Meta
Journal des traducteurs
Translators' Journal

META

Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot

Joseph Dichy

Volume 42, Number 2, juin 1997

Lexicologie et terminologie II (1) et Traduction et post-colonialisme en Inde
Translation and Postcolonialism: India (2)

URI: https://id.erudit.org/iderudit/002564ar
DOI: https://doi.org/10.7202/002564ar

See table of contents

Publisher(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (print)
1492-1421 (digital)

Explore this journal

Cite this article

Dichy, J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta*, *42*(2), 291–306. https://doi.org/10.7202/002564ar

Article abstract

This paper outlines: 1) the general structure of the lexical data base that is the fundamental requirement for effective electronic processing of the Arabic lexicon, and which is based on a declarative morphological approach; and 2) the methodology used to construct this base, that is, the design and implementation of lexical data acquisition interfaces, which must be ergonomie and compatible with type of programming to be carried out.

POUR UNE LEXICOMATIQUE DE L'ARABE : L'UNITÉ LEXICALE SIMPLE ET L'INVENTAIRE FINI DES SPÉCIFICATEURS DU DOMAINE DU MOT*

JOSEPH DICHY
CRTT – Université Lumière Lyon-2, France

Résumé

Cet article à pour objet de présenter dans leurs grandes lignes : 1) la structure générale de la base de données lexicales qui est la condition sine qua non *de tout traitement automatique opérationnel de l'arabe, et dont la conception est fondée sur une approche déclarative de la morphologie ; 2) la méthodologie de réalisation de cette base ; il s'agit de la conception et de l'implémentation d'interfaces de saisie des données lexicales. Ces dernières doivent être ergonomiques et compatibles avec les traitements envisagés.*

Abstract

This paper outlines: 1) the general structure of the lexical data base that is the fundamental requirement for effective electronic processing of the Arabic lexicon, and which is based on a declarative morphological approach; and 2) the methodology used to construct this base, that is, the design and implementation of lexical data acquisition interfaces, which must be ergonomic and compatible with type of programming to be carried out.

INTRODUCTION

Cette étude s'inscrit dans un ensemble de travaux échelonnés sur une dizaine d'années, qui présentent ma contribution et celle d'autres membres de l'équipe au programme de recherche *SAMIA* («Synthèse et Analyse Morphosyntaxiques Informatisées de l'Arabe») ainsi qu'à la conception de la base de données lexicale *DIINAR* («DIctionnaire INformatisé de l'ARabe»)[1]. Il y aura donc inévitablement des renvois, et aussi, pour des raisons de clarté, certaines redites.

L'exposé comporte trois parties :

- Dans la première sont présentées les grandes lignes du cadre théorique d'une lexicomatique de l'arabe conçue en fonction de traitements opérant dans le domaine du mot graphique.
- Dans la seconde seront exposés, à partir de cette conception de la lexicomatique le schéma général de l'*unité lexicale simple* (par opposition à l'unité lexicale composée et au phrasème), et le concept fondamental d'inventaire *fini* et *exhaustif* des *spécificateurs linguistiques* relatifs à un *domaine d'extension* donné (mot, syntagme nominal ou verbal, phrase[2]). Le concept de spécificateurs relève en propre de la linguistique informatique : associés aux unités lexicales d'une base de données, ces derniers «gèrent» les relations qui lient ces unités aux autres morphèmes (grammaticaux ou lexicaux) présents dans le domaine d'extension considéré. Cette «gestion» opère tant en synthèse qu'en analyse (les spécificateurs seront dits compatibles avec ces deux modes de traitement[3]).
- La dernière partie portera sur l'inventaire des spécificateurs morphosyntaxiques du domaine du mot graphique en arabe auquel nous sommes parvenus. La conception

informatique de ces derniers a fait l'objet d'une première présentation dans Hassoun (1987) ; les bases du raisonnement linguistique qui ont conduit à l'établissement d'un inventaire exhaustif sont développées dans Dichy (1990, 1993 et 1995).

Mot sera pris, par une extension assez fréquente en linguistique informatique, pour *mot graphique*[4]. La structure de cette unité sera brièvement rappelée, notamment en regard des problèmes redoutables posés à l'analyse automatique par les écritures sémitiques dotées d'un système alphabétique ne notant pas les voyelles brèves, la gémination des consonnes, etc., dans les textes courants (graphie dite couramment *non-vocalisée*). Il s'agit principalement des systèmes d'écriture du phénicien, de l'araméen, de l'hébreu, du syriaque et de l'arabe, seul étudié ici. Mais le modèle de traitement du niveau du mot associé à une base de données lexicale proposé pour ce dernier pourrait fort bien s'avérer utilisable dans les langues du domaine sémitique partageant un système d'écriture similaire.

La transcription de l'arabe (représentation graphémique) est explicitée en annexe.

1. VERS UNE LEXICOMATIQUE DE L'ARABE

1.1. La notion de lexicomatique

Lexicomatique — «mot-valise» construit à partir de «lexico-» et «automatique» — sera utilisé d'une manière plus restrictive que celle proposée par les organisateurs de ce colloque (bien que suggérée par le titre de ce dernier, qui associe «lexicomatique» et «dictionnairique»). On distingue aujourd'hui classiquement la *lexicologie* (étude théorique du lexique) et la *lexicographie* (étude du travail des auteurs de dictionnaires ; théorie de ce travail, en liaison avec la production de dictionnaires, en un sens qui inclut la «dictionnairique» — Quemada 1987 ; Mel'čuk, Clas & Polguère 1995 : 26-27). Le terme de *lexicomatique* ajoute à cette distinction la dimension de l'utilisation de l'outil informatique, à la fois comme support d'enregistrement des données et comme moyen ouvrant de nouvelles perspectives heuristiques et méthodologiques (notamment en ce qui concerne la lexicologie ou la lexicographie basées sur les analyses de corpus). Je propose d'en faire une utilisation qui présenterait l'avantage, si elle était retenue par l'usage, de ramener la lexicomatique dans le paradigme étymologiquement motivé de *lexicologie* («théorie du lexique») et de *lexicographie* («écriture» de celui-ci) : ce dernier terme met en rapport le lexique avec le support du livre, tandis que *lexicomatique* l'inscrit dans la «révolution informatique». D'où le schéma suivant :

LEXICOLOGIE	—> **LEXICOGRAPHIE** (y compris DICTIONNAIRIQUE) [N.B. : support scriptural]
	—> **LEXICOMATIQUE** [N.B. : support informatique]

La question est, dès lors, celle de l'apport de l'ingénierie linguistique à l'étude du lexique. Ainsi recentrée sur les relations entre le lexique et l'informatique, la *lexicomatique* peut être envisagée à deux niveaux (nullement exclusifs l'un de l'autre) :

- au sens d'«ingénierie de l'automatisation du lexique», ce terme est à mettre en parallèle, comme dans le schéma ci-dessus, avec la lexicographie et la dictionnairique ;
- dans une acception plus large, issue de la réflexion menée, dans une perspective d'intelligence artificielle, sur les relations entre l'organisation des connaissances chez l'agent cognitif humain et l'agent cognitif informatique[5], la lexicomatique intéresse directement

la lexicologie théorique (le domaine dit «d'application» opère une boucle sur le domaine théorique auquel il est rattaché).

1.2. Trois points de méthodologie

La démarche présentée ici à partir de l'arabe se situe principalement au niveau de cette deuxième acception. La lexicomatique offre en effet une excellente occasion de réfléchir sur l'organisation des connaissances linguistiques et sur la place occupée dans ces dernières par le lexique. Plus précisément, la lexicomatique définit un «point de vue d'observateur» à partir duquel des questions longtemps considérées comme marginales se trouvent reposées aux sciences du langage d'une manière nouvelle, et à mes yeux cruciale. Je mentionnerai trois points.

1) Le traitement automatique des données linguistiques est tenu de préciser ses entrées et ses sorties, ce qui conduit inéluctablement — tout comme en neurolinguistique ou en psycholinguistique — à *distinguer les démarches en synthèse* (chez le sujet humain, en production) *ou en analyse* (chez le sujet, en compréhension), *de l'oral ou de l'écrit*. Le problème, dans le cadre de la lexicomatique, est celui de la compatibilité des bases de données lexicales, et notamment des spécificateurs associés à leurs unités, avec les différents traitements envisagés. C'est ce que j'ai appelé la *contrainte de compatibilité connaissances-processus* (Dichy 1995) ; v. ci-dessous § 1.3.
2) La *relation entre lexique et grammaire*, c'est-à-dire entre les unités lexicales et leur agencement en unités syntagmatiques plus étendues, conduit à associer à chaque unité lexicale de la base de données des traits relevant de différents niveaux d'analyse linguistique (limités dans la présente étape de la recherche, pour des raisons qui seront explicitées plus loin, aux domaines du mot et de la phrase). On appellera ces traits des *spécificateurs*. Cette notion, qui relève en propre de la *linguistique informatique*, ne va toutefois pas de soi : encore faut-il, pour en déterminer le contenu avec une précision suffisante, disposer d'une définition formalisée : (a) de l'«unité lexicale» dans la langue concernée, (b) des «unités syntagmatiques plus étendues» au sein desquelles cette dernière est appelée à s'insérer et (c) des grammaires qui assurent cette insertion. La manière dont ces questions sont envisagées ici dans le domaine du mot graphique en arabe est l'objet de la deuxième partie.
3) L'un des principaux problèmes posés par les spécificateurs est celui du caractère nécessairement fini de leur inventaire, pour une application donnée. Cette contrainte apparaît, à première vue, comme une faiblesse inéluctable : à ce caractère fini semble s'opposer une impossibilité de parvenir à une description exhaustive. Or il n'en est rien : on peut en effet parvenir à une description exhaustive des données d'un *domaine d'extension* tel que le mot ou la phrase (§ 1.4.2), ce qui permet, en fonction de contraintes qu'il convient bien entendu de préciser, de *faire correspondre un inventaire fini de spécificateurs à un traitement exhaustif des données de ce «domaine»*. L'enjeu théorique de ce troisième point en fait l'un des principaux objets de cette communication. J'y reviendrai donc en conclusion.

1.3. Dissymétrie des traitements en synthèse et en analyse, ici, de l'écrit

Les entrées / sorties sont, comme déjà indiqué, constituées le plus souvent de textes écrits ; or la structure particulière des écritures sémitiques alphabétiques (arabe, hébreu, syriaque...) entraîne une importante dissymétrie entre génération et reconnaissance :

- le traitement en analyse doit être en mesure d'admettre pour entrées des formes non-vocalisées, si l'on veut prendre en compte les textes courants ;
- le traitement en synthèse, en revanche, peut avoir deux types de sorties. Les sorties en graphie non-vocalisée, plus faciles à réaliser, correspondent à l'usage courant, mais des

sorties vocalisées peuvent être requises par les nécessités liées à certaines applications : enseignement assisté par ordinateur, indexation documentaire, vérification et correction orthographiques de textes entièrement ou partiellement vocalisés, etc. Certains textes arabes classiques sont en effet édités en graphie entièrement vocalisée, et beaucoup d'éditions comportent une «vocalisation» partielle des textes, destinée à lever un certain nombre d'ambiguïtés de première lecture.

> Par exemple, un signe-voyelle diacritique *u* placé au besoin sur la première consonne d'un verbe à l'accompli situé en tête de phrase indique qu'il s'agit d'une forme «passive».

1.4. Les relations entre grammaires et lexique

Grammaire et lexique sont, comme on le sait, étroitement interconnectés. On se souvient que Chomsky (1965) formulait une «règle d'insertion lexicale» associant aux catégories lexicales des traits binaires du type [± N] («nom»), [± V] («verbe»), mais aussi [± animé], [± commun], etc. Dans des perspectives différentes, les travaux sur le français de M. Gross, de I. Mel'čuk ou les recherches réalisées dans le cadre des grammaires d'unification (Abeillé 1993), proposent des inventaires de traits morphosyntaxiques et sémantiques validant des relations à divers niveaux entre les éléments de la phrase[6].

Sur le plan formel, les relations entre grammaires et lexique sont un sous-ensemble des relations entre morphèmes, au sein de la phrase (ou d'une unité de moindre extension, telle que le mot). La distinction classique entre morphèmes grammaticaux (M^G) et morphèmes lexicaux (M^L) permet de considérer trois schémas fondamentaux de contextualisation :

(1) $<M^G - M^G>$ (2) $<M^L - M^G>$ (3) $<M^L - M^L>$

Il va de soi que les textes mêlent généralement ces possibilités. Les relations qui posent le plus de problèmes tant à la description linguistique qu'au traitement automatique des langues sont celles qui font intervenir des morphèmes lexicaux, du fait que ces derniers, à la différence des morphèmes grammaticaux, relèvent d'inventaires non finis. Les schémas ci-dessus correspondent ainsi à trois cas de relations de contextualisation :

(1) < unités d'inventaires finis (M^G) – unités d'inventaires finis (M^G) >
(2) < unités d'inventaires non finis (M^L) – unités d'inventaires finis (M^G) >
(3) < unités d'inventaires non finis (M^L) – unités d'inventaires non finis (M^L) >

La formalisation des relations entre unités relevant de ces deux types d'inventaires est cruciale, tout particulièrement pour l'étude des processus de reconnaissance des textes (compréhension humaine ou analyse automatique). Elle joue un rôle central dans la constitution de l'inventaire fini des spécificateurs morphosyntaxiques associés au noyau lexical du mot graphique en arabe et présentés dans la 3e section.

1.4.1. Une différence fondamentale entre les bases de données lexicales et la structure du lexique chez l'agent cognitif humain

Les relations ci-dessus sont elles-mêmes de nature différente (voir, par ex., Lyons 1978/90). Dans le système de définitions formalisé qu'il propose pour la linguistique, I. Mel'čuk représente le concept de *signe linguistique* comme : «le triplet ordonné :

$$X = < Y ; Z ; W >$$

où Y est le signifiant de Z, Z est le signifié de Y, et W est le syntactique de la paire <Y ; Z>» (1982a : 40 ; 1982b : 108).

Le *syntactique* est chez Mel'čuk, «une propriété constituante des signes des langues naturelles» dont la définition comporte celle de l'arbitraire non seulement de la relation signifiant-signifié (comme chez Saussure), mais encore de la combinatoire des signes : les relations décrites dans le syntactique d'un signe ne sont en effet déductibles ni de son signifiant ni de son signifié, et ne peuvent, par conséquent, être prédites par des règles.

> Un exemple élémentaire est celui du genre des noms dans des langues comme le français, le russe ou l'allemand. En arabe, certains noms ne relèvent pas, comme ≤?uMM≥, «mère», des «féminins sémantiques», et ne portent pas de marque morphologique entraînant le genre féminin ; ex. ≤DaAR≥, «maison» ou ≤HaRB≥, «guerre».

En linguistique informatique, cependant, les *spécificateurs* associés aux unités lexicales sont soumis, quelle que soit la nature des informations qu'ils renferment, à une contrainte particulière : ils doivent nécessairement comprendre des éléments qui, dans la définition du signe de Mel'čuk, sont déductibles du signifié (par ex., le trait [non transitif] associé à certains verbes), du signifiant (par ex., la conjugaison de verbes comportant une consonne *w* ou *y*, etc.), ou du syntactique (par ex., le fait que ≤Ta^{c}aALa≥, «viens», n'existe qu'à l'impératif, ou que dans ≤?a^{c}aASa≥ «faire vivre», de racine =^{c}YS, les règles de transformation morphophonologiques habituelles s'appliquent, mais non dans ≤?a^{c}YaLa≥, «avoir une famille à nourrir» de racine =^{c}YL, etc.).

Comme dans la morphologie à deux niveaux (pour le français, v. B. Fradin 1994), il y a une présentation «linéaire» des informations. Cela est dû au caractère particulier de la représentation du signe linguistique dans une base de données lexicale. Les unités enregistrées dans cette dernière comportent en effet une structure :

Signe $_{\text{[en linguistique informatique]}}$ **= < chaîne de caractères ; spécificateurs >,**

à la différence du signe linguistique, qui associe, comme on l'a vu, signifiant, signifié et syntactique. On notera au passage qu'un nombre non négligeable de processus phonologiques sont déductibles de la chaîne de caractères, *i.e.* de la manifestation graphique du signifiant. Il n'en reste pas moins que l'on est en présence d'une réduction, qui répond aux caractéristiques du texte tel qu'il s'offre (en mode analyse) à la reconnaissance automatique : les informations y sont de nature exclusivement intratextuelle, ce qui entraîne leur réduction à un ensemble de relations contextuelles (relations «d'indice à indice»). Il s'agit là d'une différence fondamentale entre le traitement de l'information lexicale par l'agent cognitif informatique d'une part, et humain de l'autre. Elle est cependant loin d'être irréductible : on peut en effet classer les spécificateurs lors de leur saisie en fonction du domaine dont ils relèvent (morphophonologie, syntaxe, sémantique...) au moyen d'interfaces adaptées aux systèmes de connaissances de l'«expert humain[7]».

1.4.2. Les domaines d'extension des relations contextuelles

On saisit l'importance de la question du caractère fini ou non fini des informations que l'on peut associer aux unités lexicales. Afin d'apporter un commencement de solution à ce problème, j'ai posé la convention selon laquelle un *formant* est un morphème dont on spécifie le *domaine d'extension* des contraintes liées aux relations contextuelles déductibles de ses signifiant, signifié et syntactique. Une telle spécification exprime explicitement le fait que les *relations* entre les morphèmes (leurs *relations contextuelles*) ont pour limites le domaine de l'unité considérée : on définit ainsi des *formants de mot* ou *de phrase*, etc., selon les domaines contextuels pris en compte (Dichy 1990, 1993).

La limitation aux domaines du mot et de la phrase (ce qui inclut au besoin le syntagme verbal ou nominal) était due, dans un tout premier temps, à un souci de faisabilité des traitements envisagés. Mais très vite, l'extrême fécondité de ce cadre de travail s'est

manifestée : à la différence du «domaine» du texte, le mot et la phrase peuvent être l'objet d'une représentation formelle, comme on le verra ici pour l'arabe. (Il faut rappeler, naturellement, que le mot doit être redéfini pour chaque langue et que l'importance de cette unité complexe en arabe et dans d'autres langues sémitiques, ne se retrouve pas partout.) L'idée consiste à aller jusqu'au bout des domaines où il est possible de rencontrer des inventaires finis de spécificateurs, et de construire les bases de données correspondantes, réalisant ainsi un progrès déterminant.

La convention posant les morphèmes comme formants de mot ou de phrase ajoute à la définition du signe due à Mel'čuk une contrainte formelle, celle du *domaine d'extension des relations contextuelles*. Ces dernières (déductibles des trois composantes du signe) sont de «portée» différente, selon l'unité considérée. L'intérêt théorique de cette modification est la suivante : nombre de théories ou de descriptions linguistiques — y compris le modèle Sens-Texte de Mel'čuk — sont orientées, de manière diversement explicite, par un raisonnement en génération[8]. Or, la dissymétrie des traitements (chez l'agent cognitif informatique) ou des processus (chez l'agent cognitif humain) en mode génération et en mode reconnaissance, entraîne la nécessité de concevoir des modèles linguistiques compatibles avec l'un et avec l'autre, c'est-à-dire des modèles soumis à la *contrainte de compatibilité connaissances-processus* (Dichy 1995). La détermination du domaine d'extension est issue de préoccupations relatives à la reconnaissance, venant s'ajouter aux formalismes et aux informations nécessaires à la génération. Elle joue un rôle central dans la théorisation des spécificateurs du domaine du mot présentés plus loin, ainsi que dans la démarche qui conduit à un inventaire fini de ces derniers et à un traitement exhaustif des données dans un domaine précis.

Les spécificateurs morphosyntaxiques pourront donc, selon le cas, opérer dans les limites :

- du mot (par convention : *M-spécificateurs*, correspondant aux *formants du mot*) ;
- de la phrase (*P-spécificateurs*, correspondant aux *formants de phrase*).

Les M- et les P-spécificateurs devront nécessairement être basés sur une définition formelle, respectivement, du mot et de la phrase[9] dans la langue concernée. Considérons donc l'unité-mot.

2. LA THÉORIE DU MOT GRAPHIQUE EN ARABE

Le *mot graphique en arabe* comporte une structure d'objet complexe. D. Cohen (1961/70) appelait *mot maximal* l'unité décomposable en : *proclitique*(s), *préfixe*, *base*, *suffixe*(s), *enclitique*(s) — ci-après : *PCL, PRF, BAS, SUF, ECL* (terminologie actualisée ; cf. Desclés *et al.* 1983). On trouvera dans le tableau ci-dessous un exemple simplifié.

La *base*, pour la partie du lexique qui relève du système dérivationnel propre aux langues sémitiques[10], s'analyse en racine (*RAC*) et schème (*SCH*), au sens que prennent ces termes dans le domaine sémitique. On notera toutefois qu'un sous-ensemble important des noms ne peut être analysé ainsi. Ces noms correspondent à des *pro-bases* (*PBA*).

Ex. : ≤YaASiMiYN≥, «jasmin» ; ≤?iBRaAHiYM≥, «Abraham», etc.

Bases et pro-bases sont le *noyau lexical* du mot graphique (ou *formant-noyau, Fn*), les autres constituants étant des *extensions* (ou *formants-extensions, Fe*).

En faisant usage des *frontières faible* et *forte* de morphème (respectivement «#» et «+»), on peut représenter le mot ainsi :

Représentation «classique» en constituants immédiats			mot maximal			
			mot minimal			
Exemple (sommaire) :	## PCL	# PRF	+ {BAS OU PBA)	+ SUF	# ECL	##
	## Li	# Ta	+ KTuB	+ uW	# Hu	##
	«pour que»	pronom 2e pers.	«écrire» inachevé	plur. masc. (subjonct.)	«lui» pronom complém., accusatif	
Représentation faisant apparaître la saillance du noyau lexical			Fn / \ Fe — Fe			

Le schéma du mot

■ Pour la représentation en constituants immédiats (Desclés *et al.* 1983), lire «##» comme *frontière de mot*; le critère empirique permettant de distinguer la frontière «+» (pré- ou suffixation) de la frontière «#» (enclise : pro- ou enclitiques) est celui de la pause potentielle (J. Lyons) : en l'absence du PRF ou du SUF auquel elle est liée par une frontière «+», la BAS — ou la PBA — ne peut constituer une forme libre minimale (Bloomfield). En revanche, elle peut, de ce point de vue, «se passer» des ECL et des PCL. Pour une définition de chacun des termes cités dans ce paragraphe, voir Dichy & Hassoun (1989 «Lexique de définitions» : 265-276) et Dichy (1990 : ch. X).

2.1. Relations contextuelles et formants de mots

La différence entre les deux représentations ci-dessus est considérable. Dans le modèle de D. Cohen (1961/71), les relations entre les morphèmes inclus dans le mot demeuraient conçues sur un mode distributionnel et concaténatoire.

La représentation introduite dans Desclés *et al.* (1983) et développée dans Hassoun (1987), Dichy & Hassoun (1989), ajoutait notamment l'idée fondamentale d'un dictionnaire des bases associé aux traitements en synthèse et en analyse, la distinction entre ces deux modes de traitement et la prise en compte, dans cette double perspective, des processus morphophonologiques de transformation.

La représentation faisant apparaître la saillance du noyau lexical constitue un nouveau saut qualitatif : le Fn est un formant de mot, *i.e.* un formant dont les relations contextuelles sont spécifiées dans le domaine de cette unité. Il s'agit dès lors de décrire le système de connaissances correspondant à l'ensemble des relations entre les unités du dictionnaire associé aux traitements du niveau du mot. Au sein de cette unité, les relations sont, au niveau le plus abstrait, de deux types :

(a) <Fe − Fe> et (b) <Fn − Fe>

Le second (b) constitue le champ des spécificateurs du noyau lexical, ce qui ouvre la voie à la définition de l'unité lexicale en arabe présentée ci-dessous.

2.2. Le schéma de la grammaire des formants du mot

Une *grammaire des formants du mot* issue de cette conception comporte deux sortes de *relations contextuelles* (Dichy 1987) :

- la *relation d'ordre* au sens strict régit la position respective des formants, qui, dans le cadre du mot, est strictement prédictible. On dira que cette unité peut être représentée au moyen d'un vecteur ordonné.

  Ex. : La préposition proclitique ≤Li#≥, «pour» est placée avant l'article, et après les coordonnants ≤Wa#≥ ou ≤Fa#≥ ;

- les *relations de collocation morphophonologiques et syntaxiques* «gèrent» les incompatibilités et cooccurrences, ainsi que les modifications morphophonologiques entraînées par la présence d'un formant.

  Ex. : L'article ≤? aL#≥ est incompatible avec les marques casuelles de l'indétermination, mais aussi avec un très grand nombre de noms propres. Le pronom clitique ≤#Hu≥ est réalisé ≤#Hi≥ après une voyelle ≤i≥.

2.3. L'évolution des modèles de représentation des phénomènes du niveau du mot en arabe

Si l'on demeure ici dans un modèle «item et agencement» (comme dans les démarches computationnelles liées à la morphologie à deux niveaux — Fradin 1994), la différence avec le modèle de D. Cohen et ceux qui en découlent plus ou moins directement réside essentiellement dans la représentation des systèmes de connaissances relatifs au mot graphique dans une perspective d'intelligence artificielle (Dichy 1993) : traitement déclaratif d'ensembles complexes de données associées à chacune des unités du lexique (Hassoun 1987), bases de données relationnelles soumises à la contrainte générale de compatibilité avec la reconnaissance et la génération (Dichy 1990).

La différence ne se situe pas seulement au niveau de la structure générale des systèmes de traitement envisagés, elle comporte également des aspects descriptifs (en linguistique) et de pure faisabilité (en linguistique informatique). Les modèles issus de celui de D. Cohen ont tous le mérite de souligner l'importance de ce qui correspond, ici, au schéma des relations <Fe – Fe>. Mais ils ne prennent pas en compte les relations <Fn – Fe>, c'est-à-dire le jeu des spécificateurs morphosyntaxiques associés aux Fn, sans lesquels il est exclu de parvenir à éliminer la production d'un grand nombre de formes exclues de la langue en génération, ou un niveau très élevé de «bruits» en analyse.

Ex. : ≤? aNna≥, «geindre», est intransitif. En mode synthèse, *≤? aNna#Hu≥, «geindre–lui», où ≤#Hu≥ serait un pronom proclitique complément d'objet, doit donc pouvoir être exclu (au moyen du M-spécificateur «verbe intransitif») ; en mode analyse, seule devra être retenue la forme homonyme ≤? aNna≥, conjonction «que», suivie du pronom ≤#Hu≥, «il» ou «lui».

À cela s'ajoute, dans le cadre du modèle SAMIA, la possibilité de produire, en génération, des formes en graphie vocalisée et de reconnaître, en mode analyse, des mots graphiques non vocalisés.

2.4. Structure de l'*unité lexicale simple* (*UL*) en arabe

En théorie de l'écriture, on peut interpréter la structure <Fe – Fn – Fe> du mot graphique en arabe comme un reflet de sa fonction logographique : le mot est en effet l'unité perceptive susceptible d'inclure un noyau lexical (nom, verbe) *au plus*[11]. Il n'existe pas en arabe, à un petit nombre d'exceptions près, de composition de radicaux (au sens où l'on entend ce terme dans les langues indo-européennes), comme dans *métro+pole, théo+sophie,* etc. : la composition doit recourir à des procédés syntaxiques, *i.e.* à des phrasèmes.

Tout se passe comme si le système d'écriture constituait les mots incluant un noyau lexical en unités ayant pour fonction d'en manifester la présence : toutes les relations associées aux *formants-extensions* (Fe) de cette unité complexe convergent, en compréhension, vers la détermination de la catégorie syntaxico-sémantique du *formant-noyau* (Fn). Cette analyse «en faisceaux» des textes en écriture arabe non vocalisée n'est possible qu'en raison de la présence d'informations relatives aux relations de collocation morphophonologiques et syntaxiques associées aux signifiant, signifié et syntactique des Fn dans la mémoire lexicale des sujets (Dichy 1990 : ch. IX et X) — informations représentées, dans le modèle de traitement automatique auquel il est fait ici référence, par un jeu de spécificateurs morphosyntaxiques régissant les relations de collocation entre le Fn et les Fe.

Or, une précision essentielle doit être apportée : en ce qui concerne les noms (et à la différence des verbes), l'unité lexicale simple ne coïncide pas toujours avec le *formant-noyau* (Fn). 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 indépendante. Une *unité lexicale simple* (*UL*) est donc constituée :

- soit d'un Fn (on écrira : UL = <Fn>) ; c'est le cas de toutes les bases verbales et des bases nominales dépourvues de Fe lexicalisé (Fel).

  Ex. : ≤MaKTaB≥ (plur. ≤MaKaATiB≥) «bureau», où l'on a : UL = <Fn = MaKTaB> ;

- soit d'un ensemble UL = <Fn,Fel>, où Fel peut inclure plus d'un formant, et dont l'ordre séquentiel est assigné par la grammaire des formants du mot (*relation d'ordre* au sens strict).

  Ex. : ≤MaKTaBa&≥ (plur. ≤MaKTaBaAT≥), «bibliothèque» ou «librairie», constitue une UL distincte de la précédente, analysable ainsi : UL = <Fn = MaKTaB, Fel = +a&>. Dans le nom propre <?aL#**T**aAHiR≥, l'article d'«excellence» ≤?aL#≥ est un Fel.

2.4.1. Le fléchage entre unités lexicales

Un deuxième trait définitoire vient s'ajouter au précédent : celui du *fléchage entre unités lexicales*, que l'on peut représenter sous la forme : UL_1 <—> UL_n (où $UL_n = UL_2$, UL_3, UL_4...).

Ce trait reflète, au départ, une propriété de la morphologie, notamment, des langues sémitiques : le passage d'une forme nominale au singulier à son correspondant au pluriel (dans un nombre important de cas, il y a plusieurs pluriels), ou de la forme du verbe à l'accompli à celle de l'inaccompli, du verbe au nom de procès (*masdar*), etc. Ce passage peut avoir lieu, selon la terminologie traditionnelle :

- soit par *«dérivation externe»*, *i.e.* par suffixation ;

  Ex. : ≤MuDaRriS≥, «enseignant» a pour pluriel masculin : ≤MuDaRriS+uWNa≥ (pluriel construit avec le suffixe ≤+uWNa≥). Le participe actif ≤KaATiB≥, «en train d'écrire» ou «ayant écrit» a de même un plur. masc. ≤KaATiB+uWNa≥ ;

- soit par *«dérivation interne»*, *i.e.* par une modification du schème de la base, la racine demeurant inchangée.

  Ex. : UL_1 = ≤**H**aDiY**T**≥, «moderne», de racine √**HDT** et de schème ≤R^1aR^2iYR^3≥ (tel que R^1 = **H**, R^2 = D et R^3 = **T**), a pour pluriels UL_2 = ≤**H**uDa**T**aA?≥ et UL_3 = ≤**H**iDaA**T**≥, de même racine, mais de schèmes différents, respectivement, ≤R^1uR^2aR^3aA?≥, et ≤R^1iR^2aAR^3≥.

On notera que dans les cas de polysémie ou, plus rarement, d'homonymie de la forme correspondant à l'entrée-vedette du dictionnaire, des pluriels distincts sont susceptibles d'être associés à des sens différents : on aura alors, dans la base de données lexicale, des UL distinctes. (Le même phénomène est observable pour la relation entre les verbes de la forme simple et les noms de procès — *masdar-s.*)

> Ex. : En français, on a, pour *ciel,* deux pluriels de sens en partie différents : *cieux* et *ciels.* En arabe, l'unité lexicale $UL'_1 = \leq$ **H**aDiYT$\underline{\text{T}}\geq$, «discours rapporté, récit, conversation (...)», est reliée par fléchage aux pluriels $UL'_2 = \leq ?$ a**H**aADiYT$\underline{\text{T}}\geq$ et $UL'_3 = \leq$ **H**iD$\underline{\text{T}}$aAN$\geq$. Cette unité doit être distinguée de l'$UL_1 = \leq$ **H**aDiYT$\underline{\text{T}}\geq$ présentée dans l'exemple précédent.

On dira qu'il y a *fléchage* entre deux UL (ou plus) lorsque l'une d'entre elles est supplétive de l'autre, c'est-à-dire lorsque l'une d'entre elles remplace l'autre dans les paradigmes morphosyntaxiques d'une manière qui ne peut être prédite au sens strict par des règles opérant en génération, ou qui présente un caractère d'opacité en reconnaissance.

> Ex. : Ainsi, comme on vient de le voir avec l'exemple de $\leq$ **H**aDiYT$\underline{\text{T}}\geq$, pour un même schème de singulier $\leq R^1aR^2iYR^3\geq$ (au demeurant fort fréquent), plusieurs schèmes de pluriels coexistent dans la langue. On ne peut considérer les fléchages que l'on vient de rencontrer comme prédictibles, ni en mode synthèse ni en mode analyse. (Il faudrait pour cela que l'on puisse constater une relation bijective entre schèmes du singulier et du pluriel, ce qui ne se vérifie quasiment jamais en arabe sur l'ensemble d'une série[12].)

Cette définition donne au concept de fléchage une extension plus large que celle de la «dérivation interne». Deux types de cas, non compris dans cette dernière, sont en effet inclus :

- le fléchage entre une unité lexicale coïncidant avec le formant-noyau, et une UL comportant un noyau et un formant-extension lexicalisé, *i.e.* le fléchage de type :

  $UL_1 = <Fn> <\!\!\longrightarrow UL_2 = <Fn,Fel>$ ou, inversement : $UL_1 = <Fn,Fel> <\!\!\longrightarrow UL_2 = <Fn>$.

  Ex. : $\leq$KaATiB$\geq$, «secrétaire», «écrivain» (UL = <Fn>) a pour l'un de ses deux pluriels : $\leq$KaTaB+a&$\geq$, où Fn = $\leq$KaTaB$\geq$ et Fel = $\leq$+a&$\geq$. Inversement, on a : $\leq$ **H**aQiYB+a&$\geq$, «bagage», «valise», (Fn = $\leq$ **H**aQiYB$\geq$; Fel = $\leq$+a&$\geq$) ; pluriel $\leq$ **H**aQaA?iB$\geq$, de structure UL = <Fn> ;

- le fléchage entre bases de racines différentes.

  Ex. : $\leq$?iMRa?+a&$\geq$, «femme», de racine =MR?, a pour pluriel $\leq$NiSaA?$\geq$, de racine =NSW (le fléchage met en relation un Fn_1 relevant d'une racine R_1 avec un Fn_2, de racine R_2). Ce phénomène de supplétion est par ailleurs bien connu : cf. en français : *je vais <—> nous allons*, ou des «paires» telles que *cerf <—> biche ; lièvre (ou bouquin) <—> hase.*

Les relations de fléchage entre unités lexicales telles qu'elles se dégagent du jeu de critères formels dont le schéma est présenté ici fera, bien entendu, l'objet d'une étude séparée.

2.4.2. Fléchages et relations contextuelles

Fléchages et relations contextuelles jouent, au sein de la partie de la grammaire de formants traitant de l'insertion de l'unité lexicale dans le mot, des rôles complémentaires. Le pluriel d'une UL nominale sera ainsi constitué :

- soit de son Fn du singulier, suivi des suffixes du pluriel : le pluriel est alors régi par les *relations contextuelles.*

  Ex. : $\leq$SaAKiN$\geq$, «qui habite» (participe «actif» ; cf. en français, le participe présent «habitant»), pluriel : $\leq$SaAKiN+uWN$\geq$ (avec le suffixe du masc. $\leq$+uWN$\geq$) ;

■ soit d'un Fn de même racine, mais de schème différent (avec ou sans Fel associé) : le pluriel est, ici, obtenu par *fléchage*.

Ex. : La même forme ≤ SaAKiN ≥, quand elle devient un nom à part entière (*i.e.* lorsqu'elle cesse d'être un déverbal), a pour pluriels ≤ SuKKaAN ≥ et ≤ SaKaN+a& ≥ lorsqu'elle prend le sens d'«habitants» (avec *-s* ; cf. également le pluriel nominal *résidents*, la forme participiale étant orthographiée *-ant*), et ≤ SaWaAKiN ≥, lorsqu'elle signifie «segment-lettre quiescent»[13].

3. LES SPÉCIFICATEURS MORPHOSYNTAXIQUES DU DOMAINE DU MOT

3.1. Les deux types de M-spécificateurs

Les *spécificateurs morphosyntaxiques* du domaine du mot (M-spécificateurs) associés aux bases nominales et verbales relèvent en conséquence de deux grands types :

■ l'un «horizontal» (de nature syntagmatique), correspond à l'ensemble des *relations de collocation morphophonologiques et morphosyntaxiques* entre le formant-noyau lexical et les formants-extensions (par ex., les verbes intransitifs sont incompatibles avec les pronoms enclitiques correspondant à des compléments d'objet) ;
■ l'autre «vertical» (de nature paradigmatique), est celui des *fléchages*, eux-mêmes répartis en deux catégories : (a) UL_1 <—> UL_n, comme ci-dessus, mais aussi : (b) BAS_1 <—> BAS_n, *i.e.* entre bases relevant de racines soumises à des transformations (voir ci-dessous, S-4.1).

3.2. Le schéma des *spécificateurs morphosyntaxiques du domaine du mot*

Considérons maintenant la liste fermée des *M-spécificateurs*. Je ne la reproduirai pas ici intégralement, me contentant d'en exposer le schéma et d'analyser ce qui relève en propre de la production et/ou de la reconnaissance. On peut considérer cinq grands types, numérotés ci-dessous de S-0 à S-4 («S» pour «spécificateur»).

S-0 - La catégorie lexicale majeure : bases de noyau lexical soit nominal, soit verbal. (Dans le domaine du mot, cette répartition, par ailleurs traditionnelle dans les sciences médiévales arabes du langage, est suffisante.)

S-1 - La structure de l'*unité lexicale*, qui peut être, comme indiqué, de la catégorie soit UL = < Fn >, soit UL = < Fn,Fel >. Ce spécificateur n'affecte que les bases de noyau nominal (les UL verbales étant toujours du type UL = < Fn >).

S-2 - Des variables affectant la catégorie lexicale majeure, pertinentes pour le traitement du niveau du mot : le paradigme de conjugaison de la base verbale concernée (achevé, inachevé dit «indicatif», «subjonctif» ou «apocopé», impératif, etc.) ; la catégorie de la base nominale, en termes de : sing. (masc., fém.) ; pluriel (masc., fém.) ; collectif ; nom propre, nom commun...

S-3 - Les *relations de collocation*, elles-mêmes réparties en deux sous-catégories :

- les collocations faisant intervenir des formants-extensions lexicalisés (Fel) ;
- les collocations corrélées à la structure morpho- ou sémantico-syntaxique de l'UL. Citons :

• <u>*Bases verbales*</u> (formulation abrégée de : «*bases à noyau lexical verbal*») :

a) Le «numéro de module» identifié par Abu Al-Chay (1988), ou par Ammar et Dichy (à paraître), qui désigne le type du verbe en fonction de son modèle de conjugaison.
b) Le caractère transitif (T) ou non (nT) du verbe, c'est-à-dire sa compatibilité, ou son incompatibilité avec un pronom complément (nécessairement, en arabe, enclitique).
c) La règle générale qui interdit en arabe la cooccurrence d'un complément d'objet de la première ou de la deuxième personne et d'un sujet à la même personne, doit

être levée avec un petit nombre de verbes (traditionnellement: *?af^{c}a:l al-qulu:b* et *?af^{c}a:l al-**hiss***, «verbes de sentiment» et «de perception»).

Ex.: ≤?a+**D**uNn+u#NiY≥ («je me crois»); ≤Ta+RaA#Ka≥ («tu te vois», au sens «mental» du verbe voir).

• *Bases nominales* (formulation abrégée de: «*bases à noyau lexical nominal*»):

- La liste des suffixes (SUF) de cas compatibles avec la base considérée. Cette catégorie de spécificateurs prend en charge les déclinaisons dites «diptotes», etc.

S-4 - Les *fléchages*, qui ont été présentés ci-dessus. Ils peuvent être répartis en deux types généraux.

S-4.1 - Le premier concerne les modifications phonologiques affectant certaines bases en fonction des suffixes ou préfixes qui leur sont associés. Il s'agit du fléchage:

$$BAS_1 <\!\!\longrightarrow BAS_n.$$

• Bases verbales:

Les paradigmes de conjugaison des verbes de racines anomales comportent plusieurs réalisations de la base verbale, chaque réalisation étant compatible avec un sous-ensemble déterminé de la liste des suffixes.

Ex.: Le verbe ≤MaDda≥, «tendre», «étendre», comporte, à l'achevé, deux bases:
BAS_1 = ≤MaDaD+≥ (en graphie non-vocalisée: ≤MDD≥), comme dans ≤MaDaD+Tu≥, ≤MaDaD+Na≥, etc., (suffixes ne commençant pas par une voyelle);
BAS_2 = ≤MaDd+≥ (en graphie non vocalisée: ≤MD≥), comme dans: ≤MaDd+aT≥, ≤MaDd+uW≥, etc., (suffixes commençant par une voyelle).

Chacune de ces bases anomales doit être enregistrée dans le dictionnaire du système, avec l'indication de la liste des suffixes (du «vecteur-suffixe») avec laquelle elle est compatible. L'enregistrement des BAS_2, BAS_3, etc., se fait par génération automatique à partir de la seule saisie de la forme qui désigne verbe dans la grammaire traditionnelle (à partir de Ammar & Dichy, à paraître; v. thèse en cours de N. Gader, Lyon 2). Une telle démarche est due aux grandes difficultés que l'on rencontre lorsque l'on cherche à écrire des règles opérant en reconnaissance, à partir des règles phonologiques, dont le formalisme concorde parfaitement avec la simulation de processus opérant en production: il faut, en effet, clairement distinguer entre la *prédictibilité par les règles en mode synthèse* et le *caractère reconnaissable en mode analyse* (Dichy 1995).

• Bases nominales:

Il s'agit de la modification susceptible d'affecter les bases nominales associées au suffixes ≤+iYY≥ du nom-adjectif de relation. Celle-ci n'est prédictible par des règles (au sens strict) que pour un sous-ensemble du lexique, les grammaires scolaires présentant de longues listes d'exceptions.

Ex.: Dans ≤MaDiYN+a&≥, «ville» et ≤MaDaN+iYY≥, «urbain», la règle de modification du schème s'applique; dans: ≤TaBiY^{c}+a&≥, «nature» et ≤TaBiY^{c}+iYY≥, «naturel», par ailleurs de schème identique, la règle ne s'applique pas. (On peut démontrer que ce phénomène ne dépend nullement de la structure phonologique des noyaux lexicaux.)

S-4.2 – Le deuxième type de fléchage, dont nous avons donné des exemples lors de la présentation de cette notion (§ 2.4.1.), est le fléchage $UL_1 <\!\!\longrightarrow UL_n$.

• Bases verbales : On a principalement :

■ la relation entre l'*achevé* et l'*inachevé* de la forme simple (non augmentée) du verbe (problème de la voyelle de la deuxième radicale) ;

■ la relation entre le verbe et ses dérivés nominaux immédiats, notamment : le nom de procès (*masdar*), dont la forme n'est pas prédictible pour le verbe «non augmenté», ainsi que les participes «actif» (*? ism al-fa:ᶜil*) et «passif» (*? ism al-mafᶜu:l*).

Ex. : Pour le verbe ≤ HaZiNa ≥ «être triste», le lexique n'atteste pas le participe «actif» *≤ HaAZiN ≥, «attristé». On a en revanche une forme adjectivale (*sifa musabbaha*) ≤ HaZiYN ≥. Pour le verbe de sens voisin ≤ Ba ? iSa ≥, «être désespéré», la forme adjectivale attestée ≤ BaA ? iS ≥ («désespéré»), est du même schème ≤ $R^1aAR^2iR^3$ ≥ que le participe actif.

Ces exemples montrent que le lexique doit contenir les informations permettant :

(1) d'attester l'existence d'un participe actif ou d'une forme adjectivale, et
(2) d'indiquer, pour la forme attestée, le schème correspondant.

Si (1) peut être déduit de la valeur sémantique du verbe, (2) relève au sens strict du syntactique, tel qu'il a été défini par Mel'čuk. Les spécificateurs morphosyntaxiques portant sur le point (1) suppléent, comme indiqué (§ 1.4.1.), l'absence de signifiant — au sens que prend ce terme chez l'agent cognitif humain — dans une base de données informatisées.

• Bases nominales :

Il s'agit principalement de relations relevant, comme on l'a vu, de la «dérivation interne», telles que : sing. <—> plur., masc. <—> fém., collectif <—> singulatif, etc., auxquelles il faut ajouter des fléchages faisant intervenir deux racines.

CONCLUSION : LE CARACTÈRE FINI DES M-SPÉCIFICATEURS ET L'INVENTAIRE EXHAUSTIF DES DONNÉES DU NIVEAU DU MOT

Un postulat sous-tend, généralement sur le mode implicite, toute entreprise lexicomatique : celui de la projection de traits en inventaires finis sur les unités en inventaire non fini du lexique. Les entrées d'une base de données lexicales étant de structure <chaîne de caractères, spécificateurs>, le principal problème est, comme indiqué, de dresser un inventaire de spécificateurs qui soit à la fois *fini* et *exhaustif* dans un domaine d'extension contextuel donné.

Par *exhaustif* il faut entendre : (1°) satisfaisant à la principale exigence de la génération (classiquement : engendrer l'ensemble des formes correctes et elles seules), et (2°) qui ne produise, en reconnaissance, que les ambiguïtés permises par la langue (ambiguïtés morphosyntaxiques de construction) dans le domaine d'extension considéré. Dans Dichy (1993 : 70-71), j'ai cherché à montrer que le nombre d'ambiguïtés permises par la langue dans un domaine donné dépend, d'une part, de l'extension de ce dernier (un domaine étroit produisant plus d'ambiguïtés qu'un domaine étendu), et d'autre part, de la richesse du système de connaissances de l'agent cognitif (humain ou informatique) qui procède à la reconnaissance des données linguistiques. En d'autres termes, pour l'analyse automatique du mot, plus le système est capable d'attribuer à une même entrée de valeurs attestables dans la langue concernée, plus il peut être considéré comme performant. Il en ira de même au niveau de la phrase, étant entendu qu'une phrase donnée comporte en principe moins d'interprétations différentes possibles qu'un mot (notamment un mot arabe non vocalisé...).

Or, au schéma des spécificateurs morphosyntaxiques du domaine du mot graphique (M-spécificateurs) présenté ici correspond un inventaire à la fois *exhaustif* et *fini*. Cela a été rendu possible par le raisonnement de type ensembliste suivant : le mot graphique

comprend en arabe deux catégories d'unités, respectivement des formants-noyaux (Fn) et des formants-extensions (Fe), les premières d'inventaires non finis, les secondes d'inventaires finis ; or la structure perceptive de cette unité (§ 2.4.) la contraint à ne comporter qu'une unité lexicale (un Fn) *au plus* ; il convient donc d'*envisager les M-spécificateurs comme la «projection» de traits en nombre fini,* parce que «gérant» un ensemble de relations contextuelles associées à des unités d'inventaires finis (les Fe du mot), *sur chacune des unités en nombre non fini du lexique,* c'est-à-dire *sur chacun des Fn de l'arabe* (lexèmes composés et phrasèmes provisoirement exclus).

Ce raisonnement découle de la structure de l'unité-mot en arabe, qui vérifie le schéma de contextualisation <morphèmes lexicaux — morphèmes grammaticaux> (§ 1.4.). L'étape suivante est celle du domaine de la phrase (synthèse de l'arabe écrit vocalisé, et analyse de l'écrit non vocalisé). Les problèmes y sont d'une autre échelle : s'ajouteront à ce schéma des relations <morphèmes lexicaux — morphèmes lexicaux>. Il faudra donc, si l'on veut parvenir à une analyse syntaxique et au parsage de phrases arabes en écriture non vocalisée, envisager des formalismes relevant de la famille des grammaires d'unification (notamment des grammaires affixales étendues). Mais d'ores et déjà, la base de données lexicale réalisée permet, pour la première fois, une reconnaissance de mots écrits non vocalisés.

ANNEXE SUR LA TRANSCRIPTION DES CARACTÈRES ARABES

La transcription adoptée ici pour des raisons de «portabilité» informatique est librement empruntée à A. Roman (1990, notamment). Les emphatiques et la constrictive vélaire ***ha: ?*** sont en caractères gras ; le soulignement distingue, lorsqu'il y a lieu, la constrictive de l'occlusive correspondante, ou un phonème d'un phonème «voisin» (ainsi, *s* notera le son correspondant au *ch* français : il ne s'agit pas d'une transcription phonémique au sens strict !).

Le présent travail portant sur le traitement automatique de l'écrit, les exemples sont en *représentation graphémique* (Dichy & Hassoun 1990), c'est-à-dire en *translittération* en caractères latins des caractères arabes. Placée entre signes —, cette dernière note au moyen de lettres minuscules les *graphèmes diacritiques* (correspondant à des signes diacritiques habituellement omis dans l'écriture courante), et par des lettres majuscules les *graphèmes de base* (= lettres présentes dans le corps du mot) — sauf pour les signes «spéciaux» *?*, *ᶜ* et *&*. On a :

(1) **VOYELLES BRÈVES** : *a*, ≤a≥ ; *u*, ≤u≥ ; *i*, ≤i≥. (2) **VOYELLES LONGUES** : *? alif* = *a:*, ≤aA≥ (*? alif-madda* = *? a:*, ≤Ã≥) ; *wa:w* = *u:*, ≤uW≥ ; *ya: ?* = *i:*, ≤iY≥. (Une voyelle longue est analysée par l'écriture arabe comme une séquence <voyelle+graphème d'allongement>.) (3) **CONSONNES** (par ordre alphabétique) : *hamza* = *?*, ≤?≥ ; *ba: ?* = *b*, ≤B≥ ; *ta: ?* = *t*, ≤T≥ ; *ṯa: ?* = *ṯ*, ≤Ṯ≥ ; *ji:m* = *j*, ≤J≥ ; ***ha: ?*** = **h**, ≤**H**≥ ; *xa: ?* = *x*, ≤X≥ ; *da:l* = *d*, ≤D≥ ; *ḏa:l* = *ḏ*, ≤Ḏ≥ ; *ra: ?* = r, ≤R≥ ; *za:y* = *z*, ≤Z≥ ; *si:n* = *s*, ≤S≥ ; *ši:n* = *š*, ≤Š≥ ; ***sa:d*** = ***s***, ≤**S**≥ ; ***da:d*** = ***d***, ≤**D**≥ ; ***ta: ?*** = ***t***, ≤**T**≥ ; ***ḍa: ?*** = ***ḍ***, ≤**Ḍ**≥ ; *ᶜayn* = *ᶜ*, ≤*ᶜ*≥ ; *ġayn* = *ġ*, ≤Ġ≥ ; *fa: ?* = *f*, ≤F≥ ; *qa:f* = *q*, ≤Q≥ ; *ka:f* = *k*, ≤K≥ ; *la:m* = *l*, ≤L≥ ; *mi:m* = *m*, ≤M≥ ; *nu:n* = *n*, ≤N≥ ; *ha: ?* = *h*, ≤H≥ ; *wa:w* = *w*, ≤W≥ ; *ya: ?* = *y*, ≤Y≥. (4) **MORPHOGRAMMES** : *? alif maqsu:ra* = *ÿ*, ≤Ÿ≥ ; *ta: ? marbu:ta* = *&*, ≤&≥ (cf. pour ce dernier, D. Cohen 1961 / 70).

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV^es^ Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. Le programme **SAMIA** met en œuvre, à Lyon, une collaboration entre le Centre de Recherche en Terminologie et Traduction (CRTT – Univ. Lumière Lyon-2 ; J. Dichy et X. Lelubre) et le Centre de Recherche en Sciences de l'Information (CERSI - École Nat. Sup. des Sciences de l'Information et des Bibliothèques ; M. Hassoun). Dans le prolongement de ce programme a été conçue et réalisée la première version de la base de données lexicale **DIINAR**, limitée aux traitements du niveau du mot et incluant les spécificateurs morphosyntaxiques présentés ici. Cette base est elle-même le fruit d'une coopération entre les deux partenaires ci-dessus, et, à Tunis, l'Institut de Recherche en Sciences de l'Informatique et des Télécommunications (IRSIT ; S. Ghazali et A. Braham). Une version de DIINAR étendue aux spécificateurs du niveau de la phrase doit être réalisée en collaboration avec l'Université de Nimègue (Institut de langues et cultures du Moyen-Orient, TCMO ; E. Ditters).
2. Le domaine du texte pose des problèmes de délimitation qu'il n'est pas possible d'envisager ici. Il n'est en outre pas situé, du point de vue de sa définition, sur le même plan que le mot et phrase. Ci-après, § 1.4.2.
3. Les spécificateurs ne correspondent pas aux catégories et aux classifications de la linguistique descriptive : ils reprennent celles-ci, mais dans une autre perspective, dont on verra par ailleurs la fécondité, y compris

sur le plan descriptif. En particulier, les spécificateurs, pour des raisons qui seront présentées plus loin, comprennent, côte à côte, des relations relevant de différents niveaux de l'analyse linguistique (morphologie, syntaxe, sémantique).

4. Voir, par exemple, Fradin (1994). Cette commodité de langage est due au fait que les entrées et les sorties des traitements demeurent encore aujourd'hui le plus souvent limitées à l'écrit.
5. Pour la comparaison entre l'agent cognitif humain et l'agent cognitif informatique, v. notamment Le Ny (1989). Pour une réflexion sur la simulation des connaissances linguistiques dans le cadre du programme SAMIA, cf. dans une perspective d'enseignement assisté par ordinateur (EAO) : Dichy & Hassoun (1989) ; Lelubre (1993) ; Dichy (1993), et pour les aspects cognitifs de la simulation des connaissances linguistiques dans le domaine du lexique : Dichy (1995).
6. Pour un état des travaux à la date de publication, en compréhension du langage dans une perspective d'intelligence artificielle, v. G. Sabah (1988 / 89), notamment vol. 1.
7. L'expert humain mobilise en effet l'ensemble des systèmes de connaissance dont il dispose. Des spécificateurs syntaxiques tels que le régime des verbes font appel le plus souvent à des informations sémantiques. Cet aspect fondamental de la méthodologie de la conception des spécificateurs ne pourra être présenté ici plus en détail. Les spécificateurs linguistiques associés aux unités lexicales sont toujours le résultat d'une saisie expert humain —> base de données, qui doit être à la fois modélisée et traduite en interfaces «intelligentes».
8. Cela, même lorsqu'il ne s'agit pas de modèles «génératifs», au sens large du terme.
9. Pour une définition formelle de la phrase avec une application à l'arabe, v. A. Roman (1990). Pour l'analyse automatique des phrases dans cette langue, v. Ditters (1992).
10. C'est-à-dire, pour la *totalité* des verbes et des dérivés verbo-nominaux immédiats (nom verbal, participes «actif» et «passif»), ainsi que pour une partie importante des noms (Dichy 1990).
11. Le mot graphique est dit ici *susceptible* d'inclure un noyau lexical en raison de la présence d'un sous-ensemble de mots graphiques ne comportant pas un tel noyau : il s'agit principalement de mots-outils, d'inventaire limité ; ex. : ≤ ? iN ≥, «si (conditionnel)» ; ≤ FiY ≥, «dans», ≤ Li# ? aNna#Hu ≥, «parce-que-il», etc.
12. Cette observation (et quelques autres) infirme la possibilité de recourir à des transformations faisant intervenir des constituants de nature prosodique pour le traitement automatique du pluriel brisé en arabe. Les analyses, par ex., de McCarthy & Prince (1990) (encore récemment citées par B. Fradin 1994 : 39-41), même en corrigeant les erreurs portant sur les faits qu'elles contiennent, ne présentent pas un niveau suffisant d'adéquation aux données pour servir de base au traitement automatique.
13. *I.e.* prononcé sans voyelle (ce qui est le cas en finale de syllabe). Dans les sciences médiévales arabes du langage, un même terme, ***harf***, désigne, notamment, le segment (phonologique ou non) et la lettre, d'où, ci-dessus «segment lettre» (Dichy 1990).

RÉFÉRENCES

ABEILLÉ, A. (1993) : *Les nouvelles syntaxes. Grammaires d'unification et analyse du français*, Paris, Armand Colin.

ABU AL-CHAY, N. (1988) : *Un système expert pour l'analyse et la production des verbes arabes dans une perspective d'enseignement assisté par ordinateur*, Thèse en sciences de l'information, Université Lyon-1.

AMMAR, S. et J. DICHY (à paraître) : *Al-ka:mil fi: t-tasri:f, 12.000 verbes arabes, formes et emploi.*

COHEN, D. (1961 / 70) : «Essai d'une analyse automatique de l'arabe», 1961, *TA information*, *in* D. Cohen, *Études de linguistique sémitique et arabe*, Paris, Mouton, 1970.

DESCLÉS, J.-P. *et al.* (1983) : *Conception d'un synthétiseur et d'un analyseur morphologique de l'arabe, en vue d'une utilisation en enseignement assisté par ordinateur*, Rapport rédigé à la demande du Ministère français des affaires étrangères (sous-direction de la politique linguistique).

DICHY, J. (1987) : «The SAMIA Research Program, Year Four, Progress and Prospects», *Processing Arabic Report*, n° 2, T.C.M.O., Université catholique de Nimègue, pp. 1-26.

DICHY, J. (1990) : *L'écriture dans la représentation de la langue : la lettre et le mot en arabe*, Thèse d'État en linguistique, Université Lumière Lyon-2.

DICHY, J. (1993) : «Knowledge-system Simulation and the Computer-aided Learning of Arabic Verb-form Synthesis and Analysis», *Processing Arabic Report*, 6 / 7, T.C.M.O., Université Catholique de Nimègue, pp. 67-84, 92-95.

DICHY, J. (1995) : «Simulation informatique du langage naturel et compatibilité connaissances-processus : la modélisation des connaissances du niveau du mot en arabe», *Journée scientifique sur la génération et l'analyse morphologique de l'arabe*, Rabat, le 20 juin 1995, Institut d'études et de recherches pour l'arabisation. (à paraître)

DICHY, J. et M. O. HASSOUN (dir.) (1989) : *Simulation de modèles linguistiques et enseignement assisté par ordinateur de l'arabe — travaux SAMIA I*, Paris, Conseil international de la langue française.

DITTERS, E. (1992) : *A Formal Approach to Arabic Syntax: The Noun Phrase and the Verb Phrase*, PhD, Catholic University of Nijmegen.

FRADIN, B. (1994) : «L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie», B. Fradin (dir.), *La morphologie computationnelle, TAL (Traitement automatique des langues)*, 34 (2), Paris, Association pour le traitement automatique des langues, pp. 9-48.

HASSOUN, M. O. (1987) : *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*, Thèse d'État en sciences de l'information, Université Lyon-1.

LELUBRE, X. (1993) : «Courseware for the Theory and Practice of Arabic Conjugation», *Processing Arabic Report*, 6 / 7. T.C.M.O., Université catholique de Nimègue, pp. 85-89, 92-95.

LE NY, J.-F. (1989) : *Science cognitive et compréhension du langage*, Paris, PUF.

LYONS, J. (1978 / 90) : *Sémantique linguistique*, 1re édition 1978, traduction française de J. Durand et D. Boulonnais, Paris, Larousse, 1990.

McCARTHY, J. J. and A. S. PRINCE (1990) : «Foot and Word in Prosodic Morphology: the Arabic Broken Plural», *Natural Language and Linguistic Theory*, 8 (2), pp. 209-283.

MEL'ČUK, I. A. (1982a) : *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology*, München, Wilhem Fink.

MEL'ČUK, I. A. (1982b) : «Élaboration d'un langage formel pour la morphologie», C. Bertaux, J.-P. Desclés, D. Dubarle *et al.*, *Linguistique et mathématiques. Peut-on construire un discours cohérent en linguistique ?*, Berne, Peter Lang, pp. 99-119.

ROMAN, A. (1990) : *Grammaire de l'arabe*, Paris, PUF, coll. «Que sais-je ?».

SABAH, G. (1988 / 89) : *L'intelligence artificielle et le langage*, vol. 1 : *Représentation des connaissances* (1988), vol. 2 : *Processus de compréhension* (1989), Paris, Hermès.