The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program

Ramzi ABBÈS
SII[a] / SILAT[b],
ENSSIB[c],
17-21, bd. du 11 nov. 1918,
69623 Villeubanne Cedex, France
abbes@enssib.fr

[a] SII: *Systèmes d'Information et Interfaces,* research centre, ENSSIB

Joseph DICHY
ÉLISA[d] / SILAT[b],
Université Lumière-Lyon 2,
86, rue Pasteur
69365 Lyon Cedex 07, France
joseph.dichy@univ-lyon2.fr

[d] ÉLISA: *Épistémologie, Linguistique, Ingénierie et Sémiologie de l'Arabe*, research centre, Lumière-Lyon 2 University

Mohamed HASSOUN
SII[a] / SILAT[b],
ENSSIB[c],
17-21, bd. du 11 nov. 1918,
69623 Villeubanne Cedex, France
hassoun@enssib.fr

[c] ENSSIB: *École Nationale Supérieure des Sciences de l'Information et des Bibliothèque.*

[b] SILAT: *Systèmes d'information, Linguistique, Ingénierie de l'Arabe et Terminologie,* research group common to the Lyon 2 Lumière University and ENSSIB. (*Silat* is Arabic for "link", "relation".)

Abstract

This paper is a contribution to the issue – which has, in the course of the last decade, become critical – of the basic requirements and validation criteria for lexical language resources in Standard Arabic. The work is based on a critical analysis of the architecture of the DIINAR.1 lexical database, the entries of which are associated with grammar-lexis relations operating at word-form level (i.e. in morphological analysis). Investigation shows a crucial difference, in the concept of 'lexical database', between *source program* and *generated lexica*. The source program underlying DIINAR.1 is analysed, and some figures and ratios are presented. The original categorisations are, in the course of scrutiny, partly revisited. Results and ratios given here for basic entries on the one hand, and for generated lexica of inflected word-forms on the other. They aim at giving a first answer to the question of the ratios between the number of lemma-entries and inflected word-forms that can be expected to be included in, or generated by, a Standard Arabic lexical dB. These ratios can be considered as one overall language-specific criterion for the analysis, evaluation and validation of lexical dB-s in Arabic.

Keywords: Arabic lexical databases – Arabic script – word-formatives grammar – lemma-entries – morphosyntactic specifiers.

1 Introduction

In the present state of the art in the development of software and language resources in Arabic, there is an urgent need for evaluation and validation criteria based on solid analytic grounds: there exists nowadays a subsequent number of Arabic lexical databases, and more are under completion.

Existing lexical dB-s are not always, for the time being, available as such to researchers and/or developers, because they are usually embedded in software (such as a morphological analyser or a parser), and are still very difficult to make use of independently. It is to be expected, though, that the issue of availability will be overcome in a reasonably near future, and that a number of Arabic lexical databases will be found on the market, or on catalogues such as, in Europe, that of ELRA[1], and in the USA, that of LDC[2]. The on-going European project NEMLAR is presently working on the availability of language resources including lexical databases[3]. As a result, the crucial question of the quality and consistency of these databases should be met as soon as possible.

[1] European Language Resources Association, 55, rue Brillat-Savarin – 75013 Paris, France.

[2] Linguistic Data Consortium, University of Pennsylvania, 3600 Market Street, Suite 810, Philadelphia, PA 19108, USA.

[3] NEMLAR (Network for Euro-Mediterranean LAnguage Resources) is coordinated by Pr. Bente Maegaard, Center for Sprogteknologi (CST), Copenhagen. E-mail and site: nemlar@cst.dk, www.nemlar.org.

One of the criteria for the evaluation and validation of a lexical database for Arabic is both quantitative (how many?) and qualitative (what of, precisely?). In this paper, which refers to previous work on the processing of Arabic and the related lexical resources[4], we will try and give evidence on the structure of a lexical database, founded on an analysis of the DIINAR.1 database[5]. Quantitative results are only interesting if they can be interpreted in such a way as to yield information on the actual structure and categories of the lexicon of the language under consideration. We will endeavour to show that a quantitative and qualitative analysis of the lexical categories incorporated in DIINAR.1 can be interpreted with this respect. Moreover, the investigation leads to proposing a more consistent organisation of lexical information and relations, which should be included in future versions of DIINAR.

2 The type of lexical dB required by the automatic analysis of Arabic

What are the fundamental requirements of a lexical database in Arabic? The first challenge to be met upon endeavouring to build language resources in Arabic is that of the structure of the writing system of the language (Dichy, 1990), the two main features of which are: non diacriticized script in standard texts (§ 2.1) and the structure of the word-form (§ 2.2). The combined effect of these features entails the need for a lexical database that includes a subsequent number of grammar-lexis relations (§ 2.3). Such a dB is to be considered as a *sine qua non* condition of high-level and elaborate Arabic NLP.

2.1 Non diacriticized writing

It is well-known that Arabic script belongs to a group of Semitic writings originating from ancient Phoenician alphabets, such as Hebrew, Aramaic or Syriac. Phonographic translation is basically restricted to the notation of consonants and “long vowels”. In the course of time, these writing systems have developed additional diacritic symbols, mainly for the needs of the oral reading of sacred texts (*Bible, New Testament, Koran*). Arabic writing has thus been provided with a sophisticated system of diacritical marks (comparable to the Massora diacritics which were later devised for the Hebrew Bible). Standard writing nevertheless disregards these symbols. This results in a high degree of homography, accounting for the multiple analyses encountered in a majority of single words by morphological analysers (which are, needless to recall, bound to consider every word off-context).

2.2 “Nucleus” and “extensions”: a quick recall of the structure of word-forms in written Arabic

Unlike automatic recognition software, human readers are, of course, able to combine semantic, syntactic and morphological analyses. They are helped in their reading of Arabic written utterances by another major feature of the writing system: the very regular structure of the word-form. This structure has been introduced and extensively described previously (Desclés ed., 1983; Dichy 1984, 1990; Hassoun, 1987 – after the pioneering work of Cohen, 1961/70), and is only recalled here for the sake of clarity.

Word-forms in Arabic can be described on the whole as consisting of a *nucleus formative* (henceforth *NF*) to which *extension formatives* (henceforth *EF*) are added, either to the left or to the right (Dichy, 1997). Ante-positioned EF-s are abbreviated as *aEF*, and post-positioned ones as *pEF*. The nucleus formative, usually called *stem*, can be represented in terms of prosodic or non-concatenative morphology (after J. McCarthy’s original and much discussed insights, 1981). In Semitic morphology, the stem is considered, according to a somewhat recent, but very widely followed tradition, as a compound of *root* and *pattern*. One must keep in mind, though, that many nouns cannot be analysed in such a way: they are referred to as *quasi-stems* (Dichy & Hassoun, eds., 1989).

Arabic word-forms consist of:

– *proclitics (PCL)*, which include mono-consonantal conjunctions, e.g. *wa*-, ‘and’ , *li*-, ‘in order to’, or prepositions, i.e. *bi*-, ‘in, at’ or ‘by’, etc.;

– a *prefix (PRF).* The category, after D. Cohen’s representation of the word-form, only includes

[4] The research and development work referred to in the SILAT research group goes back to the 1980ies and has been going on since (Desclés et alii 1983, Dichy & Hassoun, eds. 1989, Dichy 1984/89, 1987, 1993, 1997, 2000, Lelubre 1993, Braham 1998, Braham & Ghazali 1998). It includes a number of doctoral dissertations (Hassoun 1987, Abu Al-Chay 1988, Dichy, 1990, Gader 1992, Ghenima 1998). For further developments, see: Ezzahid 1996, Labed & Lelubre 1997, Abbas 1998, Dichy & Hassoun 1998, Ammar & Dichy 1999a et b, Abbès 1999, Dichy 1998, 2001a et b, Ghazali & Braham 2001, Lelubre 2001, Ouersighni 2002, Zaafrani 2002, Dichy & Fargali, 2003.

[5] **DIINAR.1** (*DIctionnaire INformatisé de l’Arabe*), Arabic acronym **Ma‘âlî (**“Mu‘jam al-‘Arabiyya l-’âlî”), is a comprehensive Arabic Language dB operating at word-form level (morphological analysis or generation). It has been completed in close cooperation, in Tunis by IRSIT (now SOTETEL-IT - A. Braham and S. Ghazali), and in France by ENSSIB (M. Hassoun) and the Lumière-Lyon 2 University (J. Dichy). See Dichy, Braham, Ghazali & Hassoun, 2002.

the prefixes of the imperfective, e.g., *ya-*, prefixed morpheme of the 3rd person;

- a *stem*, which can be represented in terms of a ROOT (an ordered triple of consonants, or, by extension of the system, a quadruple) and a PATTERN (roughly: a template of syllables, the consonants of which are the triple of the ROOT to which monoconsonantal affixes are added). The stem *takabbar*, 'to be haughty', thus consists of the 3-consonant ROOT /k-b-r/ and of the PATTERN /taR^1aR^2R^2aR^3/, where R^1, R^2 and R^3 stand for 'radical consonant 1, 2, 3', and are instantiated by the triple of the ROOT (R^1=*k*, R^2=*b*, R^3=*r*). Nouns that cannot be analysed in ROOT and PATTERN are conventionally referred to as *quasi-stems*, e.g.: *'ismâ'îl*, 'Ishmael', *yûnîskû*, 'UNESCO', *kahramân*, 'amber';
- *suffixes (SUF)*, such as verb endings, nominal cases, the nominal feminine ending *-at*, etc.;
- *enclitics (ECL)*. In Arabic, enclitics are complement pronouns.

In the table below two apparently equivalent representations of the structure of the Arabic word-form are given. The main difference between them lies in the fact that (2) aims at highlighting the relations between nucleus and extension formatives (NF and EF-s), featuring a triangle (in boldface below). The rules governing the relations between morphemes embedded in the word-form are included in a *word-formatives grammar* (henceforth *WFG* – Dichy, 1987, 1997). These rules, and the features they involve, are distributed along these three relations, a great number of which are related to the lexical nucleus, and have to rely upon the finite set of grammar-lexis relations operating at word-level (formalised in Dichy, 1990).

<table>
<tr><td><b>(1) 'Traditional' representation of the word-form (D. Cohen, 1961/70, Desclés, ed., 1983)</b></td><td>maximal
______word-form______
| |
minimal
__word-form__
| |
##PCL # PRF +STEM+ SUF # ECL##
STEM = <ROOT¤¤PATTERN></td></tr>
<tr><td><b>(2) Nucleus-extensions representation (Dichy, 1997)</b></td><td><b>NF</b>
<b>/   \</b>
<b>aEF — pEF</b>
/   \   /   \
PCL PRF   SUF ECL</td></tr>
</table>

Table 1: Structure of the word-form in Arabic

2.3 Word-formatives grammar (WFG) and word-level grammar-lexis relations

Complex as it may appear, the above structure is regular, and remains, up to a certain point, recognisable from a psychological stand. It is, subsequently, very restrictive: Arabic word-forms include one lexical stem and one only[6]. In fact, the word-formatives grammar (WFG) accounts for the regular structure of the word-form.

Rules involving word-formatives (the above nucleus and extension formatives, NF and EF) are based on three fundamental types of relations (Dichy, 1987): **⇛** 'entails', **≠>** 'excludes', ****** 'is compatible with' (or 'admits'), the third of which is attached to the opposed pair of the first two as an 'elsewhere' relation of a special kind, directly connected to ambiguity in language analysis processes (Dichy, 2000). In generation, all 'compatibility' (or 'admit') relations can in fact be rewritten in terms of 'entail' or 'exclude' rules associated with specific sets of word-formatives. 'Compatibility' relations are mostly useful in the formalisation of recognition rules, when ambiguity is at stake[7]. The formal structure of the WFG thus includes relations of the three types above, which are, in turn, involved in either one of the two following combination schemes:

- **EF ↔ EF combinations**, such as PCL → SUF rules, e.g.:
  PCL = *bi* **⇛** SUF = {*i, in, a, an, îna, î, ayni, ay*}
  which can be phrased as: 'the proclitic preposition *bi#* entails one of the indirect (or genitive) case suffixes'. Other rules will point to a given case suffix in a given utterance.

- **NF ↔ EF combinations**, such as STEM → SUF rules, e.g.:
  STEM = 'diptote' **⇛** SUF = {*u, a, i*}
  which can be phrased as: 'a stem whose declension is diptote entails case-endings belonging to the listed set'. (Diptote stems may also be compatible with dual or plural suffixes, which is taken into account in another rule.)

Another type of relation to be encoded in a lexical database is:

[6] A few exceptive compound items exist, but they are kept marginal by the structure of the language, for the obvious reasons hinted at here, unlike what has happened in Modern Hebrew, as opposed to the Biblical and Medieval state of the language (Kirtchuk, 1997).

[7] Automatic recognition and generation are not to be considered as reverse processes. Evidence from Arabic is given in Desclés, ed., 1983; Dichy, 1984, 1997, 2000.

- **NF ↔ NF linking combinations**, which have to be encoded whenever the morphological variation is not rule-predictable (cf. Melčuk’s concept of *syntactic*, 1982). This is the case in a majority of singular ↔ ‘broken plural’ links in nouns or adjectives, as well as in ‘perfective’ ↔ ‘imperfective’ (*mâ**d**î* ↔ *mu**d**âri‘*) links, in verbs belonging to ‘simple’ PATTERNS (*al-fi‘l al-mujarrad*).

In an Arabic lexical dB, lexical entries (NF-s or STEMS in the above representation) need to be associated with ***morphosyntactic specifiers*** ensuring their insertion in word-forms, and their morphological variation (conjugation or declension). Morphosyntactic specifiers, in other words, account for:

– grammar-lexis relations, i.e. *NF ↔ EF combinations*;
– morpho-lexical variation, i.e. *NF ↔ NF linking combinations*.

Lexical entries thus ‘entail’, ‘exclude’ or ‘admit’ a number of grammatical morphemes listed in the various fields of the word-form as word-formatives, either on a non regular basis, or on the basis of rules founded on semantic features that cannot be deduced from the formal structure of the morpheme. As shown in Dichy (1997), morphosyntactic specifiers make up formally, in a lexical database, for information associated in the speaker’s memory to various levels of linguistic analysis (morpho-phonological, syntactic or semantic features).

This structure has often been disregarded in the elaboration of Arabic lexical databases on the assumption that the representation of lexical entries as a mere combination of PATTERN and ROOT (plus a number of suffixes) is sufficient. This is definitely *not* the case: evidence recalled in this paragraph (also in Hassoun & Dichy, eds. 1989, Dichy, 1997, Dichy & Fergaly, 2003) show that grammar-lexis relations operating at word-form level can only be taken into account if information is associated to whole stems (or nuclei), or to stem+suffix ‘frozen’ compounds. These relations cannot be predicted on the sole basis of patterns.

The description of the WFG outlined in this paragraph has led to the elaboration of *exhaustive and finite sets of morphosyntactic specifiers* liable to be associated to the *non finite lexical entries* of an Arabic database (Dichy, 1997). These sets have been associated with the entries of the DIINAR.1 Arabic Language database. The WFG has been on the other hand implemented in the related generation and analysis source programs.

Another lexical LR including morphosyntactic information at word level is the lexicon elaborated and completed by Timothy Buckwalter, which has been used in the finite-state morphological analyser elaborated at the European Xerox Research Centre (Meylan, France)[8].

3 A few figures and ratios from DIINAR.1: generated lexica vs. source lexicon

In the previous section, we outlined the structure of the WFG and the information associated with lexical entries in the source program of the DIINAR.1 database.

It is essential to note that the expression *lexical database* is ambiguous, i.e. that it is liable to refer, either:

– to a *source program* drawing on lists of *basic* lexical or grammatical items (related to a grammar of the kind outlined in the previous section),
– or a set of *generated lexica*, the items of which can be either *basic* (as in the source program) or *combined*, i.e. resulting from the combination of basic items, according to the rules of the word-formatives grammar.

Software relying partly or entirely on morphological analysis may, or may not, need all the information outlined in section 2. They draw on lexica generated by the source program associated with the dB (Hassoun, 1987). Generated lexica can be restricted to a subset of information, as in a spelling checker (Gader, 1992), or extended to all available information, as in a parser (Ouersighni, 2002) or in an interactive language teaching software (Zaafrani, 2002). In the current section, we will examine the architecture of the DIINAR.1 database, from the standpoint of the relation between the figures of the basic entries included (§ 3.1 and 3.2), and that of the inflected word-forms (§ 3.3).

3.1 The basic figures of the DIINAR.1 source program

The total number of lemma-entries in the DIINAR.1 database is : 121,522.

This includes 445 tool-words belonging to various grammatical categories (e.g.: prepositions, conjunctions, etc.) and the prototype of a proper names database of 1,384 entries. Both types of entries are associated with a particular word-formatives grammar, and with their own subsets of morpho-syntactic specifiers.

The main parts of the database include:

[8]. Beesley, 1998, 2001, Beesley and Karttunen, 2003. Also: Buckwalter, 2002.

Nouns, including adjectives	29,534
[Broken plural nominal forms]	[9,565]
Verbs	19,457
Deverbals:	
- infinitive forms (*masdar*)	23,274
- active participles (*’ism al-fâ‘il*)	17,904
- passive participles (*’ism al-maf‘ûl*)	13,373
- ‘analogous adjectives’ (***s**ifa mu<u>s</u>abbaha*)	5,781
- ‘nouns of time & place’ (*’ism al-makân wa-z-zamân*)	10,370
Total number of deverbals	[70,702]
Subtotal of lemmas	119,693

Table 2: Number of lemmas and items belonging to main major lexical categories

3.2 Comments and critical analysis

(1)

Table 2 features two ratios of general interest for the structure of the Arabic general Lexicon:

– The ratio between broken plural nominal forms (which are not counted as lemmas[9]) and nouns and adjectives is roughly of one to four.

– Deverbals appear to be 3.6 more numerous than verbs.

(2)

The above categorisation follows that of traditional Arabic grammar. Two sub-categorisations should, nevertheless be revisited for linguistic consistency reasons:

– Adjectives (although they can appear as nouns in many syntactic structures) should be isolated. This will be needed, of course, in parsing – even in ‘shallow parsing’. Adjectives in Arabic can be identified through syntactic tests.

– ‘Nouns of time and place’ (*’asmâ’^{u} l-makân wa-z-zamân*) should not, in future versions of DIINAR, remain in the ‘deverbal’ category. They are in fact (except for the earliest stages in the development of the language) inserted in syntactic structures as full nouns.

(3)

It is to be noted, on the other hand, that (except for ‘nouns of time and place’) DIINAR.1 is very consistent in distinguishing between nouns and deverbals: deverbals re-used as nouns, and showing full nominal features appear, in the dB, twice (as ‘deverbals’ *and* as ‘nouns’, with their related morphosyntactic specifiers), e.g.:

- *sâkin*, plur. *sâkinûn, sâkinât,* ‘dwelling’, ‘inhabiting’, is a deverbal, e.g.:
  *Na**h**nu sâkinûna madînat*^{a} *al-’iskandariyya* = ‘We live in Alexandria’.
- *sâkin*, plur. *sukk■n* (broken plural form), ‘inhabitant’, is a full noun (appearing in the first line of Table 2), e.g.:
  *Na**h**nu sukkân*^{u} *madînat*^{i} *l-’iskandariyya* = ‘We are the inhabitants of Alexandria’.

(4)

The number of roots in DIINAR.1 is 6,546, it being understood that a great many nouns cannot be analysed in ROOT and PATTERN. (On the other hand, *all* the verbs and deverbals of the language can – Dichy, 1984/89.)

3.3 The DIINAR.1 lexica of inflected word-forms

The number of combined proclitics (which are effectively in use in Modern Standard Arabic), suffixes, prefixes and enclitics is shown in the tables below:

Proclitics (combined)	64
Prefixes	8
Suffixes (combined)	67
Enclitics	13

Table 3: Total number of extension formatives (EF-s)

	Associated with nouns	Associated with verbs	Common to both types
Proclitics	44	13	7
Prefixes	0	8	0
Suffixes	11	42	0
Enclitics	1	1	11

Table 4: EF-s associated with nominal and/or verbal stems

It is easy to imagine, on the basis of the above table, that one could generate huge figures through multiplying the number of extension formatives among themselves, then multiplying the result by the number of nouns and/or verbs. In order to avoid ‘over-powerful’ inflation of data, a consistent database needs to be filtered through (a) a word-formatives grammar and (b) morphosyntactic specifiers associated to stems.

[9]. ‘Broken plural’ forms are related to a singular noun-form lemma. Links between singular and plural forms, in the dB, are described as *NF* ↔ *NF linking combinations* (see § 2.3).

The overall figures for inflected forms lexica generated by the DIINAR.1 can be broken down as shown in Table 5:

	a Number of nuclei or stems	**b** Number of inflected forms	**b/a** ratio
Verbs	19,457	3,060,716	157.3
Deverbals	70,702	2,909,772	41.15
Nouns and adjectives (+broken plurals)	39,099	1,781,316	45.55
Gramma-tical words	445	---	---
Proper names	1,384	11,403	8.23
Total figure and ratio	131,087	7,774,938	59.31

Table 5: Inflected word-forms, i.e., 'minimal word-forms' (see Table 1)

3.4 The fundamental ratio between lemma-entries and inflected word-forms

High as they may seem, the above figures are not over-powerful, and result from stem-by-stem filtering of information through morphosyntactic specifiers and the associated word-formatives grammar.

One can also compare the ratio between the total number of stems and that of inflected forms to what can be found in another language, which is equally known to be a highly inflected one. The Xerox Spanish Lexical Transducer contained, in 1996 over 46,000 base-forms, and generated over 3,400,000 inflected word-forms (Beesley & Karttunen, 2003, p. xvii). The ratio between inflected forms and base-forms in the Xerox Spanish database was then of around 74 to one. In the DIINAR.1 dB, the same ratio is of just under 60 to one, which can be considered as reasonable.

The question of how many 'maximal word' forms can be correctly generated remains to be introduced and discussed in a further paper.

4 The rationale beyond ratios: towards a first set of validation criteria for Arabic lexica

The ratios considered in the present paper are divided in two general categories:

- The category encountered in § 3.2 involves NF ↔ NF linking combinations (§ 2.3):
  (a) The ratio between the number of noun lemmas (in general vocabulary) and that of 'broken plurals' is of 1 'broken plural' for every 4 nouns.
  (b) The overall ratio between verbs and deverbals gives an average of 3.6 deverbals for one verb.

- The ratios given in § 3.3 and 3.4 consider the number of basic entries, such as nouns, verbs, deverbals, etc., and the inflected forms generated through the rules of the WFG and the grammar-lexis relations specifiers included in the dB. In nouns, the relatively high ratio of 45.55 is due to the combination of case-endings with other suffixes. In proper names, case-endings are limited, because they do not vary according to definiteness or indefiniteness, and also because some categories of proper names are in addition not liable to be followed by the relative suffix –*iyy*).

In this contribution, the numbers of lemma-entries reflect the state of the DIINAR.1 database, which is likely to be modified, in the course of time, through eliminating lemmas corresponding to words that have fallen out of use or through adding new entries. Ratios, on the other hand, reflect the word-formatives grammar as well as the overall structure of the sets of morpho-syntactic specifiers associated to lexical entries. They are, on he whole, to remain stable. It is therefore reasonable to consider that they should be added to the language-specific parts of a check-list devised for the evaluation and validation of Arabic lexical resources, or of multilingual lexica including Arabic.

5 Acknowledgements

The present work is presented with the support of the NEMLAR Euro-Mediterranean project (see note 3).

The generated lexica of DIINAR.1 are to be made available through the European Language Resources Association (ELRA), and The Evaluation and Language resources Distribution Agency (ELDA), Paris, http://www.elda.fr.

References

Wijdan Abbas Mekki. 1998. Définition et description des unités linguistiques intervenant dans l'indexation automatique des textes en arabe, Doct. Dissert., ENSSIB/Université Lyon 2.

Ramzi Abbès. 1999. *Conception d’un prototype de concordancier de la langue arabe,* Mémoire de DEA en Sciences de l’information et de la communication, ENSSIB.

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.* Doct. Dissert., Université Lyon 1.

S. Ammar. & J. Dichy. 1999a. *Les verbes arabes*, Paris, Hatier (collection Bescherelle - Original introduction in French).

— 1999b. *Al-’Af‘âl al-‘arabiyya,* Paris, Hatier (collection Bescherelle - Original Arabic introduction).

Kenneth Beesley. 1989/91. “Computer Analysis of Arabic Morphology: A two-level approach with detours.” In Bernard Comrie and Mushira Eid, eds., 1991. *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics,* Amsterdam, John Benjamins: 155-172.

— 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse: 1-8.

Kenneth Beesley & Lauri Karttunen. 2003. *Finite State Morphology.* CSLI Publications, Stanford, California.

Abdelfattah Braham & Salem Ghazali. 1998. Qâ‘idatu l-bayânât al-mu‘jamiyya al-‘arabiyya, ’aw mas̲rû‘ Mu‘jam al-‘Arabiyya l-’âliyy, ‘Ma‘âlî-DIINAR’, **h**asîla wa-’âfâq. *Al-Majalla l-’Arabiyya li-l-’ulûm,* 32: 14-23.

Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0. <http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49 >

David Cohen. 1961/70. “Essai d'une analyse automatique de l'arabe”. *T.A. informations,* 1961. Reprod. in D. Cohen. 1970. *Études de linguistique sémitique et arabe*. Paris, Mouton: 49-78.

Jean-Pierre Desclés, ed. 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 sous la direction de J.-P. Desclés, à la demande du Ministère français des Affaires étrangères (sous-direction de la Politique linguistique).

Joseph Dichy. 1984/89. “Vers un modèle d’analyse automatique du mot graphique non-vocalisé en arabe”, in Dichy & Hassoun, 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*. State Doct. Dissert. thèse d’État (en linguistique), Université Lumière-Lyon 2.

— 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., 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, spring 1997, Québec, Presses de l’Université de Montréal: 291-306.

— 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.* Special issue of: *Revue Tunisienne de Sciences Sociales,* 117: 93-107.

— 2000. “Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects”. Proceedings of the *ACIDA’ 2000* conference, Monastir (Tunisia), 22-24 March 2000, *Corpora and Natural Language Processing* vol.: 55-60.

— 2001a. “Une première classification des verbes arabes en fonction de leur structure d’arguments”. A. Fassi Fehri, ed., Actes du colloque international *Génération Systématique de la langue et Traduction automatique,* (Rabat, 15-17 novembre 1999). Special issue of: *Recherches Linguistiques,* IERA, May 2001, vol. 2 : 39-70.

— 2001b. “On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases”. *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse: 23-30.

J. Dichy, A. Braham, S. Ghazali, M. Hassoun. 2002. La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l’Arabe, version 1), Proceedings of the *International Symposium on The Processing of Arabic,* Tunis (La Manouba University), 18-20 April 2002.

J. Dichy & Ali Fargaly. 2003. Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? *IXth Machine Translation Summit* (New-Orleans, Sept. 23-27, 2003), Proceedings of the Workshop on *Machine Translation for Semitic Languages: Issues and Approaches*: 1-8

J. Dichy & M.O. 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.

J. Dichy & M.O. Hassoun. 1998. Some aspects of the DIINAR-MBC research programme". In A. Ubaydly, ed., 1998: 2.8.1-6.

Everhard Ditters, ed. 1986-1995. *Processing Arabic Report* 1 (1986), 2 (1987), 3 (1988), 4 (1989), 5 (1990), 6/7 (1993), 9 (1995), Institute for the Languages and Cultures of the Middle-East., Nijmegen University.

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.* Doct. diss. Université Lyon 2.

Bernard Fradin. 1994. L'approche à deux niveaux en morphologie computationnelle et les développements récents de la morphologie", in B. Fradin, ed., *Morphologie computationnelle*, *T.A.L.*, 35, 1994-2, Paris, ATALA: 9-48.

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.

Salem Ghazali & Abdelfattah Braham. 2001. "Dictionary Definitions and Corpus-Based Evidence in Modern Standard Arabic". In *ACL 39th 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.* Doct. dissert., ENSSIB/Université Lyon 2.

Mohamed Hassoun. 1987. *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application*. State doct. dissert., Université Lyon 1.

Pablo Kirtchuk. 1997. Renouvellement grammatical, renouvellement lexical et renouvellement conceptuel en sémitique, in C. Boisson & Ph. Thoiron, eds, 1997, *Autour de la dénomination*, Presses Universitaires de Lyon: 41-69.

Lamia Labed & Xavier Lelubre. 1997. "DIINAR-TOPT: conception d'une base de données terminologique Arabe/français dans le domaine de l'optique". In *JST'97: 1ères JST FRANCIL 1997: L'ingénierie de la langue : de la recherche au produit,* Avignon, 15-16 avril 1997, Aupelf-Uref/Francil: 523-8.

Xavier Lelubre. 1993. "Courseware for the theory and practice of Arabic conjugation". *Processing Arabic Report,* 6/7, TCMO, Nijmegen University: 85-89 and 92-95.

— 2001. "A Scientific Arabic Terms Data Base: Linguistic Approach for a Representation of Lexical and Terminological Features". In *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse: 66-72.

Igor Melčuk. 1982. *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology,* München : Wilhem Fink Verlag.

Riadh Ouersighni. 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.* Doct. dissert., ENSSIB/Université Lyon 2.

SAMIA [Research Group]. 1984. "Enseignement Assisté par Ordinateur de l'arabe. Simulation à l'aide d'un modèle linguistique, la morphologie". *Actes du Colloque "E.A.O. 84"*, (Lyon, 4-5 septembre 1984), Paris, Agence de l'informatique: 81-96.

Ahmad Ubaydly, ed. 1998. *Proceedings of the 6th International Conference and Exhibition on Multilingual Computing (ICEMCO 98),* Centre of Middle Eastern Studies, University of Cambridge.

Riadh Zaafrani. 2002. *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère.* Doct. dissert., ENSSIB/Université Lyon 2.