pour objectif une meilleure homogénéisation des mots-outils et un évitement des ambiguïtés polycatégorielles, nous avons supprimé également de la sous-catégorie des adverbes certaines entrées qui ne sont en fait pas des adverbes invariables. Il s'agit de certains mots lexicaux ou noms propres qui ne sont qu'occasionnellement adverbes ; des mots lexicaux ou des noms propres ayant leur comportement propre au sein de leurs catégories respectives, mais pouvant être utilisés adverbialement par occasion. Parmi ces mots, l'on trouve les jours de la semaine, les mois de l'année, les quatre saisons, les quatre points cardinaux ou encore certains mots lexicaux tels que :

أَمَدَ – آنَاءَ – سَاعَةَ – زَمَانَ – شَهْرَ – عَامَ – مُدَّةَ – نَاحِيَةَ – وَقْتَ – يَوْمَ

Nous avons décidé également de supprimer de la liste des adverbes, les démonstratifs de lieu هَهُنَا – هُنَا – هُنَاكَ – هُنَالِكَ et de les mettre dans la sous-catégorie des démonstratifs qui fait partie de la catégorie des noms-outils.

La sous-catégorie des prépositions a également été allégée de trois éléments : حَاشَـا – خَلاَ – عَـدَا qui sont en même temps des particules d'exception (ou plus exactement des verbes à l'accompli assimilés à des particules وَاقِعَةٌ مَوْقِعَ الحَرْف *wâqiÝat-un mawqiÝa l-Îarf* ou fonctionnalisés مَنْقُولَةٌ عَنِ الفِعْلِيَّةِ إِلَى الحَرْفِيَّة *manqûlat-un Ýani l-fiÝliyyati Þilâ l-Îarfiyya*) et de ce fait, elles ne régissent pas toujours le cas indirect ; tout au contraire, pour deux de ces trois particules à savoir خَلاَ – عَدَا le mot excepté est le plus souvent au cas direct et très rarement au cas indirect. L'inverse pour la troisième particule حَاشَا .

4.2.2.2. Ce qui est en plus par rapport à DIINAR

Ce que nous avons ajouté à nos catégories lexicales au niveau des mots-outils et qui ne fait pas partie de la base de données des mots-outils de DIINAR, ce sont deux sous-catégories : celle des annectifs et celle des noms de nombre.

Par ailleurs, nous avons ajouté quelques mots-outils qui faisaient défaut à certaines sous-catégories de DIINAR. Il s'agit de زِهْ à la sous-catégorie des interjections, de لَقَدْ à la particule d'attente ou de certitude, de نَحْوَ – كُلَّمَا – قُبَالَةَ – دُونَ – خِلاَلَ تُجَاهَ

بَيْنَ – à celle des adverbes, de هكذا à celle des circonlocutions, de سِيَّمَا (qui est considérée comme assimilée à une particule d'exception شبه استثناء) à celle des particules d'exception, de إِذًا à celle des particules de réponse, de كَيْفَمَا à celle des particules de condition, de لكنْ à celle des coordonnants, de كَمَا à celle des particules de *maÒdarité* et enfin, de اللَّهُمَّ à la sous-catégorie des particules d'appel.

4.3. Quelques difficultés de catégorisation

Aussi claire que soit la définition des catégories lexicales et aussi nettes que soient leurs frontières, il est toujours des cas où l'ambiguïté polycatégorielle persiste assidûment dans ce type d'entreprise qui est la catégorisation, exigeant du linguiste une prise de décision appropriée à chacun des cas de figures. Les frontières incertaines, dans les textes arabes, surtout classiques, entre le participe actif et le nom construit sur le même schème par un procédé de substantivation en est un exemple. Celles entre le *maÒdar* et le nom obtenu par substantivation de ce dernier en est un autre, etc. Il y a même des cas où l'ambiguïté est monocatégorielle. Ces incertitudes, dues en fait, à la non-exclusivité totale des partitions lexicales, ont pour conséquence de perturber l'opération de catégorisation si l'on n'y prend pas garde et si l'on n'arrête pas les décisions adéquates aux différents problèmes et difficultés soulevés par ces cas d'homographie globale.

Nous présentons ci-après quelques unes des difficultés que nous avons rencontrées lors de notre pratique de la catégorisation ainsi que les décisions que nous avons arrêtées pour sortir des incertitudes engendrées par les différents types d'ambiguïtés. Notre souci majeur étant d'affecter, dans la mesure du possible, à chaque vocable une seule et unique catégorie :

- Comme nous l'évoquions plus haut, l'une des ambiguïtés récurrentes réside dans la difficulté de savoir si une unité lexicale est un *maÒdar* (nom d'action) ou un nom "substantivé". Quand il est vraiment malaisé de trancher, notre décision était de considérer l'aspect morphologique de l'unité lexicale c'est-à-dire de ne tenir compte que de l'origine du schème sans prendre en considération, ou très peu, d'éventuelles substantivation ou translation de sens. Ainsi, dans la phrase suivante, n'est-il pas facile de décider (hors contexte historique de l'époque de

ʔawʔîdî) de la catégorie du mot الحكومة puisqu'il peut être interprété comme [le gouvernement] (nom augmenté → code 26) ou comme [le fait d'arbitrer] (*maÒdar* primitif → code 23). C'est dans cette dernière catégorie et avec le code 23 que nous l'avons enregistré.

مُتحرّيًا للحقّ في الحكومة غيْر مُسترَقّ بالتقليد ولا مخدوع بالإلْف ولا مُسخَّر بالعادة [236]

La même décision a été prise pour le mot خِلاَفَة (*maÒdar* primitif → code 23).

↳ La même décision de considérer le schème morphologique, est prise dans les cas semblables à l'exemple suivant où les unités lexicales sont classées dans la catégorie des "Participes passifs" (→ code 42). Contrairement à la phrase 1 où l'unité lexicale مفهـوم est clairement identifiée comme un participe passif [compris], la phrase 2 ne permet pas de trancher d'une façon aussi nette et où l'unité lexicale est plutôt interprétée comme [le concept] :

1 - ولكنّه مفهوم باللغة [237]

2 - وكَما يَتَغَيَّر المفهوم باختلاف الأفعال [238]

↳ Pareillement pour l'unité lexicale حـديث qui, de part son schème morphologique, est originellement une *Òifa mušabbaha* (phrase 3) pouvant, dans certains contextes, avoir une valeur nominale (phrase 4). Nous avons donc pris la décision, surtout dans les cas où la distinction est difficile à faire, de classer ces unités lexicales dans la catégorie des "*Òifa mušabbaha*" (→ code 45) plutôt que dans celle des "Noms augmentés" (→ code 26).

3 - والحسّ شديد اللَهَج بالحادِث والمحدَث والحديث لأنّه قريب العهْد بالكوْن [239]

---

[236] *Al-ÞImtâÝ wa-l-MuÞânasa*, p. 78

[237] *idem*, p. 115

[238] *idem*, p. 102

[239] *idem*, p. 23

4 - ألا ينتدب مِنكُم إنسان لمناظَرة متّى في حديث المنطق [240]

↳ En revanche, la décision prise pour les cas précédents n'a pas été suivie dans l'exemple suivant où, suite à une décision, réfléchie, que nous avons prise lors de la lemmatisation de certains mots construits sur le schème du participe actif au féminin *fâŶila* فَفَاعِلَة sous le lemme de la même forme (c'est-à-dire au féminin) du fait de leur forte lexicalisation en tant que noms "substantivés" féminins, nous avons décidé de classer des mots tels que (نوادر) نـادِرة \ (روائح) رائِحـة \ (روادف) رادِفـة \ (عـوارض) عارِضـة \ (نـوائب) نائِبـة \ (حـوادث) حادِثـة etc., dans la catégorie des "Noms augmentés" (→ code 26) et non dans celle des "Participes actifs" (→ code 41). En effet, si ces mots avaient été considérés comme des participes actifs, ils auraient été lemmatisés sous leur forme du masculin et, ce faisant, ils auraient perdu leur valeur exclusivement nominale.

↳ Le nom de nombre ordinal أوّل [premier] (→ code 5262) présente une homographie globale avec l'élatif [début] (→ code 44). Sur les 85 occurrences de أوّل , le seul cas, dans notre corpus, où cette unité lexicale pourrait être un élatif est dans la phrase ci-après : nous avons décidé de ne pas tenir compte de cette distinction vu que le contraste au niveau sémantique entre les deux acceptions n'est pas très manifeste, sans parler de la fréquence insignifiante (1/85).

وكذا في حِدْثان ما وَلِيَ الأمير أيْ في أوّل زمانه [241]

↳ En revanche, la distinction entre le nom de nombre ordinal ثانٍ [deuxième] (→ code 5262) et le participe actif ثـانٍ [ployant/fléchissant] (→ code 41),

[240] *idem*, p. 108

[241] *Al-ÞImtâŶ wa-l-MuÞânasa*, p. 25

comme dans la phrase suivante, est primordiale. Ainsi les 15 occurrences de la première acception ont-ils été classées dans la sous sous-catégorie des "Nombres ordinaux" et l'unique occurrence de ثانٍ dans le sens de [ployant/fléchissant] a été, elle, classée comme participe actif avec le code 41.

أنّكَ قَدْ بلغت الغاية وادِع القلْب ومَلكتَ المكانة ثاني العِنان [242]

↳ Nous avons enregistré un cas qui accuse une certaine particularité au niveau de la distribution catégorielle, et où le singulier et le pluriel sont classés dans deux catégories différentes. En effet, certains adjectifs de relation الأسماء المنسوبة tels que عربيّ [arabe], تركيّ [turque], فارسيّ [persan], etc. sont classés, au singulier, en tant que tels dans la catégorie des "Adjectifs de relation" (→ code 46), alors que leurs pluriels respectifs عَرَب [Arabes], أَتْرَاك [Turques], فُرْس [Persans], etc. sont classés plutôt dans la catégorie des "Noms de Tribus, Groupes et Nations" (→ code 63) de la catégorie de base des "Noms propres".

↳ Une autre difficulté que nous avons rencontrée résidait dans la subtilité à savoir si certaines unités lexicales construite sur le schème *faʕîl* فَعِيل sont des intensifs ou des *Òifa mušabbaha*. En effet, l'unité lexicale عَلِيم *ʕalîm* [(très) savant], par exemple, est ambiguë car elle peut aussi bien être un intensif comme une *Òifa mušabbaha*. C'est en fin de compte, dans la catégorie des "Intensifs" (→ code 43) que nous l'avons consciemment classée. Ce n'est pas le cas, en revanche, de l'unité lexicale كَرِيم *karîm* [généreux] qui, elle, a été classée, en toute connaissance de cause, dans la catégorie des "*Òifa mušabbaha*" (→ code 45).

---

[242] *idem*, p. 6

↳ Enfin, une attention particulière doit être prêtée à certains mots-outils qui ont des homographes appartenant à plus d'une sous-catégorie. Le mot-outil, par exemple, مَـن [qui] est à la fois relatif (→ code 523) et interrogatif (→ code 521). مَتى [quand] est à la fois adverbe (→ code 528) et interrogatif ou encore , أَيّ [quel/lequel] qui est aussi relatif et interrogatif, etc. Le cas de ما est encre plus frappant, il peut être interrogatif, relatif, particule de *maÒdarité* (→ code 5191), nom de condition (→ code 522) ou particule de négation (→ code 513). Chacun de ces mots-outils doit soigneusement être classé dans la catégorie lexicale adéquate ; c'est ce que nous nous sommes efforcé de faire.