

**SYRIAN COMPUTER SOCIETY**

**Committee of Damascus**

**ARABIC CONTENT ON THE INTERNET WORKSHOP**

**Damascus 13-14 April 2008**

**SOCIÉTÉ SYRIENNE D'INFORMATIQUE**

**Comité de Damas**

**ATELIER SUR LES CONTENUS ARABES SUR LA TOILE**

**Damas-Syrie 13-14 Avril 2008**

**Traitement de l'arabe écrit et Web arabe : l'apport de  
l'équipe lyonnaise SILAT (Systèmes d'information,  
Ingénierie, Linguistique arabes et Terminologie)**

**Par**

**Mohamed HASSOUN\*, Joseph DICHY\*\* et Ramzi ABBES\*\***

\*ENSSIB-Lyon (École nationale supérieure des Sciences de l'Information et des Bibliothèques)

\*\*Université Lumière Lyon 2 et ICAR (Interaction, Corpus, Apprentissages, Représentations – CNRS/Univ. Lyon 2-ENS-LSH)

**English title:** Arabic NLP and the Arabic Web: the contribution of the SILAT (Systèmes d'information, Ingénierie, Linguistique arabes et Terminologie) research group in Lyon

**English abstract :**

The SILAT (“Systèmes d'information, Ingénierie, Linguistique Arabes et Terminologie”) research group in Lyon has been working on Computational Arabic in a multilingual perspective since the early 1980ies. The group is headed by Joseph Dichy (Université Lumière Lyon 2 and ICAR-CNRS Lab), Mohamed Hassoun (ENSSIB-Lyon) and Xavier Lelubre (Université Lumière Lyon 2 and ICAR-CNRS Lab). The methodology is based on linguistic analysis as well as computational expertise. The research has been centered on word-form (or morphological) generation and analysis, with special reference to the analysis of unvowelled Arabic writing.

The research has shown that both analysis and generation of Arabic word-forms could only be performed with success through a comprehensive lexical database, in which entries are associated with *morphosyntactic specifiers*. The DIINAR.1 lexical resource, which has been completed in the 1990ies, is described. A number of analysis and generation software have been built. The last-borne is the MorphArab analyser, which is integrated in the AraConc concordance software.

The last part of this contribution concentrates on the issue of information retrieval and queries in Arabic. It shows that surface analysis cannot bring satisfactory results, due to the structure of Arabic writing (unvowelled script; the agglutinative structure of words). Examples are given.

In order to meet the issues related to the development of Arabic contents on the Web, the tools built by the SILAT Lyon research group can bring a momentous contribution.

# Sommaire

## 1- Introduction

1.1- Présentation de l'équipe (recherche et formation)

1.2- Historique sommaire et réalisations

## 2- Traitement automatique de la langue arabe : réalisations et perspectives

2.1- La ressource linguistique DIINAR.1 et la démarche sous-jacente

2.1.1- La conception et la réalisation de la ressource lexicale DIINAR.1

2.1.2- Description générale

2.1.3- Quelques statistiques générales

2.1.3a. *La base de données des verbes et des déverbaux*

2.1.3b. *La base de données des noms*

2.1.3c. *Le prototype d'une base de données des noms propres*

2.1.3d. *Une base de données des mots outils*

2.2. La base de données terminologique OPTAR

## 3- L'analyse morphologique de l'arabe

3.1- Le mot graphique en arabe (rappel)

3.2. MorphArab, un Analyseur morphologique automatique de l'arabe

3.3. AraConc : un concordancier électronique de l'arabe

## 4. La langue arabe sur le web

4.1. Couverture des moteurs de recherche

4.2. Dissymétrie de l'indexation et de la recherche en langue arabe

4.3. Recherche d'information en langue arabe

4.4. Apport de l'analyse linguistique pour la recherche d'information en langue arabe

4.5. Insuffisance du parcours de surface ; mots homographes

4.5.1. Relations entre le singulier et le pluriel des noms

4.5.2. Relations entre le masculin et le féminin des noms

4.5.3. Pratiques d'écriture courantes

4.5.3a. *Hamza et 'alif*

4.5.3b. *Yâ' et 'alif maqṣūra*

4.5.3c. *Le caractère ' (kashida)*

4.5.3d. *L'absence des signes de vocalisation*

4.5.3e. *Ta marbouta ة et ha ه*

## 5. Conclusion

Principales références et thèses liées à DIINAR.1 ou au groupe SILAT

# 1- INTRODUCTION

## 1.1- Présentation de l'équipe (recherche et formation)

Équipe pluridisciplinaire formée de linguistes, d'informaticiens et de spécialistes des SIC (Sciences de l'Information et de la Communication), le groupe de recherche inter-établissements « **Systèmes d'information, Ingénierie, Linguistique Arabes et Terminologie** » (**SILAT**) accueille depuis le début des années 1990, annuellement, 10 à 15 étudiants en DEA et doctorants appartenant, respectivement à l'École Nationale Supérieure de Sciences de l'Information et des Bibliothèques (ENSSIB, Villeurbanne) et à l'université Lumière-Lyon 2. Ce groupe est co-dirigé par Joseph Dichy, professeur de linguistique arabe à Lyon 2, Mohamed Hassoun, professeur de sciences de l'information à l'ENSSIB et Xavier Lelubre, maître de conférences directeur de recherches en linguistique et terminologie de l'arabe à Lyon 2. Il donne lieu à un séminaire de recherche doctorale. Une douzaine de thèses ont été soutenues au cours de la décennie écoulée dans le cadre d'une co-direction entre J. Dichy et M. Hassoun, soit en linguistique arabe, soit en sciences de l'information (voir le site, encore en construction à l'heure actuelle <http://.silat.univ-lyon2.fr>).

L'équipe a créé, depuis 2006, une formation de niveau Master 2 intitulée « **Systèmes d'Information Multilingues et Ingénierie Linguistique** » (**SIMIL**), commune à l'Université Lumière-Lyon 2 (où elle est intégrée au Master de Langues étrangères appliquées, LEA) et à l'ENSSIB (où elle est intégrée au Master de Systèmes d'Information et des Bibliothèques, SIB). Cette formation, qui est co-dirigée par Mohamed Hassoun (ENSSIB) et Xavier Lelubre (Lyon 2), et a pour objet d'accueillir des étudiants ayant une formation en langues d'une part, et en informatique ou en sciences de l'information d'autre part. Elle est la seule formation de ce type en France et sans doute en Europe, qui soit centrée sur les langues arabe, anglaise et française.

La démarche de notre équipe est caractérisée par un équilibre entre linguistes et informaticiens, qui résulte de longues années de collaboration dans le domaine du traitement automatique de l'arabe. Les premiers ont acquis une expérience de la formalisation et de l'élaboration de représentations et de règles susceptibles d'être utilisées par des informaticiens, les seconds ont acquis une expertise réelle des problèmes linguistiques et des données et structurations complexes du langage naturel.

## 1.2- Historique sommaire et réalisations

Nous travaillons, depuis le début des années 1980 sur des problématiques relevant du traitement automatique de l'arabe (TAL arabe), dans une perspective unilingue comme dans une perspective plurilingue incluant le français et l'anglais. Nous avons réalisé un ensemble de ressources pour le traitement de l'arabe sur lesquels nous avons bâti une collection d'outils informatiques pour le traitement de la langue écrite.

Notre approche concerne essentiellement :

- les bases de données lexicales, notamment, dans le domaine du **vocabulaire général**, la ressource lexicale **DIINAR.1**, « **Dictionnaire INformatisé de l'Arabe, version 1** », et en **terminologie scientifique**, la base de données **OPTAR**, qui porte sur la terminologie scientifique dans le domaine de l'optique,
- la génération et l'analyse automatiques de l'arabe, tant au niveau du mot qu'à celui de la phrase,
- l'analyse et l'indexation de corpus arabes,

- l'enseignement de l'arabe avec les nouvelles technologies de l'information et de la communication (TIC) [SAMIA, 1984], [Zaafrani, 1997, 1998a et 1998b, 2002].

Les deux premiers points seront développés ci-dessous. En ce qui concerne le « Web arabe », ces outils permettent aujourd'hui l'utilisation des techniques du TAL pour la recherche d'information (RI) et l'analyse approfondie des corpus arabes, au moyen d'outils de fouille de texte [Abbès et Dichy, 2008]. Dernièrement nous avons lancé une thèse sur la classification et la recherche en utilisant des méthodes d'apprentissage automatique. Lors des derniers stages et contrats post-doctoraux réalisés dans notre laboratoire nous avons développé de premières approches de l'utilisation de grammaires de surfaces pour l'extraction des entités nommées et pour l'utilisation d'algorithmes de mesure de distance entre chaînes pour l'identification de toutes les variantes d'une entité nommée arabe en translittération.

La principale caractéristique de notre approche est, comme on le verra, d'être compatible avec les deux démarches de la génération et de la reconnaissance automatique de l'arabe. Cette double perspective renforce considérablement les possibilités de fouille de texte et de recherche d'information en arabe.

L'équipe SILAT est, ou a été présente dans plusieurs projets ou réseaux européens portant sur la langue arabe (DIINAR-MBC, coordonné par nous, ALMA, NEMLAR, MEDAR).

## **2- TRAITEMENT AUTOMATIQUE DE LA LANGUE ARABE : REALISATIONS ET PERSPECTIVES**

### **2.1- La ressource linguistique DIINAR.1 et la démarche sous-jacente**

#### ***2.1.1- La conception et la réalisation de la ressource lexicale DIINAR.1***

La ressource linguistique DIINAR.1 a été conçue et réalisée en commun sur deux sites : à Lyon (Joseph Dichy, université Lumière-Lyon 2, aspects linguistiques, et Mohamed Hassoun, ENSSIB, aspects informatiques) et à Tunis (IRSIT « Institut de Recherche en Sciences de l'Informatique et des Télécommunications de Tunis », Abdelfattah Braham, université de la Manouba et Salem Ghazali, Institut Supérieur des Langues de Tunis). La saisie des données a eu lieu pour l'essentiel à Tunis dans les années 1992 à 1997, avec des interfaces de filtrage et de saisie-consultation conçues à Lyon et expérimentées et améliorées en commun. Elle constitue l'une des ressources qui ont été à la base du projet européen DIINAR-MBC. La diffusion-valorisation de cette ressource est actuellement proposée via ELDA ("The Evaluation and Language resources Distribution Agency", Paris – [www.elda.org](http://www.elda.org)).

**DIINAR-MBC** ("Dictionnaire INformatisé de l'ARabe, Multilingue et Basé sur Corpus") est un projet soutenu par la Commission européenne (projet n° 961791 du programme de Coopération avec les Pays Tiers et les Organisations Internationales – INCO-DC). La durée de ce projet, qui s'est achevé en décembre 2000, était de 30 mois. La coordination scientifique a été assurée par J. Dichy, Université Lumière-Lyon 2 (avec la participation de X. Lelubre), assisté par EZUS-Lyon 1 pour les aspects administratifs et de gestion. DIINAR-MBC avait pour autres partenaires l'École Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB, France - M. Hassoun), l'Electronics Research Institute (ERI, Égypte - N. Hegazi), l'Institut d'Etudes et de Recherche pour l'Arabisation (IERA, Maroc - A. Fassi-Fehri), l'Institution Régionale des Sciences Informatiques et des Télécommunications (IRSIT, Tunisie - A. Braham, S. Ghazali) et l'Université de Nimègue (Pays-Bas - E. Ditters). **Résultats principaux** : un analyseur morphosyntaxique de haut niveau de performance [OUERSIGHNI, 2002] ; un ensemble d'interfaces et de lexiques (dont un prototype de lexique bilingue arabe-français et arabe-anglais) ainsi que des procédures de traitement et

d'indexation des données textuelles ou des corpus. Un corpus de 10 millions de mots a été compilé à l'université de Nimègue ainsi qu'à l'IRSIT (Tunis).

### 2.1.2- Description générale

La base de données DIINAR.1 est une ressource lexicale très importante pour une application de traitement automatique de la langue arabe. Elle est très riche en informations syntaxiques et morphologiques (pour le modèle linguistique voir notamment [Dichy, 1984, 1990, 1997], [Dichy, Braham, Ghazali, Hassoun, 2002], [Dichy et Hassoun, 2005] pour les aspects informatiques, voir notamment [Hassoun, 1987], [Abu Al Chay, 1988], [Gader, 1992] [Ghenima, 1998]). Sur le plan informatique DIINAR.1 répond à la définition d'un modèle relationnel :

- définition des relations entre les tables,
- contrainte d'intégrité des données,
- limitation de la redondance des données,
- évolution rapide du contenu et de la structure,
- maintenance facile du contenu et de la structure,

La plupart des informations sont codées :

- pour garantir leur sécurité,
- pour limiter sa taille à un volume raisonnable et assurer un accès rapide (dans la mesure où nous réduisons le temps d'exécution des requêtes).

Trois traits généraux caractérisent en propre la base de données DIINAR.1, et la distinguent d'autres bases de données existantes :

(1) Les entrées lexicales y sont associées à des *spécificateurs morpho-syntaxiques* qui garantissent une insertion conforme au fonctionnement de la langue de ces entrées dans l'unité mot (dont le schéma est rappelé plus loin) [Dichy, 1987, 1997, 2000] [Dichy et Farghaly, 2007].

(2) L'ensemble des sorties des lexiques générés à partir de DIINAR.1 correspond à des *unités effectives de la langue*. Les spécificateurs morphosyntaxiques opèrent comme un filtre qui empêchent par exemple la génération de formes correspondant à un verbe intransitif (tel que *nâma* نام "dormir") avec des pronoms complément (ainsi : \**nâma-ka* نامك "il t'a dormi", qui n'existe pas, n'est pas généré). De même, les entrées de DIINAR.1 ont été effectuées à partir de mots effectivement attestés par les dictionnaires arabes, en non par "génération aveugle" à partir d'une liste de racines et d'une liste de schèmes. ([Hassoun, 1987], [Dichy, 1987], [Dichy et Hassoun, éd., 1989]. Pour des données chiffrées, voir [Abbès, Dichy et Hassoun, 2004 et 2005].)

(3) Les règles de formation du mot graphique, ainsi que les spécificateurs morphosyntaxiques associés aux entrées de la base sont conçus pour répondre aux besoins des deux démarches asymétriques de la génération (ou de la synthèse) et de l'analyse [Desclés, éd. 1983], [Dichy, 1984, 1987, 1997], [Bouché, Dichy et Hassoun, 1984]. On dira que DIINAR.1 a été conçue pour être compatible avec les deux processus de la synthèse et de l'analyse (selon le *principe de compatibilité connaissances-processus* [Dichy, 1990]).

### 2.1.3- Quelques statistiques générales

La ressource linguistique DIINAR.1 est composée de plusieurs base de données (verbale, nominale, noms propres et mot-outils) dont voici le bilan total [Abbes, Dichy et Hassoun, 2004a].

	Noyaux ou lemmes	LEXIQUES GÉNÉRES des mots minimaux (formes fléchies)
Verbes	19 457	3 050 715
Dérivés nominaux immédiats (déverbaux)	70 702	2 909 772
Noms (pluriels brisés inclus)	39 099	1 781 316
Noms propres (prototype)	1 384	11 403
Mots outils (chiffres provisoires)	445	11 731
<b>Total</b>	<b>131 087</b>	<b>7 764 937</b>

Figure 1: **Composition de DIINAR.1 (lemmes et formes fléchies)**

#### 2.1.3a. La base de données des verbes et des déverbaux

Elle comporte 19.457 verbes, l'ensemble de ces verbes se conjuguent en suivant l'un des 125 modèles de conjugaison existant (pour une explication de cette modélisation, voir [Dichy, 1993], [Ammar et Dichy, 1999]).

À chaque verbe nous associons une liste de déverbaux (la définition de ce terme correspond, dans les choix de DIINAR.1, à la liste ci-dessous. Elle sera revue dans les versions ultérieures, dans lesquelles les « noms de temps et de lieu » devront, pour un état post-classique de la langue arabe, être intégrés aux noms, et les « adjectifs analogues », être insérés dans une base de données consacrée aux adjectifs). Chaque déverbal suit un modèle de déclinaison. Voici un tableau récapitulatif de la composition de la base des verbes et des déverbaux.

	Noyaux ou lemmes	LEXIQUES GÉNÉRES des mots minimaux (formes fléchies)
Verbes	19 457	3 050 715
Dérivés nominaux immédiats (déverbaux) au total :	[70 702]	[2 909 772]
Noms de procès (مصدر)	23 274	418 815
Participes actifs (اسم الفاعل)	17 904	939 816
Participes passifs (اسم المفعول)	13 373	722 142
Noms de temps et de lieu (اسم الزمان والمكان)	5 781	290 178
Adjectifs analogues (صفة مشبهة)	10 370	538 821
<b>Total</b>	<b>90 159</b>	<b>5 960 487</b>

Figure 2: **Composition de la base de données verbale (déverbaux inclus)**

### 2.3.1b. La base de données des noms

La base de données des noms comprend un total de 39.099 unités. Le nombre de modèles de déclinaison peut être ramené à 11.

Chaque nom est associé à un formant-extension au moins. Cette combinaison peut produire de nouvelles entrées de la base en lexiques générés. Citons pour exemple l'ajout du *ياء النسبة* *yâ' al-niba'* « le *yâ'* de relation », suivi des suffixes de cas (إعراب). Voici un tableau récapitulatif.

	Noyaux ou lemmes	LEXIQUES GÉNÉRÉS des mots minimaux (formes fléchies)
Noms	29 494	265 428
Pluriels Brisés	9 565	86 085
Noms masculins "par nature"	20	180
Noms féminins "par nature"	22	198
Noms avec formants-extension lexicalisé [Dichy, 1990, 1997]	---	1 429 425
<b>Total</b>	<b>39 099</b>	<b>1 781 316</b>

**Figure 3: Composition de la base de données des noms**

### 2.3.1c. Le prototype d'une base de données des noms propres

Cette base comporte les noms propres qui n'existent pas dans la base des noms. Plusieurs des noms propres existant en arabe correspondent également à des formes fléchies de noms (ainsi le prénom : *Husâm* حسام, correspond aussi à « fil de l'épée ») ou de déverbaux (*Kâtib* كاتب, prénom ou participe actif de *kataba*, « écrire »). Ces derniers cas sont signalés dans la base de données des noms par le spécificateur « nom propre ».

Cette base est actuellement à l'état de prototype. Elle ne comporte qu'un échantillon de 1.384 entrées, à partir desquelles on peut générer 11.403 mots minimaux (à partir d'une grammaire du nom propre réalisée par J. Dichy et non publiée à la date d'aujourd'hui).

### 2.3.1d. Une base de données des mots outils

C'est la dernière réalisation en date de la boîte à outils de DIINAR.1. Elle n'est pas tout à fait finie et elle comporte pour l'instant 445 mot outils, qui seront réduits, une fois éliminés les doublons dus à la sur-catégorisation des grammaires arabes médiévales à partir desquelles nous avons travaillé, à un nombre situé autour de 200. Comme les noms ou les verbes, les mots-outils se combinent avec des clitiques (qui sont eux-mêmes des mots-outils liés). Les possibilités théoriques de combinaison des mots-outils et des clitiques donnent 11.731 formes différentes. Ce chiffre, qui est indicatif, sera bien évidemment revu.

## 2.2. La base de données terminologique OPTAR

Parallèlement aux travaux qui ont donné lieu à la ressource linguistique DIINAR.1, Xavier Lelubre a réalisé une base de données terminologique de l'arabe dans le domaine scientifique de l'optique (Lelubre, 1992). Le nom de cette base de données est **OPTAR**.

OPTAR comporte environ 5 000 entrées en langue arabe, extraites de corpus, et indexées. Elles sont accompagnées de leurs correspondants en français et en anglais, et associées à une ontologie du domaine. Aux termes sont également associés un jeu de *spécificateurs terminologiques*, qui incluent des informations morphologiques, syntaxiques et phraséologiques [Lelubre, 2001, 2002].

Il s'agit, en terminologie, d'une base de données pilote, qui doit, dans une deuxième étape, être associée à DIINAR.1, et constituer avec elle une hyperbase [LABED et LELUBRE, 1997].

## 3- L'ANALYSE MORPHOLOGIQUE DE L'ARABE

Si l'on se limite à l'analyse morphologique (en laissant de côté ici les aspects liés à la génération [GHENIMA, 1998], [ZAAFRANI, 2002]), on peut observer que toute entreprise d'analyse morphologique de l'arabe peut être ramenée à une reconnaissance des composants du mot – c'est le volet *segmentation* –, et à l'identification du rôle de chacun des composants – c'est le volet *étiquetage*. Une autre fonction est celle de l'attestation de l'appartenance du mot à la langue [Gader, 1992]. Cette entreprise se heurte à deux difficultés majeures : il faut premièrement identifier avec précision les frontières des morphèmes et deuxièmement s'assurer de la validité de la segmentation en se référant à une ressource lexicale. La ressource lexicale DIINAR.1 a été conçue pour répondre à ce besoin.

### 3.1- Le mot graphique en arabe (rappel)

L'analyse morphologique en arabe s'intéresse, comme les autres langues, aux formants du mot. Étant donné la richesse des informations incluse dans le mot graphique, un modèle linguistique de celui-ci a été élaboré, et a servi de base, d'un point de vue linguistique, à l'élaboration de DIINAR.1 [Dichy et Hassoun, 2005].

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), enclitiques), 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 de la même famille que l'arabe<sup>1</sup>, s'analyse en une *racine (RAC)* et un *schème (SCH)*. On notera toutefois, qu'un sous-ensemble important des noms ne peut être analysé ainsi. Ces noms correspondent à des *pro-bases (PBA)*.

Ex. : *yâsimîn*, “jasmin”; *'Ibrâhîm*, “Abraham”, etc.

---

<sup>1</sup> 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].

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*).

On peut représenter le mot ainsi :

<p><b>Représentation “classique” en constituants immédiats</b></p> <p>Exemple (sommaire) :</p>	<pre> graph TD     MM[mot maximal] --- PCL[PCL]     MM --- ECL[ECL]     MM --- MM2[mot minimal]     MM2 --- PRF[PRF + {BAS ou PBA}]     MM2 --- SUF[+SUF]     PRF --- Li[Li]     PRF --- Ta[Ta]     SUF --- KT[B + KT]     SUF --- uW[+uW]     Li --- Li2["pour que"]     Ta --- Ta2[pronom 2e pers.]     KT --- KT2["écrire inaccompli (subjunctif)"]     uW --- uW2[plur. masc.]     ECL --- Hu[Hu]     Hu --- Hu2["lui pronom complément"]     </pre>
<p><b>Représentation faisant apparaître la saillance du noyau lexical</b></p>	<pre> graph TD     Fn[Fn] --- Fe1[Fe]     Fn --- Fe2[Fe]     Fe1 --- Fe3[Fe]     </pre>

Figure 4: Le schéma du mot (d’après [Dichy, 1997])

**Légende du schéma :**

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 : 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. 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 et Hassoun, éd. 1989, “Lexique de définitions”, p. 265-76] et [Dichy 1990, chap. X].

**3.2. MorphArab, un Analyseur morphologique automatique de l’arabe**

Pour procéder à l’analyse ou à la génération des mots graphiques, le système doit être assisté d’un lexique pour éviter le analyse ou la génération d’unités théoriquement possibles mais inexistantes dans la langue. Par la suite l’analyseur doit donner la liste des traits associés au mot en entrée [Dichy, 1984]. Une ressource lexicale est, comme indiqué, pour cela nécessaire [Hassoun, 1987].

Sur le plan informatique, MorphArab, le dernier-né des analyseurs morphologiques conçus à partir de DIINAR.1 [Gader, 1992], [Ghenima, 1998], [Zaafrani, 2002], [Ouersighni, 2002], a été développé en Orienté Objet pour répondre à la composition et la décomposition du mot graphique en parties autonomes et indépendantes [Abbès, 2004]. Il a été choisi de construire un analyseur morphologique modulable pouvant se greffer sur toute application de traitement automatique faisant appel à une telle ressource.

Entièrement développé en orienté objet, ce programme informatique est donc une collection de classes indépendantes et d’objets. AraMorph identifie la plupart des traits

morpho-syntaxiques inclus dans le mot graphique. Certains traits augmentent considérablement le nombre de propositions. La moins ambiguë des marques morpho-syntaxiques est la racine, qui fédère le plus grand nombre d'analyses ([Abbès, 2004]; pour une approche psycholinguistique, voir [Grainger et al., 2003]).

### 3.3. AraConc : un concordancier électronique de l'arabe

L'analyseur MorphArab est intégré dans un autre logiciel, le concordancier électronique AraConc [Abbès, 2004]. La réalisation d'une concordance est une entreprise laborieuse et complexe. Elle nécessite un travail préalable très important notamment la collecte des textes et l'administration de certains pré-traitements toujours dépendants de la nature du corpus.

Un mot arabe dans une concordance est indissociable de son analyse qui reste le seul moyen de distinguer les lexèmes homographes (mots différents partageant une même réalisation graphique). La résolution des problèmes d'analyses multiples reviendra le plus souvent à l'utilisateur. Envisager des interactions à posteriori avec l'utilisateur exige la mémorisation de la position de chaque terme. Ainsi la concordance finale tournera autour du triplet : mot – analyse – position.

La complexité du problème est augmentée par le nombre et la dimension des traits morpho-syntaxique susceptibles d'être portés par un mot graphique arabe donné, situé entre deux espaces blancs ou marqueurs comparables. Cette richesse rend le mot plus ambigu et atteste définitivement de l'échec de toutes les techniques de parcours de surface des mots en vue de l'extraction des contextes en arabe [Abbès et Dichy, 2008].

AraConc peut fusionner plusieurs fichiers physiques dans une seule et même concordance. La position correspond donc à un couple comprenant (a) les coordonnées du mot objet de la requête dans le fichier du texte auquel il appartient et (b) et la localisation de ce fichier sur le disque.

## 4. LA LANGUE ARABE SUR LE WEB

Les outils que l'on vient de décrire constituent des atouts majeurs pour le traitement de l'arabe sur le Web.

### 4.1. Couverture des moteurs de recherche

Les robots des moteurs de recherche (*spiders* ou *crawlers*) parcourent les sites de la toile, à intervalles réguliers. **L'exploration** est indépendante de l'alphabet, elle dépend surtout des performances en termes de couverture de chacun des moteurs. La problématique de recherche d'information dépend de deux facteurs : le premier concerne l'indexation des pages et le second est lié à la recherche dans les index ou dans les pages elles mêmes.

**L'indexation** des pages web se fait, pour l'essentiel des moteurs de recherche, par l'une ou la combinaison des méthodes suivantes :

- La récupération des balises « méta » contenant les mots clés décrivant le contenu des pages et proposés par le créateur du site.
- La récupération du contenu de la balise « titre », il est d'ailleurs recommandé de donner des titres différents à chacune des pages web du site pour avoir un maximum de chance de sortir parmi les résultats du moteur.
- Pour les ressources jugées importantes les robots peuvent indexer tout le contenu de la page.

La **recherche** est la partie *secrète* des robots. Les algorithmes pondèrent les pages en fonction d'un ensemble de critères comme la position du mot dans la page (titre, paragraphe, lien hypertexte) en fonction de l'historique ou encore de la nature de la ressource.

Les webmasters de leur côté positionnent les mots pertinents pour leurs sites dans les endroits stratégiques pour le robot. Cet aspect n'a pas de pertinence par rapport aux langues utilisées, mais il est pertinent selon le traitement *linguistique* possible dans chacun des moteurs, en fonction de la maîtrise de la lemmatisation implémentée dans le moteur, de la dérivation, ou d'ontologies permettant de relier les mots de la recherche aux mots-clés proches sémantiquement ou faisant partie de la même famille morphologique ou lexicale.

## 4.2. Dissymétrie de l'indexation et de la recherche en langue arabe

Si l'on considère les schémas classiques de l'indexation d'une part et de la recherche de l'autre, on constate une dissymétrie dans le traitement de l'écrit arabe selon ces deux démarches, en raison de l'absence des voyelles dans les textes arabes courants et de la nature agglutinante de cette langue.

Par exemple, lors de l'indexation d'un document, on peut indexer le verbe « écrire » (كُتِبَ), le nom « livre » (كِتَابٌ) et le nom de procès (مصدر) "fait d'écrire" (كُتِبَ) sous une seule et même entrée كُتِبَ, car ils ne sont généralement pas vocalisés dans le texte. Il en est de même pour le mot شعر dont les différentes formes vocalisées possibles ont des significations différentes (« sentir », « poésie », « cheveux », etc.). Autre exemple : le mot graphique علم peut désigner plusieurs sens (« drapeau », « science », « connaître », etc.).

Quelles que soient les précisions apportées à la recherche (même si on note le mot entièrement vocalisé), le moteur ne pourra pas distinguer ces formes, car elles ne sont pas vocalisées dans les textes. Par conséquent, les mots de l'index de sont pas non plus vocalisés.

L'agglutination consiste, en figure simplifiée, en l'augmentation de la forme minimale du mot par des proclitiques (pour signifier l'interrogation, la ressemblance, la liaison, etc.) ou des enclitiques (pour rajouter notamment des pronoms). Dans les trois exemples suivants, divers cas d'agglutination affectent les mots auxquels renvoie la réalisation graphique كاتب : (1) « Kateb Yassine » كاتِبِ يَسِينِ, (2) « est-ce que Kateb Yassine...? » أَكَاتِبِ يَسِينِ, (3) « j'écris à Yassine » أَكَاتِبِ يَسِينِ.

## 4.3. Recherche d'information en langue arabe

Une grande partie des requêtes sur le web, indépendamment des langues, concernent des entités nommées tels que des noms propres. Nos tests sur un échantillon de 2850 requêtes arabes sur un annuaire [M. Boualem et al., 2001] nous ont permis de constater que 94,2% des requêtes concernent des formes nominales, 3,30% concernent des formes verbales et 2,5%, des mots grammaticaux. Bien entendu, ces valeurs peuvent être modifiées si nous prenons en compte le contexte non vocalisé des requêtes. En effet, en dehors de quelques formes verbales et de mots grammaticaux non ambigus comme لعل, متى, اخترع, on retrouve beaucoup de formes ambiguës comme نزل, رقص, طلب. Notons aussi que les formes verbales rencontrées ne sont pas fléchies.

Mots graphiques		
Formes Verbales	Mots outils	Formes nominales
3,30%	3%	94,20%

Figure 5: Catégories morphosyntaxiques des requêtes de l'échantillon

La particularité des noms propres arabes est qu'ils sont souvent des dérivées de formes verbales (participe actif, participle passif, etc.). كاتب est à la fois l'écrivain et aussi un nom propres comme pour Kateb Yassine. Toutefois, la recherche par كاتب renvoie essentiellement écrivain. Voici quelques exemples de recherche avec des noms :

- Soit la recherche du mot *katab* كَتَب sur Google. Les premiers résultats concernent « les livres » (en arabe *kutub*). Est-ce une question de « ranking » (rang donné à l'objet de la requête) ou de priorité donné aux noms ? Nous constatons en tout état de cause que **la voyellation du mot clé n'a aucune influence sur la recherche**.
- Soit une recherche autour des entités nommées *Jamâl al-dîn al-'Afghânî* جمال الدين الأفغاني et *al-za'im Jamâl 'Abd al-NâSir* جمال عبد الناصر الزعيم, "le leader Jamal Abdel-Nasser". La recherche par *Jamâl* جمال renvoie en premier les résultats concernant le nom homonyme *beauté*, au lieu du prénom. Nous trouvons 5 340 000 réponses pour جمال, 737 000 pour جمال الدين, 70 700 pour جمال الدين الأفغاني. Pour la recherche de Nasser nous obtenons 805 000 pour *Jamâl 'Abd* جمال عبد, 293 000 pour *Jamâl 'Abd al-NâSir* جمال عبد الناصر et 253 000 pour *al-za'im Jamâl 'Abd al-NâSir* جمال عبد الناصر الزعيم. Parallèlement, une recherche avec le mot *al-za'im* الزعيم, "le leader" donne 2 100 000 réponses. Nous trouvons parmi les premiers résultats des blogs d'amateurs de foot, des informations sur la pièce de théâtre de Adel Imam (742 000 pour الزعيم عادل). Le premier résultat concernant جمال عبد الناصر arrive en trentième position. Nous constatons donc une grande faiblesse dans le traitement réservé aux entités nommées.

Au niveau de notre corpus de travail nous avons réparti les entités nommées identifiées automatiquement selon les trois catégories représentées ci-dessous :

Noms propres		
Pays	Noms/prénoms	villes
74,75%	23,41%	1,87%

Figure 6: Catégories des entités nommées dans notre corpus

#### 4.4. Apport de l'analyse linguistique pour la recherche d'information en langue arabe

En réalité, la recherche d'information est une tâche dépendante de la langue et son succès est donc lié aux langues des documents et à la manière dont les moteurs prennent en compte les caractéristiques de la langue concernée.

Les caractéristiques qui ont le plus d'impact sur la précision des moteurs de recherche concernent principalement la structure morphologique des mots et les variations morphologiques d'un même mot, d'où l'importance accordée par les moteurs de recherche à la lemmatisation et à la troncature.

L'apport de la lemmatisation reste discutable pour l'amélioration des performances des systèmes de recherche d'information dans les documents anglais [D. Harman 1991, 1995], car les règles de formations des mots en anglais sont relativement limitées et systématiques. Les langues à morphologie complexe comme l'arabe [J. Dichy, 1990], présentent un défi aux systèmes de recherche puisque le nombre de règles morphologiques est important. Nous allons montrer dans ce qui suit l'insuffisance des traitements de surface et l'apport de la lemmatisation à la recherche d'information en arabe.

#### 4.5. Insuffisance du parcours de surface ; mots homographes

Les moteurs de recherche utilisent les parcours de surface pour l'identification des mots. Or le mot graphique en arabe présente un caractère complexe. L'arabe est une langue flexionnelle où les familles morphologiques (dérivées d'une même racine) peuvent atteindre une taille assez importante. Nous trouvons souvent des formes graphiques proches ou semblables mais n'appartenant pas à la même famille morphologique.

Voyons, à titre d'exemple, ce que donne une recherche de surface des dérivées du mot قال: La requête « \*قال\* » donne près de 146 formes dans un corpus, ce qui dépasse largement les possibilités dérivationnelles de ce mot. En effet, la requête renvoie des termes comme ، الانتقال ، الاعتقال ، استقالة ، العقال ، مقاليد اقاتهم ، التقاليد ، برتقالية ، اثقالاً ، وقالياً ، الأقاليم ، قالب ... En plus de ce bruit considérable, beaucoup de formes de ce mot restent muettes et il est nécessaire de les rechercher à travers d'autres requêtes : c'est le cas de toutes les formes déclinées de l'inaccompli يقول et des autres déverbaux. Voici quelques exemples :

Item	Translittération (non vocalisée)	Traduction
اقالتهم	iqâlthm	(le) fait de les avoir licenciés
مقاليد	mqâlyd	rênes (par exemple : du pouvoir)
العقال	al'qâl	le cordon qui tient la kéfiyyé
استقالة	istqâlt	démission
اعتقالهم	i'tqâlhm	leur arrestation
الانتقال	alintqâl	Le transport
قالب	qâlb	moule
الأقاليم	alaqâlym	les régions (pays ou contrées)
وقالياً	wqâlbâ	et un moule (accusatif – منصوب)
اثقالاً	âtqâlâ	poids (accusatif – منصوب)
برتقالية	brtqâlyt	[de couleur] orange (fém.)
التقاليد	altqâlyd	les traditions
برتقالة	brtqâlt	orange
...		...

Figure 7: Quelques résultats de la requête « \*قال\* »

Pour l'exemple de recherche du mot « سماء », (“ciel”), Google renvoie 594 000 pour سماء en appliquant le principe de la complétion. Les résultats contiennent aussi 279 000 pour أسماء, (“noms”) parmi lesquelles nous trouvons aussi الأسماء (les rares cas du pluriel de سماء sont renvoyés dans des titres comme أسماء السماوات). Cet écart sémantique et morphologique entre les mots de la recherche provient de l'application des règles de lemmatisation des langues à caractères latins sur l'arabe.

Bien entendu, l'ambiguïté « naturelle » de la langue arabe, à tous les niveaux linguistiques, vient s'ajouter pour complexifier tous ces problèmes, déjà nombreux, en recherche d'information en langue arabe.

La lemmatisation se définit par l'identification d'une forme canonique correspondant à différentes formes flexionnelles ou dérivationnelles d'un mot donné (dérivation du pluriel, du singulier, du féminin ou du masculin, etc.). L'application de la lemmatisation en recherche d'information en langue arabe ne donne pas toujours les résultats escomptés car le système régissant la dérivation en arabe est plus complexe et ne se résume pas, souvent, à une simple suffixation.

Toutefois la lemmatisation est indispensable en arabe en raison du caractère agglutinant de la langue. Le mot graphique en arabe est augmenté par les proclitiques (de coordination, d'interrogation, marque de futur, de détermination, de préposition...) et les enclitiques (pronoms compléments).

Dans ce qui suit, nous allons présenter les difficultés de la lemmatisation concernant deux aspects de la morphologie nominale : (1) le passage du singulier au pluriel, et (2) le passage du masculin au féminin. Nous montrerons l'insuffisance des techniques de suffixation à travers des exemples.

De leur côté, les utilisateurs introduisent toujours des mots minimaux, c'est-à-dire sans clitiques, sauf en ce qui concerne le proclitique de détermination qui permet souvent de lever les ambiguïtés entre les noms et les verbes.

Détermination	
Indéterminé	Déterminé
61,03%	37,97%

**Figure 8: Pourcentage de recours à l'article avec des noms**

Étant donné le nombre considérable de réponses renvoyés pour les requêtes avec l'article ou avec divers clitiques d'une manière générale, nous n'avons pas pu vérifier si google procède à la lemmatisation ou pas. En tout cas tous les résultats retournés contiennent la chaîne de caractères recherchée.

#### **4.5.1. Relations entre le singulier et le pluriel des noms**

Soit le nom au pluriel كتابات. La procédure de lemmatisation "classique" qui considère que le pluriel est obtenu par la suffixation de ات au nom singulier pour l'obtention du pluriel, mais ne dispose pas de la possibilité de consulter une base de donnée incluant des spécificateurs morphosyntaxiques telle que DIINAR.1, donnera le singulier كتاب pour كتابات, alors que le lemme exact est كتابة.

L'obtention du singulier à partir du duel peut aussi poser quelques difficultés. Pour obtenir le singulier correspondant au mot فتاتان il est nécessaire de retirer le suffixe ان, mais cela ramène le mot à une forme incorrecte qui est فتات, alors que la bonne orthographe est فتاة. Nos tests sur le moteur Google ont montré son insuffisance dans le traitement de phénomènes linguistiques tels que la différentiation entre les terminaisons ت et ة.

Pour ce qui est du pluriel en langue arabe, un autre phénomène s'ajoute pour complexifier la lemmatisation : il s'agit de l'existence d'un type de pluriel dit "pluriel brisé", qui n'obéit pas à des règles, exemples : رجال-رجل, pour "homme-hommes" et نساء-نسوة-امرأة pour "femme-femmes".

Du côté des usagers nous avons extrait les statistiques suivantes concernant l'utilisation du nombre dans les mots clés :

Nombre				
Singulier	Duel	Pluriel		
74,21%	1,77%	24,02%		
		Régulier du masculin	Régulier du féminin	brisé
		71,09%	21,29%	7,52%

**Figure 9: Recours des utilisateurs au nombre dans les mots clés**

Il existe par ailleurs un ensemble de noms au duel d'un point de vue morphologique, mais qui désignent en réalité des noms au singulier, par exemple : le nom de pays البحرين ou encore le prénom محمدين (c'est ce qui est désigné, dans DIINAR.1 comme des formants-extension lexicalisés [Dichy, 1997]).

#### 4.5.2. Relations entre le masculin et le féminin des noms

La règle de suffixation peut-être utilisée aussi pour l'obtention du féminin ou du masculin. Généralement, la règle appliquée consiste à ajouter au masculin le suffixe ة pour l'obtention du féminin, mais cette règle non plus n'est pas systématique. Ainsi, les mots ائارة ou دراسة comportent le suffixe -at de la "chose générale" [A. Roman, 1990], qui peut être confondu avec la marque -at du féminin. Ils n'admettent naturellement pas de masculin. Au niveau graphique nous trouvons aussi un petit nombre de noms masculins se terminant par la lettre ة, -at, qui peut également confondues avec la marque du féminin, exemple : *xalifa* خليفة, "calife".

Par ailleurs, le genre peut aussi être rendu par des mots ayant des racines différentes. Ainsi pour : *rajul* رجل "homme" – *HiSân* حصان "cheval" – *walad* ولد "enfant" – *jamal* جمال "chameau" – 'ab أب "père".

Du côté des usagers nous avons extrait les statistiques suivantes concernant l'utilisation du genre dans les mots clés :

Genre					
Masculin	Féminin				
50,13%	49,84%				
	avec marque		sans marque		Autres
	ayant masculin	Sans masculin	ayant masculin	Sans masculin	Féminin d'un pluriel masculin
	47,11%	11,69%	16,81%	1,01%	23,38%

**Figure 10: Le recours au genre dans les mots clés**

#### 4.5.3. Pratiques d'écriture courantes

La recherche d'information en langue arabe doit aussi faire face à des difficultés supplémentaires dues aux habitudes orthographiques des auteurs, dont l'impact sur la recherche d'information n'est pas négligeable. Pour éviter de faire un recensement des "fautes d'orthographe" sur le web, nous avons mené l'étude sur un corpus journalistique contemporain. Nous donnerons des statistiques relatives aux pratiques d'écriture chez les auteurs arabes contemporains. Pour notre étude, nous avons choisi de travailler sur un corpus de deux millions de mots de la presse écrite. Des analyses détaillées seront publiées dans les Journées internationales d'Analyse statistique des Données Textuelles [Abbes et Dichy, 2008].

#### 4.5.3a. Hamza et 'alif

Les scripteurs confondent souvent la *hamza* (أ – إ) et le 'alif en début de mot. On trouve par exemple, dans le corpus de deux millions de mots, 26 923 fois إلى 'ilâ, « à », et 2 089 fois إلى. On trouve également 33 901 ان indifférencié, contre 50 569 أن (conjonctions 'an ou 'anna) et 759 إن (conjonctions 'in ou 'inna). Les estimations auxquelles nous étions parvenus dans cette étude indiquent que le taux des items renfermant cette seule confusion s'élève à 5,79% de l'ensemble des items ou encore à 6,76% des mots. Nos tests sur le moteur Google nous ont montré qu'il renvoie indifféremment toutes les orthographes de la *hamza* au début du mot.

#### 4.5.3b. Yâ' et 'alif maqṣûra

A la *hamza-'alif* initiale s'ajoute une autre confusion, située pour celle-ci en fin de mot, entre ي (lettre yâ' finale) et ى (ou 'alif maqṣûra). Le mot نادي, nâdî, « club », par exemple, peut être noté نادى, ce qui correspond à l'usage des typographes égyptiens (mais peut aussi être lu comme nâdâ, « convier, convoquer »). De même, pour cette confusion, le moteur Google renvoie systématiquement les résultats avec ي et ى, même si le mot de la requête est écrit entre guillemets. Par exemple, la recherche du mot الأولي renvoie en tête des pages contenant الأولى. Notons qu'il s'agit d'un mot qui cumule les deux difficultés.

#### 4.5.3c. Le caractère '-' (kashida)

Les typographes font un usage fréquent du caractère '-' (appelé *kashida*), qui permet l'allongement du trait au milieu des mots, pour une meilleure lisibilité, pour limiter les espaces blancs sur une ligne justifiée ou même parfois pour des raisons purement calligraphiques. Ce caractère, ne faisant pas partie de l'alphabet arabe, est souvent une source de confusion pour les systèmes de traitement de la langue arabe. Le moteur Google semble éliminer ce caractère dans le mot de la requête.

#### 4.5.3d. L'absence des signes de vocalisation

L'absence des signes de vocalisation dans les textes – à laquelle les lecteurs arabes sont accoutumés –, constitue une source de difficulté considérable pour l'analyse automatique de l'arabe. Certains signes diacritiques relatifs à la base (ou noyau lexical) sont indispensables pour la détection du sens du mot. Ils sont par conséquent indispensables pour le choix du mot pertinent dans la recherche d'information, particulièrement en l'absence de contexte. Les analyses peuvent en effet reconnaître dans un même item plusieurs patrons (وزن), voire plusieurs combinaisons de patrons et de racines.

#### 4.5.3e. Ta marbouta ة et ha ه

Nous avons remarqué essentiellement sur le Web, une confusion fréquente entre les lettres ة et ه en fin de mot. Google semble avoir transformé la fréquence en règle et semble renvoyer exactement les mêmes résultats pour les requêtes مكتبة et مكتبة.

## 5. CONCLUSION

La recherche d'information et la fouille de textes en langue arabe constitue un défi majeur. L'indexation automatisée des textes et la construction de protocoles de requêtes centrés sur la langue arabe sont appelés à jouer, dans le développement du Web arabe, un rôle de premier plan. Nous avons cherché, dans ce travail, à présenter des outils susceptibles de contribuer très fortement au travail collectif auquel cet atelier organisé par le comité de Damas de la Société syrienne d'informatique nous convie.

## Principales références et thèses liées à DIINAR.1 ou au groupe SILAT

- Wijdan, ABBAS-MEKKI, 1998. *Définition et description des unités linguistiques intervenant dans l'indexation automatique des textes en arabe*. Thèse de doctorat, Lyon, ENSSIB/Université Lyon 2.
- Ramzi ABBES. 2002. "AraFreq: un outil pour le calcul de fréquences de mots arabes", in A. Braham, ed. *Colloque international sur le traitement automatique de l'arabe – Proceedings of the International Symposium on The Processing of Arabic* (Avril 18-20, 2002). Université de la Manouba, Tunis (en Arabe, Français et Anglais).
- 2004. *La conception et la réalisation d'un concordancier électronique pour l'arabe*. Thèse de doctorat en sciences de l'information, Lyon, ENSSIB/INSA.
- Ramzi ABBES et Joseph DICHY. 2008. "Extraction automatique de fréquences lexicales en arabe et analyse d'un corpus journalistique avec le logiciel AraConc et la base de connaissances DIINAR.1", in : Serge Heiden & Bénédicte Pincemain, *Proceedings of JADT 2008, 9<sup>th</sup> International Conference on Textual Data statistical Analysis*, Lyon 12-14.03.2008, Presses Universitaires de Lyon, 2 vol. : 31-44.
- Ramzi ABBÈS, Joseph DICHY, Mohamed HASSOUN. 2004. "The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program", in *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages- COLING 2004 – University of Geneva, 28<sup>th</sup> August 2004* : 15-22.
- Ramzi ABBÈS, Joseph DICHY, Mohamed HASSOUN. 2005. "Morpho-lexical ambiguities in the recognition of written Arabic word-forms, evidence from the DIINAR.1 lexical resource", colloque ACIDCA-ICMI'05 (International Conference on Machine Intelligence), Tozeur (Tunisie), 5-7 novembre 2005.
- Najim ABU AL-CHAY. 1988. *Un Système expert pour l'analyse et la production des verbes arabes dans une perspective d'Enseignement Assisté par Ordinateur*. Thèse de doct., Université Lyon 1.
- Sam AMMAR et Joseph DICHY. 1999a. *Les verbes arabes*, Paris, Hatier (collection Bescherelle).
- 1999b. الأفعال العربية (*Al-'af'â al-'arabiyya*). Paris, Hatier (collection Bescherelle – introduction originale en arabe).
- Abdelfattah BRAHAM et Salem GHAZALI. 1998. قاعدة البيانات المعجمية العربية أو مشروع معجم العربية الآلي (معالي – DIINAR)، حصيلة وآفاق – المجلة العربية للعلوم – ع ٣٢ – ١٤-٢٣
- (“Qâ'idatu l-bayânât al-mu'jamiyya al-'arabiyya, 'aw maṣrû' Mu'jam al-'Arabiyya l-'âliyy, *Ma'âli-DIINAR, hasîla wa-'âfâq*”, *Al-Majâlla l-'Arabiyya li-l-'ulûm*, n°32, déc. 1998, pp. 14-23).
- Jean-Pierre DESCLES, dir. 1983. (H. Abaab, J.-P. Desclés, J. Dichy, D.E. Kouloughli, M.S. Ziadah). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur*, Rapport rédigé à la demande du Ministère des Affaires étrangères.
- Joseph DICHY. 1984/89. "Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe", in Dichy et Hassoun, eds., 1989: 92-158.
- 1987. "The SAMIA Research Program, Year Four, Progress and Prospects". *Processing Arabic Report 2*, T.C.M.O., Nijmegen University: 1-26.
- 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.
- 1993a. "Knowledge-system simulation and the computer-aided learning of Arabic verb-form synthesis and analysis". *Processing Arabic Report 6/7*, T.C.M.O., Nijmegen University: 67-84, 92-95.
- 1997. "Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot". *Meta 42*, printemps 1997, Québec, Presses de l'Université de Montréal: 291-306. [www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf](http://www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf)
- 1998. "Mémoire des racines et mémoire des mots : le lexique stratifié de l'arabe". T. Baccouche, A. Clas et S. Mejri, eds., *La Mémoire des mots*. Numéro spécial de la *Revue Tunisienne de Sciences Sociales*, 117: 93-107.

- 2000. “Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects”. Proceedings of *ACIDA' 2000* (Monastir, Tunisia, 22-24.03.00), *Corpora and Natural Language Processing* volume: 55-60.
  - 2001a. “On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases”. *ACL 39<sup>th</sup> Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect*, Toulouse: 23-30. <http://www.elsnet.org/arabic2001/dichy.pdf>
  - 2002a. “Arabic lexica in a cross-lingual perspective”. *Proceedings of ARABIC Language Resources and Evaluation: Status and Prospects*, A Post Workshop of LREC 2002, Paris: ELRA.
  - 2002b. *Structure et évolution de la dérivation lexicale en arabe : sens et forme des verbes et des dérivés nominaux les plus immédiats*, Cours de préparation au CAPES d’arabe, session 2002, question de linguistique, 70 pages.
  - 2004. “Six Basic Criteria for the Assessment and Validation of Arabic Processing Software and Lexical Language Resources”, Proceedings of the *NEMLAR International Conference on Arabic Language Resources and Tools*, Le Caire, 22-23 Sept. 2004, distrib. Paris: ELDA: 94-101.
  - 2005a. “Spécificateurs engendrés par les traits [±animé], [±humain], [±concret] et structures d’arguments en arabe et en français”, in Henri Béjoint et François Maniez, dir., *De la mesure dans les termes*, Actes du colloque organisé en hommage à Philippe Thoiron, Université Lumière Lyon 2, 23-25 septembre 2004, Presses Universitaires de Lyon, p. 151-181.
  - 2005b. “The crucial role of language resources (LRs) in the assessment of written Arabic NLP applications including recognition.” Key-note address on Text recognition aspects in the Special session on ‘Linguistic Information Integration in Arabic Character and Text Recognition’ », colloque *ACIDCA-ICMI'05 (International Conference on Machine Intelligence)*, Tozeur (Tunisie), 5-7 novembre 2005.
- Joseph DICHY, Abdelfattah BRAHAM, Salem GHAZALI, Mohamed HASSOUN. 2002. “La base de connaissances linguistiques DIINAR.1 (Dictionnaire INformatisé de l’Arabe, version 1)”, in Abdelfattah Braham, ed. *Colloque international sur le traitement automatique de l’arabe – Proceedings of the International Symposium on The Processing of Arabic* (Avril 18-20, 2002). Université de la Manouba, Tunis (en Arabe, Français et Anglais).
- Joseph DICHY et Ali FARGALY. 2007. “Grammar-lexis relations in the computational morphology of Arabic”. In Abdelhadi Souidi, Guenter Neumann and Antal Van den Bosch, eds., *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, Dordrecht : Kluwer/Springer (series on Text, Speech, and Language Technology), chapter 7, p. 115-140.
- Joseph DICHY et Mohamed HASSOUN, eds. 1989. *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l’arabe - Travaux SAMIA I*. Paris, Conseil International de la Langue Française.
- Joseph DICHY et Mohamed HASSOUN. 1998. “Some aspects of the DIINAR-MBC research programme”. In A. Ubaydly, ed., 1998. In Ahmad Ubaydly, ed. *Proceedings of the 6th International Conference and Exhibition on Multilingual Computing (ICEMCO 98)*, Centre of Middle Eastern Studies, University of Cambridge: 2.8.1-6.
- 2005. « The DIINAR.1-« معالي » Arabic Lexical Resource, an outline of contents and methodology », in *The ELRA Newsletter*, Vol. 10, n°2, April-June 2005, pp. 5-10.
- Samia EZZAHID. 1996. *Méthodologie d’élaboration d’une base de données lexicale de l’arabe (vocabulaire général) d’après la théorie Sens-Texte d’Igor Mel’cuk*. Thèse de doct., Université Lyon 2.
- Lynne FRANJIE. 2003. *Etude sémantique et traductologique de verbes arabes dans les dictionnaires bilingues : le Larousse (arabe-français) et le H. Wehr (arabe-anglais)*, Thèse de doct., Université Lyon 2.
- Nabil GADER. 1992. *Conception et réalisation d’un prototype de correcteur orthographique de l’arabe*. Mémoire de DEA en Sciences de l’information et de la communication, ENSSIB/Université Lumière-Lyon 2.

- Salem GHAZALI et Abdelfattah BRAHAM. 2001. "Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic". In *ACL 39<sup>th</sup> Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect*, Toulouse: 51-57.
- Malek GHENIMA. 1998. *Analyse morpho-syntaxique en vue de la voyellation assistée par ordinateur des textes écrits en arabe*. Thèse de doct., ENSSIB/Université Lyon 2.
- Jonathan GRAINGER, Joseph DICHY, Mohamed EL-HALFAOUI, Mohamed BAMHAMED. 2003. "Approche expérimentale de la reconnaissance du mot écrit en arabe", in Jean-Pierre Jaffré, éd., *Dynamiques de l'écriture : approches pluridisciplinaires*, revue *Faits de langue*, n°22 : 77-86.
- Mathieu GUIDERE (sous la direction de), Marie-Hélène AVRIL, Salam BAZZI-HAMZE, Amal EL SABBANE, Lynne FRANJIE, Mathieu GUIDERE, Xavier LELUBRE, Rita MOUCANNAS-MAZEN, Hoda MOUCANNAS-MEHIO, Manar ROUCHDY, Camilia SOUBHI, *Kalimât .- Le vocabulaire arabe*. Ellipses, Paris, 2003, 477p
- Mohamed HASSOUN. 1987. *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*. Thèse d'État, Université Lyon 1.
- Lamia LABED et Xavier LELUBRE. 1997. "DIINAR-TOPT: conception d'une base de données terminologique Arabe/français dans le domaine de l'optique", in *JST'97: L'ingénierie de la langue: de la recherche au produit*, Avignon, 15-16/04/1997, AUPELF-UREF/Francil: 523-8.
- Xavier LELUBRE. 1985. "Projet pour un didacticiel de conjugaison de verbes arabes", Ministère de l'éducation nationale.
- 1992. *La terminologie arabe contemporaine de l'optique: faits – théories – évaluation*, thèse en Linguistique, Université Lyon 2. 1993.
- 1997. "Terminologie scientifique : entre le phraséologisme et l'unité terminologique complexe", in Boisson C. et Thoiron P., eds (1997), *Autour de la dénomination*, Presses Universitaires de Lyon, p. 221-39.
- 2001. "A Scientific Arabic Terms Data Base: Linguistic Approach for a Representation of Lexical and Terminological Features". In *ACL 39<sup>th</sup> Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect*, Toulouse: 66-72.
- 2002. "La reconnaissance d'unités terminologiques complexes candidates par patron de surface", in A. Braham, ed., *Colloque international sur le traitement automatique de l'arabe – Proceedings of the International Symposium on The Processing of Arabic* (April 18-20, 2002). University of Manouba, Tunis.
- Sylver Aboubakar MINKO MI-NSEME. 2003. *Modélisation des expressions figées en arabe en vue de la constitution d'une base de données lexicale*. Thèse de doct., Univ. Lyon 2.
- Riadh OUERSIGHNI. 2001. "A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts". In *ACL 39<sup>th</sup> Annual Meeting. Workshop on Arabic Language Processing: Status and Prospect*, Toulouse, pp. 66-72. <http://www.elsnet.org/arabic2001/ouersighni.pdf>
- 2002. *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 doct., ENSSIB/Univ. Lyon 2
- SAMIA (groupe de recherche). 1984. "Enseignement Assisté par Ordinateur de l'arabe: simulation à l'aide d'un modèle linguistique – la morphologie." In *E.A.O. 1984*, Paris, Agence de l'Informatique, pp. 81-96. [V. Bouché, Dichy et Hassoun, 1984 ci-dessus]
- Riadh ZAAFRANI. 1997. "Morphological analysis for an Arabic Computer-aided learning system", Proceedings of *DIALOGUE'97, International Conference on computational linguistics and its applications*, June 10-15, 1997, Yasnaya Polyana, Russia.
- 1998a. "Al-Mu'allim 2 Software : An Arabic Computer Learning System Using Conceptual Sentence Generation", *Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, April 16-19, 1998, Cambridge, England.
- 2002. *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère*. Thèse de doct., ENSSIB/Université Lyon 2.