Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?

Joseph Dichy* - Ali Farghaly**

*Université Lumière-Lyon 2,
Lyon, France
joseph.dichy@univ-lyon2.fr

**SYSTRAN Software Inc.
San Diego (CA), USA
farghaly@systransoft.com

Abstract

Machine translation engines draw on various types of databases. This paper is concerned with Arabic as a source or target language, and focuses on lexical databases. The non-concatenative nature of Arabic morphology, the complex structure of Arabic word-forms, and the general use of vowel-free writing present a real challenge to NLP developers. We show here how and why a *stem-grounded lexical database,* the items of which are associated with *grammar-lexis specifications* – as opposed to a root-&-pattern database –, is motivated both linguistically and with regards to efficiency, economy and modularity. Arguments in favour of databases relying on stems associated with grammar-lexis specifications (such as DIINAR.1 or the Arabic dB under development at SYSTRAN), rather than on roots and patterns, are the following: (a) The latter include huge numbers of rule-generated word-forms, which do not actually appear in the language. (b) Rule-generated lemmas – as opposed to existing ones – are widely under-specified with regards to grammar-lexis relations. (c) In a Semitic language such as Arabic, the mapping of grammar-lexis specifications that need to be associated with every lexical entry of the database is decisive. (d) These specifications can only be included in a stem-based dB. Points (a) to (d) are crucial and in the context of machine translation involving Arabic.

Keywords: MT, multilingual lexical databases, Arabic morphology, Semitic roots and patterns, stem-based lexicons, morphosyntactic specifiers, grammar-lexis specifications, NLP and MT feasibility.

1 Introduction

It is widely acknowledged to-day that machine translation (MT) requires lexical databases, which need to be multilingual on the one hand, and language-oriented on the other. The creation of a multilingual database is a challenging and costly process. It has to provide proper mapping across languages efficiently and economically. This paper is concerned with some aspects of the building of multilingual lexical databases including Arabic, i.e. with language-specific lexical and morphosyntactic structures, which appear to be crucial in the context of MT research and development (R&D).

In relation with well-known features of Semitic languages, the question of the grounds on which an Arabic lexical database should be built is unavoidable. By no means a new question, it remains a tricky one. Pioneering works (such as D. Cohen, 1961/70, Hlal, 1979) suggested ROOT-&-PATTERN grounded analysis of fully vowelled Arabic script. In the 1980s, work from Desclés et al. (1983), Dichy (1984/89), SAMIA (1984), Hassoun (1987), Dichy & Hassoun, eds. (1989), showed that only a stem-grounded database, the entries of which are associated with grammar-lexis specifications, provide NLP applications with sufficient feasibility conditions. This is especially true if the recognition of Arabic standard vowel-free writing is involved (Abbas-Mekki, 1998, Ghenima, 1998, Ouersighni, 2001, Zaafrani, 2002).

The Arabic computational lexicon of T. Buckwalter also takes the stem level into account (Buckwalter, 1990, Beesley, 2001, Maamouri & Cieri, 2002). Recent research on a Lexeme-Based Morphological treatment of Arabic (after Aronoff, 1994 and Beard, 1995) proposed a computational approach according to which "only the stem is morphologically relevant in that realization rules act on it" (Soudi, Cavalli-Sforza, Jamari, 2001).

Other research could also be recalled here [1]. The issue, nevertheless, does not appear to be closed: a number of publications, as well as R&D projects propose, even today, root-&-pattern based processing of Arabic texts.

Another crucial issue lies behind the discussion: this paper focuses on the precise reasons why and how *stem-grounded lexical databases,* the entries of which are associated with *grammar-lexis specifications,* should be recommended in Arabic NLP applications, with special reference to MT.

2 Morpho-lexical relations in Arabic

Arabic ROOTS consist massively of ordered sequences of three consonants, which are traditionally considered as representative of a semantic field. Related nouns, verbs and adjectives are considered as generated through processes of vocalization and affixation, forming a syllabic PATTERN. McCarthy's works (e.g. 1981) present evidence that the combination of ROOTS and PATTERNS in linguistic units is both non-concatenative and sensitive to constraints and rules (the second point going back to Medieval Arabic linguistics).

Ordering of elements in PATTERNS is as crucial as it is in ROOTS. Farghaly (1994) proposes underspecified lexical entries, which are shown in figure 1, where PATTERNS are exemplified – for short – disregarding consonantal affixes.

ROOT	VOCALISM	STEM
[k,t,b	a-a	V, perfect]
[k,t,b	aa-i	N]
[k,t,b	-u-	V, imperfect]
[D,r,b	-a-	Masdar]

Fig. 1: Under-specified Arabic Lexical Entries

PATTERNS are in turn traditionally considered as linguistic units with given meanings and functions. In D. Cohen's elegant phrasing, (1961/70: 50), ROOTS and PATTERNS are considered as defining the meaning of lexical entries in Arabic. Roughly, nouns, verbs and adjectives are seen as resulting from the combination of (a) the "general meaning" of a given ROOT, and (b) a "specific meaning" associated with a PATTERN. The whole lexicon of the language could then be generated using two databases, to which rules accounting for constraints of various natures would be added.

This view – presented here in simplified terms – can be described as the *ROOT-&-PATTERN paradigm*, which goes back to Cantineau (1950). It is still widely shared by many NLP researchers and developers. In this section, we propose a number of critical remarks leading to an alternative approach (§ 2.1 to 2.4).

2.1 ROOT-&-PATTERN representation is only valid for a subset of the lexicon

A substantial subset of nouns is not subject to analysis in terms of ROOT and PATTERN (Dichy, 1984/89; Hassoun, 1987).

- Ancient and medieval Arabic examples:
?ismâ'îl (إسماعيل), "Ishmael", *nâranj* (نارنج), "orange", *sunûnû* (سنونو), "sparrow", ***sirât*** (سراط), "path, way";
- Modern standard Arabic examples:
fusfât or *fu**s**fât* (فسفات - فصفات), "phosphate", *naylûn* or *nîlûn* (نيلون), "nylon".

2.2 ROOT-&-PATTERN representation is essentially valid for verbs and deverbals

A PATTERN can be roughly defined as a template of syllables (including affixed consonants and long vowels), in which the ordered consonants of the ROOT occupy specified positions.

E.g.: The stem *mudarris* (مدرّس), "teacher" consists of (a) the 3-consonant ROOT /d-r-s/ and (b) the PATTERN /muR^1aR^2R^2iR^3/, where R^1, R^2 and R^3 stand for 'radical consonant 1, 2, 3' (R^1 = /r/, R^2 = /d/, R^3 = /s/). Note that R^2 is doubled, and that the pattern-affix *mu-* is based on the mono-consonantal root /m/ [2]. Both features must be included in the formal definition of PATTERNS.

Any number of consonants cannot be a ROOT. The criterion is that of prosodic non-concatenative derivation (known in traditional Semitic studies as "internal" derivation): only tri-consonantal (3-C), and by extension of the morphological structure of the language, quadri-consonantal (4-C) sequences, can be ROOTS (Dichy, 1984/89). In other words, only 3-C and 4-C sequences can enter such derivation as verb ↔ infinitive form, e.g.:

*wa**s**al* (وصل), "to arrive" ↔ *wu**s**ûl* (وصول),"arrival" or "arriving", both forms sharing (a) the ROOT /w-**s**-l/ and

[1] See, e.g., Ditters, ed. 1986-1995 (issues of *Processing Arabic Report*), Ubaydly, ed., 1998, Fassi-Fehri, ed., 2001, or Braham, ed., 2002.

[2] Mono-consonantal roots in Semitic languages were disclosed by A. Roman (1990). For lack of space, they are not taken into account in the conventional use of the term ROOT in this paper.

(b) a coded link of derivation involving two PATTERNS: /R^{1}aR^{2}aR^3/ ↔ /R^{1}uR^{2}ûR^3/,

or, in nouns, singular ↔ "broken" (or "internal") plural, e.g.:

***t**arîq* (طريق), "way, road", sing. ↔ ***t**uruq* (طرق), plural. ROOT: /**t**-r-q/; PATTERNS: /R^{1}aR^{2}îR^3/ ↔ /R^{1}uR^{2}uR^3/.

Form-to-form derivational relations essentially operate in the domain of verbs and deverbals – i.e. verbo-nominal derivatives, such as infinitive forms (*ma**s**dar* – مصدر) and active or passive participles (*?ism al-fâ'il, ?ism al-maf'ûl* – اسم الفاعل والمفعول). It is essential to note that *all Arabic verbs and all deverbals can be analysed in terms of ROOT and PATTERN* (Dichy, 1984/89, 1997). [3]

As a consequence of § 2.1 and 2.2, the *ROOT-&-PATTERN paradigm* already appears to be doubly mistaken: extending its representation to the entire lexicon (a) leaves a large number of lexical entries un-represented, and (b) does not sufficiently take into account its own effective domain of validation.

2.3 Arabic word-form structures entail complex grammar-lexis relations

Arabic word-forms can be roughly considered as equivalent to graphic words. The structure of word-forms in Arabic (Cohen, 1961/70, Desclés et al., 1983, SAMIA, 1984, Hassoun, 1987, Dichy & Hassoun, 1989), comprises [4]:

– *proclitics (PCL),* which include mono-consonantal conjunctions, prepositions, etc.,
  e.g.: *wa-*, "and", *li-*, "in order to"; *bi-*, "in, at" or "by"...
– a *prefix (PRF).* The category, after D. Cohen's representation of the word-form, only includes the prefixes of the imperfective, e.g.:
  ya-, prefixed morpheme of the 3rd person masc. sing.;
– a *stem*. Stems divide into two general types:
  • *Type 1 stems* can be represented in terms of a ROOT and a PATTERN, e.g.:
  The stem *takabbar* (تكبّر), "to be haughty", consists of the 3-C ROOT /k-b-r/ and of the PATTERN /taR^{1}aR^{2}R^{2}aR^3/.
  • *Type 2 stems,* which can be described as *quasi-stems,* cannot be analysed in terms of ROOTS and PATTERNS, but are liable to be included in a word-form, e.g.:
  wa-bi-barnâmaj-i-him (وببرنامجهم), "and by (or through) their program", word-for-word: "and-by-program-indirect case ending /i/-them, masculine plural". No verb or "broken plural" can be derived from the 5-consonant sequence */b-r-n-m-j/* [5];
– *suffixes (SUF),* such as verb endings, nominal case-endings, the nominal feminine ending *-at*, etc.;
– *enclitics (ECL).* In Arabic, enclitics are complement pronouns.

Stems and quasi-stems can be described as the *nucleus formative* (NF) of the word-form, and the other morphemes mentioned above, as *extension formatives* (EF). This convention is particularly useful in highlighting the two overall fields of the *word-formative grammar*, which divides into (1) *EF-EF rules* and (2) *NF-EF rules* (Dichy, 1997).

(1) *EF-EF rules* purely belong to the grammar of the language, e.g.:

• If the proclitics include the preposition *bi-* or *li-*, then the case-ending suffixes are that of the indirect case.
• The proclitic article *?al-* excludes undetermined case endings known as *tanwîn.*

(2a) *NF-EF rules* are correlated to NF categories and sub-categories. They pertain partly to grammar, e.g.:

• the proclitic article *?al-* is exclusively compatible with adjectives and common nouns;
• the proclitic morpheme *sa-*, which denotes the future of verbs, is only compatible with imperfective verb stems;

and for a greater part to grammar-lexis relations:

e.g.: enclitic pronouns are associated with verbs according to selection features such as

<+ human vs. – human complements>

(متعد إلى العقلاء ~ إلى غير العقلاء). One can say, for example: *qara?tu-hu* (قرأته), "I read it", but not **qara?tu-hum*, as the plural masculine pronoun *–hum* only refers to human complements, which is excluded by the grammar-lexis relations associated with the verb *qara?a* in the Arabic lexicon. Unlike *qara?a*, the corresponding French verb *lire* can be used in the phrase *lire quelqu'un*, meaning "to read someone's writings". In English, *to read someone* means either "reading other people's thoughts or feelings" "understanding them" or "receiving their message well", which does not translate *qara?a,* "to read something written". Selection features associated with Arabic *qara?a,* French *lire* and English *read* are not the same.

(2b) A large set of *NF-EF rules* involves "frozen" or "lexicalized" relations between nucleus and extension formatives, as opposed to compositional relations, e.g.:

• The word *jâmi*^{c}*a&* (جامعة) can be analysed either as:

[3] Experimental evidence as to the psychological effect of root and pattern in word recognition is given for Hebrew in Frost et al. (1997, 2000), and for Arabic, in Grainger et al. (2003).

[4] Sampson (1985: 90-1) analyses graphic word-forms in Hebrew. Not surprisingly, word-form structure analyses in that other Semitic language are akin to the one presented here for Arabic.

[5] Note that there exists a de-nominal derivation process, through a reduction of the 5-consonantal sequence *barnâmaj* to a 'productive' 4-C root */b-r-m-j/*, on the basis of which the "internal" plural *barâmaj* and the verb *barmaja*, "to program" could be derived.

(a) the active participle *jâmi^c* "bringing together", "collecting", to which the fem. suffix *–a&* is added, or as:
(b) a lexicalized compound including the meanings of the active participle and the suffix *–a&* of the *res generalis*, "the thing that..." (Roman, 1990). The whole compound, which includes a semantic addition (Dichy, 2002), means "university".
In (a), the relation between *jâmi^c* and *–a&* is simply compositional. In (b), it is clearly frozen or lexicalized (deriving from "the thing that brings together").
• The word *xârijiyy* (خارجي), in a similar way, admits two analyses:
(a) With a compositional relation between its two formatives, respectively, *xârij*, "standing outside" (active participle of *xaraja,* "to go out") and *–iyy*, the suffix of relative nouns and adjectives, it means "exterior", "outside (adjective)".
(b) It has also come to mean "Kharidji" (an early sect of Islam, deriving from "those who walked out" [on Ali]). In this case, the relation between the two formatives is lexicalized (or frozen). Note that the morphological compound *xârijiyy* has a "broken" (or internal) plural *xawârij* (خوارج).

The two types of NF-EF relations account for a finite *and* exhaustive set of grammar-lexis relations, which operate in the domain of the Arabic word-form. They have been formalized in Hassoun (1987) and Dichy (1987, 1990, 1997), and implemented in the DIINAR.1 language database [6]. They have also been extended to scientific terminological units (Lelubre, 2001).

These relations are not connected with PATTERNS. They are not predictable on the sole basis of ROOTS and PATTERNS, and can only be associated with actual lexical entries, which can only be identified in a stem-based lexicon.

2.4 Derivational and Semantic irregularities in the Arabic Lexicon

The widely held assumption that the Arabic ROOT represents a semantic field is based on the belief that all words formally generated from a specific ROOT share some common meanings. Let us revisit one of the most quoted examples. It is traditionally said that all stems that contain the ordered radicals /k-t-b/ relate to the semantic field of writing.

E.g.: *katab* (كتب), "to write", *kâtib* (كاتب), "writer", *maktab* (مكتب), "writing place", "office, desk", *maktaba&* (مكتبة), "library, bookshop", etc.

Not surprisingly, new meanings are likely to occur. For example, Form II verbs, – which are formally generated by the pairing (in the Semitic tradition, the gemination) of the second radical (R^2) of the corresponding Form I verb –, are assumed to have either an iterative or a causative meaning. But many verbs of Form II are not derived from Form I and do not have transparent semantic relations when this is expected, e.g.:

Related to the root /k-t-b/, the verb *kattaba* (كتّب) means:
(1) "to make someone write", which corresponds (for short) to an iterative and causative semantic relation with Form I *kataba*, and
(2) "to form or deploy into squadrons", which derives from the noun *katîba* (كتيبة), "[armoured or cavalry] squadron", the original meaning of which is: "phalanx". The semantic link requires a diachronic mode of explanation: the phalanx is the part of the finger that is used for holding the writing instrument (stylus, calamus…). It has also been used to refer – through another trope of metonymic nature – to a part of the Roman legion (hence "squadron"). The second meaning can by no means be considered as transparent to present-day standard speakers.

In addition, unpredictable derivational links are very often found, e.g.:

Two "nouns of time and place" (*?ism al-makân wa-z-zamân* – اسم المكان والزمان) are related to the verb *kataba*: (1) *maktab* (مكتب), "writing place, desk, office" and (2) *maktaba&* (مكتبة), "copying place, library". The second answered the social need, which probably appeared around the end of the VIIth c., for what was then a new lexical entry. As a result, *maktaba&* has an external plural *maktabât* (مكتبات). By comparison, *madrasa&* (مدرسة), "school" (ROOT: /d-r-s/) has a regular "broken" plural form *madâris* (مدارس). The external plural *maktabât* is due to the fact that the regular or predictable "broken" plural form *makâtib* (مكاتب) was already associated with the singular *maktab* (مكتب), "desk".

Form-to-Form semantic and derivational relations are in fact much more complex than what may seem. Another significant example is:

[6] **DIINAR.1** (*DIctionnaire INformatisé de l'Arabe*), Arabic acronym **Ma'âlî** (*Mu'jam al-'Arabiyya l-'âlî* – **معالي** معجم العربية الآلي), is a comprehensive Arabic Language dB of around 130,000 lemmas, comprising approximately 20,000 verbal entries, 79,000 deverbal items, 29,000 nominal entries (to which 10,000 related "broken plural" items are attached), 1,000 proper names and 450 grammatical tool-words (each of which is associated with a specific grammar). The resource valso includes the clitics and affixes of the language. Entries are by no means mere lists of items: each lexical unit is associated with *morphosyntactic specifiers*. The set of specifiers accounting for the grammar-lexis specifications operating at word-form level is both *finite* and *exhaustive*. Specifiers also include links between morphologically related items such as verb-deverbal(s) or singular-plural, etc. DIINAR.1 has been completed in close cooperation at IRSIT (A. Braham and S. Ghazali) in Tunis, and in France at ENSSIB (Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques – M. Hassoun) and the Lumière-Lyon 2 University (J. Dichy). See Dichy, Braham, Ghazali & Hassoun, 2002.

- Form II *sakkana* (سكّن) "to calm, to appease (someone)", pertaining to the ROOT /s-k-n/, has a meaning which is both iterative (referring to the process of appeasing) and causative. It is related to Form I *sakana* (سكَن), "to be or become still, tranquil, peaceful". The infinitive form (*ma**s**dar*) of *sakana* is *sukûn* (سكون).
- Form IV *?askana* (أسكن), "to give or allocate living quarters", "to settle" or "lodge (someone)", also has a causative meaning related to a Form I verb *sakana.* But this is no longer the same verb as above: the meaning is "to live, to dwell", and the related infinitive form is *sakan* (سَكَن).

Figure 2 below shows other examples where the traditionally anticipated semantic correspondence does not hold:

/?amara/	Form I	To order
/ta?âmara/	Form VI	To conspire
/?axad̲a/	Form I	To take
/?axad̲a/	Form III	To hold against

Figure 2: Semantic irregularities

The interesting point about the examples given here is that these semantic links can be explained through analogy in the aftermath, but could not have been predicted in advance. This is due to the fact that, in order to answer lexical needs, the language resorts to the very rich set of virtual relations offered by the morphological system, in a way that mixes "frozen" and rule-governed relations (Dichy, 2002). As a result, it does not seem possible to consider the morphology of Arabic as a "regular language".

The semantic irregularities noted here suggest that a stem-based dictionary may capture relevant morphological, syntactic and semantic information more adequately, and will not result in overgeneration such as that observed in ROOT-based morphological systems.

3 The ROOT, PATTERN and rule based lexicon of the Xerox analyzer

The fact that rules operating in the composition of Arabic word-forms draw so heavily on grammar-lexis relations, – i.e. that they rely on information associated with lexical entries in the system of the language –, represents a challenge to NLP developers. In this section, we focus on the Xerox Arabic morphological analyzer because it is accessible on the web for everyone to test. Furthermore, it is very well documented in several publications and is based on solid and innovative finite-state technology. Beesley (2001) specifies the function of a morphological analyzer as identifying and separating the “component morphemes of the input word”. In his development of the Arabic morphological analyzer/generator at Xerox Research Centre, Europe, Beesley (2001) takes up the idea that Arabic words consist of at least two building blocks: the ROOT and the prosodic template (McCarthy, 1981). In addition to the theoretical argument for the psychological reality of the Arabic ROOT (see recent references above), Beesley cites an important practical motivation, which is that most Arabic dictionaries are organised around the ROOT, including the famous Hans Wehr-Cowan Arabic-English dictionary.

The Xerox Arabic Morphological Analyzer/Generator was built using Finite-State theory and techniques that have been extended to suit other Semitic languages (Beesley and Karttunen, forthcoming). It is implemented as a finite-state transducer, which applies to input, strings. These strings can be either surface words or lexical analyses of the surface words [7].

Several processes apply in the generation and analysis of Arabic words. First, the process of "interdigitation" or the "merging" of ROOTS and PATTERNS to form stems. Second, alternation rules apply to perform deletion, epenthesis, assimilation, gemination and metathesis. Third, rules for short vowels and other diacritics are relaxed to allow for variations in the way Arabic words are written.

Xerox has several lexicons. The first is a lexicon of ROOTS, which contains 4,930 entries. Each ROOT-entry is manually coded and associated with PATTERNS. The second is a dictionary of PATTERNS, which includes about 400 entries. The manual association of ROOTS and PATTERNS produces about 90,000 Arabic stems. When these stems combine with possible prefixes, suffixes and clitics by composition, 72 million abstract words are generated (Beesley, 2001, p. 7).

[7] The approach relies on previous research, including Buckwalter's lexicon presently used at LDC (Maamouri & Cieri, 2002), and a contribution to Two-level Morphology (Beesley, 1989/91). See also Kiraz, 1994, 1998 (although the review of "Arabic Computational Morphology in the West" is incomplete). In France, a very interesting morphological analyzer based on an original conception of finite-state transducers has been developed by Jaccarini (1997).

4 Stem-based Arabic Lexicons

A multilingual database should require, as primary design criteria, functionality, modularity, efficiency and ease of development. We claim that stem-based lexicons, compared to ROOT-based ones, are more intuitive to build (Farghaly and Senellart, 2003), more efficient, and easier to develop and extend.

First, unlike the entries of ROOT-&-PATTERN grounded databases, in a stem-based dictionary, *all the lemmas are actual lexical units*. They are not abstract or virtual items. A purely ROOT-&-PATTERN generated dictionary would include over 2 million stems (the Xerox lexicons comprise about 5,000 ROOTS and 400 PATTERNS), against 90,000 at Xerox and 130,000 stems in the DIINAR.1 database.

Second, in a stem-based lexicon, the entries of which are associated with word-form grammar-lexis specifications, *rule-governed combination with prefixes, suffixes, proclitics and enclitics only generates existing word-forms* that can actually be found in Modern Standard Arabic oral or written corpora. This is not the case of the 72 million word-forms generated from the 90,000 stems of the Xerox lexicon, which are clearly virtual or abstract units. In the DIINAR.1 lexical database, only 6.2 million *existing* word-forms are generated from the approximately 130,000 stem-based entries (Ouersighni, 2001).

Third, in a stem-based morphological analyzer and/or generator, the process of generating stems from underlying ROOTS is eliminated altogether [8]. Consider the lexical analysis in the ROOT-based morphology of Xerox (Beesley, 2001) as presented in Figure 3:

Upper	[ktb&CaCaC]	+Verb	+Form1+Perf+ Act+At+3p +Fem+Sing
Lower	katabat		

Figure 3: Lexical analysis of the surface form /katabat/

The analysis is excellent. It is extremely valuable from both the linguistic and technical perspectives. However, it also includes information that is not needed in multilingual applications such as machine translation or information retrieval. For example, the information in the second column of the upper side about the ROOT and prosodic template will not be used in such systems – albeit its effective usefulness in an interactive Arabic language teaching/learning software (e.g. Zaafrani, 2002), or in a purely morphological analyzer, such as the Xerox internet demonstrator. A stem-based approach is free either to retain or to eliminate such information.

On the other hand, relevant morphological, syntactic and semantic information need to be associated to lexical entries, *which applies only at the stem level* and not at the ROOT level. This includes, as shown above, sub-categorisation frames and features, argument-structures, etc., pertaining to grammar-lexis relations.

In the perspective of machine translation, it is essential that a multilingual database including Arabic provides for the three fundamental criteria above.

4.1 Arabic lexical dB-s based on stems associated with grammar-lexis specifications are crucial in the context of MT

The example of Arabic *qara?a* vs. English *to read* and French *lire* mentioned in § 2.3, is a good illustration of how argument-structure can vary from one language to another. The lexical database must include, associated with the verbal entry *qara?a,* a grammar-lexis specification code that excludes human direct objects:

> Arabic grammar-lexis relations reject the word-form *qara?a-hum* (قرأهم*), because the enclitic 3rd person plur. masc. pronoun *-hum* can only refer to human objects.

Such examples illustrate generation constraints. These are crucial in MT with Arabic as a target language.

In the recognition of Arabic texts, including MT with Arabic as a source language, one may assume at first that such words-forms are not likely to occur. But this does not close the discussion, as there are many cases of word-forms that are "virtually ambiguous", i.e. that remain ambiguous unless

[8] This does not mean that the concepts of ROOT and PATTERN should be abandoned. Information associated with ROOTS remains essential in the general architecture of an Arabic lexical dB (Hassoun, 1987, Dichy, 1997), in a way that can only be hinted at here. See § 2.2 above.

grammar-lexis specifications are used as filters, e.g.:

• *Wa**s**ala* (وصل), meaning "to arrive" (infinitive form *wu**s**ûl* – وصول) does not take an indirect object. One says: *wa**s**ala ?ilâ makân* (وصل إلى مكان), "to arrive at a [given] place", but not * *wa**s**ala-hu,* *"to arrive it".

• The fact is, the language admits a word-form *wa**s**ala-hu* (وصله), which is related to another verb *wa**s**ala,* "to connect, join", the infinitive form of which is ***s**ila&* (صلة).

• The verb *?ixtalafa* (اختلف) admits two argument-structures, related to two different meanings. (1) If the subject is specified as human and "plural" (the agent corresponds to a "plural subject" as in *tanâfasa,* "to rival, to compete"), the meaning is: "to disagree, to quarrel, to dispute". (2) With an unspecified subject, the verb means "to differ, to vary".

These examples illustrate the fact that a stem-grounded database with grammar-lexis specifications is much more likely to be compatible with MT requirements. They also underline the need for a multilingual database to concentrate on the actual specifications of the Arabic lexicon, in the present state of research and development in machine translation systems taking Arabic as a source of target language.

Transliteration conventions

The transliteration includes no special character, for the sake of portability. 'Emphatic' (pharyngalized) consonants as well as the voice-less pharyngeal ***h**â'* are in **boldface**. Underlining is used to distinguish constrictive consonants from their occlusive counterpart, or from a 'neighbouring' phoneme (this is not a phonetic transcription system!). 'Long' vowels bear a circumflex accent. Here is a short presentation:

- **Short vowels**: a, u, i.
- **Long vowels**: *'alif* = â; *wâw* = û; *yâ'* = î.
- **Consonants** (in Alphabetic order): *hamza* = ?; *bâ'* = b; *tâ'* = t; *tâ'* = t; *jîm* = j; ***h**â'* = **h**; *xâ'* = x; *dâl* = d; *dâl* =d; *râ'* = r; *zây* = z; *sîn* = s; *sîn* = s; ***s**âd* = **s**; ***d**âd* = **d**; ***t**â'* = **t**; ***d**â'* = **d**; ^{c}*ayn* = ^{c}; *gayn* = g; *fâ'* = f; *qâf* = q; *kâf* = k; *lâm* = l; *mîm* = m; *nûn* = n; *hâ'* = h; *wâw* = w; *yâ'* = y.
- **Morphogram**: *tâ' marbû**t**a* = +a&.

References

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

Aronoff, Mark.1994. *Morphology by itself: Stems and Inflectional Classes*. MIT Press, Cambridge, Mass.

Beard, Robert. 1995. *Lexeme-Morpheme Based Morphology: A General Theory of Inflection and Word Formation.* State University of New-York Press, Albany, N.-Y.

Beesley, Kenneth. 1989/91. Computer Analysis of Arabic Morphology: A two-level approach with detours. In Comrie, Bernard and Eid, Mushira, eds. 1991. *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics,* Amsterdam, John Benjamins, pp. 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, pp. 1-8.

Beesley, Kenneth and Karttunen Lauri, forthcoming. *Finite State Morphology.* CSLI Publications, Stanford, California.

Braham Abdelfattah, ed. 2002. *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 (in Arabic, French and English).

Buckwalter, Timothy. 1990. Lexicographic notation of Arabic noun PATTERN morphemes and their inflectional features. In *Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English* (September 5-7, 1990).

Cantineau, Jean. 1950. Racines et schèmes. In *Mélanges W. Marçais,* Paris, G.P. Maisonneuve, pp. 119-124.

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

Desclés, Jean-Pierre, 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,* Paris, Rapport scientifique au Ministère des Affaires étrangères.

Dichy, Joseph. 1984/89. Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe. In Dichy & Hassoun, eds., 1989, pp. 92-158.

— 1987. The SAMIA Research Program, Year 4: Progress and Prospects. In Ditters, ed., 1987, pp. 1-26.

— 1997, Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. In *Meta* 42, Québec, Presses de l'Université de Montréal, pp. 291-306.

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

— 2002. Sens des schèmes et sens des racines en arabe: le *principe de figement lexical* (PFL) et ses effets sur le vocabulaire d'une langue sémitique, in L. Panier et S. Rémi-Giraud, eds., *La polysémie,* Presses Universitaires de Lyon.

Dichy, Joseph & Hassoun Mohamed, eds. 1989. *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe - Travaux SAMIA I.* Paris, Conseil International de la Langue Française.

Dichy. Joseph & Hassoun Mohamed. 1998. Some Aspects of the DIINAR-MBC Research Programme. In Ubaydly, ed., 1998, pp. 2.8.1-5.

Dichy, Joseph, Braham, Abdelfattah, Ghazeli Salem & Hassoun Mohamed. 2002. La base de connaissances linguistiques DIINAR.1 (Dictionnaire Informatisé de l'Arabe – version 1). In Braham, ed., 2002, pp. 45-56.

Ditters, Everhard, ed. 1986-1995. *Processing Arabic Report* 1 (1986), 2 (1987), 3 (1988), 4 (1989), 5 (1990), 6/7 (1993), 9 (1995), T.C.M.O., Catholic University of Nijmegen (Netherlands).

Farghaly, Ali. 1994. Discontinuity in the Lexicon: A Case from Arabic Morphology. *International Conference on Arabic Linguistics*, The American University in Cairo, Cairo, Egypt.

Farghaly, Ali and Senellart, Jean. 2003. Intuitive Coding of the Arabic Lexicon. *Proceedings of the IXth MT Summit*, New Orleans.

Fassi-Fehri, Abdelkader, ed. 2001. *Generation, Systematicity and Machine Translation*, Rabat, Mans̲ûrât Ma^{c}had ad-Dirâsât wa-l-?ab**h**âṯ li-t-Ta^{c}rîb, 2 vol. (In Arabic, English and French).

Frost R., Forster K. & Deutsch A.. 1997. What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 23, pp. 829-856.

Frost R., Deutsch A. & Forster K. 2000. Decomposing morphologically complex words in a non linear morphology. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26, pp. 751-65.

Ghenima, Malek. 1998. *Un système de voyellation de textes arabes,* PhD, Lyon, ENSSIB/Université Lyon 2.

Grainger, Jonathan, Dichy, Joseph, El-Halfaoui, Mohamed and Bamhamed, Mohamed. 2003 (under press). Approche expérimentale de la reconnaissance du mot écrit en arabe. In *Dynamiques de l'écriture : approches pluridisciplinaires*, special issue of *Faits de langue (FDL)*, Jean-Pierre Jaffré, ed.

Hassoun Mohamed. 1987. *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application,* PhD (thèse d'État), Université Lyon 1.

Hlal, Yahya. 1979. *Méthode d'apprentissage pour l'analyse morphosyntaxique (expérimentée dans le cas de l'arabe et du français)*, PhD (thèse d'État), Université de Paris-Sud.

Jaccarini, André. 1997. *Grammaires modulaires de l'arabe. Modélisations, mise en œuvre informatique et strategies.* PhD, Université de Paris IV-Sorbonne.

Kiraz, George Anton. 1994. *Computational Analysis of Arabic Morphology.* PhD, University of Cambridge.

— 1998. Arabic Computational Morphology in the West. In Ubaydly, ed., 1998, pp. 2.2.1-11.

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

Maamouri, Mohamed & Cieri, Christopher. 2002. Resources for Arabic Natural Language Processing at the Linguistic Data Consortium. In Braham, ed., 2002, pp. 125-146.

McCarthy, John. 1981. A Prosodic Theory of Nonconcatenative Morphology. *Linguistic Inquiry*, 12(3), pp. 373-418.

Ouersighni, Riadh. 2001. A major offshoot of the DIINAR-MBC Project: AraParse, a morpho-syntactic analyser of unvowelled Arabic texts. *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse, pp. 9-16.

Roman, André, 1990. *Grammaire de l'arabe.* Paris: P.U.F. (Que sais-je? series).

SAMIA research group. 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.

Sampson, Geoffrey. 1985. *Writing systems.* Stanford University Press.

Soudi, Abdelhadi, Cavalli-Sforza, Violetta, Jamari, Abderrahim. 2001. A Computational Lexeme-Based Treatment of Arabic Morphology. *ACL 39th Annual Meeting.* Workshop on *Arabic Language Processing; Status and Prospect,* Toulouse, pp. 155-62.

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

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